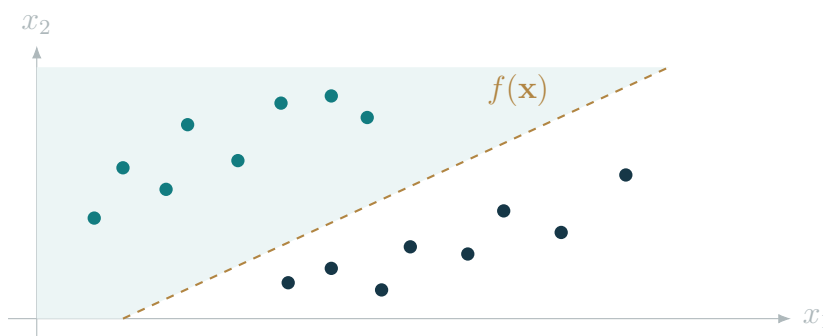


جزوه درسی و آموزشی

داده کاوی

از داده خام تا الگو و مدل قابل اعتماد



پیش پردازش و طبقه بندی | رگرسیون
خوشه بندی | تحلیل وابستگی

استاد درس: دکتر شیرین شعاعی
گردآورنده جزوه: احسان شهبازی

راهنمای مطالعه

این کتاب آموزشی، محتوای علمی هفت مجموعه اسلاید درس داده‌کاوی را به فارسی توضیح می‌دهد و مباحث را با بخش‌های مرتبط از ویرایش دوم کتاب Introduction to Data Mining بسط می‌دهد. ترتیب فصل‌ها از ترتیب اسلایدهای ارسالی پیروی می‌کند؛ هدف، پیوند شهود، فرمول، محاسبه و کاربرد است.

در هر مبحث ابتدا مسئله و زبان آن معرفی می‌شود، سپس سازوکار روش، فرض‌ها و محدودیت‌ها و مثال عددی می‌آید. پرسش‌های خودآزمایی با پاسخ کوتاه برای سنجش فهم‌اند. در سایت، آزمایشگاه‌ها کنار بخش مرتبط و همچنین در صفحه مستقل آزمایشگاه در دسترس‌اند. محاسبات شبیه‌سازی‌ها آموزشی و قابل تکرارند.

دقت کنید | قراردادهای مطالعه

$X \in \mathbb{R}^{n \times d}$ ماتریس داده با n مشاهده و d ویژگی است؛ x_i یک مشاهده و y_i برچسب یا پاسخ آن است. داده آموزش برای برآزش، اعتبارسنجی برای انتخاب تنظیمات و آزمون برای ارزیابی نهایی است. همه پارامترهای پیش‌پردازش نیز فقط روی آموزش برآزش می‌شوند. اعداد و مثال‌های افزوده، مثال آموزشی این جزوه‌اند مگر آنکه صریحاً به مثال اسلاید یا کتاب اشاره شود.

منابع هر فصل با شماره صفحه PDF مشخص شده‌اند؛ شماره چاپی کتاب نیز جداگانه ذکر می‌شود. در نسخه کتاب ارسالی، صفحه PDF معمولاً یک واحد بیشتر از شماره چاپی است. اطلاعات اداری اسلایدهای مقدماتی، مانند برنامه جلسات و ایمیل افراد، در متن علمی تکرار نشده‌اند. مباحث خارج از دامنه اسلایدها به عنوان فصل تازه به درس نسبت داده نمی‌شوند.

برای شروع، فصل اول را برای صورت‌بندی مسئله بخوانید و پیش از مدل‌سازی فصل دوم را انجام دهید. فصل سوم زبان مشترک آموزش و ارزیابی است؛ بدون آن، مقایسه مدل‌های فصل‌های بعد معنای قابل اعتمادی ندارد. فصل‌های ششم و هفتم نیز به داده آزمون و اعتبار الگو حساس‌اند، هرچند الزاماً برچسب هدف ندارند.

فهرست مطالب

i	راهنمای مطالعه
۱	۱ داده‌کاوی؛ از مسئله تا کشف الگو
۱	۱۰۱ داده، الگو و تصمیم چه تفاوتی دارند؟
۲	۲۰۱ چرخه کشف دانش و نقش هر مرحله
۲	۳۰۱ چهار وظیفه اصلی و معنای برجسب
۳	۴۰۱ نمایش داده، ویژگی و هدف
۳	۵۰۱ نخستین الگوریتم: نزدیک‌ترین همسایه‌ها
۴	۶۰۱ چالش‌های داده‌کاوی و پرسش‌های پایانی
۵	۷۰۱ از فاصله عددی تا پیش‌بینی؛ یک اجرای دستی
۵	۸۰۱ خودآزمایی و تمرین
۷	۲ پیش‌پردازش؛ ساختن نمایش درست از داده
۷	۱۰۲ کیفیت داده و معنای نوع ویژگی
۷	۲۰۲ مقدار گم‌شده؛ نبودن داده هم اطلاعات دارد
۸	۳۰۲ عدم توازن طبقه‌ها و نمونه‌سازی
۸	۴۰۲ مقیاس‌بندی و جلوگیری از نشت اطلاعات
۹	۵۰۲ پیش‌بین کاذب و طراحی خط لوله
۹	۶۰۲ رمزگذاری دسته‌ها، گسسته‌سازی و تاریخ
۱۰	۷۰۲ از متن به بردار؛ فراوانی و TF-IDF
۱۱	۸۰۲ شباهت جا‌کارد، کسینوسی و معنای صفر مشترک
۱۲	۹۰۲ انتخاب ویژگی؛ جست‌وجوی یک زیرمجموعه
۱۲	۱۰۰۲ تحلیل مؤلفه‌های اصلی و کاهش بُعد
۱۳	۱۱۰۲ نمونه‌گیری و پوشش فضای داده
۱۳	۱۲۰۲ اجرای یک خط لوله بدون نشت؛ گام به گام

۱۴		۱۳.۲ خودآزمایی و تمرین
۱۶		۳ طبقه‌بندی؛ درخت، تجزیه و ارزیابی
۱۶		۱.۳ از مثال‌های برجسب‌دار تا تصمیم برای داده تازه
۱۶		۲.۳ درخت تصمیم: پرسش‌های ساده، مرزهای مستطیلی
۱۷		۳.۳ ناخالصی جینی و محاسبه یک تقسیم
۱۸		۴.۳ پیدا کردن آستانه ویژگی پیوسته
۱۸		۵.۳ بیش‌برازش، سادگی و تجزیه مدل‌ها
۱۹		۶.۳ ارزیابی: کدام اشتباه و با چه هزینه‌ای؟
۲۱		۷.۳ امتیاز، آستانه و ROC
۲۱		۸.۳ ارزیابی معتبر، انتخاب ابرپارامتر و نشت اطلاعات
۲۳		۹.۳ خودآزمایی و تمرین
۲۶		۴ بیز، حاشیه و شبکه‌های عصبی
۲۶		۱.۴ قاعده بیز و استقلال شرطی
۲۷		۲.۴ هموارسازی، ویژگی پیوسته و داده گمشده
۲۸		۳.۴ هندسه ماشین بردار پشتیبان
۲۸		۴.۴ حاشیه نرم، زیان لولایی و هسته
۲۹		۵.۴ پرسپترون و محدودیت یک مرز خطی
۲۹		۶.۴ شبکه چندلایه، زیان و پس‌انتشار
۳۰		۷.۴ معرفی خانواده‌های عمیق و مرز دامنه درس
۳۰		۸.۴ یک گام واقعی پس‌انتشار، با عدد و مشتق
۳۱		۹.۴ خودآزمایی و تمرین
۳۳		۵ رگرسیون؛ پیش‌بینی پاسخ عددی
۳۳		۱.۵ مسئله چیست؟
۳۳		۲.۵ رگرسیون نزدیک‌ترین همسایه
۳۴		۳.۵ معیارهای خطا و دام همبستگی
۳۴		۴.۵ درخت رگرسیون و درخت مدل
۳۵		۵.۵ رگرسیون خطی و کمترین مربعات
۳۵		۶.۵ منظم‌سازی ridge و lasso
۳۶		۷.۵ درونیابی، برون‌یابی و خط مبنا
۳۶		۸.۵ ایزوتونیک و الگوریتم ادغام ناقضان مجاور

۳۷	چندجمله‌ای و رگرسیون محلی	۹۰۵
۳۷	شبکه عصبی و موازنه اریبی/واریانس	۱۰۰۵
۳۸	چرا میانگین، پاسخ کمترین مربعات است؟	۱۱۰۵
۳۸	خودآزمایی و تمرین	۱۲۰۵
۴۰	خوشه‌بندی و کشف ناهنجاری	۶
۴۰	هدف و نوع خوشه	۱۰۶
۴۰	k-means: الگوریتم و دلیل میانگین	۲۰۶
۴۱	k-medoids و تمایز میانه از medoid	۳۰۶
۴۱	silhouette: سنجش انتساب هر نقطه	۴۰۶
۴۲	DBSCAN و اتصال نواحی پرتراکم	۵۰۶
۴۲	سلسله‌مراتبی: محتوای مکمل تصریح شده در درس	۶۰۶
۴۳	فاصله مناسب، بخشی از تعریف مسئله است	۷۰۶
۴۳	مسئله و روش‌های مقاوم	۸۰۶
۴۴	روش آماری، فاصله‌ای و چگالی نسبی	۹۰۶
۴۴	خوشه‌محور و جنگل جداسازی	۱۰۰۶
۴۵	اجرای دستی؛ چه چیزی در هر گام تغییر می‌کند؟	۱۱۰۶
۴۵	خودآزمایی و تمرین	۱۲۰۶
۴۷	تحلیل وابستگی و کشف زیرگروه	۷
۴۷	سؤال درست و داده تراکنشی	۱۰۷
۴۸	مثال عددی مرجع، یکسان با اسلاید و کتاب	۲۰۷
۴۹	چرا جست‌وجوی خام جواب نمی‌دهد؟	۳۰۷
۴۹	Apriori، منطق و اجرای واقعی	۴۰۷
۵۰	تولید قواعد و هرس درست اطمینان	۵۰۷
۵۰	جذایت، نرخ پایه و Lfit	۶۰۷
۵۱	بسط کنترل‌شده کتاب: نمایش فشرده و FP-growth	۷۰۷
۵۲	شباهت و معیارهای تکمیلی	۸۰۷
۵۲	ویژگی‌های اسمی، پیوسته و کشف زیرگروه	۹۰۷
۵۳	خودآزمایی و تمرین	۱۰۰۷
۵۷	واژه‌نامه، نقشه منابع و مسیر تمرین	آ
۵۷	واژه‌نامه کلیدی	۱.آ

۵۸		تطبیق فصل‌ها با منابع	۲.آ
۵۸		مسیر پیشنهادی تمرین و پروژه	۳.آ
۵۹		چک‌لیست فهم مفهومی	۴.آ

فصل ۱

داده‌کاوی؛ از مسئله تا کشف الگو

منبع مطالعه: اسلاید مقدمه، صفحات PDF ۱۹ تا ۵۸؛ کتاب، فصل ۱، صفحات چاپی ۲۴ تا ۳۳ و فصل ۶، بخش ۳، ۶، ۳.

۱.۱ داده، الگو و تصمیم چه تفاوتی دارند؟

ثبت شدن هزاران خرید، به‌تنهایی توضیح نمی‌دهد که چرا مشتریان خرید را ترک می‌کنند. یک جدول بزرگ فقط مجموعه‌ای از مشاهده‌هاست. وقتی رابطه‌ای تکرارشونده میان مشاهده‌ها پیدا می‌کنیم، به یک الگو می‌رسیم؛ وقتی آن الگو را در زمینه مسئله تفسیر می‌کنیم، می‌تواند مبنای تصمیم شود. داده‌کاوی Data Mining فرایند کشف الگوهای معتبر، تازه، بالقوه مفید و قابل فهم در داده است. هریک از این چهار صفت مهم است: رابطه‌ای که تنها حاصل تصادف باشد معتبر نیست؛ رابطه بدیهی ممکن است اطلاعات تازه ندهد؛ رابطه دقیق ولی بی‌ربط به تصمیم، فایده عملی کمی دارد؛ و خروجی نامفهوم را به‌سختی می‌توان ارزیابی یا به‌کار گرفت.

تعریف | بازیابی با یادگیری یکسان نیست

پرسش «مشتریان با خرید بیش از یک میلیون را پیدا کن» یک شرط از پیش معلوم را روی داده اجرا می‌کند. پرسش «با رفتار ماه گذشته، چه کسی احتمالاً ماه آینده خرید می‌کند؟» نیازمند یادگیری رابطه‌ای از داده و تعمیم آن به موارد ندیده است. اجرای سریع یک پرس‌وجوی پایگاه داده، به‌خودی‌خود داده‌کاوی نیست.

کتاب داده‌کاوی را بخشی از فرایند گسترده‌تر کشف دانش از پایگاه داده، یا KDD، معرفی می‌کند. داده خام ممکن است ناقص، سوگیرانه یا نامتناسب با الگوریتم باشد. حتی پس از ساخت مدل، باید اعتبار، قابلیت تفسیر و اثر استفاده از آن بررسی شود. بنابراین کیفیت نتیجه حاصل کل زنجیره است، نه صرفاً انتخاب یک الگوریتم مشهور.

مثال ۱۰۱ | سه پرسش درباره یک فروشگاه

«چه کالاهایی دیروز فروخته شدند؟» توصیف مستقیم داده است. «چه کالاهایی معمولاً با هم خریده می‌شوند؟» تحلیل وابستگی است. «درآمد هفته آینده چقدر است؟» رگرسیون یا پیش‌بینی سری زمانی است، بسته به نحوه صورت‌بندی و وابستگی زمانی. یک جدول واحد می‌تواند ورودی چند مسئله باشد؛ نوع مسئله را پرسش و خروجی مورد نیاز تعیین می‌کند.

۲۰۱ چرخه کشف دانش و نقش هر مرحله

چرخه با فهم مسئله آغاز می‌شود: تصمیم‌گیرنده کیست، خروجی چه زمانی لازم است، خطا چه هزینه‌ای دارد و هنگام پیش‌بینی چه اطلاعاتی واقعاً در دسترس خواهد بود؟ سپس داده انتخاب و بررسی اولیه می‌شود. تعداد سطرها، نوع ویژگی‌ها، مقدارهای گم‌شده، توزیع هدف و منشأ داده را بررسی می‌کنیم. این مرحله را کاوش یا پروفایل‌سازی داده می‌نامند.

در پیش‌پردازش، خطاها و تکرارها را بررسی، نوع داده را درست و نمایش مناسب را تولید می‌کنیم. در مدل‌سازی، پارامترهای یک روش را با داده آموزش برآورد می‌کنیم. در ارزیابی، می‌پرسیم آیا مدل روی داده ندیده و با معیار متناسب با هدف کار می‌کند؟ در تفسیر، بررسی می‌کنیم چه چیزی آموخته شده و چه فرضی لازم است. در استقرار، مدل در یک فرایند واقعی استفاده و عملکرد آن در زمان پایش می‌شود. چارچوب CRISP-DM نیز بر فهم کسب‌وکار، فهم داده، آماده‌سازی، مدل‌سازی، ارزیابی و استقرار تأکید دارد.

نتیجه کلیدی | بازگشت به عقب بخشی از کار است

چرخه داده‌کاوی یک خط مستقیم نیست. اگر مدل روی یک زیرگروه ضعیف باشد، ممکن است به جمع‌آوری داده بیشتر، اصلاح تعریف هدف یا نمایش بهتر ویژگی‌ها نیاز داشته باشیم. افزایش پیچیدگی مدل همیشه نخستین پاسخ مناسب نیست.

فرض کنید هدف، تشخیص ایمیل ناخواسته است. باید پیش از دیدن مدل روشن کنیم که ایمیل متعلق به زمان ارسال است یا پاسخ کاربر پس از دریافت هم در ویژگی‌ها آمده است. متن ایمیل را به ویژگی عددی تبدیل می‌کنیم، مدل را روی ایمیل‌های گذشته آموزش می‌دهیم و آن را روی ایمیل‌های مستقل می‌سنجیم. اگر واژگان فرستندگان تغییر کنند، عملکرد آینده می‌تواند افت کند؛ پایش و بازآموزی ادامه چرخه‌اند.

۳۰۱ چهار وظیفه اصلی و معنای برچسب

در یادگیری نظارت‌شده Supervised Learning، برای مثال‌های آموزشی، پاسخ مطلوب داریم. در طبقه‌بندی Classification، پاسخ از مجموعه‌ای از طبقه‌هاست؛ مثلاً «هرزنامه/عادی». در رگرسیون Regression، پاسخ عددی است؛ مثلاً قیمت یک خانه. هدف مشترک این است که رابطه ورودی و پاسخ را برای نمونه جدید بیاموزیم، نه اینکه پاسخ‌های آموزش را حفظ کنیم.

در خوشه‌بندی Clustering، برچسب گروه‌ها از قبل به الگوریتم داده نشده است. الگوریتم بر اساس یک تعریف شباهت، ساختاری در داده پیدا می‌کند. گروه‌های حاصل لزوماً با طبقه‌بندی انسانی یکی نیستند. در تحلیل وابستگی Association Analysis، به دنبال اقلام یا رخدادهایی هستیم که با هم مشاهده می‌شوند؛ معمولاً خروجی به شکل مجموعه پرتکرار یا قاعده است.

وظیفه	خروجی	نمونه پرسش
طبقه‌بندی	برچسب گسسته	آیا این پیام ناخواسته است؟
رگرسیون	مقدار عددی	هزینه سفر چقدر خواهد شد؟
خوشه‌بندی	گروه‌بندی بدون برچسب هدف	مشتریان چه تیپ‌های رفتاری دارند؟
وابستگی	مجموعه یا قاعده هم‌وقوعی	کدام کالاها با هم خریداری می‌شوند؟

توصیفی بودن با بی‌نیازی از ارزیابی برابر نیست. خوشه‌هایی که با تغییر کوچک مقیاس از بین می‌روند یا قواعدی که فقط در چند خرید ظاهر شده‌اند، ممکن است قابل اتکا نباشند. همچنین «پیش‌بینی» به معنای کشف رابطه علی نیست: رابطه‌ای که برای پیش‌بینی مفید است الزاماً نشان نمی‌دهد با مداخله روی یک متغیر، دیگری تغییر خواهد کرد.

۴.۱ نمایش داده، ویژگی و هدف

معمولاً هر سطر یک شیء یا مشاهده و هر ستون یک ویژگی Feature است. برای داده جدولی می‌نویسیم

$$X = (x_{ij}) \in \mathbb{R}^{n \times d}, \quad \mathbf{y} = (y_1, \dots, y_n).$$

این نمایش عددی به آن معنا نیست که داده اولیه حتماً عددی بوده است؛ واژه‌ها، رنگ‌ها و تاریخ‌ها باید با رعایت معنایشان به نمایش قابل استفاده تبدیل شوند. «شناسه دانشجو» ممکن است عدد باشد ولی فاصله دو شناسه معنای آموزشی ندارد. در مقابل، فاصله دو اندازه‌گیری قد معمولاً معنا دارد.

کتاب میان داده رکوردی، تراکنشی، سند، گراف و داده مرتب مانند سری زمانی تفاوت می‌گذارد. برای تراکنش، هر سطر یک مجموعه از اقلام است. در سند، هر مؤلفه می‌تواند وزن یک واژه باشد. در گراف، رابطه میان اشیا نیز بخشی از داده است. اگر داده زمانی باشد، جابه‌جا کردن تصادفی سطرها می‌تواند مسئله واقعی را عوض کند؛ هنگام طراحی آزمون باید ترتیب و گروه‌بندی رعایت شود.

پیش از ساخت مدل، واحد مشاهده را دقیق تعریف کنید. اگر هر بیمار چند ردیف داشته باشد، تقسیم ردیف‌ها میان آموزش و آزمون ممکن است بیمار یکسان را در هر دو قرار دهد. در این حالت باید تقسیم بر اساس بیمار انجام شود. این نکته از همان ابتدا بر معنی «نمونه جدید» اثر می‌گذارد.

۵.۱ نخستین الگوریتم: نزدیک‌ترین همسایه‌ها

ایده k-Nearest Neighbors یا k-NN این است که نمونه جدید احتمالاً به نمونه‌های مشابه خود شبیه است. برای پیش‌بینی، فاصله آن تا داده‌های آموزش را حساب می‌کنیم، k فاصله کوچک‌تر را برمی‌گزینیم و در طبقه‌بندی

رأی اکثریت می‌گیریم:

$$\hat{y}(\mathbf{x}) = \operatorname{argmax}_c \sum_{i \in N_k(\mathbf{x})} \mathbf{1}\{y_i = c\}.$$

برای داده‌ عددی، یک انتخاب معمول فاصله اقلیدسی است:

$$d(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}.$$

این فاصله یک انتخاب مدل‌سازی است و باید با معنای ویژگی‌ها سازگار باشد. در رأی وزن‌دار، همسایه نزدیک‌تر سهم بیشتری دارد؛ اگر وزن معکوس فاصله انتخاب شود باید تکلیف فاصله صفر را جداگانه روشن کرد.

مثال ۲۰۱ | رأی‌گیری و نقش مقیاس

پنج ایستگاه نزدیک، به ترتیب برچسب‌های «آفتابی، ابری، ابری، آفتابی، ابری» دارند. برای $k = 5$ رأی اکثریت «ابری» است. اگر فقط نزدیک‌ترین ایستگاه را بگیریم، پاسخ «آفتابی» می‌شود. تغییر k می‌تواند پاسخ را تغییر دهد؛ انتخاب آن باید با اعتبارسنجی انجام شود. اگر فاصله را از سن و درآمد بسازیم، اختلاف چند هزار واحد درآمد می‌تواند اختلاف چند سال سن را تقریباً حذف کند. بنابراین همسایه‌ها تابع نمایش و مقیاس داده‌اند.

k کوچک، مرز تصمیم جزئی و حساس به نویز می‌سازد؛ k بسیار بزرگ می‌تواند تفاوت محلی طبقه‌ها را محو کند. عدد فرد برای مسئله دوطبقه‌ای احتمال تساوی رأی را کم می‌کند، ولی جایگزین قاعده روشن برای شکستن تساوی نیست. آموزش این روش عمدتاً ذخیره داده است، اما پیش‌بینی ساده یک مشاهده حدود $O(nd)$ هزینه دارد. ساختمان‌های جست‌وجو در ابعاد مناسب می‌توانند کمک کنند، ولی در ابعاد زیاد مفهوم نزدیکی دشوارتر می‌شود.

آزمایشگاه پیشنهادی. در آزمایشگاه نزدیک‌ترین همسایه‌ها، k ، محل نقطه پرسش و نویز داده را تغییر دهید. پیش از دیدن پاسخ، برچسب را حدس بزنید و بعد بررسی کنید کدام همسایه‌ها رأی داده‌اند.

۶.۱ چالش‌های داده‌کاوی و پرسش‌های پایانی

بزرگی داده فقط تعداد سطرها نیست. تعداد زیاد ویژگی‌ها، سرعت ورود، نوع‌های متفاوت و توزیع داده میان چند منبع نیز مهم‌اند. زیاد شدن ویژگی‌ها فضای جست‌وجو را بزرگ و نقاط را در فضا پراکنده می‌کند. داده فراوان ولی نامرتب یا سوگیرانه لزوماً مدل خوبی نمی‌سازد. داده‌کاوی از آمار برای سنجش عدم قطعیت، از یادگیری ماشین برای تعمیم، از بهینه‌سازی برای برازش و از پایگاه داده برای ذخیره و پردازش کارآمد استفاده می‌کند.

در گزارش هر پروژه، این زنجیره را قابل پیگیری بنویسید: پرسش، واحد مشاهده، داده قابل دسترس هنگام پیش‌بینی، نمایش ویژگی، تقسیم داده، مدل پایه، معیار ارزیابی و نتیجه آزمون مستقل. اگر نتوانید معیار موفقیت را پیش از دیدن نتیجه بیان کنید، مقایسه روش‌ها مبهم می‌ماند.

۷.۱ از فاصله عددی تا پیش‌بینی؛ یک اجرای دستی

فرض کنید نقطه پرسش $q = (2, 1)$ و چهار مشاهده آموزش به شکل زیر باشند:

مشاهده	مختصات و برجسب	فاصله تا پرسش
A	$(1, 1), y = 0$	1
B	$(2, 3), y = 1$	2
C	$(4, 1), y = 1$	2
D	$(0, 0), y = 0$	$\sqrt{5}$

برای نمونه، فاصله D برابر $\sqrt{(2-0)^2 + (1-0)^2} = \sqrt{5}$ است. پس با یک همسایه، برجسب صفر و با سه همسایه، برجسب یک پیش‌بینی می‌شود. برابری فاصله B, C در این پرسش رأی سه‌تایی را مبهم نمی‌کند، چون هر دو انتخاب می‌شوند؛ اگر مرز انتخاب از میان هم‌فاصله‌ها بگذرد باید قرارداد ثابتی داشته باشیم. فاصله فقط می‌گوید کدام مشاهده نزدیک‌تر است؛ برجسب با رأی‌گیری به دست می‌آید. این دو مرحله را از هم جدا کنید.

برای استفاده واقعی، ابتدا ویژگی‌ها و تعریف فاصله را تعیین و تبدیل لازم را فقط از آموزش یاد می‌گیریم؛ سپس k را با اعتبارسنجی انتخاب می‌کنیم. نقطه پرسش نباید در محاسبه میانگین و مقیاس آموزش شرکت کند. پس از پیش‌بینی نیز لازم است بدانیم همسایه‌ها از چه گروهی آمده‌اند: نزدیکی عددی، به‌تنهایی دلیل علی یا تضمین کیفیت نیست.

آزمایشگاه پیشنهادی. در آزمایشگاه، نقطه پرسش را در مرز دورنگ بگذارید؛ پاسخ یک و چند همسایه را مقایسه کنید و نام همسایه‌های رأی‌دهنده را از جدول بررسی کنید.

۸.۱ خودآزمایی و تمرین

پاسخ را ابتدا با بیان خودتان بنویسید؛ سپس توضیح را مقایسه کنید.

۱. نوع مسئله را از روی خروجی تعیین کنید

گروه‌بندی مشتریان بدون برجسب، پیش‌بینی مبلغ خرید و پیش‌بینی خرید/عدم خرید چه تفاوتی دارند؟
پاسخ و توضیح: به ترتیب خوشه‌بندی، رگرسیون و طبقه‌بندی‌اند. در دو مورد آخر پاسخ آموزشی داریم، ولی نوع خروجی متفاوت است. یکسان بودن جدول ورودی، وظیفه را یکسان نمی‌کند.

۲. بازیابی یا یادگیری؟

آیا پیدا کردن دانشجویی با یک شماره مشخص، داده‌کاوی است؟
پاسخ و توضیح: بازیابی مستقیم است؛ شرط را از قبل می‌دانیم و رابطه‌ای برای تعمیم به مشاهده ندیده نمی‌آموزیم. پیش‌بینی خطر افت تحصیلی از ویژگی‌های قبلی، مسئله یادگیری متفاوتی است.

۳. رأی همسایه‌ها را حساب کنید

در مثال این فصل، خروجی برای یک و سه همسایه چیست؟ اگر هر چهار همسایه رأی دهند چه مسئله‌ای پیش می‌آید؟

پاسخ و توضیح: برای یک همسایه صفر و برای سه همسایه یک است. با چهار همسایه رأی‌ها برابر می‌شوند؛ باید قاعده شکستن تساوی را از قبل تعیین کرد. عوض کردن آن پس از دیدن پاسخ درست، ارزیابی را مخدوش می‌کند.

۴. چرا آزمون نهایی محل انتخاب نیست؟

بیست مقدار k را روی آزمون می‌سنجیم و بهترین را گزارش می‌کنیم. مشکل چیست؟

پاسخ و توضیح: نتیجه آزمون وارد انتخاب مدل شده است. بهترین امتیاز از میان تلاش‌های متعدد معمولاً خوش‌بینانه است؛ تنظیم را با اعتبارسنجی انتخاب و فقط نتیجه نهایی را روی آزمون مستقل گزارش می‌کنیم.

۵. نشت پاسخ را تشخیص دهید

برای پیش‌بینی قیمت خانه، ستون قیمت فروش را هم در ورودی قرار می‌دهیم. دقت عالی چه معنایی دارد؟

پاسخ و توضیح: پاسخ از قبل در اختیار مدل است؛ این نتیجه کیفیت پیش‌بینی واقعی را نشان نمی‌دهد. باید فقط اطلاعاتی را وارد کرد که در لحظه پیش‌بینی در دسترس‌اند. همین بررسی درباره ویژگی‌های پس از رخداد نیز لازم است.

۶. همراهی چه زمانی علت نیست؟

خریداران یک کالا بیشتر کالای دیگری می‌خرند. آیا تخفیف کالای اول حتماً فروش دومی را افزایش می‌دهد؟

پاسخ و توضیح: خیر؛ ویژگی مشتری، فصل یا عامل مشترک می‌تواند هر دو خرید را توضیح دهد. الگوی مشاهده‌ای برای پیش‌بینی مفید است، اما اثر مداخله به طراحی علیّ، آزمایش یا فرض‌های اضافی نیاز دارد.

فصل ۲

پیش‌پردازش؛ ساختن نمایش درست از داده

منبع مطالعه: اسلاید پیش‌پردازش، صفحات PDF ۳ تا ۵۹؛ کتاب، فصل ۲، بخش‌های ۲،۱ تا ۲،۴ و فصل ۶، بخش عدم توازن طبقه‌ها.

۱.۲ کیفیت داده و معنای نوع ویژگی

الگوریتم، معنای واقعی یک ستون را از نام آن نمی‌فهمد. ابتدا برای هر ویژگی نوع، واحد، دامنه معتبر، روش اندازه‌گیری و زمان دسترس‌پذیری را ثبت کنید. ویژگی اسمی Nominal فقط هویت گروه را بیان می‌کند؛ ویژگی ترتیبی Ordinal ترتیب دارد ولی فاصله رتبه‌ها الزاماً برابر نیست؛ ویژگی فاصله‌ای مبدأ قراردادی دارد، مانند دمای سلسیوس؛ ویژگی نسبتی صفر معنادار دارد، مانند طول. مجاز بودن جمع، نسبت یا فاصله به این معنا بستگی دارد.

کتاب خطای اندازه‌گیری را از خطای گردآوری جدا می‌کند. نویز مؤلفه تصادفی خطاست؛ سوگیری اندازه‌گیری می‌تواند همه مقادارها را در یک جهت جابه‌جا کند. تکرارپذیری بالا به‌تنهایی صحت را تضمین نمی‌کند: ترازویی که همیشه دو کیلوگرم بیشتر نشان می‌دهد، ممکن است اندازه‌گیری‌های بسیار نزدیک به هم ولی نادرست تولید کند.

داده پرت Outlier الزاماً خطا نیست. خرید بسیار بزرگ می‌تواند ورود اشتباه یک صفر یا یک مشتری واقعی و مهم باشد. حذف خودکار تمام نقاط نادر ممکن است همان پدیده‌ای را پاک کند که می‌خواستیم کشف کنیم. ابتدا قيود قطعی دامنه، ناسازگاری واحد، سطرهای تکراری و منشأ ثبت را بررسی کنید. حدهایی مانند «دمای مجاز» فقط در زمینه مشخص معنا دارند.

۲.۲ مقدار گم‌شده؛ نبودن داده هم اطلاعات دارد

میان «صفر»، «پرسیده نشده» و «قابل اعمال نیست» تفاوت وجود دارد. صفر کردن همه فقدان‌ها می‌تواند فاصله‌ها، میانگین‌ها و مرز تصمیم را عوض کند. حذف سطر، جای‌گذاری میانگین یا میانه، مقدار پرتکرار، نشانه فقدان و مدل پیش‌بینی مقدار گم‌شده، گزینه‌هایی با فرض‌های متفاوت‌اند. میانگین معمولاً واریانس را کم و تراکم مصنوعی در مرکز ایجاد می‌کند؛ میانه در برابر چند مقدار بسیار بزرگ مقاوم‌تر است، ولی ساختار چندمتغیره را بازسازی نمی‌کند.

در MCAR فقدان مستقل از داده‌هاست؛ در MAR پس از شرط کردن بر اطلاعات مشاهده‌شده می‌توان سازوکار فقدان را توضیح داد؛ در MNAR خود مقدار مشاهده‌نشده نیز در احتمال فقدان نقش دارد. این دسته‌بندی توضیح می‌دهد چرا یک جای گذاری یکسان برای همهٔ ستون‌ها، راه‌حل عمومی نیست. یک نشانگر دودویی برای «این مقدار گمشده بود» می‌تواند بخشی از اطلاعات را حفظ کند، ولی سوگیری گردآوری را خودبه‌خود برطرف نمی‌کند.

مثال ۱۰۲ | جای گذاری می‌تواند رابطه را وارونه کند

در مثال اسلاید، برای ۴۰ نفر قند خون بالا ثبت شده و ۳۰ نفرشان عارضه دارند: نسبت $30/40 = 0.75$ است. برای ۲۰۰ نفر قند خون نامعلوم است و ۲۰ نفر عارضه دارند. اگر همهٔ مقادیرهای نامعلوم را «بالا» قرار دهیم، گروه «بالا» به ۲۴۰ نفر با ۵۰ عارضه تبدیل می‌شود و نسبت $50/240 \approx 0.208$ خواهد شد. دادهٔ تازه‌ای دربارهٔ قند خون کشف نکرده‌ایم؛ صرفاً با فرض جای گذاری، معنی گروه را عوض کرده‌ایم.

در جای گذاری مبتنی بر همسایه، افراد مشابه بر اساس ویژگی‌های موجود پیدا می‌شوند و از مقدارهایشان برای برآورد استفاده می‌شود. فاصله باید با فقدان و مقیاس سازگار باشد. تعداد همسایه و وزن‌دهی، تنظیمات مدل‌اند و باید در اعتبارسنجی انتخاب شوند. میانگین یا مدل جای گذاری را فقط روی آموزش برازش کنید.

۳.۲ عدم توازن طبقه‌ها و نمونه‌سازی

اگر از هزار نمونه فقط ده نمونه مثبت باشند، مدلی که همیشه منفی بگوید دقت 0.99 دارد، ولی هیچ مثبت را پیدا نمی‌کند. ابتدا ماتریس درهم‌ریختگی و هزینهٔ خطاها را بررسی کنید. کم‌نمونه‌گیری طبقهٔ بزرگ داده را کنار می‌گذارد؛ پیش‌نمونه‌گیری با تکرار، اطلاعات مستقل تازه تولید نمی‌کند و می‌تواند به حفظ نمونه‌ها بینجامد. وزن طبقه، آستانهٔ تصمیم و معیار مناسب نیز راه‌های مقابله با این مسئله‌اند.

در SMOTE، از یک نمونهٔ اقلیت و یکی از همسایه‌های همان طبقه، نقطه‌ای در پاره‌خط میان آن‌ها می‌سازیم:

$$\mathbf{x}_{\text{new}} = \mathbf{x} + \lambda(\mathbf{x}_{\text{neighbor}} - \mathbf{x}), \quad 0 \leq \lambda \leq 1.$$

در نسخهٔ متعارف، یک ضریب مشترک برای مختصات یک نقطه استفاده می‌شود؛ انتخاب ضریب مستقل برای هر مؤلفه الزاماً همان نمونه‌سازی روی پاره‌خط نیست. نمونهٔ مصنوعی می‌تواند مرز طبقه را خراب کند، به‌ویژه اگر همسایه نامناسب یا دادهٔ آمیختهٔ عددی و اسمی داشته باشیم. بازنمونه‌گیری فقط داخل دادهٔ آموزش هر دور انجام می‌شود؛ آزمون باید توزیع واقعی کاربرد را بازتاب دهد.

۴.۲ مقیاس‌بندی و جلوگیری از نشت اطلاعات

فاصلهٔ دو فرد با سن‌های ۲۵ و ۳۰ و درآمدهای ۷۰۰۰۰ و ۸۵۰۰۰ تقریباً فقط از اختلاف درآمد اثر می‌گیرد. استانداردسازی و مقیاس‌بندی کمینه-بیشینه دو نمایش متفاوت می‌سازند:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad u_{ij} = \frac{x_{ij} - a_j}{b_j - a_j}.$$

در این روابط، $\mu_j, \sigma_j, a_j, b_j$ از آموزش به دست می‌آیند. برای ویژگی ثابت، تقسیم بر صفر باید با قاعده‌ای مشخص، مثلاً خروجی صفر و ثبت ویژگی ثابت، مدیریت شود. آزمون می‌تواند خارج از بازه آموزش باشد؛ در این صورت تبدیل کمینه-پیشینه ممکن است مقداری خارج از $[0, 1]$ بدهد و این تناقضی ندارد.

دقت کنید | برازش تبدیل هم بخشی از آموزش است

اگر میانگین، واژگان متن، مؤلفه‌های PCA یا ویژگی‌های منتخب را با کل داده محاسبه کنیم، اطلاعات اعتبارسنجی یا آزمون وارد فرایند یادگیری شده است. ترتیب درست عبارت است از تقسیم داده، برازش تبدیل روی آموزش و اعمال همان تبدیل روی داده ندیده. در اعتبارسنجی متقابل، این کار داخل هر دور تکرار می‌شود.

استانداردسازی، توزیع را الزاماً نرمال نمی‌کند؛ تنها مکان و مقیاس را تغییر می‌دهد. درخت تصمیم نسبت به تبدیل یکنوازی یک ویژگی رفتاری متفاوت با فاصله‌محورها دارد، زیرا می‌تواند آستانه را جابه‌جا کند، اما روش‌هایی مانند همسایه، SVM، PCA و رگرسیون با جریمه به مقیاس حساس‌اند. برای داده پرت، مقیاس‌بندی مقاوم با میانه و دامنه میان‌چارکی می‌تواند مناسب‌تر باشد.

آزمایشگاه پیشنهادی. در آزمایشگاه مقیاس، فاصله و همسایه نزدیک را در حالت خام، استاندارد و کمینه-پیشینه مقایسه کنید. بررسی کنید چرا تبدیل محور و تغییر تعریف فاصله باید از هم متمایز باشند.

۵.۲ پیش‌بین کاذب و طراحی خط لوله

ویژگی‌ای که از پاسخ مشتق شده، یا فقط پس از رخداد هدف در دسترس است، پیش‌بین معتبر زمان استفاده نیست. رتبه همان دانشگاه برای پیش‌بینی «دانشگاه برتر»، نتیجه نهایی آزمون برای پیش‌بینی قبولی پیش از آزمون و وضعیت پرداخت آینده برای پیش‌بینی عدم پرداخت نمونه‌های نشت‌اند. همبستگی بالا نشانه بررسی است، نه حکم قطعی کاذب بودن؛ زمان تولید و معنای ویژگی تعیین‌کننده‌اند.

یک Pipeline پیش‌پردازش و مدل را به واحدی تبدیل می‌کند که در هر دور اعتبارسنجی از نو برازش می‌شود. ColumnTransformer برای اعمال تبدیل متفاوت روی ستون‌های عددی، اسمی و متنی است. ترتیب پیشنهادی می‌تواند «جای‌گذاری و مقیاس‌بندی عددی؛ جای‌گذاری و رمزگذاری گروه‌ها؛ اتصال خروجی‌ها؛ مدل» باشد. انتخاب ویژگی تحت نظارت نیز باید داخل همین فرایند ارزیابی قرار گیرد.

۶.۲ رمزگذاری دسته‌ها، گسسته‌سازی و تاریخ

برای ویژگی ترتیبی، کد عددی باید ترتیب را حفظ کند؛ باین حال نسبت کد ۸ به ۴ الزاماً معنای «دو برابر بودن کیفیت» ندارد. برای رنگ‌های اسمی، عددهای ۱، ۲ و ۳ فاصله و ترتیب ساختگی می‌سازند. در رمزگذاری تک‌داغ One-Hot، هر گروه یک نشانگر دارد. در مدل خطی با عرض از مبدأ، نگه داشتن همه ستون‌های تک‌داغ می‌تواند وابستگی خطی ایجاد کند؛ حذف یک گروه مرجع یا منظم‌سازی، راه‌های معمول مدیریت آن‌اند. تکلیف گروه ناشناخته در آزمون باید مشخص باشد.

گسسته‌سازی، عدد را به بازه تبدیل می‌کند. در پهنای مساوی، طول بازه‌ها برابر است؛ در فراوانی مساوی، تعداد مشاهده‌ها تا حد امکان نزدیک است. تکرار مقدارها ممکن است مانع تقسیم دقیق به تعدادهای مساوی شود. مرزها نباید دو مقدار برابر را بی‌دلیل به گروه‌های متفاوت بفرستند. این تبدیل می‌تواند نویز را کم کند، ولی اطلاعات دقیق درون بازه را از بین می‌برد.

مثال ۲۰۲ | دو روش بازه‌بندی

برای داده مرتب 0, 4, 12, 16, 16, 18, 24, 26, 28، سه بازه با پهنای ۱۰، تعدادهای 3, 4, 2 می‌دهند. با هدف سه مشاهده در هر گروه می‌توان مرزهای ۱۴ و ۲۱ را برگزید و سه گروه سه‌تایی ساخت. این مثال بدون شکستن تساوی مقدارهای ۱۶ امکان‌پذیر است، ولی چنین امکانی برای همه داده‌ها وجود ندارد.

برای تاریخ، نخست قالب و منطقه زمانی و سپس ویژگی کاربردی مانند روز هفته، فصل یا فاصله تا یک رویداد را استخراج کنید. ساعت ۲۳ و ۰ نزدیک‌اند؛ یک کد عددی خام این نزدیکی را نشان نمی‌دهد. نمایش چرخه‌ای با سینوس و کسینوس، برای دوره T ، می‌تواند چنین باشد:

$$\left(\sin \frac{2\pi t}{T}, \cos \frac{2\pi t}{T} \right).$$

استفاده از متغیر زمانی باید با هدف پیش‌بینی سازگار باشد؛ اطلاعات آینده را نمی‌توان صرفاً با تغییر قالب، به ویژگی مجاز تبدیل کرد.

۷.۲ از متن به بردار؛ فراوانی و TF-IDF

در پاک‌سازی متن، نویسه‌ها یک‌دست و متن به واحدهای مناسب شکسته می‌شود. برای فارسی، تفاوت «ی/ی»، «ک/ک» و نیم‌فاصله مهم است. حذف نشانه‌گذاری یا واژه توقف همیشه بی‌ضرر نیست: حذف «نه» می‌تواند معنی جمله را عوض کند. ریشه‌یابی شکل‌های صرفی را نزدیک می‌کند، ولی ممکن است واژه‌های متفاوت را نیز ادغام کند.

در کیسه واژه‌ها Bag of Words، ترتیب واژه‌ها نادیده گرفته می‌شود و هر سند با برداری از وقوع، شمارش یا فراوانی واژه‌ها نمایش داده می‌شود. نمایش دودویی تنها حضور را می‌سنجد؛ شمارش به طول سند حساس است. فراوانی نرمال‌شده و وزن‌دهی TF-IDF اثر طول و واژه‌های بسیار رایج را تغییر می‌دهند:

$$w_{ij} = \text{tf}_{ij} \log \frac{N}{\text{df}_j}.$$

N تعداد سندهای آموزش و df_j تعداد سندهای حاوی واژه j است. در قرارداد هموارشده رایج، وزن معکوس سند برابر است با

$$\text{idf}_j = \log \frac{1 + N}{1 + \text{df}_j} + 1.$$

پس هنگام مقایسه دو پیاده‌سازی باید قرارداد TF، هموارسازی و نرمال‌سازی بردار را مشخص کرد؛ همه خروجی‌های موسوم به TF-IDF عدد یکسانی نمی‌دهند.

مثال ۳.۲ | وزن واژه عمومی و اختصاصی

در سه سند، واژه «داده» در هر سه و «خوشه» فقط در یکی آمده است. در قرارداد بدون هموارسازی، وزن IDF «داده» صفر و وزن «خوشه» برابر $\log 3$ است. در قرارداد هموارشده، وزن‌ها به ترتیب 1 و $1 + \log 2$ می‌شوند. صفر نبودن واژه عمومی در قرارداد دوم به معنی خطا نیست؛ تعریف وزن متفاوت است.

بردارهای واژه معمولاً تنک‌اند: از ده‌ها هزار مؤلفه تنها تعداد کمی ناصفر است. بردار تعبیه یا Embedding نمایش چگال و آموخته‌شده‌ای است که رابطه معنایی را به فاصله یا جهت بردار پیوند می‌دهد. مدل‌های رمزگذار مانند BERT، رمزگشا مانند مدل‌های مولد و رمزگذار-رمزگشا نقش‌های متفاوتی دارند؛ صرفاً کوتاه‌تر بودن بردار تضمین نمی‌کند که برای هر وظیفه بهتر باشد. انتخاب نمایش باید با داده و ارزیابی همان مسئله سنجیده شود.

۸.۲ شباهت جا‌کارد، کسینوسی و معنای صفر مشترک

برای بردار دودویی نامتقارن، نبودن هم‌زمان هزار واژه دلیل خوبی برای مشابه بودن دو سند نیست. جا‌کارد، حضورهای مشترک را نسبت به اجتماع حضورها می‌سنجد:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{M_{11}}{M_{11} + M_{10} + M_{01}}.$$

برای دو مجموعه تهی، این کسر تعریف ندارد و قرارداد کاربرد باید جداگانه مشخص شود. ضریب تطابق ساده، صفرهای مشترک را هم می‌شمارد و برای داده تنک می‌تواند شباهت گمراه‌کننده بسازد. برای بردارهای وزن‌دار غیرصفر، کسینوس زاویه به کار می‌رود:

$$\cos(\mathbf{x}, \mathbf{z}) = \frac{\mathbf{x}^T \mathbf{z}}{\|\mathbf{x}\|_2 \|\mathbf{z}\|_2}.$$

ضرب بردار در عدد مثبت، کسینوس را تغییر نمی‌دهد؛ جمع یک ثابت با همه مؤلفه‌ها عموماً آن را تغییر می‌دهد. کسینوس برای بردارهای نامنفی بین صفر و یک است، ولی برای بردار عمومی می‌تواند منفی باشد. برای بردار صفر باید قاعده مشخصی در نظر گرفت.

مثال ۴.۲ | کنترل محاسبه کسینوس

برای $\mathbf{x} = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$ و $\mathbf{z} = (1, 0, 0, 0, 0, 0, 1, 0, 2)$ ضرب داخلی 5 و مربع طول‌ها 42 و 6 است. بنابراین

$$\cos(\mathbf{x}, \mathbf{z}) = \frac{5}{\sqrt{42}\sqrt{6}} \approx 0.315.$$

طول بردار دوم $\sqrt{6} \approx 2.449$ است. محاسبه طول‌ها و ضرب داخلی به صورت جدا، راه ساده‌ای برای پیدا کردن خطاهای عددی است.

۹.۲ انتخاب ویژگی؛ جست‌وجوی یک زیرمجموعه

ویژگی کم‌تغییر، نامرتب یا تکراری ممکن است محاسبه را زیاد و تعمیم را ضعیف کند. حذف نام و شناسه یا ستون کاملاً ثابت، یک بررسی اولیه است؛ ولی کم بودن تغییر همیشه به معنی بی‌اهمیتی نیست، به‌ویژه برای رخداد نادر. همبستگی خطی نزدیک صفر نیز وابستگی غیرخطی را رد نمی‌کند.

روش فیلتر Filter با معیاری مانند اطلاعات متقابل یا آزمون مناسب، ویژگی‌ها را مستقل از اجرای مکرر مدل هدف رتبه‌بندی می‌کند. روش پوششی Wrapper زیرمجموعه را با اجرای مدل و سنجش اعتبارسنجی ارزیابی می‌کند. روش تعبیه‌شده Embedded انتخاب را در خود آموزش انجام می‌دهد. فیلتر معمولاً ارزان‌تر است؛ پوششی به مدل و تقسیم داده حساس و پرهزینه‌تر است. ارزیابی همهٔ زیرمجموعه‌های d ویژگی، 2^d حالت دارد.

در انتخاب پیش‌رونده از مجموعه‌تهی شروع می‌کنیم و هر بار بهترین افزودن را می‌پذیریم. در حذف پس‌رونده از همهٔ ویژگی‌ها شروع می‌کنیم. این جست‌وجوهای حریصانه بهترین جواب جهانی را تضمین نمی‌کنند. برای مثال، در الگوی XOR هر ویژگی به‌تنهایی ممکن است بی‌اطلاع باشد، اما دو ویژگی با هم کاملاً مفید باشند؛ انتخاب پیش‌رونده‌ای که بدون بهبود تک‌ویژگی متوقف می‌شود، این تعامل را از دست می‌دهد. روش‌های حذف بازگشتی و انتخاب مبتنی بر اهمیت مدل نیز باید با اعتبارسنجی مناسب انجام شوند.

۱۰.۲ تحلیل مؤلفه‌های اصلی و کاهش بُعد

انتخاب ویژگی، ستون‌های موجود را برمی‌گزیند؛ PCA ستون‌های تازه‌ای می‌سازد که ترکیب خطی ستون‌های قبلی‌اند. پس از مرکز کردن داده، مؤلفهٔ اول جهت واحدی است که واریانس تصویر داده روی آن بیشینه باشد:

$$\mathbf{v}_1 = \underset{\|\mathbf{v}\|_2=1}{\operatorname{argmax}} \mathbf{v}^T S \mathbf{v}, \quad S = \frac{X_c^T X_c}{n-1}.$$

مؤلفه‌های بعدی همین هدف را تحت قید عمود بودن بر قبلی‌ها دنبال می‌کنند. با تجزیهٔ ویژهٔ S ، بردارهای ویژه جهت‌ها و مقدارهای ویژه واریانس‌ها را می‌دهند. برای r مؤلفهٔ نخست:

$$Z = X_c V_r, \quad \hat{X} = Z V_r^T + \mu, \quad \operatorname{EVR}_j = \frac{\lambda_j}{\sum_{\ell} \lambda_{\ell}}.$$

برگرداندن میانگین در رابطهٔ بازسازی، سطر به سطر انجام می‌شود. علامت هر بردار ویژه دلخواه است؛ برعکس شدن آن مفهوم مؤلفه یا بازسازی را تغییر نمی‌دهد. مؤلفه‌ها ناهمبسته‌اند، ولی بدون فرض اضافی الزاماً مستقل نیستند.

مثال ۵.۲ | PCA دو متغیر همبسته

اگر $S = \begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$ باشد، مقدارهای ویژه ۷ و ۱ و جهت اول $(1, 1)/\sqrt{2}$ است. یک مؤلفه $7/8 = 87.5\%$ واریانس را نگه می‌دارد. این جهت «بیشترین تغییر داده» را نشان می‌دهد، نه لزوماً «بهترین جهت برای پیش‌بینی برچسب». اگر برچسب در جهت کم‌واریانس جدا شود، حذف آن جهت می‌تواند به طبقه‌بندی آسیب بزند.

مقیاس ویژگی‌ها و داده‌های پرت بر PCA اثر دارند. انتخاب تعداد مؤلفه صرفاً با درصد واریانس، جای ارزیابی وظیفه نهایی را نمی‌گیرد. ابتدا میانگین و جهت‌ها را از آموزش یاد بگیرید و همان‌ها را روی آزمون اعمال کنید. آزمایشگاه پیشنهادی. در آزمایشگاه PCA، جهت تصویر را بچرخانید و واریانس نگه‌داشته شده و خطای بازسازی را ببینید. سپس آن را با جهت مؤلفه اول مقایسه کنید.

۱۱.۲ نمونه‌گیری و پوشش فضای داده

کاهش تعداد سطرها با کاهش تعداد ستون‌ها متفاوت است. نمونه تصادفی ساده احتمال انتخاب برابر می‌دهد؛ نمونه طبقه‌بندی شده، سهم گروه‌ها را کنترل می‌کند. حفظ نسبت گروه‌ها برای برآورد توزیع مهم است، در حالی که برابر کردن تعداد گروه‌ها ممکن است برای آموزش یک مسئله نامتوازن هدف دیگری داشته باشد. «نمونه متوازن» و «نمونه نماینده توزیع» مترادف نیستند.

در روش Kennard–Stone، ابتدا دورترین زوج انتخاب می‌شود؛ سپس در هر گام، برای هر نامزد فاصله‌اش تا نزدیک‌ترین عضو منتخب را می‌سنجیم و نامزدی با بزرگ‌ترین این فاصله کمینه را برمی‌گزینیم. این روش پوشش هندسی را تقویت می‌کند، ولی می‌تواند نقاط پرت را بیش از حد انتخاب و نسبت گروه‌ها را تغییر دهد. محاسبه همه فاصله‌های زوجی نیز هزینه حافظه و زمان دارد.

نمونه‌گیری ماه دسامبر برای پیش‌بینی همه سال یا نظرسنجی تلفنی تنها در ساعت کاری می‌تواند سوگیری انتخاب بسازد. هیچ الگوریتم پیچیده‌ای از داده‌ای که گروه مهمی را اصلاً نمی‌بیند، اطلاعات آن گروه را خلق نمی‌کند. منشأ نمونه و جمعیت هدف را همراه نتیجه گزارش کنید.

۱۲.۲ اجرای یک خط لوله بدون نشت؛ گام به گام

در چهار رکورد آموزش، سن‌ها 20, 40, missing, 60 هستند و سن یک مراجعه‌کننده تازه ۸۰ است. سه مقدار مشاهده شده میانگین ۴۰ دارند؛ مقدار گمشده را با همین میانگین جایگزین می‌کنیم. اکنون آموزش شامل 20, 40, 40, 60 است. اگر استانداردسازی با واریانس مخرج n انجام شود:

$$\begin{aligned}\mu_{\text{train}} &= 40, \\ \sigma_{\text{train}} &= \sqrt{\frac{(20-40)^2 + 0 + 0 + (60-40)^2}{4}} = \sqrt{200}, \\ z_{\text{new}} &= \frac{80-40}{\sqrt{200}} \approx 2.8284.\end{aligned}$$

ترتیب این خط لوله «جایگذاری، سپس مقیاس‌بندی» است؛ پارامترهای هر مرحله از خروجی مرحله قبل روی آموزش یاد گرفته می‌شوند. اگر فرد تازه را در میانگین و واریانس وارد کنیم، نمایش آموزش به پرسش وابسته می‌شود. در اعتبارسنجی چندبخشی، تمام این مراحل باید برای هر بخش از نو فقط با آموزش همان دور برارز شوند.

در تبدیل کمینه-بیشینه، آموزش کمینه ۲۰ و بیشینه ۶۰ دارد؛ برای فرد تازه مقدار $(80-20)/(60-20) = 1.5$ به دست می‌آید. پس بازه صفر تا یک، ویژگی داده آموزش است و تضمینی درباره هر داده آینده نیست. بریدن

مقدار در یک، تصمیم جداگانه‌ای است که اطلاعات فاصله از محدوده آموزش را از بین می‌برد. برای ستون ثابت هم مخرج صفر است و باید قرارداد، مثلاً تبدیل به صفر و ثبت ثابت بودن ستون، روشن باشد.

چرا کاهش بعد همیشه بهبود پیش‌بینی نیست؟

PCA جهت دارای بیشترین واریانس ورودی را پیدا می‌کند. فرض کنید یک ویژگی با پراکندگی زیاد تقریباً مستقل از هدف و ویژگی کوچک دیگری تعیین‌کننده هدف باشد. نگه داشتن فقط مؤلفه پرواریانس ممکن است اطلاعات پیش‌بینی را حذف کند. بنابراین تعداد مؤلفه‌ها را همراه با عملکرد خط لوله روی اعتبارسنجی بررسی می‌کنیم. واریانس توضیح داده‌شده، معیار بازسازی ورودی است؛ صحت طبقه‌بندی یا خطای هدف را مستقیماً اندازه نمی‌گیرد.

۱۳.۲ خودآزمایی و تمرین

۱. تبدیل آموزش را روی مشاهده تازه اعمال کنید

در مثال سن‌ها، مقدار استاندارد شده ۸۰ و مقدار کمینه-بیشینه آن چیست؟ آیا دومی خطاست؟
پاسخ و توضیح: به ترتیب تقریباً ۲.۸۲۸۴ و ۱.۵ است. دومی خطای محاسباتی نیست؛ مشاهده خارج از دامنه آموزش است. برای محاسبه آن نباید بیشینه را با استفاده از مشاهده تازه دوباره برازش کرد.

۲. استاندارد با نرمال یکسان است؟

آیا استانداردسازی یک توزیع دوقله‌ای را نرمال می‌کند؟

پاسخ و توضیح: خیر؛ فقط مکان و مقیاس را تغییر می‌دهد. شکل چندقله‌ای باقی می‌ماند. میانگین صفر و واریانس یک، شرط کافی برای نرمال بودن نیست.

۳. چرا واژگان و IDF هم از آموزش‌اند؟

اگر پیش از تقسیم داده واژگان و وزن‌های TF-IDF را بسازیم، چه اطلاعاتی نشت می‌کند؟

پاسخ و توضیح: حضور و فراوانی سندی واژه‌های آزمون در انتخاب نمایش اثر می‌گذارد. واژگان و IDF باید از اسناد آموزش یاد گرفته شوند؛ اسناد تازه با همان نگاشت تبدیل می‌شوند و واژه ندیده قرارداد مشخصی دارد.

۴. همبستگی و استقلال را جدا کنید

اگر X حول صفر متقارن و $Y = X^2$ باشد، آیا همبستگی صفر می‌تواند با وابستگی همراه باشد؟
پاسخ و توضیح: بله؛ با وجود وابستگی تابعی کامل، تحت وجود گشتاورهای لازم $E[X^3] = 0$ و کوواریانس صفر می‌شود. همبستگی خطی، همه انواع وابستگی را نمی‌سنجد.

۵. نمونه متوازن یا نماینده؟

در جامعه‌ای با یک درصد مثبت، برای آموزش تعداد دو طبقه را برابر می‌کنیم. آیا آزمون را هم باید برابر کنیم؟

پاسخ و توضیح: اگر هدف، عملکرد در همان جامعه است، آزمون باید نماینده نرخ واقعی باشد. نمونه‌سازی آموزش تصمیمی جداست؛ متوازن کردن آزمون، درستی هشدار و برآورد فایده عملی را عوض می‌کند.

۶. PCA و پوشش هندسی را تفسیر کنید

آیا نکه داشتن ۹۵ درصد واریانس، حفظ ۹۵ درصد اطلاعات هدف را تضمین می‌کند؟ آیا Kennard-Stone الزاماً نسبت گروه‌ها را حفظ می‌کند؟

پاسخ و توضیح: هیچ کدام تضمین نمی‌شود. PCA معمولی برچسب را نمی‌بیند و تنوع هندسی هم با نمایندگی توزیع یکی نیست؛ نقاط نادر یا پرت می‌توانند بیش از سهمشان انتخاب شوند. هر دو انتخاب را نسبت به هدف نهایی ارزیابی کنید.

فصل ۳

طبقه‌بندی؛ درخت، تجمیع و ارزیابی

منبع مطالعه: اسلاید طبقه‌بندی ۱، PDF ۳ تا ۶۰؛ طبقه‌بندی ۲، PDF ۴۲ تا ۶۱؛ کتاب فصل ۳، صفحات چاپی ۱۳۴ تا ۱۹۳، و بخش‌های ۶، ۱۰ و ۱۱، ۰۶.

۱.۳ از مثال‌های برچسب‌دار تا تصمیم برای داده تازه

در طبقه‌بندی (Classification) داده آموزشی شامل زوج‌های (x_i, y_i) است؛ x_i بردار ویژگی‌ها و y_i برچسب گسسته است. الگوریتم باید تابعی مانند $f(x)$ بسازد تا برای یک مشاهده تازه برچسب پیش‌بینی کند. تفاوت با رگرسیون در نوع خروجی است، نه در عددی بودن ورودی: پیش‌بینی قیمت خانه رگرسیون است و پیش‌بینی «پرخطر/کم‌خطر» بودن وام طبقه‌بندی. طبقه‌بندی دوکلاسه می‌تواند «تقلب/سالم» و چندکلاسه می‌تواند نوع گیاه باشد. نمونه‌های آموزشی هرگز تعریف کامل مفهوم نیستند؛ مدلی که فقط درخت‌های سبز و بزرگ دیده ممکن است گیاه بزرگ گلدانی را هم درخت تشخیص دهد.

گردش کار صحیح از تعریف مسئله و واحد مشاهده شروع می‌شود؛ سپس داده برچسب‌دار به بخش‌های مناسب جدا می‌شود، پیش‌پردازش فقط از آموزش آموخته می‌شود، مدل آموزش می‌بیند، تنظیمات با اعتبارسنجی انتخاب می‌شوند و آزمون مستقل کیفیت تعمیم را می‌سنجد. هدف صرفاً حفظ داده‌های گذشته نیست. «یادگیری مشتاق» (Eager) پیش از دریافت پرسش مدل صریح می‌سازد؛ درخت تصمیم نمونه روشن آن است. روش تنبل (Lazy) مانند kNN بیشتر نمونه‌ها را نگه می‌دارد و بخش مهم محاسبات را به زمان پیش‌بینی موکول می‌کند. تنبل بودن به معنی بی‌هزینه بودن کل سامانه نیست و مشتاق بودن نیز تفسیرپذیری را تضمین نمی‌کند؛ شبکه عصبی مشتاق است ولی معمولاً توضیح انسانی ساده ندارد.

۲.۳ درخت تصمیم: پرسش‌های ساده، مرزهای مستطیلی

درخت تصمیم (Decision Tree) زنجیره‌ای از آزمون‌هاست. ریشه نخستین پرسش است، هر گره داخلی آزمون ویژگی، هر یال یک پاسخ و هر برگ تصمیم نهایی است. پیش‌بینی با عبور از ریشه به برگ انجام می‌شود. در

مثال اسلاید، ابتدا وضعیت بازپرداخت مالیات پرسیده می‌شود؛ سپس وضعیت تأهل و در نهایت درآمد مشمول مالیات. برای رکورد «Refund=No، متأهل، درآمد ۸۰ هزار» مسیر به برگ Cheat=No می‌رسد و نیازی نیست همه ویژگی‌ها در هر پیش‌بینی بررسی شوند.

اگر هر گره فقط یک ویژگی را با آستانه مقایسه کند، مرزها موازی محورند؛ $x < 0.43$ یک خط عمودی و $y < 0.33$ خط افقی می‌سازد. ترکیب تعداد زیادی چنین مستطیل‌هایی می‌تواند مرز خمیده را تقریب بزند، اما ممکن است درخت بزرگ شود. این محدودیت مربوط به درخت تک‌ویژگی متداول است؛ همه انواع درخت در جهان الزاماً چنین نیستند (کتاب در پایان ۶۰۳.۳ درخت مایل را نیز نام می‌برد).

الگوریتم هانت (Hunt) تقسیم بازگشتی حریصانه دارد: اگر همه برجسب‌ها یکی باشند برگ می‌سازیم؛ وگرنه آزمون‌های ممکن را مقایسه می‌کنیم، آزمون دارای بیشترین کاهش ناخالصی را انتخاب و داده‌ها را میان فرزندان تقسیم می‌کنیم؛ همین کار برای هر فرزند تکرار می‌شود. اگر ویژگی‌ها یکسان ولی برجسب‌ها متناقض باشند تقسیم ممکن نیست و برگ با رأی اکثریت برجسب می‌گیرد. اگر یک شاخه هیچ مشاهده‌ای ندارد می‌توان اکثریت والد را به کار برد. این روش بهینه سراسری را تضمین نمی‌کند؛ بهترین پرسش محلی الزاماً بهترین درخت نهایی نیست.

برای ویژگی اسمی با k مقدار، تقسیم چندشاخه یک شاخه برای هر مقدار می‌سازد؛ تقسیم دودویی همه مقادیر را در دو زیرمجموعه قرار می‌دهد و تعداد تقسیم‌های غیرتکراری غیرتهی $2^{k-1} - 1$ است. برای ویژگی ترتیبی باید ترتیب حفظ شود؛ {کوچک، متوسط، بزرگ} معتبر است، ولی {کوچک، بزرگ} متوسط ترتیب را به هم می‌زند. برای ویژگی پیوسته معمولاً دو شاخه $A \leq v$ و $A > v$ کافی است؛ مقدار دقیق برابری باید صریح باشد تا هیچ رکوردی گم یا دوبار شمرده نشود.

۳.۳ ناخالصی جینی و محاسبه یک تقسیم

اگر $p_j(t)$ سهم طبقه j در گره t باشد، ناخالصی جینی چنین تعریف می‌شود:

$$G(t) = 1 - \sum_{j=1}^c p_j(t)^2.$$

گره خالص جینی صفر دارد؛ بیشینه برای c طبقه هم‌احتمال $1 - 1/c$ است. در مسئله دوطبقه‌ای $G = 2p(1-p)$ و بیشینه ۰.۵ است. برای شمارش‌های (۱، ۵) مقدار $10/36 \approx 0.2778$ و برای (۲، ۴) مقدار $16/36 \approx 0.4444$ به دست می‌آید. این معیار با ضریب جینی نابرابری درآمد یک تعریف ندارد. پس از شکستن، اندازه فرزندان در وزندهی ضروری است:

$$G_{\text{split}} = \sum_{r=1}^m \frac{n_r}{n} G(t_r), \quad \Delta G = G(t) - G_{\text{split}}.$$

برای والد (۶، ۶) و فرزندان (۵، ۲) و (۱، ۴)، جینی والد ۰.۵، جینی فرزندان ۲۰/۴۹ و ۸/۲۵ و جینی وزنی $13/35 \approx 0.37143$ است. پس سود خلوص $9/70 \approx 0.12857$ می‌شود. استفاده از میانگین ساده دو فرزند، سهم جمعیت هر شاخه را نادیده می‌گیرد.

کتاب معیار آنتروپی و سود اطلاعاتی را نیز بسط می‌دهد:

$$H(t) = - \sum_j p_j(t) \log_2 p_j(t), \quad 0 \log 0 = 0.$$

کاهش آنتروپی، سود اطلاعاتی و کاهش جینی، سود خلوص است. یک شناسه یکتا می‌تواند همه برگ‌های آموزش را خالص کند، ولی به‌تنهایی تعمیم را بهتر نمی‌کند. نسبت سود، سود اطلاعاتی را نسبت به آنتروپی اندازه شاخه‌ها می‌سنجد و یکی از راه‌های کنترل تمایل به تقسیم‌های پرشاخه است.

۴.۳ پیدا کردن آستانه ویرگی پیوسته

داده‌ها را بر حسب ویرگی مرتب می‌کنیم، شمار کلاس‌های چپ و راست را نگه می‌داریم، و هنگام عبور آستانه فقط شمارش رکورد جابه‌جاشده را به‌روز می‌کنیم. مرتب‌سازی $O(n \log n)$ و پیمایش پس از آن $O(n)$ است؛ اسکن مجدد کل داده برای هر آستانه $O(n^2)$ می‌شود. نقاط میانی میان مقادیر متمایز کافی‌اند و مقادیر تکراری باید به صورت گروهی جابه‌جا شوند. در مثال درآمدهای ۶۰،۷۰،۷۵،۸۵،۹۰،۹۵،۱۰۰،۱۲۰،۱۲۵،۲۲۰ با برچسب‌های No, No, No, Yes, Yes, Yes, No, No, No, No حذف نامزدهای بین برچسب‌های یکسان در حالت استاندارد بدون محدودیت اضافه محاسبات را کم می‌کند؛ با وزن‌ها یا قیود خاص باید احتیاط کرد.

ابهام مهم مثال هانت اسلاید ۱۲: گفته شده Refund خالص‌ترین تقسیم اولیه است، اما در جدول نامزدهای پیوسته خود اسلاید ۲۸، بهترین آستانه درآمد $G_{\text{split}} = 0.300$ دارد و از Refund با ۰.۳۴۳ بهتر است. بنابراین روایت هانت مثال روند ساخت است، نه اثبات اینکه آن ریشه در میان تمام آستانه‌ها بهینه است. این دو مثال، هدف و مجموعه نامزدهای یکسانی را اثبات نمی‌کنند.

۵.۳ بیش‌برازش، سادگی و تجزیه مدل‌ها

دو درخت در مثال اعتبار مالی همه پنج رکورد آموزش را درست پیش‌بینی می‌کنند. اولی می‌پرسد بدهی از ۵۰۰ بیشتر است یا نه؛ دومی فقط نام John و Alice را حفظ می‌کند. عملکرد آموزش مساوی است ولی شواهدی نداریم که حفظ نام‌ها برای مشتری جدید مفید باشد. اصل تیغ اُکام می‌گوید میان توضیح‌هایی با کیفیت مشابه، توضیح ساده‌تر را ترجیح دهیم؛ این اصل یک راهنمای انتخاب است و تضمین همیشگی برتری هر مدل کوچک نیست.

کم‌برازش یعنی مدل حتی رابطه اصلی را نمی‌گیرد؛ درخت تک‌برگ برای مسئله پیچیده مثال آن است. با افزایش ظرفیت ابتدا خطای آموزش و اعتبارسنجی هر دو کم می‌شوند؛ پس از نقطه‌ای آموزش باز هم بهتر ولی اعتبارسنجی بدتر می‌شود. بیش‌برازش یعنی یادگیری جزئیات تصادفی نمونه به جای ساختار قابل تعمیم. نویز، نمونه کم، ویرگی‌های نامرتب زیاد و ظرفیت بالا زمینه آن را فراهم می‌کنند. عمق زیاد علامت قطعی نیست؛ عمق را نسبت به مقدار داده و پیچیدگی واقعی مسئله باید سنجید. در نمودار آموزشی می‌توان خطای آزمون را برای تشریح پدیده نشان داد، اما تنظیم واقعی مدل باید با اعتبارسنجی انجام شود.

سه راه مورد تأکید اسلاید عبارت‌اند از داده آموزشی بیشتر، پیش هرس و تجزیه. داده بیشتر وقتی مفید است که نماینده جمعیت هدف باشد؛ کپی چندباره همان رکورد اطلاعات تازه نمی دهد. پیش هرس با max_depth ، حداقل اندازه برگ یا حداقل کاهش ناخالصی رشد را محدود می کند. توقف بسیار زود نیز خطر کم‌برازش دارد: در XOR ممکن است تقسیم اول سود اندک بدهد ولی تقسیم بعدی لازم باشد. کتاب پس هرس را نیز اضافه می کند: نخست درخت بزرگ ساخته و سپس زیردرخت‌های بی‌فایده با برگ یا شاخه ساده‌تر جایگزین می شوند؛ این بسط از کتاب است نه موضوع مفصل اسلاید.

تجزیه Ensemble رأی چند مدل پایه را ترکیب می کند. اگر همه یک اشتباه کنند، رأی گیری کمی ندارد. برای M رأی دهنده مستقل، با M فرد و احتمال خطای یکسان q ، احتمال خطای اکثریت برابر است با

$$\sum_{i=(M+1)/2}^M \binom{M}{i} q^i (1-q)^{M-i}.$$

این مدل ساده نقش تنوع خطاها را روشن می کند؛ در کاربرد، استقلال کامل معمول نیست و این فرمول تضمین عملکرد جنگل تصادفی نیست.

در Bagging از داده آموزش نمونه‌گیری با جایگذاری می کنیم و برای هر نمونه یک درخت جدا می سازیم. یک نمونه بوت استرپ n تایی فقط حدود ۶۳٫۲٪ رکورد متمایز دارد، زیرا احتمال حداقل یک بار انتخاب رکورد $1 - (1 - 1/n)^n$ است. حدود ۳۶٫۸٪ برای همان درخت خارج از کیسه می مانند. جنگل تصادفی علاوه بر بوت استرپ، در هر گره تنها یک زیرمجموعه تصادفی ویژگی را نامزد تقسیم می کند؛ این کار غلبه همیشگی یک ویژگی قوی را می شکند و همبستگی درخت‌ها را کاهش می دهد. «استقلال درخت‌ها» در اسلاید توصیف ایده آل است؛ جنگل استقلال دقیق آماری تضمین نمی کند. p خیلی کوچک درخت را ضعیف می کند، p بزرگ جنگل را شبیه Bagging می کند. تعداد درخت‌ها و اندازه زیرمجموعه ویژگی‌ها ابرپارامترند. درخت تنها معمولاً تفسیرپذیرتر است؛ جنگل پیچیده‌تر و اغلب قوی‌تر، ولی برتری همیشگی ندارد.

بوستینگ مدل را افزایش می سازد؛ مدل تازه برای کاهش زیان مجموعه موجود آموزش می بیند:

$$F_t(\mathbf{x}) = F_{t-1}(\mathbf{x}) + \eta f_t(\mathbf{x}).$$

در معرفی XGBoost در اسلاید، تابع هدف شامل زیان و جریمه پیچیدگی درخت است:

$$\text{Obj} = L + \Omega, \quad \Omega(f) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^T w_j^2.$$

T تعداد برگ و w_j امتیاز برگ است. برای سه برگ با امتیازهای $-1, 0.1, 2$ ، جریمه $3\gamma + 2.505\lambda$ است؛ برای هر دو ضریب برابر یک، مقدار ۵.۵۰۵ می شود. ساختار گسسته درخت مانند وزن‌های پیوسته شبکه با یک گام معمول گرادیان حرکت نمی کند؛ گرادیان زیان برای ساخت درخت بعدی استفاده می شود. این با ساخت درخت‌های جدا در جنگل تصادفی متفاوت است.

۶.۳ ارزیابی: کدام اشتباه و با چه هزینه‌ای؟

ابتدا کلاس مثبت را روشن تعریف می کنیم؛ «مثبت» به معنی خوب یا مطلوب نیست، در تشخیص تقلب همان تقلب مثبت است. TP مثبت واقعی درست شناسایی شده، FN مثبت واقعی از دست رفته، FP منفی واقعی با هشدار

مثبت و TN منفی واقعی درست تشخیص داده شده است. در این جزوه سطر حقیقت و ستون پیش بینی است. حرف اول True/False درستی پیش بینی و حرف دوم Positive/Negative خروجی مدل را توصیف می کند. با قرارداد سطر حقیقت و ستون پیش بینی، $n = TP + TN + FP + FN$ است. صحت کلی و نرخ خطا چنین اند:

$$\text{Accuracy} = \frac{TP + TN}{n}, \quad \text{Error} = 1 - \text{Accuracy}.$$

در مثال $TP = 25, FN = 4, FP = 6, TN = 15$ ، صحت $40/50 = 0.8$ است. باید آن را با مینا سنجید: پیش بینی همیشگی طبقه اکثریت در همین داده $29/50 = 0.58$ صحت دارد. درصد بدون اندازه آزمون، خط مینا و هزینه خطا، قضاوت کاملی نیست.

در داده ای با ۹۹۹۰ منفی و ۱۰ مثبت، پیش بینی همیشگی منفی صحت ۹۹٫۹ درصد دارد ولی هیچ مثبت را نمی یابد. درستی هشدار و بازیابی دو سؤال متفاوت اند:

$$P = \text{Precision} = \frac{TP}{TP + FP}, \quad R = \text{Recall} = \frac{TP}{TP + FN}.$$

اولی سهم هشدارهای صحیح و دومی سهم مثبت های واقعی کشف شده است. در مثال پیشین $P = 25/31 \approx 0.80645$ و $R = 25/29 \approx 0.86207$ است.

میانگین هارمونیک درستی و بازیابی، F_1 ، به کم بودن یکی از دو معیار حساس است:

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN}.$$

در مثال بالا $F_1 = 50/60 \approx 0.83333$ است. صورت عمومی صحیح چنین است:

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R} = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}.$$

$\beta > 1$ بازیابی را مهم تر می کند. برای $TP = 1, FP = 0, FN = 99$ ، درستی کامل ولی بازیابی 0.01 و $F_1 = 2/101$ است. اگر خرج یک معیار صفر باشد باید قرارداد گزارش روشن باشد؛ صفر قراردادی و مقدار تعریف نشده از نظر مفهومی یکسان نیستند.

هزینه تصمیم از شمار درست و غلط فراتر می رود. در اسلاید ۵۸، هزینه های TP, FN, FP, TN به ترتیب ۰، ۱، ۱۰۰، -۱ هستند. هزینه منفی پاداش است. مدل اول شمارش های (162, 38, 160, 240) و مدل دوم (155, 45, 5, 395) را با همین ترتیب دارد:

$$\text{Accuracy}_1 = \frac{402}{600} = 0.67,$$

$$\text{Cost}_1 = -162 + 3800 + 160 = 3798,$$

$$\text{Accuracy}_2 = \frac{550}{600} \approx 0.9167,$$

$$\text{Cost}_2 = -155 + 4500 + 5 = 4350.$$

بنابراین با این هزینه ها مدل اول مطلوب تر است، با اینکه صحت پایین تری دارد. هزینه کل $\sum_{a,p} n_{a,p} c_{a,p}$ است؛ تقسیم آن بر تعداد مشاهده ها، هزینه متوسط را می دهد. این دو مدل روی تعداد مثبت و منفی یکسان مقایسه شده اند.

۷.۳ امتیاز، آستانه و ROC

بسیاری مدل‌ها به جای برچسب فقط، امتیاز مثبت بودن $s(x)$ می‌دهند. قانون $s(x) \geq \tau$ برچسب مثبت می‌سازد؛ با کاهش τ افراد بیشتری مثبت اعلام می‌شوند و Recall هرگز کاهش نمی‌یابد، ولی Precision الزاماً یکنواخت تغییر نمی‌کند. عبارت اسلاید ۵۶ «آستانه بالاتر همیشه Precision بیشتر» یک روند معمول است، نه قضیه برای هر نمونه. آستانه را روی اعتبارسنجی متناسب با هدف یا هزینه انتخاب می‌کنیم، نه با آزمون نهایی.

در نمودار ROC، نرخ مثبت واقعی در برابر نرخ مثبت کاذب رسم می‌شود:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

از آستانه بالاتر از همه امتیازها در (۰،۰) شروع و امتیازها را نزولی بررسی می‌کنیم. عبور از یک مثبت واقعی گام عمودی $1/n_{\text{positive}}$ و عبور از یک منفی واقعی گام افقی $1/n_{\text{negative}}$ می‌دهد. اسلاید ۵۹ «پیش‌بینی درست/نادرست» می‌گوید؛ دقیق‌تر و صحیح‌تر «برچسب واقعی مثبت/منفی هنگام جاروب امتیاز مثبت» است. اگر امتیازها مساوی‌اند باید همزمان پردازش شوند؛ ترتیب دلخواه میان هم‌امتیازها AUC را تحریف می‌کند. نقطه ایده‌آل (۰،۱)، پیش‌بینی همگی مثبت (۱،۱) و خط انتظار تصادفی قطر است.

مثال ۱۰.۳ | محاسبه ROC با چهار مشاهده

امتیازهای نزولی 0.9, 0.8, 0.7, 0.4 و برچسب‌های +, -, +, - را در نظر بگیرید. نقاط ROC به ترتیب چنین‌اند:

$$(0, 0), (0, 0.5), (0.5, 0.5), (0.5, 1), (1, 1).$$

مساحت دوزنقه‌ها 0.75 است؛ از چهار جفت مثبت-منفی، سه جفت ترتیب درست دارند. در آستانه 0.7، شمارش‌ها $TP = 2, FP = 1, FN = 0, TN = 1$ و درستی 2/3 و بازیابی یک‌اند. AUC کیفیت رتبه‌بندی را خلاصه می‌کند، نه کالیبراسیون یا احتمال صحیح بودن تک پیش‌بینی‌ها. برای عدم توازن شدید، منحنی درستی-بازیابی نیز اهمیت دارد.

۸.۳ ارزیابی معتبر، انتخاب ابرپارامتر و نشت اطلاعات

معیار ارزیابی «چه عددی بسنجیم» را تعیین می‌کند، روش ارزیابی «این عدد را از کدام داده و چه آزمایشی به دست آوریم» را. Holdout یک بخش مستقل را برای آزمون نگار می‌گذارد؛ نسبت ۸۰/۲۰ یا ۲/۳-۱/۳ قانون همگانی نیست. آموزش کوچک مدل را ضعیف می‌کند و آزمون کوچک برآورد را پرنوسان. در نامتوازن، نمونه‌گیری طبقه‌بندی شده (Stratified) نسبت کلاس‌ها را تقریباً حفظ می‌کند، نه اینکه تعداد مثبت و منفی را مساوی کند. برای داده‌های زمانی یا چند رکورد از یک شخص، جداسازی زمانی/گروهی بر تصادفی‌سازی ساده مقدم است؛ این قید تکمیلی برای جلوگیری از نشت است.

در Holdout Repeated هر بار تقسیم تصادفی تازه و امتیازها میانگین گرفته می‌شوند؛ آزمون‌ها ممکن است همپوشانی داشته باشند و برخی نمونه‌ها هرگز آزمون نشوند. در k-fold داده به k بخش تقسیم می‌شود و هر بار یکی آزمون و بقیه آموزش است؛ هر مشاهده یک بار آزمون و $k-1$ بار آموزش می‌شود. با اندازه‌های برابر، میانگین خطای fold ها مناسب است؛ برای اندازه‌های متفاوت، خطای شمارشی کل تقسیم بر تعداد کل دقیق‌تر است. اگر از

F۱ استفاده می‌کنیم، میانگین F۱ fold ها و F۱ حاصل از تجميع پيش بيني‌هاي out-of-fold لزوماً يکي نيستند؛
روش گزارش بايد صريح باشد.

پارامتر از داده یاد گرفته می‌شود: وزن شبکه، میانگین بیز و محل تقسیم درخت. ابرپارامتر قواعد آموزش یا معماری را تعیین می‌کند: عمق، η, γ, C ، تعداد لایه. انتخاب ابرپارامتر نیز نوعی یادگیری از داده اعتبارسنجی است و نباید از آزمون نهایی انجام شود. جست و جوی شبکه‌ای تمام حاصل ضرب دگارتی مقادیر را می‌آزماید؛ جست و جوی تصادفی از توزیع‌های مشخص نمونه می‌گیرد و وقتی فقط چند بعد حساس اند اغلب پوشش مؤثرتری می‌دهد. برتری همیشگی تصادفی بر شبکه‌ای ادعا نمی‌شود. بهینه‌سازی تکاملی ترکیب‌های مناسب را نگه و تغییر می‌دهد؛ بهینه‌سازی بیزی یک مدل جانشین از امتیاز بر حسب ابرپارامتر می‌سازد و با در نظر گرفتن عدم قطعیت نامزد بعدی را پیشنهاد می‌کند. واژه «بیزی» اینجا ارتباطی با انتخاب طبقه‌بند Bayes Naïve ندارد.

اگر اعتبارسنجی ساده ناپایدار باشد، CV داخلی برای انتخاب تنظیم به کار می‌رود. اگر کل داده باید هم در انتخاب و هم در برآورد تعمیم به صورت منظم استفاده شود، CV تودرتو (Nested) لازم است: در هر fold بیرونی، آزمون بیرونی کاملاً کنار می‌ماند؛ فقط دادهٔ آموزش بیرونی به CV داخلی می‌رود؛ تنظیم برنده روی کل آموزش بیرونی آموزش مجدد می‌بیند و روی آزمون بیرونی سنجیده می‌شود. میانگین امتیازهای بیرونی کیفیت «رویهٔ انتخاب و آموزش» را برآورد می‌کند، نه الزاماً یک مدل خاص با تنظیم ثابت. پس از گزارش ارزیابی می‌توان برای مدل قابل استفاده تنظیم را روی کل داده انتخاب و مدل نهایی را آموزش داد؛ این مرحله آزمون تازه‌ای تولید نمی‌کند.

برای ۱۰۰ تنظیم و ۵ بخش، تعداد آموزش‌ها در سه شیوه چنین است:

CV : $5 \times 100 + 1 = 501$,

حالت آخر ۵ بخش بیرونی و ۵ بخش داخلی دارد. این شمارش، برازش مجدد هر بخش بیرونی را دارد و آموزش مدل نهایی روی تمام داده را شامل نمی‌شود. تعداد بیشتر بدون طراحی مستقل می‌تواند فقط بیش‌برازش انتخاب را بیشتر کند.

دقیق نیست. در داده نامتوازن باز نمونه‌گیری فقط آموزش است؛ نسبت آزمون باید مسئله واقعی را باز نمای کند. لزوم Nested به وابستگی انتخاب و ارزیابی مربوط است؛ «تکرارپذیری فقط با یک split ممکن است» درست نیست و نسخه/داده/seed را برای CV هم می‌توان ثبت کرد.

سه سطح داده در اعتبارسنجی تودرتو

در یک دور بیرونی، آزمون بیرونی را کنار می‌گذاریم. برای هر تنظیم نامزد، چند دور داخلی روی داده باقی‌مانده اجرا می‌شود؛ هر دور داخلی آموزش، جایگذاری، مقیاس‌بندی و مدل خاص خود را دارد. پس از انتخاب تنظیم، مدل از ابتدا با کل داده غیرآزمون همان دور بیرونی برازش و فقط یک بار روی آزمون بیرونی سنجیده می‌شود. آزمون بیرونی نباید برای توقف زودهنگام یا انتخاب ویژگی استفاده شود. با ۵ دور بیرونی، ۴ دور داخلی و ۱۰ تنظیم:

$$N_{\text{fit}} = 5(4 \times 10 + 1) = 205.$$

این عدد شامل بازبرازش هر دور بیرونی است. میانگین امتیازهای بیرونی، کیفیت رویه انتخاب را توصیف می‌کند؛ اختلاف جزئی امتیازها بدون توجه به تغییرپذیری داده، دلیل قطعی برتری نیست.

۹.۳ خودآزمایی و تمرین

۱. جینی یک تقسیم را محاسبه کنید

والد با شمارش (5, 5) به دو فرزند (4, 1) و (1, 4) تقسیم می‌شود. جینی وزنی و سود چقدر است؟ پاسخ و توضیح: جینی والد 0.5 و جینی هر فرزند $0.32 = (1/5)^2 - (4/5)^2 = 1 - 1$ است. اندازه‌ها برابرند، پس جینی وزنی 0.32 و سود 0.18 است. با اندازه‌های نابرابر، میانگین ساده جای وزن‌دهی را نمی‌گیرد.

۲. صحت زیاد را نقد کنید

در ۱۰۰۰ مشاهده با ۱۰ مثبت، مدلی همیشه منفی می‌گوید. صحت و بازیابی چیست؟ پاسخ و توضیح: صحت $0.99 = 990/1000$ ولی بازیابی صفر است. چون هیچ هشدار مثبتی صادر نشده، درستی هشدار مخرج صفر دارد و تعریف نشده است؛ نباید آن را با درستی کامل اشتباه گرفت.

۳. آستانه و AUC را جدا کنید

با ثابت بودن امتیازها، کاهش آستانه چه اثری بر بازیابی و AUC دارد؟ پاسخ و توضیح: بازیابی کاهش نمی‌یابد، زیرا هیچ مثبت قبلاً کشف شده‌ای از مجموعه هشدارها حذف نمی‌شود. AUC از تمام آستانه‌ها و ترتیب امتیازها حاصل می‌شود و با حرکت نقطه جاری تغییر نمی‌کند. درستی هشدار الزاماً یکنواخت نیست.

۴. تجزیه چه زمانی کمکی نمی‌کند؟

اگر همه درخت‌ها روی دقیقاً همان مشاهده‌ها خطا کنند، رأی اکثریت چه کمکی می‌کند؟

پاسخ و توضیح: برای آن خطاهای مشترک کمکی نمی‌کند. تنوع خطاها مهم است؛ جنگل تصادفی با بوت‌استرپ و نامزدی تصادفی ویژگی‌ها همبستگی را کاهش می‌دهد، اما استقلال کامل یا برتری همیشگی را تضمین نمی‌کند.

۵. داده‌های هر دور را مشخص کنید

در طرح ۵ بیرونی، ۴ داخلی و ۱۰ تنظیم، چند برازش داریم؟ آیا آزمون بیرونی برای توقف زودهنگام مجاز است؟

پاسخ و توضیح: با بازبرازش هر دور بیرونی، ۲۰۵ برازش داریم. استفاده از آزمون بیرونی برای توقف یا تنظیم، استقلال آن را از بین می‌برد؛ این تصمیم‌ها باید در داده آموزش بیرونی انجام شوند.

۶. خطای آموزش کافی است؟

خطای آموزش با افزایش عمق از ۱۰ درصد به صفر می‌رسد، ولی اعتبارسنجی از ۱۵ به ۲۵ درصد می‌رود. چه چیزی را تغییر می‌دهید؟

پاسخ و توضیح: این الگو با بیش‌برازش سازگار است؛ عمق یا حداقل اندازه برگ را با اعتبارسنجی انتخاب می‌کنیم و نشد و نمایندگی تقسیم را بررسی می‌کنیم. آموزش را برای بهتر شدن امتیاز آزمون نهایی تکرار نمی‌کنیم.

۷. واحد درآمد چه اثری بر درخت دارد؟

چرا تبدیل ریال به میلیون ریال معمولاً تقسیم‌های درخت تک‌ویژگی را عوض نمی‌کند؟

پاسخ و توضیح: تبدیل مثبت یکنوا ترتیب مقادیر را حفظ می‌کند و آستانه‌ها با آن تبدیل می‌شوند؛ اما فاصله نسبی ویژگی‌ها در kNN می‌تواند عوض شود.

۸. رأی و انتخاب تصادفی ویژگی

چهار رأی مثبت و سه منفی چه خروجی‌ای می‌دهند؟ آیا سهم رأی، احتمال کالیبره است؟ آیا زیرمجموعه ویژگی‌های نامزد در کل یک درخت ثابت است؟

پاسخ و توضیح: خروجی مثبت است؛ سهم ۴/۷ تضمین کالیبراسیون ندارد. در جنگل متعارف، ویژگی‌های نامزد در هر گره دوباره نمونه‌گیری می‌شوند، نه فقط یک‌بار برای کل درخت.

۹. محاسبه معیارها از چهار شمارش

شمارش‌های $TP = 8, FN = 2, FP = 4, TN = 86$ را در نظر بگیرید.

پاسخ و توضیح:

$$\text{Accuracy} = 0.94,$$

$$\text{Precision} = \frac{2}{3},$$

$$\text{Recall} = 0.8,$$

$$F_1 = \frac{8}{11} \approx 0.7273,$$

$$\text{FPR} = \frac{4}{90} \approx 0.04444.$$

آیا مدل با صحت ۹۹ درصد که هیچ مثبت غمی گیرد حتماً بهتر است؟ به معیار و هزینه بستگی دارد؛ برای کشف مثبت قابل دفاع نیست.

۱۰. نشت و انتخاب دوباره را تشخیص دهید

چرا محاسبه مقیاس پیش از CV نشت است؟ چرا انتخاب بهترین دور بیرونی مجاز نیست؟ در طرح ۴ بیرونی، ۳ داخلی و ۲۰ تنظیم چند برازش لازم است؟

پاسخ و توضیح: آمار بخش کارگذاشته شده وارد آموزش شده است؛ انتخاب بهترین دور بیرونی نیز انتخاب بر اساس آزمون است. با بازبرازش هر دور، تعداد $4(3 \times 20 + 1) = 244$ است.

فصل ۴

بیز، حاشیه و شبکه‌های عصبی

منبع مطالعه: اسلاید طبقه‌بندی ۲، PDF ۳ تا ۴۱؛ کتاب بخش‌های ۶،۴، ۶،۷، ۶،۸ و ۶،۹، صفحات چاپی ۴۱۴ تا ۴۹۷.

۱.۴ قاعده بیز و استقلال شرطی

یک طبقه‌بند احتمالی می‌پرسد پس از دیدن ویژگی‌ها، کدام طبقه محتمل‌تر است؟ اگر C متغیر طبقه و x ویژگی‌های مشاهده‌شده باشد، قاعده بیز می‌دهد:

$$P(C = c | \mathbf{x}) = \frac{P(\mathbf{x} | C = c)P(C = c)}{P(\mathbf{x})}.$$

پیشین $P(C = c)$ فراوانی یا باور پیش از دیدن ویژگی‌ها، درست‌نمایی $P(\mathbf{x} | C = c)$ سازگاری مشاهده با طبقه و پسین، احتمال طبقه پس از مشاهده است. مخرج میان طبقه‌ها مشترک است؛ برای انتخاب طبقه، بیشینه کردن صورت کافی است، اما برای گزارش پسین باید امتیازها را نرمال کنیم.

اگر d ویژگی v داشته باشیم، جدول مشترک تا v^d ترکیب دارد. داده محدود برای برآورد همه خانه‌ها کافی نیست. بیز ساده فرض می‌کند ویژگی‌ها به شرط طبقه مستقل‌اند:

$$\begin{aligned} P(\mathbf{x} | C = c) &= \prod_{j=1}^d P(x_j | C = c), \\ q_c &= P(C = c) \prod_j P(x_j | C = c), \\ P(C = c | \mathbf{x}) &= \frac{q_c}{\sum_r q_r}. \end{aligned}$$

این فرض با استقلال بدون شرط متفاوت است. دو ویژگی می‌توانند به علت وابستگی به یک عامل مشترک همبسته باشند و پس از شرط کردن بر آن، رابطه کمتری داشته باشند. بیز ساده تعداد پارامترها را کاهش می‌دهد؛ نقض فرض لزوماً برچسب را خراب نمی‌کند، ولی احتمال‌های مدل ممکن است بیش از حد مطمئن باشند. دوبار وارد کردن یک ویژگی، شواهد مستقلی تولید نمی‌کند.

مثال ۱۰۴ | مثال گلف و محاسبهٔ پسین

در ۱۴ روز مثال اسلاید، ۹ روز پاسخ مثبت و ۵ روز پاسخ منفی است. برای ویژگی‌های Sunny, Cool, High, True، شمارش‌های شرطی می‌دهند:

$$q_{\text{yes}} = \frac{9 \ 2 \ 3 \ 3 \ 3}{14 \ 9 \ 9 \ 9 \ 9} = \frac{1}{189}, \quad q_{\text{no}} = \frac{5 \ 3 \ 1 \ 4 \ 3}{14 \ 5 \ 5 \ 5 \ 5} = \frac{18}{875}.$$

پس از نرمال‌سازی، احتمال مثبت تقریباً 0.20458 و احتمال منفی 0.79542 است. حاصل ضرب خام 1/189 خود پسین نیست. اگر همهٔ ویژگی‌ها نامعلوم باشند، مدل مبتنی بر حذف عامل‌ها تنها از پیشین‌ها استفاده می‌کند.

۲.۴ هموارسازی، ویژگی پیوسته و داده گمشده

اگر یک مقدار در یک طبقه مشاهده نشده باشد، تخمین فراوانی صفر است و کل حاصل ضرب صفر می‌شود. صفر بار مشاهده شدن در نمونه کوچک با ناممکن بودن برابر نیست. هموارسازی افزایشی برای یک ویژگی v حالتی چنین است:

$$P(X_j = a \mid C = c) = \frac{N_{j,a,c} + \alpha}{N_c + \alpha v}, \quad \alpha > 0.$$

برای لاپلاس $\alpha = 1$ است. اگر در ۱۰ رکورد یک طبقه، یک مقدار از ویژگی سه‌حالتی دیده نشده باشد، احتمال هموارشده 1/13 است. همهٔ خانه‌های همان ویژگی باید با یک قرارداد هموار شوند تا جمع احتمال‌ها یک بماند. هموار کردن فقط خانهٔ صفر، یک توزیع معتبر نمی‌سازد.

برای ویژگی پیوسته می‌توان گسسته‌سازی کرد یا یک مدل درون طبقه‌ای، مثلاً نرمال، برازش داد:

$$f_j(x \mid c) = \frac{1}{\sigma_{jc}\sqrt{2\pi}} \exp \left[-\frac{(x - \mu_{jc})^2}{2\sigma_{jc}^2} \right].$$

میانگین و واریانس باید از رکوردهای همان طبقه به دست آیند. چگالی در یک نقطه، احتمال نقطه‌ای نیست. برای نمونه $\mu = 73, \sigma = 6.2, x = 66$ چگالی حدود 0.034 است، ولی برای متغیر پیوسته احتمال دقیقاً برابر ۶۶ بودن صفر است. برآورد بیشینهٔ درست‌نمایی واریانس مخرج N_c و برآورد ناریب مخرج $N_c - 1$ دارد؛ قرارداد مثال و کد باید ثابت باشد. واریانس صفر به قاعدهٔ پایدار، مانند کف کوچک یا مدل جایگزین نیاز دارد. برای جلوگیری از زیرریز عددی، به جای ضرب تعداد زیادی احتمال کوچک، امتیاز لگاریتمی را جمع می‌کنیم:

$$\ell_c = \log P(C = c) + \sum_j \log P(x_j \mid C = c).$$

عامل ویژگی گمشده را می‌توان در پیش‌بینی حذف کرد؛ در مدل مستقل این با حاشیه‌گیری روی آن ویژگی سازگار است. اما اگر سازوکار فقدان اطلاعات مهمی دربارهٔ طبقه داشته باشد، حذف ساده آن را نادیده می‌گیرد. در آموزش نیز مخرج شمارش هر ویژگی باید با رکوردهای مشاهده‌شدهٔ همان ویژگی سازگار باشد.

آزمایشگاه پیشنهادی. در آزمایشگاه بیز ساده، شواهد هر ویژگی، اثر هموارسازی و حذف یک ویژگی گمشده را جداگانه ببینید؛ سپس جمع امتیازها را به پسین تبدیل کنید.

۳.۴ هندسه ماشین بردار پشتیبان

برای برچسب‌های $y_i \in \{-1, +1\}$ ، مرز خطی با $\mathbf{w}^T \mathbf{x} + b = 0$ تعریف می‌شود. چند مرز ممکن است آموزش را بی‌خطا جدا کنند؛ SVM مرزی می‌جوید که فاصله نزدیک‌ترین مشاهده‌ها تا آن زیاد باشد. فاصله هندسی و پهنای حاشیه در مقیاس متعارف برابرند با

$$d(\mathbf{x}, H) = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|_2}, \quad \text{margin} = \frac{2}{\|\mathbf{w}\|_2}.$$

صفحه‌های حاشیه امتیازهای $+1$ و -1 دارند. ضرب کردن هم‌زمان $\mathbf{w}, b$ در یک عدد مثبت، مرز را جابه‌جا نمی‌کند؛ بنابراین بزرگی امتیاز خام با فاصله هندسی یکی نیست. در حالت جدایی خطی کامل، مسئله حاشیه سخت چنین است:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|_2^2 \quad \text{to subject} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1.$$

برای نقاط یک‌بعدی منفی $-1, -2$ و مثبت $1, 2$ ، پاسخ $w = 1, b = 0$ است؛ نقاط -1 و 1 بردار پشتیبان و حاشیه کامل برابر 2 است. وقتی داده‌ها هم‌پوشان یا نویزی‌اند، قید سخت ممکن است شدنی نباشد.

۴.۴ حاشیه نرم، زیان لولایی و هسته

متغیر لغزش اجازه نقض حاشیه می‌دهد:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_i \xi_i, \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

اگر $0 < \xi_i < 1$ باشد، نقطه هنوز می‌تواند درست طبقه‌بندی شده باشد؛ نقض حاشیه با اشتباه برچسب یکسان نیست. برای یک نقطه مثبت با امتیاز 0.4 ، لغزش کینه 0.6 است. امتیاز منفی برای همین نقطه به معنی اشتباه در طبقه‌بندی است. صورت بدون قید از زیان لولایی استفاده می‌کند:

$$L_{\text{hinge}}(y, f) = \max(0, 1 - yf).$$

در این قرارداد، C بزرگ نقض حاشیه را پرهزینه‌تر می‌کند؛ پس نباید آن را مستقیماً «شدت بیشتر منظم‌سازی» نامید. اثر نهایی بر تعمیم با اعتبارسنجی سنجیده می‌شود.

هسته ضرب داخلی در یک فضای تبدیل‌شده را بدون ساخت صریح تمام مؤلفه‌ها محاسبه می‌کند:

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z}), \quad K_{\text{RBF}} = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|_2^2).$$

هسته چندجمله‌ای نمونه‌ای مانند $(\mathbf{x}^T \mathbf{z} + 1)^r$ دارد. در RBF، بزرگ شدن γ ناحیه اثر هر نقطه را محلی‌تر می‌کند. انتخاب C و γ همراه با مقیاس‌بندی داخل هر دور اعتبارسنجی انجام می‌شود. امتیاز تصمیم SVM خودبه‌خود احتمال کالیبره نیست. برای داده اسمی، رمزگذاری مناسب لازم است؛ عددی بودن ورودی پس از رمزگذاری با پیوسته بودن معنای اولیه آن تفاوت دارد.

آزمایشگاه پیشنهادی. آزمایشگاه حاشیه، جابه‌جایی خط و قیده‌های هندسی را بررسی می‌کند؛ حرکت دستی خط، آموزش یک حل‌کننده SVM نیست. مقدار حاشیه، تعداد خطاها و زیان لولایی از داده و خط واقعاً محاسبه می‌شوند.

۵.۴ پرسپترون و محدودیت یک مرز خطی

پرسپترون جمع وزن‌دار ورودی را با آستانه مقایسه می‌کند:

$$\hat{y} = \mathbf{1}\{\mathbf{w}^T \mathbf{x} + b > 0\}.$$

برای سه ورودی دودویی، وزن‌های 0.3 و آستانه 0.4، شرط «حداقل دو ورودی یک» را پیاده می‌کنند: یک ورودی یک، امتیاز 0.3 و دو ورودی یک، امتیاز 0.6 می‌دهد. این نمایش دستی تابع است؛ هنوز الگوریتم یادگیری اجرا نشده است.

در قرارداد برچسب صفر و یک، یک قاعده به‌روزرسانی متداول چنین است:

$$\mathbf{w} \leftarrow \mathbf{w} + \eta(y - \hat{y})\mathbf{x}, \quad b \leftarrow b + \eta(y - \hat{y}).$$

برای $\mathbf{x} = (1, 1)$, $y = 1$, $\mathbf{w} = (0.1, 0.1)$, $b = -0.5$, $\eta = 0.1$ پیش‌بینی صفر است و وزن تازه (0.2, 0.2) و عرض تازه -0.4 می‌شود. تکلیف برابری امتیاز با صفر باید روشن باشد. تضمین همگرایی استاندارد به جدایی خطی و فرض‌های متعارف وابسته است.

XOR با یک مرز خطی جدا نمی‌شود. دو واحد پنهان می‌توانند OR و AND را بسازند:

$$h_1 = \mathbf{1}\{x_1 + x_2 > 0.5\}, \quad h_2 = \mathbf{1}\{x_1 + x_2 > 1.5\}, \quad \hat{y} = \mathbf{1}\{h_1 - h_2 > 0.5\}.$$

برای چهار حالت ورودی، خروجی‌ها 0, 1, 1, 0 می‌شوند. تابع پله برای نمایش منطق مناسب است، اما فعال‌ساز متداول پس‌انتشار گرادینانی نیست.

۶.۴ شبکه چندلایه، زیان و پس‌انتشار

هر لایه از یک تبدیل آفین و یک فعال‌ساز تشکیل می‌شود:

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)} \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \quad \mathbf{a}^{(l)} = g(\mathbf{z}^{(l)}), \quad \mathbf{a}^{(0)} = \mathbf{x}.$$

بدون فعال‌ساز غیرخطی، ترکیب هر تعداد لایه خطی باز هم یک نگاشت خطی یا آفین است. سیگموئید و ReLU دو انتخاب آشنا هستند:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \text{ReLU}(z) = \max(0, z).$$

در سیگموئید، $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ در ناحیه اشباع کوچک است. ضرب این مشتق‌ها در زنجیره لایه‌ها می‌تواند گرادینان را کوچک کند. ReLU در نیمه مثبت مشتق یک دارد، ولی این به معنی حذف همیشگی تمام مشکلات گرادینان نیست.

برای خروجی دودویی سیگموئیدی، آنتروپی متقاطع یکی از زیان‌هاست:

$$L = -y \log a - (1 - y) \log(1 - a), \quad \frac{\partial L}{\partial z} = a - y.$$

پس‌انتشار با قاعده زنجیره‌ای گرادینان پارامترهای همه لایه‌ها را محاسبه می‌کند؛ گرادینان کاهشی از این گرادینان برای تغییر وزن استفاده می‌کند:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} L.$$

این دو مرحله مترادف نیستند. در شبکه، وزن‌ها میان همه نمونه‌ها مشترک‌اند؛ نباید اندیس نمونه را به معنی شبکه مستقل برای هر نمونه گرفت. مقدار اولیه، نرخ یادگیری، اندازه دسته، معماری و منظم‌سازی بر نتیجه اثر دارند. مسئله آموزش معمولاً غیرمحدب است؛ رسیدن به یک نقطه بهینه محلی تضمین بهترین جواب جهانی نیست. کتاب، حذف تصادفی واحدها در آموزش یا Dropout، پیش‌آموزش و تنظیم دقیق را نیز توضیح می‌دهد. حذف تصادفی، به معنی حذف ثابت یک زیرشبکه در زمان استفاده نیست. در رگرسیون، خروجی نهایی می‌تواند خطی و نامحدود باشد؛ در طبقه‌بندی باید خروجی و زیان با معنای برچسب سازگار باشند.

۷.۴ معرفی خانواده‌های عمیق و مرز دامنه درس

در معرفی اسلاید، CNN از الگوهای محلی، کانولوشن و اشتراک وزن استفاده می‌کند؛ هر کانولوشن الزاماً اندازه فضایی را کم نمی‌کند. مدل زبانی پیش‌آموخته از متن فراوان نمایش می‌آموزد و می‌تواند برای وظیفه‌ای خاص تنظیم شود. مدل مولد، خروجی تازه مانند متن یا تصویر تولید می‌کند. آموزش روی دستور و پاسخ، شکل استفاده از مدل را تغییر می‌دهد. یادگیری درون‌بافت از نمونه‌های موجود در ورودی با تنظیم دقیق که وزن‌ها را تغییر می‌دهد متفاوت است.

نام‌های BERT، T5 و خانواده‌های مولد در اسلاید مثال‌اند؛ این معرفی، رتبه‌بندی روز محصولات یا آموزش تفصیلی معماری Transformer نیست. برای این درس، مهم است بتوانید نوع ورودی، نوع خروجی، تابع زیان و نقش پیش‌آموزش را توضیح دهید و تشخیص دهید کدام پارامتر از داده آموخته و کدام تنظیم پیش از آموزش انتخاب می‌شود.

۸.۴ یک گام واقعی پس‌انتشار، با عدد و مشتق

برای روشن شدن تفاوت «محاسبه گرادیان» و «به‌روزرسانی»، یک شبکه با ورودی $x = 1$ ، وزن پنهان $w = 0.5$ ، بایاس پنهان صفر، وزن خروجی $v = 2$ و بایاس خروجی $b = -1$ در نظر بگیرید. هر دو فعال‌ساز سیگموئید و پاسخ مطلوب $y = 1$ است. گذر رو به جلو:

$$\begin{aligned} z_h &= wx = 0.5, & h &= \sigma(z_h) \approx 0.622459, \\ z_o &= vh + b \approx 0.244919, & a &= \sigma(z_o) \approx 0.560925. \end{aligned}$$

با آنتروپی متقاطع، خطای خروجی و گرادیان‌ها به کمک قاعده زنجیره‌ای به دست می‌آیند:

$$\begin{aligned} \delta_o &= a - y \approx -0.439075, \\ \frac{\partial L}{\partial v} &= \delta_o h \approx -0.273306, \\ \frac{\partial L}{\partial w} &= \delta_o v h(1 - h)x \approx -0.206368. \end{aligned}$$

برای نرخ یادگیری 0.1، وزن‌ها به $v \approx 2.027331$ و $w \approx 0.520637$ می‌رسند. گرادیان بایاس خروجی δ_o و گرادیان بایاس پنهان $\delta_o v h(1 - h)$ است. همه این گرادیان‌ها با وزن‌های پیش از به‌روزرسانی محاسبه می‌شوند؛ استفاده از v تازه هنگام محاسبه گرادیان w ، همان گام گرادیان کاهشی نیست. این مثال تک نمونه‌ای است؛ برای دسته، گرادیان نمونه‌ها با قرارداد جمع یا میانگین ترکیب می‌شود.

تعبیر ملهوس ترفند هسته

برای بردار دوبعدی، تبدیل $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ را در نظر بگیرید. ضرب داخلی دوبردار تبدیل شده برابر $(x^T z)^2$ است. مثلاً برای $x = (1, 2)$ و $z = (3, 4)$ داریم:

$$K(x, z) = 11^2 = 121 = 1 \times 9 + (2\sqrt{2})(12\sqrt{2}) + 4 \times 16.$$

پس هسته راه محاسبه شباهت فضای تبدیل شده است؛ خودش الگوریتم انتخاب مرز یا یک احتمال نیست. غیرخطی شدن مرز در فضای اولیه با خطی بودن آن در فضای ویژگی سازگار است.

۹.۴ خودآزمایی و تمرین

۱. شواهد گمشته و صفر را یکسان نگیرید

در بیز ساده، اگر همه ویژگی‌ها نامعلوم باشند چه می‌ماند؟ اگر یک مقدار مشاهده شده در یک طبقه شمارش صفر داشته باشد چه می‌شود؟

پاسخ و توضیح: با حذف عامل ویژگی‌های نامعلوم، فقط پیشین طبقه‌ها باقی می‌ماند. شمارش صفر یک مقدار مشاهده شده، بدون هموارسازی امتیاز آن طبقه را صفر می‌کند. نامعلوم بودن با مشاهده یک رخداد با احتمال تخمینی صفر تفاوت دارد.

۲. هموارسازی معتبر بسازید

در طبقه‌ای با ۱۰ نمونه، ویژگی سه‌حالتی شمارش‌های ۰، ۴، ۶ دارد. احتمال‌ها با لاپلاس چیست؟ پاسخ و توضیح: احتمال‌ها $1/13, 5/13, 7/13$ می‌شوند و جمعشان یک است. افزودن یک فقط به خانه صفر و نگه داشتن مخرج ۱۰، توزیع معتبر نمی‌سازد.

۳. نقض حاشیه با خطا فرق دارد؟

برای یک نقطه مثبت با امتیاز $f(x) = 0.4$ ، زیان لولایی و صحت برچسب چیست؟ پاسخ و توضیح: زیان $\max(0, 1 - 0.4) = 0.6$ است، اما امتیاز مثبت است و برچسب درست می‌ماند. نقطه داخل حاشیه است. در قرارداد این فصل، C بزرگ‌تر این نقض را پرهزینه‌تر می‌کند.

۴. چرا ده لایه خطی XOR را حل نمی‌کند؟

ترکیب $W_2(W_1x + b_1) + b_2$ چه نوع تابعی است؟

پاسخ و توضیح: برابر $(W_2W_1)x + (W_2b_1 + b_2)$ ، یعنی دوباره آفین است. بدون فعال‌ساز غیرخطی، تعداد لایه‌ها به‌تنهایی ظرفیت ساخت مرز XOR را ایجاد نمی‌کند.

۵. پس انتشار دقیقاً چه کاری می‌کند؟

آیا پس انتشار نرخ یادگیری را انتخاب می‌کند؟ در مثال عددی این فصل، چرا هنگام محاسبه گرادیان پنهان از وزن خروجی تازه استفاده نمی‌کنیم؟

پاسخ و توضیح: پس انتشار گرادیان را با قاعده زنجیره‌ای محاسبه می‌کند؛ بهینه‌ساز و نرخ یادگیری، گام بعد را تعیین می‌کنند. همه مؤلفه‌های گرادیان یک گام باید در یک نقطه مشترک از فضای پارامترها محاسبه شوند.

۶. هر امتیاز عددی احتمال نیست

آیا امتیاز SVM یا حاصل ضرب خام بیز را می‌توان مستقیم احتمال دانست؟

پاسخ و توضیح: امتیاز SVM به‌خودی‌خود احتمال کالیبره نیست. حاصل ضرب بیز نیز باید با جمع امتیاز تمام طبقه‌ها نرمال شود؛ حتی پس از نرمال‌سازی، کیفیت کالیبراسیون به فرض‌های مدل و داده بستگی دارد.

فصل ۵

رگرسیون؛ پیش‌بینی پاسخ عددی

منبع مطالعه: اسلاید رگرسیون، PDF ۳ تا ۵۵؛ بسط مفاهیم تعمیم و اریبی-واریانس از بخش‌های ۳٫۴ تا ۳٫۸ و ۳٫۱۰، ۶٫۳ و ۶٫۴ کتاب. توضیح ridge، lasso، ایزوتونیک و رگرسیون محلی بر اسلاید و استدلال ریاضی تکیه دارد.

۱۰۵ مسئله چیست؟

در طبقه‌بندی می‌پرسیم «این خانه گران است یا ارزان؟» و در رگرسیون می‌پرسیم «قیمت این خانه چقدر است؟». داده آموزش زوج‌های (x_i, y_i) دارد؛ بردار x_i شامل ویژگی‌هایی مانند مساحت، تعداد اتاق و فاصله تا حمل و نقل و y_i قیمت مشاهده شده است. خروجی مدل $\hat{y}_i = f(x_i)$ پیش‌بینی است و با مقدار واقعی یکی نیست. بهتر است در جزوه از n برای تعداد نمونه و p برای تعداد ویژگی استفاده شود تا نمادها قاطی نشوند. مدل با کمینه کردن خطا روی داده آموزش ساخته می‌شود، ولی هدف نهایی کم‌بودن خطا روی خانه‌ها یا شرایطی است که ندیده‌ایم. جدا کردن داده آموزش باید پیش از مقیاس‌بندی، حذف برون‌افتاده، انتخاب ویژگی و انتخاب پارامتر انجام شود. آمار مقیاس‌بندی فقط از آموزش به دست می‌آید. برای داده زمانی، تقسیم زمانی به جای تقسیم تصادفی می‌تواند ضروری باشد. این ادامه اصول بخش‌های ۳٫۵-۳٫۸ کتاب است.

۲۰۵ رگرسیون نزدیک‌ترین همسایه

برای برآورد دمای محلی که ایستگاه ندارد، دمای ایستگاه‌های نزدیک را میانگین می‌گیریم. اگر پنج دمای نزدیک ۱۸، ۲۰، ۲۱، ۲۲ و ۲۱ باشند، پیش‌بینی $102/5 = 20.4$ درجه است؛ این همان مثال اسلاید ۸ است. در حالت کلی:

$$\hat{y}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i, \quad \hat{y}_w(x) = \frac{\sum_{i \in N_k} w_i y_i}{\sum_{i \in N_k} w_i}, \quad w_i = \frac{1}{d(x, x_i) + \varepsilon}.$$

ابتدا فاصله را تعریف می‌کنیم، سپس k همسایه را پیدا می‌کنیم و در پایان مقدار عددی آنان را ترکیب می‌کنیم. برای فاصله صفر باید صریحاً خروجی نمونه‌های هم‌مکان را استفاده کرد یا قرارداد پایدار داشت. k کوچک به

جزئیات و نويز حساس تر است؛ k بزرگ نمودار را هموارتر می‌کند و ممکن است الگوی محلی را محو کند. مقیاس ویژگی‌ها مستقیماً روی انتخاب همسایه اثر می‌گذارد: «حقوق به ریال» می‌تواند «سن به سال» را تحت الشعاع قرار دهد. k را روی اعتبارسنجی انتخاب می‌کنیم. در میانگین با وزن‌های نامنفی، پیش‌بینی در بازه مقادیر هدف همسایگان می‌ماند؛ به همین دلیل مدل معمولاً برون‌یابی عددی نمی‌کند.

۳.۵ معیارهای خطا و دام همبستگی

اگر $e_i = y_i - \hat{y}_i$ ، آنگاه

$$\text{MAE} = n^{-1} \sum_i |e_i|, \quad \text{MSE} = n^{-1} \sum_i e_i^2, \quad \text{RMSE} = \sqrt{\text{MSE}}.$$

برای واقعی‌های $(2, 4, 6)$ و پیش‌بینی‌های $(3, 4, 4)$ ، خطاها $(-1, 0, 2)$ ، MAE برابر ۱، MSE برابر $5/3$ و RMSE حدود ۱٫۲۹۱ است. MAE و RMSE واحد هدف را دارند؛ MSE واحد مربع هدف را دارد. مربع کردن خطا سبب می‌شود یک خطای بزرگ نسبت به چند خطای کوچک وزن بیشتری بگیرد. به همین علت گزارش هم‌زمان MAE و RMSE می‌تواند وجود خطاهای بزرگ را آشکار کند.

همبستگی پیرسون

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}}$$

روند خطی را می‌سنجد، نه توافق عددی را. برای واقعی‌های ۱، ۲، ۳ و پیش‌بینی‌های ۱۰۰۰، ۲۰۰۰، ۳۰۰۰ همبستگی ۱ است اما خطا بسیار بزرگ است. اگر یکی از دو بردار ثابت باشد، مخرج صفر و همبستگی تعریف نشده است.

تعریف عمومی ضریب تعیین چنین است:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

برای مثال بالا $\bar{y} = 4$ ، SST برابر ۸ و SSE برابر ۵ است، پس $R^2 = 0.375$. پیش‌بینی ثابت میانگین داده ارزیابی R^2 صفر دارد و پیش‌بینی کاملاً درست R^2 یک دارد. اصلاح اسلاید ۱۲: R^2 به‌طور عمومی محدود به بازه صفر و یک نیست؛ برای مدل بد روی آزمون منفی می‌شود. نسبت $\text{SS total} / \text{SS explained}$ فقط در شرایط تجزیه مناسب، از جمله OLS دارای عرض از مبدأ روی آموزش، با تعریف عمومی برابر است. وقتی تمام y ها یکسان‌اند، SST صفر است و باید قرارداد نرم‌افزاری را جداگانه توضیح داد.

۴.۵ درخت رگرسیون و درخت مدل

ساختار سؤال‌های درخت تغییر نمی‌کند: مثلاً «مساحت کمتر از ۱۰۰ است؟». تفاوت در هدف شکستن و مقدار برگ است. در برگ با کمینه‌سازی SSE باید میانگین y های آموزش را پیش‌بینی کنیم. برای هر شکستن کاندید،

کاهش خطا را می‌سنجیم:

$$\Delta = \text{SSE}(P) - \text{SSE}(L) - \text{SSE}(R).$$

مثال مستقل: $y = (1, 2, 8, 9)$, $x = (1, 2, 3, 4)$. پیش از شکستن، میانگین ۵ و SSE برابر $16+9+9+16 = 50$ است. شکستن در ۲٫۵ دو میانگین ۱٫۵ و ۸٫۵ می‌دهد و SSE کل ۱ می‌شود؛ کاهش برابر ۴۹ است. نمودار خروجی پله‌ای یا تکه‌ای ثابت خواهد بود. عمق زیاد و برگ‌های تک نمونه‌ای می‌توانند نویز را حفظ کنند؛ محدودیت عمق و حداقل اندازه برگ ابزار کنترل‌اند.

در «درخت مدل» به جای ثابت در هر برگ، یک رگرسیون خطی محلی آموزش می‌دهیم؛ مثلاً مصرف خودروهای دیزلی و بنزینی هرکدام رابطه متفاوتی با وزن دارند. هنگام پیش‌بینی ابتدا مسیر درخت و سپس فرمول برگ اجرا می‌شود. چنین مدلی تکه‌ای خطی است، و برخلاف برگ‌های میانگین درخت رگرسیون، ممکن است برون‌یابی نکند.

۵.۵ رگرسیون خطی و کمترین مربعات

مدل ساده $\hat{y} = w_0 + w_1x$ و مدل چندمتغیره $\hat{y} = w_0 + \sum_{j=1}^p w_jx_j$ است. «خطی» در اینجا یعنی خطی نسبت به ضرایب؛ می‌توان ویژگی x^2 را هم وارد مدل خطی کرد. تابع هدف OLS برابر $\min_w \|y - Xw\|_2^2$ است؛ ستون ثابت در X عرض از مبدأ را نمایش می‌دهد. در حالت ساده:

$$\hat{w}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}, \quad \hat{w}_0 = \bar{y} - \hat{w}_1\bar{x}.$$

برای $x = (1, 2, 3)$, $y = (2, 3, 5)$, $\bar{y} = 10/3$, $\bar{x} = 2$ ، شیب $3/2$ و عرض $1/3$ است. پیش‌بینی‌ها $(11/6, 10/3, 29/6)$ و SSE برابر $1/6$ می‌شوند. در چندمتغیره اگر X رتبه ستونی کامل داشته باشد، راه حل یکتا و به صورت $(X^T X)^{-1} X^T y$ بیان می‌شود؛ در اجرا QR/SVD پایدارتر از معکوس‌گیری صریح است.

برای برازش عددی کمترین مربعات، نرمال‌بودن خطا لازم نیست. برای تفسیر آماری کلاسیک، درست‌بودن شکل میانگین شرطی، صفر بودن میانگین خطای شرطی، استقلال مناسب مشاهدات و شرایط واریانس اهمیت دارند؛ نرمال‌بودن معمولاً به استنباط دقیق کوچک‌نمونه مربوط است. همبستگی شدید ویژگی‌ها می‌تواند ضرایب را ناپایدار کند. کوچک‌بودن SSE آموزش به‌تنهایی دلیل تعمیم خوب نیست.

۶.۵ منظم‌سازی ridge و lasso

افزودن ویژگی‌ها انعطاف مدل را زیاد می‌کند. منظم‌سازی در مقابل بهبود برازش، برای ضرایب بزرگ هزینه تعیین می‌کند:

$$\text{Ridge} : \min_w \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p w_j^2, \quad \text{Lasso} : \min_w \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |w_j|.$$

عرض از مبدأ معمولاً جریمه نمی‌شود. مقیاس‌بندی ویژگی‌ها پیش از جریمه مهم است؛ در غیر این صورت اندازه ضریب به واحد اندازه‌گیری وابسته می‌شود. Ridge ضرایب را پیوسته کوچک می‌کند؛ lasso می‌تواند ضرایب را

دقیقاً صفر و نوعی انتخاب ویژگی انجام دهد. در حضور ویژگی‌های بسیار هم‌بسته، انتخاب lasso ممکن است ناپایدار باشد. $\lambda = 0$ حالت بدون جریمه است و الزاماً بیش‌برازش ایجاد نمی‌کند؛ λ خیلی بزرگ، با عرض از مبدأ آزاد، مدل را به پیش‌بینی نزدیک میانگین می‌رساند. «بدون جریمه» با «بدون نرمال‌سازی» یکسان نیست؛ تعبیر اسلاید ۲۱ باید دقیق‌تر نوشته شود.

برای شهود، مسئله تک‌ضربی $(w - 3)^2 + 2w^2$ کینه ridge برابر ۱ دارد؛ اما $(w - 3)^2 + 2|w|$ کینه lasso برابر ۲ دارد. در قرارداد کلی $(w - a)^2 + \lambda|w|$ ، پاسخ $\text{sign}(a) \max(|a| - \lambda/2, 0)$ است. ضریب $1/2$ به قرارداد تابع هدف وابسته است؛ در کتابخانه‌ها ممکن است ضرایب n یا $1/2$ متفاوت باشند.

۷.۵ درونیابی، برون‌یابی و خط مبنا

در کاربرد علمی، درونیابی/برونیابی معمولاً نسبت به ناحیه ورودی آموزش تعریف می‌شود. اسلاید ۲۳ علاوه بر آن روی خارج‌شدن مقدار پیش‌بینی از بازه هدف تأکید دارد؛ این دو مفهوم را نباید یکی گرفت. خطی می‌تواند بیرون بازه y آموزش پیش‌بینی کند؛ kNN میانگین و درخت میانگین نمی‌توانند، اما این ویژگی تضمین نمی‌کند پیش‌بینی آنان در موقعیت جدید معتبر باشد. حتی در بازه ورودی هم نقاط کم‌ترکم ممکن است نامطمئن باشند. برون‌یابی برای آینده‌نگری لازم است ولی به دوام فرض رابطه تکیه دارد.

خط مبنا باید پیش از مدل پیچیده ساخته شود. ثابت میانگین آموزش SSE را و ثابت میانه MAE را روی آموزش کینه می‌کند. در آزمون فقط همان ثابت آموخته‌شده از آموزش به کار می‌رود، نه میانگین هدف آزمون. اگر مدل از این مبنا بهتر نباشد، پیچیدگی آن توجیه تجربی ندارد.

۸.۵ ایزوتونیک و الگوریتم ادغام ناقضان مجاور

گاهی رابطه خطی نیست ولی می‌دانیم افزایشی است؛ مثلاً انتظار داریم یک امتیاز کالیبره‌شده با امتیاز خام کاهش نیابد. پس پس از مرتب‌کردن x ها، حل می‌کنیم:

$$\min_{z_1 \leq \dots \leq z_n} \sum_{i=1}^n a_i (y_i - z_i)^2, \quad a_i > 0.$$

هر مشاهده را ابتدا یک بلوک می‌گیریم. اگر مقدار بلوک چپ از بلوک راست بیشتر باشد، دو بلوک ناقض‌اند؛ آن‌ها را ادغام و میانگین وزنی‌شان را جایگزین می‌کنیم. سپس دوباره بلوک قبل را بررسی می‌کنیم و تا رفع همه نقض‌ها ادامه می‌دهیم. وزن بلوک ادغام‌شده مجموع وزن‌هاست؛ فراموش کردن وزن سبب خطا می‌شود.

مثال: y های مرتب $(1, 4, 3, 2, 5)$. ابتدا ۴ و ۳ به میانگین ۳٫۵ با وزن ۲ تبدیل می‌شوند. چون ۳٫۵ از ۲ بزرگ‌تر است، بلوک سه‌تایی میانگین $3 = (4 + 3 + 2)/3$ می‌گیرد. جواب $(1, 3, 3, 3, 5)$ و SSE برابر ۲ است. نقاط x جابه‌جا نمی‌شوند؛ مقادیر برازش تعدیل می‌شوند. برابری برای همه همسایه‌ها لازم نیست؛ شرط صحیح $z_i \leq z_{i+1}$ است. نمایش بین گره‌ها می‌تواند با قرارداد خطی یا پله‌ای انجام شود. اسلاید ۳۶ رفتار بیرون دامنه را با clip/nan/raise تفکیک می‌کند؛ این رفتار بخشی از تعریف کاربردی مدل است.

۹.۵ چندجمله‌ای و رگرسیون محلی

مدل درجه d برابر $\hat{y} = \sum_{j=0}^d w_j x^j$ است و همچنان نسبت به w خطی است. افزایش d امکان دنبال کردن خمیدگی و همچنین نویز را فراهم می‌کند. برای p ویژگی، تعداد تک‌جمله‌ای‌های تا درجه d با جمله ثابت $\binom{p+d}{d}$ است؛ برای $p = 3$ و $d = 3$ تعداد ۲۰ داریم: یک ثابت، سه درجه‌یک، شش درجه‌دو و ده درجه‌سه. این شمارش، توضیح دقیق تر اسلاید ۳۹ است. داشتن ویژگی بیشتر از نمونه به‌خودی‌خود ممنوعیت ریاضی نیست، ولی بدون منظم‌سازی/ساختار مناسب، ناپایداری و بیش‌برازش محتمل می‌شود. انتخاب درجه باید با اعتبارسنجی انجام شود. بیرون دامنه آموزش نوسان چندجمله‌ای می‌تواند شدید شود.

رگرسیون محلی برای هر x تازه همسایه‌های نزدیک را پیدا می‌کند، روی همان همسایه‌ها خط برازش می‌دهد و در x تازه آن خط را ارزیابی می‌کند. شکل عمومی کمترین مربعات وزن‌دار است:

$$\hat{\beta}(x) = \arg \min_{\beta} \sum_i K\left(\frac{x_i - x}{h}\right) (y_i - \beta_0 - \beta_1(x_i - x))^2, \quad \hat{y}(x) = \hat{\beta}_0(x).$$

نسخه اسلاید فقط k همسایه را به کار می‌گیرد. h یا k میزان هموارسازی را کنترل می‌کند. برخلاف kNN ، خط محلی روند را هم تخمین می‌زند و ممکن است از بازه هدف همسایه‌ها خارج شود. فرض شهودی این است که تابع خمیده در همسایگی کوچک تقریباً خطی است. هزینه پیش‌بینی بیشتر است، چون ارزیابی همسایه و برازش برای هر پرسش تکرار می‌شود.

۱۰.۵ شبکه عصبی و موازنه اریبی/واریانس

برای تبدیل شبکه طبقه‌بندی به رگرسیون، خروجی عددی را نگه می‌داریم و به جای خطای طبقه‌بندی، مثلاً MSE را بهینه می‌کنیم. لایه خروجی خطی محدود به بازه صفر و یک نیست؛ عبارت اسلاید ۴۵ این محدودیت را توجیه نمی‌کند. شبکه چندلایه تنها وقتی از ترکیب خطی فراتر می‌رود که میان لایه‌ها فعال‌ساز غیرخطی داشته باشیم. افزودن صرف لایه‌های خطی دوباره یک تابع خطی می‌سازد.

برای مدل داده $Y = f(x) + \varepsilon$ ، با $E[\varepsilon | x] = 0$ ، و آموزش تصادفی D ، تجزیه خطای مربعی در x چنین است:

$$E_{D,\varepsilon}[(Y - \hat{f}_D(x))^2] = \underbrace{(E_D \hat{f}_D(x) - f(x))^2}_{\text{Bias}^2} + \underbrace{E_D[(\hat{f}_D(x) - E_D \hat{f}_D(x))^2]}_{\text{Variance}} + \underbrace{\text{Var}(\varepsilon | x)}_{\text{Noise}}.$$

اریبی یعنی میانگین مدل‌های آموزش‌دیده از رابطه واقعی فاصله دارد؛ واریانس یعنی مدل با عوض شدن نمونه آموزش تغییر زیادی می‌کند. نویز کاهش‌ناپذیر تحت اطلاعات و مدل داده فعلی باقی می‌ماند؛ حذف کورکورانه نقاط غیرعادی آن را لزوماً از بین نمی‌برد. مثال حسگر خراب می‌تواند آلودگی قابل اصلاح باشد و با تصادفی بودن ذاتی یکی نیست.

منحنی یادگیری خطای آموزش و اعتبارسنجی را در برابر اندازه آموزش می‌کشد. خطای هر دو زیاد و نزدیک، نشانه احتمالی کم‌برازش است؛ خطای آموزش کم ولی فاصله بزرگ با اعتبارسنجی، نشانه احتمالی بیش‌برازش است. داده بیشتر اغلب واریانس را کم می‌کند اما شکل تابع اشتباه را اصلاح نمی‌کند. اسلاید ۵۲ امکان آستانه‌گذاری خروجی رگرسیون روی ۰.۵ برای طبقه‌بندی دودویی را مطرح می‌کند؛ خروجی رگرسیون خطی الزاماً احتمال معتبر نیست و برچسب‌گذاری عددی کلاس‌های اسمی فاصله مصنوعی ایجاد می‌کند.

۱۱.۵ چرا میانگین، پاسخ کمترین مربعات است؟

در یک برگ درخت، یک بلوک ایزوتونیک یا مدل ثابت، می‌خواهیم یک عدد c را به همه پاسخ‌ها نسبت دهیم. با زیان مربعی:

$$L(c) = \sum_{i=1}^n (y_i - c)^2, \quad L'(c) = 2nc - 2 \sum_i y_i.$$

مشتق صفر می‌دهد $c = \bar{y}$ و مشتق دوم $2n > 0$ کمینه بودن را ثابت می‌کند. با وزن‌های مثبت، همین استدلال میانگین وزن‌دار را می‌دهد. اما برای زیان قدرمطلق، یک میانه کمینه‌کننده است. انتخاب نماینده از تابع زیان می‌آید، نه از یک قاعده ثابت برای تمام الگوریتم‌ها.

برای پاسخ‌های 1, 2, 9، بهترین ثابت مربعی ۴ و مجموع مربع خطا $9 + 4 + 25 = 38$ است. میانه ۲، مجموع قدرمطلق خطای $1 + 0 + 7 = 8$ می‌دهد؛ در مقابل، مجموع قدرمطلق خطای ثابت ۴ برابر ۱۰ است. عوض کردن معیار می‌تواند بهترین مدل را هم عوض کند. وقتی RMSE و MAE اختلاف زیادی دارند، چند باقیمانده بزرگ را بررسی کنید؛ حذف آن‌ها بدون بررسی منشأ، پاسخ علمی نیست.

ارزیابی یک برازش، فراتر از رسم منحنی

باقیمانده را در برابر ورودی و پیش‌بینی رسم می‌کنیم. الگوی خمیده می‌تواند نشان دهد رابطه خطی کافی نیست؛ افزایش دامنه پراکندگی می‌تواند نشانه واریانس ناهمسان باشد؛ ترتیب زمانی باقیمانده‌ها می‌تواند وابستگی نادیده را آشکار کند. این‌ها ابزار تشخیص‌اند و به‌تنهایی اثبات یک مدل خاص نیستند. پیش‌بینی، تفسیر ضریب و استنباط آماری سه هدف متفاوت‌اند؛ برای نتیجه‌گیری احتمالی درباره ضریب به فرض‌های خطا و طراحی داده نیز نیاز داریم.

در مقایسه درجه‌ها، یک مجموعه آموزش و یک اعتبارسنجی ثابت نگه دارید. اگر داده تصادفی با هر حرکت کنترل تازه شود، معلوم نیست تغییر امتیاز از درجه آمده یا از عوض شدن نمونه. در آزمایشگاه این فصل بذرتاب است، جریمه روی عرض از مبدأ اعمال نمی‌شود و قرارداد تابع هدف، مجموع مربع خطا به‌علاوه جریمه است. بنابراین مقدار عددی جریمه را بدون توجه به قرارداد با نرم‌افزار دیگر مقایسه نکنید.

۱۲.۵ خودآزمایی و تمرین

۱. خطاها و واحدها را محاسبه کنید

برای حقیقت (2, 4, 6) و پیش‌بینی (3, 4, 4)، MAE، MSE و RMSE چیست؟

پاسخ و توضیح: خطاها (-1, 0, 2)، مقدار MAE برابر یک، MSE برابر 5/3 و RMSE برابر $\sqrt{5/3} \approx 1.2910$ است. MAE و RMSE واحد پاسخ و MSE مربع واحد پاسخ را دارد.

۲. همبستگی کامل، پیش‌بینی کامل؟

اگر $\hat{y} = 2y + 10$ باشد، چرا ممکن است همبستگی یک ولی خطا زیاد باشد؟

پاسخ و توضیح: همبستگی ترتیب و رابطه خطی را می‌سنجد و به انتقال و مقیاس مثبت حساس نیست؛ ولی اختلاف با مقدار واقعی بزرگ است. برای کالیبراسیون عددی باید خطا و خط مبنا را هم بررسی کرد.

۳. R^2 منفی را تفسیر کنید

اگر $SSE = 20$ و $SST = 8$ باشد، R^2 چیست؟

پاسخ و توضیح: $R^2 = 1 - 20/8 = -1.5$ است؛ مدل با این تعریف از پیش‌بینی ثابت میانگین همان داده ارزیابی بدتر است. اگر پاسخ‌ها ثابت باشند، SST صفر و این تعریف کسر شکل نامعتبر است.

۴. میانگین یا میانه؟

برای 1, 2, 9 کدام ثابت برای مربع خطا و کدام برای قدرمطلق خطا بهینه است؟

پاسخ و توضیح: برای مربع خطا، میانگین ۴ و برای قدرمطلق خطا، میانه ۲ است. تابع زیان تعیین می‌کند «بهترین ثابت» چه معنایی دارد.

۵. ادغام ایزوتونیک را انجام دهید

برازش غیرکاهشی 1, 4, 3, 2, 5 با وزن برابر چیست؟

پاسخ و توضیح: ۴ و ۳ ابتدا بلوک 3.5 می‌سازند؛ این بلوک با ۲ ناسازگار است و سه مقدار به میانگین ۳ ادغام می‌شوند. پاسخ (1, 3, 3, 3, 5) و مجموع مربع خطا ۲ است؛ ترتیب ورودی‌ها عوض نمی‌شود.

۶. درجه، جریمه و برون‌یابی را نقد کنید

درجه هفت روی آموزش خطای کمتری دارد ولی بیرون بازه نوسان می‌کند. آیا باید آن را انتخاب کنیم؟

پاسخ و توضیح: خطای آموزش کافی نیست. درجه و جریمه را با اعتبارسنجی و دامنه کاربرد انتخاب می‌کنیم؛ عملکرد خوب داخل بازه تضمین رفتار خارج آن نیست. استانداردسازی و ساخت ویژگی هم باید درون خط لوله آموزش باشد.

فصل ۶

خوشه‌بندی و کشف ناهنجاری

منبع مطالعه: اسلاید خوشه‌بندی، PDF ۳ تا ۶۴؛ کتاب فصل ۵، صفحات چاپی ۳۰۷ تا ۳۷۸؛ فصل ۹، صفحات ۷۲۳ تا ۷۵۱؛ بخش ۲٫۴. جنگل جداسازی از اسلایدها توضیح داده شده است.

۱.۶ هدف و نوع خوشه

در یادگیری بدون نظارت برچسب پاسخ نداریم؛ می‌خواهیم ساختار داده را کشف کنیم. شباهت درون گروه زیاد و تمایز گروه‌ها زیاد باشد، اما این تعریف به نوع فاصله و هدف تحلیل وابسته است. شش گروه کوچک مشتری ممکن است برای بازاریابی دقیق مفید باشند و ادغامشان به دو گروه برای تصمیم‌گیری مناسب‌تر باشد؛ وجود یک پاسخ بصری یگانه را نباید فرض کرد. کتاب بخش ۵٫۱ میان خوشه‌های جداشده، مرکزگرا، پیوسته، چگالی‌محور و مفهومی تفاوت می‌گذارد.

خوشه‌بندی افزای یک تقسیم در یک سطح می‌سازد؛ خوشه‌بندی سلسله‌مراتبی خانواده‌ای از گروه‌های تودرتو می‌دهد که با دندروگرام نمایش داده می‌شوند. روش‌هایی مثل DBSCAN می‌توانند بعضی نقاط را نیز اعلام کنند؛ بنابراین عبارت «همه نقاط حتماً عضو خوشه‌اند» ویژگی k -means است و برای همه روش‌ها صدق نمی‌کند. کاربردهای اسلاید شامل بخش‌بندی مشتری، تطبیق پیشنهادهای یک محصول و تشخیص بخش‌های تصویر است. مفید بودن خوشه‌ها باید با سؤال کاربرد سنجیده شود.

۲.۶ k -means: الگوریتم و دلیل میانگین

برای داده عددی، k -means مرکزهای $\mu_1, \dots, \mu_k$ و تقسیم را برای کم کردن

$$J = \sum_{j=1}^k \sum_{\mathbf{x}_i \in C_j} \|\mathbf{x}_i - \mu_j\|_2^2$$

جست‌وجو می‌کند. ابتدا k مرکز تعیین می‌کنیم. در گام انتساب هر نقطه به نزدیک‌ترین مرکز می‌رود. در گام به‌روزرسانی هر مرکز میانگین اعضای جدیدش می‌شود. دو گام را تا ثابت شدن انتساب‌ها، جابه‌جایی ناچیز مرکزها

یا رسیدن به سقف تکرار ادامه می‌دهیم. وقتی عضویت‌ها ثابت‌اند، مشتق خطای خوشه نسبت به مرکز صفر می‌شود و پاسخ میانگین است:

$$\frac{\partial}{\partial \mu_j} \sum_{x_i \in C_j} (x_i - \mu_j)^2 = 2|C_j|\mu_j - 2 \sum_{x_i \in C_j} x_i = 0.$$

در مثال یک‌بعدی $k = 2, 1, 2, 8, 9$ و مرکزهای اولیه ۱ و ۸، انتساب‌ها $\{1, 2\}$ و $\{8, 9\}$ می‌شوند. مرکزهای جدید ۱٫۵ و ۸٫۵ و SSE برابر ۱ است. برای $k = 1$ مرکز ۵ و SSE برابر ۵۰ است؛ برای $k = 4$ ، SSE صفر می‌شود. بنابراین کاهش خام SSE با افزایش k دلیل کافی برای انتخاب تعداد بیشتر خوشه نیست.

هر گام، با قراردادهای ثابت، هدف را افزایش نمی‌دهد و الگوریتم به یک جواب محلی می‌رسد؛ تضمین بهترین جواب جهانی وجود ندارد. بذره‌های متفاوت می‌توانند پاسخ‌های متفاوت بدهند. چند شروع مستقل و انتخاب کمترین SSE برای k ثابت مفید است. کتاب در چاپی ۳۲۵-۳۲۶ PDF/۳۲۶-۳۲۷ روش $k\text{-means++}$ را بسط می‌دهد: اولین مرکز تصادفی است و مرکزهای بعدی با احتمال متناسب با مربع فاصله از نزدیک‌ترین مرکز انتخاب می‌شوند. خوشه خالی را باید با قرارداد مشخص، مثلاً قراردادن مرکز روی نقطه دارای بیشترین فاصله، مدیریت کرد.

زمان تقریبی $O(nkpI)$ برای n نقطه، p ویژگی و I تکرار است. استانداردسازی ویژگی‌ها اهمیت دارد. حساسیت به نقاط دور، شکل‌های غیرکروی و گروه‌های با اندازه یا چگالی بسیار متفاوت از محدودیت‌های مهم‌اند. منحنی آرنج محل کاهش منفعت افزودن k را نشان می‌دهد ولی همیشه آرنج آشکاری وجود ندارد. اسلاید ۲۱ X-means را به‌عنوان گسترش انتخاب تعداد خوشه معرفی می‌کند؛ این روش با انتخاب ساده آرنج یکسان نیست.

۳.۶ k-medoids و تمایز میانه از medoid

اسلاید ۲۶ k-medoids را با عبارت «استفاده از میانه» توضیح می‌دهد، اما تعریف دقیق این است که نماینده باید یک مشاهده موجود باشد:

$$\text{medoid}(C) = \arg \min_{z \in C} \sum_{x \in C} d(x, z).$$

میانه هر مؤلفه در فضای چندبعدی الزاماً نقطه‌ای از داده نیست و k-medians با k-medoids متفاوت است. در یک‌بعد با فاصله مطلق، یکی از میانه‌های مشاهده‌شده می‌تواند medoid باشد. مثال 1, 3, 5, 7, 10, 9 میانگین ۲۰۵ ولی میانه و medoid برابر ۵ می‌دهد. مزیت نسبت به میانگین مربع‌خطا، مقاومت بیشتر در برابر نقاط بسیار دور است؛ ولی انتخاب نماینده می‌تواند هزینه محاسباتی بیشتری داشته باشد.

۴.۶ silhouette: سنجش انتساب هر نقطه

برای نقطه i ، a_i میانگین فاصله تا سایر اعضای خوشه خودش و b_i کمترین میانگین فاصله تا اعضای یک خوشه دیگر است:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad S = \frac{1}{n} \sum_i s_i.$$

اگر $a = 2, b = 6$ ، مقدار $2/3$ نشانهٔ انتساب مناسب است؛ اگر $a = 6, b = 2$ ، مقدار $-2/3$ نشان می‌دهد نقطه به خوشهٔ دیگر نزدیک تر است. نزدیک صفر یعنی مرز مبهم. برای خوشهٔ تک‌عضوی معمولاً $s = 0$ قرارداد می‌شود؛ تعریف به دست کم دو خوشه و کمتر از n خوشه نیاز دارد. میانگین بالا مفید است اما «مستقل از تعداد خوشه» را نباید به معنی بی‌تأثیر بودن k دانست: خود غمره با تقسیم‌بندی، مقیاس و شکل داده عوض می‌شود و گروه‌های کروی و جدا را ترجیح می‌دهد. بهتر است نمودار silhouette هر خوشه نیز کنار میانگین دیده شود.

۵.۶ DBSCAN و اتصال نواحی پرتراکم

اگر شکل داده هلالی باشد، مرکز کروی k -means ابزار مناسبی نیست. DBSCAN خوشه را ناحیهٔ پرتراکم متصل می‌بیند. همسایگی $N_\epsilon(x) = \{z : d(x, z) \leq \epsilon\}$ شامل خود نقطه است. نقطهٔ هسته حداقل MinPts عضو در این همسایگی دارد؛ نقطهٔ مرزی هسته نیست ولی همسایهٔ یک هسته است؛ بقیه نویزند. مرز فقط به خوشه متصل می‌شود و خودش اجازهٔ گسترش اتصال را نمی‌دهد.

روند آموزشی روشن: ابتدا هسته‌ها را تعیین کنیم، میان هسته‌هایی که فاصله‌شان حداکثر ϵ است یال بگذاریم، مؤلفه‌های متصل هسته‌ها را خوشه بگیریم و سپس مرزی‌ها را وصل کنیم. اگر مرزی به دو مؤلفه نزدیک باشد، قرارداد ترتیب/شکستن تساوی ممکن است بر عضویتش اثر بگذارد؛ مؤلفه‌های هسته به همین دلیل نباید از راه یک مرزی ادغام شوند. این همان بسط بخش ۵٫۴٫۲ کتاب است.

مثال $x = (0, 0.1, 0.2, 0.35, 2)$ ، $\epsilon = 0.16$ و $\text{MinPts} = 3$: نقطهٔ ۰٫۱ همسایه‌های ۰٫۲، ۰٫۱، ۰٫۰۱ و نقطهٔ ۰٫۲ همسایه‌های ۰٫۱، ۰٫۲، ۰٫۳۵ را دارند و هسته‌اند. ۰٫۳۵ مرزی و ۰٫۲ نویز است. در نتیجه یک خوشهٔ چهارعضوی می‌سازیم. زیاد کردن ϵ ممکن است خوشه‌های جدا را وصل کند؛ کم کردن آن یا زیاد کردن MinPts ممکن است نویز را زیاد کند. نمودار مرتب فاصله تا همسایهٔ k ام راهنمای ϵ است، اما قرارداد شمردن خود نقطه باید شفاف باشد. چگالی‌های بسیار متفاوت و ابعاد بالا نقاط ضعف اصلی‌اند. پیچیدگی جست‌وجوی ساده مربعی است؛ نمایهٔ مکانی در ابعاد کم می‌تواند کمک کند.

۶.۶ سلسله‌مراتبی: محتوای مکمل تصریح‌شده در درس

اسلاید ۶۴ این مبحث را به منبع مکمل ارجاع می‌دهد؛ توضیح زیر بر بخش ۵٫۳ کتاب استوار است. در روش تجعی هر نقطه ابتدا یک خوشه است؛ نزدیک‌ترین دو خوشه ادغام و فاصله‌ها به‌روز می‌شوند. ارتفاع ادغام در دندروگرام، مقدار ناهمبندی اتصال است؛ برش افقی درخت یک تقسیم می‌دهد.

linkage single کمترین فاصلهٔ دو عضو، complete بیشترین فاصله و average میانگین همهٔ فاصله‌های بین خوشه‌ای است. Ward ادغامی را برمی‌گزیند که کمترین افزایش SSE را دارد. برای دو مجموعهٔ یک‌بعدی $A = \{0, 2\}$ و $B = \{5, 9\}$ ، فاصله‌های جفتی ۵٫۹، ۳٫۷ است؛ $d_{\text{single}} = 3$ ، $d_{\text{complete}} = 9$ ، $d_{\text{average}} = 6$. پس انتخاب اتصال واقعاً ساختار خروجی را عوض می‌کند. Single می‌تواند گروه‌های کشیده را پیدا کند ولی به زنجیرهٔ نقاط نویزی حساس است؛ complete گروه‌های فشرده‌تر می‌سازد. ادغام‌های حریصانه برگشت‌ناپذیرند و حافظهٔ ماتریس فاصله می‌تواند $O(n^2)$ باشد.

۷.۶ فاصله مناسب، بخشی از تعریف مسئله است

برای p ویژگی عددی، فاصله مینکوفسکی

$$d_q(x, z) = \left(\sum_{j=1}^p |x_j - z_j|^q \right)^{1/q}, \quad q \geq 1$$

است؛ $q = 1$ منتهی و $q = 2$ اقلیدسی. برای اختلاف مؤلفه‌ها ۳ و ۴، این دو فاصله به ترتیب ۷ و ۵ هستند. ریشه در تعریف فاصله اقلیدسی ضروری است. برای صفت اسمی، برابری/نابرابری طبیعی‌تر از تفاضل عدد کدگذاری است؛ برای ترتیبی می‌توان رتبه‌ها را با قرارداد نرمال کرد. فاصله لونشتاین حداقل درج/حذف/جایگزینی حرف است؛ table و cable فاصله ۱ دارند. اسلاید ۴۰ دیگر خانواده‌های رشته‌ای را فقط معرفی می‌کند، پس تبدیل آن به فصل عمیق پردازش زبان خارج از دامنه فعلی است.

به عنوان بسط کتاب، Jaccard برای مجموعه‌ها $|A \cap B|/|A \cup B|$ و cosine برای بردارهای غیرصفر $x^T z / (||x|| ||z||)$ است. نبود مشترک یک رخداد نادر معمولاً در Jaccard پاداش نمی‌گیرد. انتخاب فاصله باید با مفهوم شباهت در کاربرد هماهنگ باشد؛ استانداردسازی به خودی خود فاصله نامناسب را مناسب نمی‌کند.

۸.۶ مسئله و روش‌های مقاوم

یک مشاهده وقتی ناهنجار است که با ساختار معمول مورد انتظار ناسازگار باشد؛ این ناسازگاری می‌تواند زمینه‌ای یا محلی باشد. دمای ۳۰ درجه در تابستان طبیعی و برای یک سردخانه غیرعادی است. داده نادر الزاماً خطا نیست؛ ممکن است همان تقلب یا رخداد مورد نظر باشد. پس حذف همه برون‌افتاده‌ها سیاست عمومی نیست. اسلاید ۴۴ تأکید دارد اگر پالایش آموزش انتخاب شد، آزمون برای ارزیابی واقعی دست‌نخورده بماند؛ آستانه‌ها نیز از آموزش یاد گرفته شوند.

حصار توکی با $IQR = Q_3 - Q_1$ و بازه $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ یک قاعده اکتشافی مقاوم است و برخلاف عبارت اسلاید ۴۶، نرمال بودن پیش شرط محاسبه آن نیست. انتخاب روش محاسبه چارک، به‌ویژه در نمونه کوچک، جواب را تغییر می‌دهد. برای داده مرتب $(0, 1, 2, 3, 4, 5, 6, 20)$ ، با میانه نیمه‌ها $Q_1 = 1.5$ و $Q_3 = 5.5$ ، پس $IQR = 4$ و حصارها -۵.۴ و ۵.۱۱ هستند؛ ۲۰ علامت می‌خورد. این مثال قرارداد را صریح دارد و از ابهام چارک اسلاید جلوگیری می‌کند.

انحراف مطلق میانه $MAD = \text{median}_i |x_i - \text{median}(X)|$ است. در مثال اسلاید $(0, 1, 1, 3, 5, 7, 42)$ ، میانه ۳ و $MAD = 2$ است؛ با قاعده خام $k = 3$ بازه -۳ تا ۹ می‌شود و ۴۲ بیرون است. MAD خام با تخمین سازگار انحراف معیار نرمال $MAD = 1.4826$ فرق دارد؛ آستانه‌ها را نباید مخلوط کرد. اگر $MAD = 0$ شد تقسیم بر آن معتبر نیست و باید قرارداد مقاوم جایگزین یا گزارش عدم امکان تعیین ثمره داشت. IQR/MAD تک‌متغیره ممکن است نقطه‌ای در «حفره میان دو گروه» را نبینند؛ مقدار عددی آن نقطه شدید نیست ولی از همه همسایگان دور است.

۹.۶ روش آماری، فاصله‌ای و چگالی نسبی

در مدل نرمال تک‌متغیره، نمره استاندارد $z = (x - \mu) / \sigma$ دوری از مرکز را نسبت به پراکندگی می‌سنجد. ارزیابی احتمال به درست بودن مدل توزیع بستگی دارد؛ برای داده‌ی چندقله‌ای یک نرمال واحد ممکن است گمراه کند. در چندبعد با ویژگی‌های هم‌بسته، فاصله‌ی مایه‌لانوویس مربعی $(x - \mu)^T \Sigma^{-1} (x - \mu)$ شکل و جهت ابر داده را لحاظ می‌کند؛ کوواریانس باید قابل‌برآورد و غیرتکین یا منظم شده باشد.

روش‌های فاصله‌ای اسلاید ۴۹ سه صورت دارند: تعداد همسایه در شعاع ثابت، فاصله تا همسایه k ام، یا میانگین فاصله تا k همسایه. فاصله بیشتر می‌تواند نمره ناهنجاری بزرگ‌تر بدهد. اما یک خوشه طبیعی کم‌تراکم نباید صرفاً به علت فاصله‌های بزرگ‌تر غیرعادی شود. اینجا مقایسه محلی کمک می‌کند.

کتاب PDF ۷۴۳ نسبت «میانگین چگالی همسایه‌ها / چگالی نقطه» را برای بهبود توضیح می‌دهد و صریحاً می‌گوید LOF دقیق از تعریف مقاوم‌تری استفاده می‌کند. برای معرفی دقیق LOF باید نوشت:

$$\text{reachdist}_k(x, o) = \max\{k\text{-dist}(o), d(x, o)\}.$$

$$\text{lrd}_k(x) = \left(\frac{\sum_{o \in N_k(x)} \text{reachdist}_k(x, o)}{|N_k(x)|} \right)^{-1},$$

$$\text{LOF}_k(x) = \frac{1}{|N_k(x)|} \sum_{o \in N_k(x)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(x)}.$$

نزدیک ۱ یعنی تراکم نقطه با همسایه‌ها سازگار است؛ مقدار خیلی بزرگ‌تر نامعمول است، ولی هر مقدار صرفاً بزرگ‌تر از ۱ حکم قطعی ناهنجاری نیست. انتخاب k و آستانه به مسئله وابسته است. پیاده‌سازی آزمایشگاه اگر فقط عکس فاصله میانگین را استفاده کند باید «چگالی نسبی ساده» نام بگیرد، نه LOF دقیق.

۱۰.۶ خوشه‌محور و جنگل جداسازی

DBSCAN مستقیماً برچسب نويز می‌دهد، اما نويز با پارامترهای انتخاب شده تعریف شده و اثبات خطای داده نیست. روش خوشه‌محور می‌تواند فاصله از مرکز، فاصله نسبی به پراکندگی همان خوشه، کوچک بودن گروه و دوری آن از گروه‌های بزرگ را ترکیب کند. اسلایدهای ۵۳-۵۵ CBLOF را با اندازه خوشه و فاصله تا خوشه بزرگ معرفی می‌کنند. کتاب بخش ۹.۵ بحث متناظر خوشه‌ای دارد، اما فرمول دقیق CBLOF را ارائه نمی‌کند؛ فرمول ساختگی به آن نسبت داده نشود. گروه کوچک می‌تواند یک جمعیت واقعی باشد و باید زمینه کاربرد را دید.

Forest Isolation به جای مدل کردن چگالی، می‌پرسد چند برش تصادفی برای جدا کردن نقطه لازم است. در هر گره یک ویژگی و سپس یک مقدار شکستن بین کمینه و بیشینه همان ویژگی تصادفی انتخاب می‌شود. این روند روی زیرگروه‌ها تکرار می‌شود و عمق مسیر نقطه ثبت می‌شود. نقطه دورافتاده اغلب با برش‌های کمتری تنها می‌شود؛ نقاط درون ناحیه انبوه به برش‌های بیشتر نیاز دارند. میانگین گیری طول مسیر در چند درخت، تصادف یک درخت را تعدیل می‌کند. نسخه آموزشی اسلاید ۵۶-۶۳ همین بهبود و احتمال‌های جداسازی یک گامه را نشان می‌دهد؛ طول مسیر کوتاه‌تر یعنی ناهنجاری بیشتر. برای توضیح این مفهوم نیازی به معرفی نمره نرمال شده‌ای که در منابع ارسالی ارائه نشده نیست. سقف عمق و نمونه‌گیری، الگوریتم عملی را سریع‌تر می‌کنند.

۱۱.۶ اجرای دستی؛ چه چیزی در هر گام تغییر می‌کند؟

برای نقاط یک‌بعدی 1, 2, 8, 9، دو مرکز اولیه ۱ و ۸ انتخاب کنید. نزدیک‌ترین مرکز، دو گروه {1, 2} و {8, 9} می‌دهد. پیش از به‌روزرسانی، مجموع مربع فاصله‌ها $0 + 1 + 0 + 1 = 2$ است. با ثابت ماندن عضویت‌ها، مرکزها به 1.5, 8.5 می‌روند و هدف به $4 \times 0.25 = 1$ کاهش می‌یابد. انتساب بعدی همان گروه‌ها را حفظ می‌کند؛ این اجرا همگراست. «یک گام» را در گزارش دقیق تعریف کنید: انتساب، به‌روزرسانی یا یک چرخه کامل؟

در مقابل، DBSCAN مرکز و میانگین نمی‌سازد. برای نقاط 0, 0.1, 0.2, 0.35, 2، حداقل همسایه سه با احتساب خود نقطه، فقط 0.1 و 0.2 را هسته می‌کند. این دو به هم متصل‌اند؛ صفر و 0.35 مرزی و ۲ نویز است. یک نقطه مرزی اگر نزدیک دو مؤلفه هسته باشد، طبق قرارداد به یکی وصل می‌شود؛ نباید دو مؤلفه را از راه آن ادغام کنیم. این تفاوت دلیل وجود دو مرحله «اتصال هسته‌ها» و «انتساب مرزی‌ها» است.

شکل نمودار را با تعریف فاصله بخوانید

در نمودار اقلیدسی، یک واحد افقی و عمودی باید طول تصویری یکسان داشته باشند تا دایره همسایگی و عمود بودن تصویر PCA درست دیده شود. برای سن و درآمد، محورهای متفاوت دارند و شکل خام به‌تنهایی نمایش فاصله استاندارد شده نیست. عدد فاصله را با قرارداد تبدیل بخوانید. در ناهنجاری هم باید روشن باشد که رنگ، آستانه واقعی است یا فقط رتبه چند نمره بزرگ‌تر؛ رتبه اول بودن لزوماً به معنی خطای داده نیست.

۱۲.۶ خودآزمایی و تمرین

۱. یک چرخه k-means را کامل کنید

برای 1, 2, 8, 9 و مرکزهای اولیه ۱ و ۸، مرکزهای تازه و SSE چیست؟ آیا SSE صفر با چهار مرکز، چهار خوشه را توجیه می‌کند؟

پاسخ و توضیح: مرکزها 1.5, 8.5 و SSE یک است. چهار مرکز می‌تواند SSE را صفر کنند، اما این فقط کاهش هدف آموزش است و تعداد گروه‌های معنادار را ثابت نمی‌کند.

۲. میانه و medoid یکسان‌اند؟

آیا میانه مؤلفه‌ای در دو بعد حتماً یکی از نقاط مشاهده‌شده است؟

پاسخ و توضیح: خیر؛ medoid باید عضو واقعی داده باشد و مجموع ناهمانندی را کمینه کند، ولی میانه مؤلفه‌ای ممکن است مشاهده نشده باشد. روش‌های مبتنی بر میانه و medoid را نباید با یک نام معرفی کرد.

۳. نقطه مرزی را به‌جای هسته استفاده نکنید

چرا در DBSCAN اتصال دو خوشه از راه یک نقطه غیرهسته درست نیست؟

پاسخ و توضیح: گسترش چگالی از هسته انجام می‌شود. نقطه مرزی به هسته دسترسی دارد ولی خودش تراکم کافی برای ادامه زنجیره ندارد؛ استفاده از آن برای ادغام می‌تواند دو ناحیه هسته جدا را نادرست یکی کند.

۴. معیار اتصال چه تفاوتی می‌سازد؟

برای دو گروه $A = \{0, 2\}$ و $B = \{5, 9\}$ ، فاصله، complete single و average چیست؟
پاسخ و توضیح: فاصله‌های جفتی 5, 9, 3, 7 هستند؛ کمینه ۳، بیشینه ۹ و میانگین ۶ است. انتخاب اتصال بر ارتفاع ادغام و گروه‌های نهایی اثر واقعی دارد.

۵. آیا IQR نرمال بودن می‌خواهد؟

برای داده دوقله‌ای، آیا محاسبه حصار IQR ممنوع است؟ آیا هر نقطه داخل حصار عادی است؟
پاسخ و توضیح: محاسبه چارک و حصار به نرمال بودن نیاز ندارد؛ ولی تفسیر حصار به شکل داده و زمینه وابسته است. یک نقطه در حفره میان گروه‌ها ممکن است داخل حصار تک متغیره باشد ولی از همسایگان دور باشد.

۶. غمزه محلی را درست تفسیر کنید

آیا هر $LOF > 1$ یا هر نقطه با رتبه بالاتر باید حذف شود؟
پاسخ و توضیح: خیر؛ LOF نزدیکی چگالی را با همسایگان مقایسه می‌کند و آستانه کاربردی جدا می‌خواهد. رتبه بالا، خطای ثبت یا بی‌ارزش بودن داده را اثبات نمی‌کند؛ ممکن است رخداد مهم یا جمعیتی واقعی باشد.

فصل ۷

تحلیل وابستگی و کشف زیرگروه

منبع مطالعه: اسلاید تحلیل وابستگی، PDF ۳ تا ۴۶؛ کتاب فصل ۴، صفحات چاپی ۲۱۳ تا ۲۷۴ و بخش‌های ۷، ۱ و ۲، ۷. FP-growth و مجموعه‌های بسته بسط کتاب‌اند؛ WRAcc و کشف زیرگروه بر اسلایدها تکیه دارند.

۱.۷ سؤال درست و داده تراکنشی

تحلیل ارتباط می‌پرسد «کدام رخدادها در یک واحد مشاهده کنار هم ظاهر می‌شوند؟» واحد می‌تواند یک سبد خرید، یک نشست وب یا مجموعه علائم یک بیمار باشد. ابتدا باید تعریف تراکنش روشن شود؛ تجزیه همه خریدهای یک مشتری با جدا نگه داشتن هر فاکتور، پشتیبانی و اطمینان متفاوتی می‌دهد (تمرین ۲ کتاب چاپی PDF/۲۹۵/۲۹۶). هر تراکنش مجموعه است؛ تکرار چند واحد شیر در مدل دودویی یک حضور شمرده می‌شود. صفرِ نخریدن هزاران کالای نامرتبط معمولاً از یک خریدن اهمیت کمتری دارد، پس داده دودویی نامتقارن است.

اگر I مجموعه d قلم و T شامل N تراکنش باشد، برای $X \subseteq I$ داریم:

$$\sigma(X) = \sum_{i=1}^N \mathbf{1}[X \subseteq t_i], \quad s(X) = \sigma(X)/N.$$

دو سوی قاعده $X \rightarrow Y$ باید ناتمی و مجزا باشند: $X \cap Y = \emptyset$. تعریف پشتیبانی و اطمینان چنین است:

$$s(X \rightarrow Y) = s(X \cup Y), \quad c(X \rightarrow Y) = \frac{s(X \cup Y)}{s(X)}.$$

اطمینان، احتمال شرطی تجربی تالی در تراکنش‌های دارای مقدم است؛ با پشتیبانی مقدم اشتباه نشود. اگر پشتیبانی مقدم صفر باشد، اطمینان تعریف ندارد. پیکان، علت یا ترتیب زمانی را ثابت نمی‌کند. همبستگی دو ستون دودویی همان ضریب فی است و اگر یکی ثابت باشد تعریف نشده می‌شود.

۲.۷ مثال عددی مرجع، یکسان با اسلاید و کتاب

برای رفع خطای املایی Break در برخی اسلایدها، همه جا Bread منظور است. نمادهای کوتاه این مثال عبارت‌اند از: B نان، M شیر، D پوشک، R نوشیدنی، E تخم‌مرغ و C نوشابه. در این نمایش آموزشی، R معادل قلم نوشیدنی منبع است.

تراکنش	اقلام
۱	$\{B, M\}$
۲	$\{B, D, R, E\}$
۳	$\{M, D, R, C\}$
۴	$\{B, M, D, R\}$
۵	$\{B, M, D, C\}$

شمارش تک‌قلم‌ها به ترتیب چنین است:

$$\sigma(B) = 4, \quad \sigma(M) = 4, \quad \sigma(D) = 4, \quad \sigma(R) = 3, \quad \sigma(C) = 2, \quad \sigma(E) = 1.$$

برای قاعده $\{M, D\} \rightarrow \{R\}$ ، شمارش مقدم ۳، شمارش تالی ۳ و شمارش اجتماع ۲ است. بنابراین:

$$\begin{aligned} s(\{M, D\} \rightarrow \{R\}) &= \frac{2}{5} = 0.4, \\ c(\{M, D\} \rightarrow \{R\}) &= \frac{2}{3} \approx 0.6667, \\ \text{Lift}(\{M, D\} \rightarrow \{R\}) &= \frac{2/3}{3/5} = \frac{10}{9} \approx 1.1111. \end{aligned}$$

قانون معکوس $\{R\} \rightarrow \{M, D\}$ نیز اطمینان $2/3$ دارد، اما این برابری اتفاقی است و اطمینان به‌طور کلی متقارن نیست.

برای قاعده $D \rightarrow R$ داریم:

$$\begin{aligned} s(D \rightarrow R) &= \frac{3}{5} = 0.6, \\ c(D \rightarrow R) &= \frac{3}{4} = 0.75, \\ \text{Lift}(D \rightarrow R) &= \frac{0.75}{0.6} = 1.25. \end{aligned}$$

در جهت $R \rightarrow D$ اطمینان یک است؛ بالابری همان 1.25 باقی می‌ماند. در نتیجه اطمینان جهت‌دار و بالابری متقارن است.

۳.۷ چرا جست و جوی خام جواب نمی دهد؟

برای d قلم، $2^d - 1$ زیرمجموعه ناتهی داریم. از یک مجموعه k عضوی، $2^k - 2$ قاعده با دو سوی ناتهی ساخته می شود. تعداد تمام قواعد ممکن برابر است با

$$3^d - 2^{d+1} + 1.$$

هر قلم در مقدم، تالی یا هیچ کدام است؛ سپس حالت های دارای یک سوی تهی کنار گذاشته می شوند. برای شش قلم، ۶۰۲ قاعده ممکن است. تطبیق خام M نامزد با N تراکنش و پهنای حداکثر w حدود $O(NMw)$ هزینه دارد. رشد نمایی، جست و جوی خام را حتی برای داده متوسط دشوار می کند.

۴.۷ Apriori، منطق و اجرای واقعی

اگر $X \subseteq Y$ باشد، هر تراکنش حاوی Y حتماً X را هم دارد، پس $s(Y) \leq s(X)$. بنابراین اگر X کم تکرار است، تمام ابرمجموعه هایش را می توان بدون شمارش حذف کرد. عکس گزاره برقرار نیست: پرتکرار بودن همه زیرمجموعه ها تضمین پرتکرار بودن اجتماع نیست.

الگوریتم: تک اقلام را بشمار؛ مجموعه پرتکرار F_1 را بساز؛ از F_k نامزد C_{k+1} با اتصال پیشوند مشترک بساز؛ هر نامزدی را که یک زیرمجموعه k عضوی کم تکرار دارد حذف کن؛ با اسکن داده پشتیبانی بازمانده ها را بشمار؛ تکرار تا خالی شدن F . ترتیب واژه نامه ای از نامزد تکراری جلوگیری می کند. حمایت ضدیکنوا شرط هرس است، نه «اعتماد به مشابه بودن نام ها».

با حداقل شمارش ۳، یعنی پشتیبانی ۶۰ درصد، مراحل برای داده مرجع چنین است:

۱. تک اقلام پرتکرار $F_1 = \{B, M, D, R\}$ هستند؛ C, E حذف می شوند.

۲. شمارش شش زوج: $BM = 3, BD = 3, BR = 2, MD = 3, MR = 2, DR = 3$.

۳. بنابراین $F_2 = \{BM, BD, MD, DR\}$ است.

۴. تنها نامزد سه تایی BMD است، چون همه زوج هایش پرتکرارند. شمارش واقعی آن ۲ است و حذف می شود.

۵. F_3 تهی و فرایند تمام می شود. در این اجرا $1 + 6 + 6 = 13$ نامزد شمرده شده، درحالی که تعداد مجموعه های ناتهی تا اندازه سه ۴۱ و تمام اندازه ها ۶۳ است.

حداقل شمارش از پشتیبانی نسبی باید $\lceil N \times \text{minsup} \rceil$ باشد؛ مثلاً $N = 5$ و $\text{minsup} = 0.5$ ، حداقل شمارش ۳ است. مرز «بزرگ تر یا مساوی» را یکسان رعایت کنیم. کاهش minsup ، افزایش تعداد اقلام، تعداد تراکنش ها و پهنای سبد می تواند هزینه را بالا ببرند. در داده متراکم الگوهای بلند بسیار زیادند و هیچ الگوریتمی هزینه خروجی نمایی را جادویی حذف نمی کند.

۵.۷ تولید قواعد و هرس درست اطمینان

از هر مجموعه پرتکرار L ، افزازهای $X \rightarrow L \setminus X$ تولید می‌شوند. شمارش زیرمجموعه‌ها از مرحله قبل موجود است. اطمینان در تغییر آزاد مقدم، ضدیکنوا نیست؛ اما با ثابت ماندن اجتماع L ، بزرگ‌تر کردن تالی، مقدم را کوچک‌تر و پشتیبانی مقدم را بزرگ‌تر یا مساوی می‌کند. صورت کسر ثابت است، پس اطمینان افزایش نمی‌یابد. در نتیجه، اگر یک تالی آستانه اطمینان را نگذرد، تالی‌های بزرگ‌تر شامل آن قابل هرس‌اند. این حکم را نمی‌توان به قواعدی با اجتماع متفاوت تعمیم داد.

نمونه $L = \{M, D, R\}$ را در آستانه پشتیبانی ۴۰ درصد در نظر بگیرید. اطمینان شش قاعده چنین است:

$$\begin{aligned} c(\{M, D\} \rightarrow R) &= \frac{2}{3}, & c(\{M, R\} \rightarrow D) &= 1, \\ c(\{D, R\} \rightarrow M) &= \frac{2}{3}, & c(M \rightarrow \{D, R\}) &= \frac{1}{2}, \\ c(D \rightarrow \{M, R\}) &= \frac{1}{2}, & c(R \rightarrow \{M, D\}) &= \frac{2}{3}. \end{aligned}$$

اگر $\text{minconf} = 0.7$ باشد، فقط $\{M, R\} \rightarrow D$ موفق است؛ نباید L را با آستانه ۶۰ درصد مثال قبل پرتکرار معرفی کرد.

۶.۷ جذابیت، نرخ پایه و Lfit

اطمینان بالا به تنهایی همراهی مثبت را نشان نمی‌دهد. معیار بالابری یا Lfit، همراهی را نسبت به نرخ پایه می‌سنجد:

$$\text{Lift}(X \rightarrow Y) = \frac{c(X \rightarrow Y)}{s(Y)} = \frac{s(X \cup Y)}{s(X)s(Y)}.$$

مقدار بیشتر از یک همراهی مثبت تجربی، مقدار یک استقلال در جدول تجربی و مقدار کمتر از یک همراهی منفی را بیان می‌کند. این معیار آزمون معنی‌داری یا سند علیت نیست؛ چند مشاهده نادر می‌توانند مقدار بسیار بالا و ناپایداری تولید کنند.

مثال اسلایدهای ۳۲ تا ۳۵. جدول چای و قهوه، با ترتیب خانه‌های $[a, b; c, d]$ ، شامل $[3, 1; 15, 1]$ و جمعیت $N = 20$ است. برای قاعده چای به قهوه داریم:

$$\begin{aligned} P(\text{coffee}) &= \frac{18}{20} = 0.9, \\ c(\text{tea} \rightarrow \text{coffee}) &= \frac{3}{4} = 0.75, \\ s(\text{tea} \rightarrow \text{coffee}) &= \frac{3}{20} = 0.15, \\ \text{Lift} &= \frac{0.75}{0.9} = \frac{5}{6} \approx 0.8333. \end{aligned}$$

دانستن چای‌نوش بودن احتمال قهوه را از ۹۰ درصد به ۷۵ درصد کاهش می‌دهد. مثال مستقل گلاب، صفحات چاپی ۲۵۹ تا ۲۶۱. جدول $[150, 50; 650, 150]$ ، با $N = 1000$ ، نرخ پایه قهوه ۸۰ درصد و اطمینان ۷۵ درصد دارد؛ بالابری ۰.۹۳۷۵ است. این را با اعداد اسلاید مخلوط نکنید. مثال چای

و عسل کتاب، جدول $[100, 100; 20, 780]$ ، نرخ پایهٔ عسل ۱۲ درصد، اطمینان ۵۰ درصد و بالابری تقریباً 4.1667 دارد: اطمینان کمتر ولی ارتباط مثبت قوی تر.

افزونی: اگر افزودن شرطی به مقدم هیچ پوشش/اطمینان جدیدی ندهد، قاعدهٔ طولانی تر ممکن است اطلاعات افزوده نداشته باشد. معیار عددی با فایدهٔ عملی، تازگی و دانش دامنه تکمیل می‌شود؛ بالاترین Lfit همیشه بهترین تصمیم نیست.

۷.۷ بسط کنترل‌شدهٔ کتاب: نمایش فشرده و FP-growth

مجموعهٔ پرتکرار بیشینه هیچ ابرمجموعهٔ پرتکراری ندارد؛ مجموعهٔ بسته هیچ ابرمجموعهٔ دارای همان پشتیبانی ندارد. همهٔ مجموعه‌های پرتکرار بیشینه بسته‌اند، ولی عکس درست نیست. بیشینه‌ها وجود الگوهای پرتکرار کوچک تر را بیان می‌کنند اما شمارش پشتیبانی آن‌ها را حفظ نمی‌کنند؛ مجموعه‌های بسته همراه شمارش می‌توانند شمارش تمام مجموعه‌های پرتکرار را بازسازی کنند.

در نمایش عمودی برای هر قلم فهرست شناسهٔ تراکنش‌ها را نگه می‌داریم و اشتراک فهرست‌ها شمارش اجتماع را می‌دهد. این به Jaccard شباهت ظاهری دارد اما هدف الگوریتم شمارش پشتیبانی است.

FP-growth به جای تولید نامزدهای همهٔ سطوح، داده را در FP-tree فشرده می‌کند: گذراول شمارش و حذف تک اقلام کم‌تکرار؛ مرتب‌سازی مشترک اقلام باقی‌مانده (معمولاً نزولی پشتیبانی با شکست تساوی ثابت)؛ گذر دوم درج هر سبد در مسیر درخت و افزایش شمارندهٔ پیشوند مشترک. جدول سرآیند و پیوند گره‌های هم‌نام دسترسی به مسیرها را آسان می‌کند. سپس برای هر پسوند، مسیرهای پیشوندی با وزن شمارندهٔ گره همان پسوند استخراج می‌شوند؛ این پایگاه الگوی شرطی است. درخت شرطی با حذف قلم پسوند و اقلام کم‌تکرار ساخته و فرایند بازگشتی تکرار می‌شود. شمارندهٔ والد را نباید به اشتباه وزن همهٔ مسیرهای شرطی گرفت. مسیر تهی برای ساخت الگوی بلندتر چیزی اضافه نمی‌کند. شمارندهٔ یک گره لزوماً پشتیبانی کلی قلمش نیست؛ ممکن است چند گره با همان برچسب وجود داشته باشد. فشرده‌سازی و کارایی به اشتراک پیشوندها وابسته است؛ مرتب‌سازی نزولی الزاماً کمینه‌ترین درخت ممکن را تضمین نمی‌کند. این بخش بسط تکمیلی کتاب است.

پایگاه شرطی را واقعاً بسازیم

در همان پنج سبد و حداقل شمارش ۳، اقلام C, E حذف می‌شوند. برای شکست تساوی، ترتیب مشترک B, D, M, R را ثابت بگیرید؛ فراوانی سه قلم اول برابر ۴ و قلم آخر ۳ است. سبدهای مرتب، به ترتیب BM ، BDR ، DMR ، $BDMR$ هستند. مسیرهای منتهی به R ، پس از حذف خود این قلم، پایگاه شرطی زیر را می‌دهند:

$$(BD, 1), \quad (BDM, 1), \quad (DM, 1).$$

وزن‌ها از شمارندهٔ گره R هر مسیر آمده‌اند. در این پایگاه، شمارش‌های شرطی $B = 2, D = 3, M = 2$ هستند؛ با حداقل شمارش ۳ فقط D می‌ماند. بنابراین الگوی همراه R که از این شاخه به دست می‌آید DR با شمارش ۳ است. حذف R از مسیر، به معنی حذف وزن تراکنش‌های منتهی به آن نیست. این اجرای کوتاه، تفاوت «شمارندهٔ یک گره» و «پشتیبانی کل یک قلم» را نیز روشن می‌کند.

۸.۷ شباهت و معیارهای تکمیلی

برای جدول دو متغیر دودویی، $a = f_{11}, b = f_{10}, c = f_{01}, d = f_{00}$ و $N = a + b + c + d$ است. چند معیار پرکاربرد عبارت‌اند از:

$$J = \frac{a}{a + b + c}, \quad \cos = \frac{a}{\sqrt{(a+b)(a+c)}},$$

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}, \quad \text{Lift} = \frac{Na}{(a+b)(a+c)}.$$

اختلاف همراهی از استقلال نیز $\text{leverage} = a/N - [(a+b)/N][(a+c)/N]$ است. در هر معیار، حالت مخرج صفر باید جداگانه مدیریت شود.

Jaccard و cosine با افزودن تراکشن‌هایی که هیچ کدام از دو قلم را ندارند تغییر نمی‌کنند، ولی بالابری تغییر می‌کند. در مثال کتاب، صفحات چاپی ۲۶۹ تا ۲۷۰، جدول $[700, 100; 100, 100]$ را به $[700, 100; 100, 1100]$ تبدیل می‌کنیم. نتایج عبارت‌اند از:

$$\cos = 0.875, \quad J = \frac{7}{9},$$

$$\text{Lift}_{\text{before}} = 1.09375, \quad \text{Lift}_{\text{after}} = 2.1875.$$

خود متن کتاب در این مثال، بالابری‌های ۰.۹۷۲ و ۱.۹۴۴ را چاپ کرده که با جدول ناسازگارند؛ از محاسبه مستقیم استفاده شود. مخرج صفر برای هر معیار باید مدیریت شود. شباهت بدون توجه به صفرهای مشترک برای سبد خرید معمولاً معنادارتر است؛ برای ویژگی دودویی مقارن، صفر نیز اطلاعات دارد.

پارادوکس سیمپسون (بسط، چاپی ۲۷۲-۲۷۴ PDF/۲۷۵-۲۷۳): در داده تجمیعی خریداران تلویزیون ۵۵٪ و غیرخریداران ۴۵٪ ابزار ورزشی می‌خرند؛ اما بین دانشجویان $4/34 < 1/10$ و بین شاغلان $98/170 < 50/86$ است. ترکیب متفاوت گروه‌ها علامت ارتباط را عوض می‌کند. تفسیر رابطه باید متغیرهای محدودش‌کننده و تعریف جامعه را در نظر بگیرد.

۹.۷ ویژگی‌های اسمی، پیوسته و کشف زیرگروه

ویژگی اسمی به اقلام «نام ویژگی=مقدار» تبدیل می‌شود. ویژگی تک مقداری مانند مرورگر نمی‌تواند هم‌زمان دو مقدار داشته باشد؛ نامزدهای شامل دو مقدار ناسازگار همان ویژگی را از ابتدا حذف کنیم. مقادیر نادر را فقط با توجه معنایی یا گروه «سایر» تجمیع کنیم. قلم بسیار رایج ممکن است قواعد بدیهی بسازد؛ حذف کورکورانه آن می‌تواند رابطه معنادار را هم از بین ببرد، پس هدف و نرخ پایه مهم‌اند.

ویژگی پیوسته را می‌توان با بازه‌های هم‌عرض یا هم‌فراوانی گسسته کرد. بازه خیلی باریک حمایت کافی ندارد؛ بازه خیلی پهن ممکن است گروه‌های متفاوت را مخلوط و اطمینان را تضعیف کند. مرز و شمول ابتدا/انتهای روشن باشد، مثلاً $[5, 10]$. مثال کتاب با سن کاربران و چت: گروه ۱۶-۲۴ شامل ۲۲ پاسخ مثبت از ۲۷ نفر در کل ۲۵۰ مشاهده است؛ $s = 22/250 = 0.088$ و $c = 22/27 \approx 0.8148$. شکستن به ۱۶-۲۰ و ۲۰-۲۴ هرکدام حمایت $11/250 = 0.044$ دارد؛ ادغام در بازه خیلی پهن ۱۲-۳۶ اطمینان $75/130 \approx 0.5769$ می‌دهد. اثر مرزبندی را می‌توان در آزمایشگاه بررسی کرد.

کشف زیرگروه میان توصیف و پیش‌بینی است: هدف، یافتن توصیف‌های قابل فهم از گروه‌هایی با رفتار متفاوت هدف T است، نه فقط کمینه کردن خطای یک طبقه‌بند جامع. زیرگروه‌ها می‌توانند هم‌پوشان باشند و لزومی ندارد همهٔ جمعیت را بپوشانند. در اسلاید ۴۱ توصیف کوتاه «خرطوم دارد» و توصیف طولانی تر فیل مطرح می‌شود؛ برای پیش‌بینی سادگی و تعمیم مهم است، برای شناخت داده توصیف‌های متعدد می‌توانند مفید باشند. الگوریتم‌های مطرح شده: استخراج قواعد از درخت هرس‌نشده و رتبه‌بندی، (EXPLORA/MIDOS) روش‌های مبتنی بر Apriori-SD، روش‌های تکاملی. این روش‌ها مسیرهای متفاوت جست‌وجوی زیرگروه‌اند؛ امتیاز و فضای توصیف، نتیجه را تعیین می‌کند.

اگر S زیرگروه و T رویداد هدف باشد، معیار صحت نسبی وزن دار چنین است:

$$\text{WRAcc}(S, T) = P(S \cap T) - P(S)P(T) = P(S)\{P(T | S) - P(T)\}.$$

امتیاز هم پوشش و هم افزایش نرخ هدف نسبت به پایه را می‌سنجد. گروه بسیار کوچک و خالص ممکن است امتیاز کمتری از گروه بزرگ تر با خلوص متوسط بگیرد. استقلال امتیاز صفر و نرخ کمتر از پایه امتیاز منفی می‌دهد. اگر $p = P(T)$ ، کران نظری $p(1 - p)$ در $S = T$ حاصل می‌شود؛ ممکن است هیچ توصیف قابل استفاده‌ای دقیقاً این گروه را بیان نکند. خود WRAcc ضدیکنوا نیست؛ کم بودن امتیاز یک مقدم، بدون کران معتبر، مجوز هرس همهٔ تخصیصی‌سازی‌های آن نیست.

مثال ۱۰۷ | کشف زیرگروه در مثال فیل

شمارش‌ها $TP = 1894, FP = 0, FN = 32, TN = 54874$ و جمعیت $N = 56800$ است. خلوص یک است و

$$P(T) = 1926/56800 \approx 0.03391,$$

$$P(S) = 1894/56800 \approx 0.03335,$$

$$\text{WRAcc} \approx 0.03221.$$

کران نظری $P(T)(1 - P(T)) \approx 0.03276$ است. ابتدا از شمارش‌های دقیق استفاده کنید و در پایان گرد کنید.

۱۰۷ خودآزمایی و تمرین

پیش از دیدن پاسخ، شمارش‌ها و استدلال خود را بنویسید.

۱. شمارش با احتمال شرطی چه فرقی دارد؟

در سبدهای مرجع، قاعدهٔ نان → شیر را بررسی کنید. پشتیبانی قاعده، اطمینان و Lift را به دست آورید و بگویید آیا دیدن نان احتمال شیر را افزایش داده است؟
پاسخ و توضیح: شمارش نان ۴، شیر ۴ و اجتماع ۳ است:

$$s = \frac{3}{5} = 0.6, \quad c = \frac{3}{4} = 0.75, \quad \text{Lift} = \frac{0.75}{0.8} = 0.9375.$$

پس با وجود اطمینان ۷۵ درصد، احتمال شیر از ۸۰ درصد به ۷۵ درصد کاهش یافته است. اطمینان را باید با نرخ پایه مقایسه کرد.

۲. آیا همه زوج‌های پرتکرار، سه‌تایی پرتکرار می‌سازند؟

در داده مرجع و با حداقل شمارش ۳، مجموعه‌های پرتکرار هر سطح و نامزد سه‌تایی را پیدا کنید. آیا پرتکرار بودن تمام زوج‌ها کافی است؟ پاسخ و توضیح:

(آ) تک‌اقلام پرتکرار $F_1 = \{B, M, D, R\}$ هستند؛ C, E حذف می‌شوند.

(ب) شمارش شش زوج: $BM = 3, BD = 3, BR = 2, MD = 3, MR = 2, DR = 3$.

(ج) بنابراین $F_2 = \{BM, BD, MD, DR\}$ است.

(د) تنها نامزد سه‌تایی BMD است، چون همه زوج‌هایش پرتکرارند. شمارش واقعی آن ۲ است و حذف می‌شود.

(ه) F_3 تهی و فرایند تمام می‌شود. در این اجرا $6 + 6 + 1 = 13$ نامزد شمرده شده، درحالی که تعداد مجموعه‌های ناتمی تا اندازه سه ۴۱ و تمام اندازه‌ها ۶۳ است.

BMD فقط در تراکنش ۴ و ۵ ظاهر می‌شود، پس شمارش ۲ دارد. اصل Apriori می‌گوید پرتکراری مجموعه، پرتکراری زیرمجموعه‌ها را لازم می‌کند؛ شرط کافی نمی‌دهد. هنگام تولید نامزد هنوز باید شمارش واقعی انجام شود.

۳. آستانه نسبی و گسستگی شمارش

در ۵ تراکنش اگر حداقل حمایت ۵۰٪ باشد، چه حداقل شمارشی لازم است؟ آیا مجموعه دارای شمارش ۲ پذیرفته می‌شود؟

پاسخ و توضیح: $[5 \times 0.5] = 3$. شمارش ۲ حمایت ۴۰٪ دارد و رد می‌شود. آستانه با گرد کردن معمولی به ۲ تبدیل نمی‌شود.

۴. تعداد قواعد

از یک مجموعه پرتکرار چهارعضوی چند قاعده می‌توان ساخت؟ چرا عدد ۱۶ صحیح نیست؟

پاسخ و توضیح: $2^4 - 2 = 14$ ؛ انتخاب مقدم تهی و مقدم کل مجموعه، یکی از دو سوی قاعده را تهی می‌کنند و کنار گذاشته می‌شوند. قواعد متفاوت می‌توانند پشتیبانی یکسان و اطمینان متفاوت داشته باشند.

۵. هرس اعتماد با شرط صحیح

از مجموعه ثابت $L = \{A, B, C, D\}$ ، قاعده $ABC \rightarrow D$ اطمینان ۰.۶ دارد و $\text{minconf} = 0.7$ است. درباره $AB \rightarrow CD$ و $A \rightarrow BCD$ چه می‌دانیم؟ آیا همین منطق برای $AB \rightarrow E$ هم قابل استفاده است؟

پاسخ و توضیح: دو قاعده اول اطمینانی حداکثر ۰.۶ دارند و قابل هرس‌اند؛ اجتماع اقلام هر سه قاعده L ثابت است و کوچک کردن مقدم مخرج را کاهش نمی‌دهد. درباره $AB \rightarrow E$ چنین نتیجه‌ای نداریم، چون اجتماع عوض شده است.

۶. معکوس قاعده

در داده مرجع $c(D \rightarrow R) = 0.75$ ولی $c(R \rightarrow D) = 1$ است. آیا تناقض است؟ بالابری دو جهت را محاسبه کنید.

پاسخ و توضیح: نه؛ جمعیت شرطی جهت اول ۴ و جهت دوم ۳ تراکنش است. بالابری در دو جهت برابر است:

$$\frac{0.75}{0.6} = 1.25, \quad \frac{1}{0.8} = 1.25.$$

نرخ پایه تفاوت جهت را خنثی کرده است.

۷. برداشتن یک قلم کم تکرار

اگر E فقط در یک تراکنش حضور دارد و حداقل شمارش ۳ است، آیا لازم است همه مجموعه های شامل E را تک تک بشماریم؟

پاسخ و توضیح: خیر؛ هر مجموعه شامل E شمارش حداکثر ۱ دارد. با شش قلم، $2^5 = 32$ مجموعه شامل E وجود دارد و می توان آن ها را یکجا با اصل ضدیکنوایی حذف کرد (در شمارش اجرای خاص، برخی ممکن است از قبل به دلیل اقلام دیگر هرس شده باشند؛ تعداد صرفه جویی را دو بار بشماریم).

۸. خلوص در برابر پوشش

در جامعه ۱۰۰ نفری با ۲۰ عضو هدف، گروه S_1 شامل ۵ نفر است و هر ۵ نفر هدف اند. گروه S_2 شامل ۲۰ نفر است و ۱۲ نفر هدف اند. کدام اطمینان و کدام WRAcc بیشتری دارد؟

پاسخ و توضیح: گروه اول پوشش ۰.۰۵ و خلوص یک دارد؛ بنابراین $WRAcc = 0.04$. گروه دوم پوشش ۰.۲ و خلوص ۰.۶ دارد و امتیازش ۰.۰۸ است. اولی خالص تر و دومی از دید این معیار جذاب تر است؛ زیرا افزایش نرخ هدف را برای جمعیت بیشتری ایجاد می کند.

۹. درخت FP و پشتیبانی

در FP-tree قلم A در دو گره با شمارنده های ۳ و ۲ ظاهر شده است. آیا می توان گفت پشتیبانی شمارشی A برابر ۳ است؟ برای ساخت پایگاه شرطی پسوند A از وزن کدام گره استفاده می کنیم؟

پاسخ و توضیح: شمارش کل A جمع ۵ است. هر مسیر پیشوندی باید با شمارنده گره A همان مسیر وزن بگیرد، نه شمارنده والد یا ریشه. در غیر این صورت تراکنش های فاقد پسوند به اشتباه شمرده می شوند.

۱۰. مجموعه بسته و پیشینه

فرض کنید شمارش های $\sigma(A) = 5$ ، $\sigma(AB) = 5$ و $\sigma(ABC) = 3$ است و حداقل شمارش ۰.۳. آیا A بسته است؟ آیا AB پیشینه است؟

پاسخ و توضیح: A بسته نیست، چون AB ابرمجموعه ای با همان شمارش دارد. AB پیشینه نیست چون ABC نیز پرتکرار است؛ ممکن است AB بسته باشد، مشروط به اینکه هیچ ابرمجموعه دیگر با شمارش ۵ نداشته باشد. بسته بودن و پیشینه بودن دو سؤال متفاوت اند.

۱۱. صفرهای مشترک

جدول دو قلم $[a, b; c, d] = [70, 10; 10, 10]$ است. ۱۰۰ تراکنش فاقد هر دو قلم اضافه کنیم. چه چیزی درباره Jaccard، cosine و Lfit تغییر می کند؟

پاسخ و توضیح: مقادیر $J = 70/90 = 7/9$ و $\cos = 70/80 = 0.875$ ثابت‌اند. جمعیت N از ۱۰۰ به ۲۰۰ می‌رسد و بالابری دو برابر می‌شود:

$$\text{Lift}_{\text{before}} = \frac{100 \times 70}{80 \times 80} = 1.09375,$$

$$\text{Lift}_{\text{after}} = \frac{200 \times 70}{80 \times 80} = 2.1875.$$

پس انتخاب معیار با تعریف جامعه و اهمیت غیاب مشترک پیوند دارد.

۰۱۲. چرا تغییر مرز گسسته‌سازی مهم است؟

گروه سن ۱۶-۲۴ در نمونه ۲۵۰ نفری، ۲۲ پاسخ مثبت از ۲۷ نفر دارد. اگر آن را به دو بازه تقسیم کنیم که هر کدام ۱۱ پاسخ مثبت دارند، در حداقل حمایت ۵٪ چه رخ می‌دهد؟

پاسخ و توضیح: گروه اصلی حمایت ۸,۸٪ دارد؛ هر زیرقاعده حمایت ۴,۴٪ دارد و رد می‌شود، حتی اگر اطمینان هر کدام بالا باشد. گسسته‌سازی صرفاً آرایش نمایش نیست؛ فضای قواعد قابل کشف را تغییر می‌دهد.

پیوست آ

واژه‌نامه، نقشه منابع و مسیر تمرین

آ.۱ واژه‌نامه کلیدی

اصطلاح	معنی و تمایز اصلی
Feature / Target Parameter / Hyperparameter Training / Validation / Test Leakage	ویژگی قابل استفاده برای پیش‌بینی، در برابر پاسخ مورد پیش‌بینی. پارامتر از داده برازش می‌شود؛ ابرپارامتر شیوه آموزش یا ظرفیت را تعیین می‌کند. برازش، انتخاب تنظیم و ارزیابی مستقل سه نقش جدا هستند. ورود اطلاعاتی به آموزش که هنگام استفاده واقعی یا ارزیابی مستقل مجاز نیست.
Accuracy / Precision	سهم کل پیش‌بینی‌های درست، در برابر سهم هشدارهای مثبت درست.
Recall / Specificity	سهم مثبت‌های کشف‌شده، در برابر سهم منفی‌های درست تشخیص داده‌شده.
Bias / Variance	خطای نظام‌مند میانگین مدل‌ها، در برابر حساسیت مدل به نمونه آموزش.
Centroid / Medoid Support / Confidence / Lift Correlation / Causation	میانگین خوشه، در برابر یک عضو واقعی با کمترین مجموع ناهمانندی. فراوانی اجتماع، احتمال شرطی تجربی و نسبت همراهی به استقلال. همراهی آماری، در برابر اثر تغییر با مداخله؛ اولی دومی را اثبات نمی‌کند.

آ.۲ تطبیق فصل‌ها با منابع

فصل	اسلاید	بسط کتاب
۱	مقدمه، ۱۹-۵۸؛ ۱-۱۸ اداری	فصل ۱، ۲۴-۳۳؛ بخش ۳، ۶، ۴۱۰-۴۱۳
۲	پیش‌پردازش، ۳-۵۹	فصل ۲، ۴۶-۱۱۹؛ بخش ۱۱، ۶
۳	طبقه‌بندی ۱، ۳-۶۰؛ طبقه‌بندی ۲، ۴۲-۶۱	فصل ۳، ۱۳۴-۱۹۳؛ بخش ۱۰، ۶، ۴۹۸-۵۱۴
۴	طبقه‌بندی ۲، ۳-۴۱	بیز: ۴۱۴-۴۲۸؛ شبکه: ۴۵۱-۴۷۷؛ SVM: ۴۷۸-۴۹۷
۵	رگرسیون، ۳-۵۵	تعمیم: ۱۶۷-۱۹۳؛ اریبی-واریانس: ۵۰۲-۵۰۴
۶	خوشه‌بندی، ۳-۶۴	۵۰۴. رگرسیون تفصیلی از اسلاید و بسط ریاضی است.
۷	وابستگی، ۳-۴۶	فصل ۵، ۳۰۷-۳۷۸؛ ناهنجاری فصل ۹، ۷۲۳-۷۵۱؛ فاصله بخش ۴، ۲
		فصل ۴، ۲۱۳-۲۷۴؛ گسسته‌سازی و اقلام اسمی بخش‌های ۷، ۲-۷

۳.۴. مسیر پیشنهادی تمرین و پروژه

۲. یک تقسیم مستقل و متناسب با زمان یا گروه بسازید. پیش‌پردازش را فقط با آموزش پیامزید و یک خط منای ساده بسازید.

۴. برای پاسخ عددی، خط پایه، رگرسیون خطی و یک مدل انعطاف پذیر را با MAE و RMSE مقایسه کنید؛ منطقه ی رونمایی را مشخص کنید.

۵. برای خوشه‌بندی، دو تعریف متفاوت خوشه، دو مقیاس‌بندی و بیش از یک شروع اولیه را بررسی کنید. خوشه‌ها را فقط با زیبایی نمودار قضاوت نکنید.
۶. برای قواعد، شمارش، پشتیبانی، اطمینان و نرخ پایه را هم‌زمان گزارش کنید. رابطه جالب را روی داده مستقل یا با تحلیل پایداری بررسی کنید.
۷. تنظیمات، بذر تصادفی، شناسه تقسیم‌ها و نسخه داده را ثبت کنید. هیچ انتخابی را بر اساس آزمون نهایی انجام ندهید.

آ.۴ چک‌لیست فهم مفهومی

اگر برای هر روش بتوانید «ورودی چیست؟ خروجی چیست؟ چه چیزی بهینه می‌شود؟ چه فرضی لازم است؟ چه تنظیمی دارد؟ چگونه ارزیابی می‌شود؟ در چه داده‌ای شکست می‌خورد؟» را پاسخ دهید، از حفظ نام الگوریتم فراتر رفته‌اید. فرمول را همراه با واحد، دامنه، حالت مخرج صفر و قرارداد مقیاس توضیح دهید. در آزمایشگاه، پیش از حرکت کنترل‌ها نتیجه را حدس بزنید و پس از مشاهده، تفاوت را با فرض‌های مدل تفسیر کنید.