دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گزارش پروژه درس داده کاوی محاسباتی

پروژه ۲

نگارش

محمدصادق قلی زاده

استاد راهنما

دکتر مهدی قطعی


بسم الله الرحمن الرحيم

صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.

دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)

به نام خدا

تعهدنامه اصالت اثر

تاریخ:

اینجانب **محمدصادق قلی زاده** متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مآخذ بلامانع است.

محمدصادق قلی زاده

امضا

نویسنده پایان نامه، در صورت تمایل میتواند برای سپاسگزاری پایان نامه خود را به شخص یا اشخاص و یا ارگان خاصی تقدیم نماید.

سپاس‌گزاری

نویسنده پایان‌نامه می‌تواند مراتب امتنان خود را نسبت به استاد راهنما و استاد مشاور و یا دیگر افرادی که طی انجام پایان‌نامه به نحوی او را یاری و یا با او همکاری نموده‌اند ابراز دارد.

محمدصادق قلی‌زاده

چکیده

این مطالعه به بررسی نقش حیاتی مهندسی ویژگی در بهینه‌سازی عملکرد الگوریتم‌های داده‌کاوی در حوزه‌های طبقه‌بندی، رگرسیون و خوشه‌بندی می‌پردازد. با استفاده از سه مجموعه‌داده‌ی معیار، سرطان پستان ویسکانسین (طبقه‌بندی)، مسکن بوستون (رگرسیون) و آیریس UCI (خوشه‌بندی)، این گزارش تأثیر هم‌خطی چندگانه و بُعد بالای داده‌ها را بر پایداری و دقت مدل‌ها تحلیل می‌کند. روش‌شناسی پژوهش در مراحل متمایز پیش می‌رود: نخست، شناسایی وابستگی‌های ویژگی‌ها از طریق تحلیل همبستگی؛ دوم، به‌کارگیری تکنیک‌های استخراج ویژگی (PCA، ICA، SVD) بر اساس واریانس تبیین‌شده؛ و سوم، پیاده‌سازی روش‌های انتخاب ویژگی (SelectKBest، RFE) به‌منظور جداسازی پیش‌بین‌های با ارزش بالا. در نهایت، کارایی این فضاهای ویژگیِ تبدیل‌شده بر روی مدل‌های مشخص یادگیری ماشین، از جمله رگرسیون خطی، SGDRegressor، K-Means، K-نزدیک‌ترین همسایه‌ها (KNN) و جنگل‌های تصادفی ارزیابی می‌شود. نتایج تجربی نشان می‌دهد که اگرچه کاهش بُعد (PCA) به‌طور قابل‌توجهی سرعت همگرایی را برای بهینه‌سازی مبتنی بر مشتق و کارایی محاسباتی را برای مدل‌های مبتنی بر فاصله (KNN، K-Means) بهبود می‌بخشد، ممکن است با پنهان‌سازی مرزهای ویژگی، عملکرد مجموعه‌های مبتنی بر درخت (جنگل‌های تصادفی) را تضعیف کند. در مقابل، انتخاب ویژگی (RFE) در حفظ قابلیت تفسیر مدل و دقت در مسائل رگرسیونی که به‌شدت تحت تأثیر عوامل اجتماعی-اقتصادی خاص هستند، برتری نشان داد. لینک گیت هاب پروژه: GitHub

واژه‌های کلیدی:

داده‌کاوی، کاهش بُعد، PCA، انتخاب ویژگی، هم‌خطی چندگانه، بهینه‌سازی مدل، رگرسیون، خوشه‌بندی، طبقه‌بندی.

فهرست مطالب

عنوان صفحه

فهرست اشکال

شکل صفحه

فهرست جداول

جدول صفحه

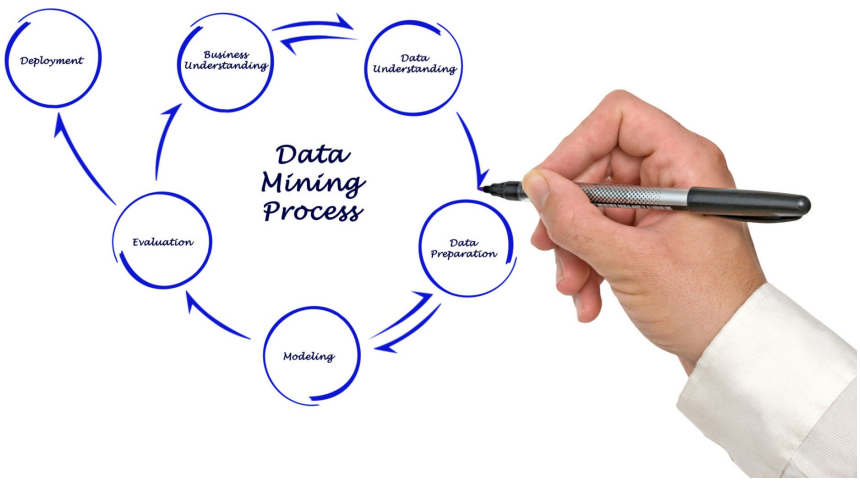
فهرست نمادها

نماد	مفهوم
$\mathbb{R}^n$	فضای اقلیدسی با بعد n
$\mathbb{S}^n$	کره یکه n بعدی
M^m	خمینه m-بعدی M
$\mathfrak{X}(M)$	جبر میدان‌های برداری هموار روی M
$\mathfrak{X}^1(M)$	مجموعه میدان‌های برداری هموار یکه روی (M,g)
$\Omega^p(M)$	مجموعه p-فرمی‌های روی خمینه M
Q	اپراتور ریچی
$\mathcal{R}$	تانسور انحنای ریمان
ric	تانسور ریچی
L	مشتق لی
Φ	۲-فرم اساسی خمینه تماسی
∇	التصاق لوی-چویتای
Δ	لاپلاسین ناهموار
∇^*	عملگر خودالحاق صوری القا شده از التصاق لوی-چویتای
g_s	متر ساساکی
∇	التصاق لوی-چویتای وابسته به متر ساساکی
Δ	عملگر لاپلاس-بلترامی روی p-فرم‌ها

فصل ۱

مقدمه

در حوزه‌ی داده‌کاوی و یادگیری ماشین، کیفیت داده‌های ورودی اغلب نقش مهم‌تری در موفقیت پیش‌بینی نسبت به پیچیدگی خودِ مدل ایفا می‌کند. مجموعه‌داده‌های دنیای واقعی معمولاً از «نفرین بُعد» رنج می‌برند؛ پدیده‌ای که داده‌های با بُعد بالا، پراکنده یا مملو از ویژگی‌های افزونه (هم‌خطی چندگانه) را توصیف می‌کند. این مسائل می‌توانند منجر به بیش‌برازش، افزایش هزینه‌های محاسباتی و ناپایداری در پیش‌بینی‌های مدل شوند.



شکل ۱-۱: داده کاوی

این گزارش یک تحلیل جامع از تکنیک‌های مهندسی ویژگی ارائه می‌دهد که با هدف کاهش این چالش‌ها طراحی شده‌اند. هدف اصلی، ارزیابی این موضوع است که کاهش فضای ویژگی چه از طریق

استخراج ویژگی (تبدیل داده‌ها به ابعاد جدید) و چه از طریق انتخاب ویژگی (حفظ بهترین ابعاد اصلی) چه تأثیری بر عملکرد الگوریتم‌های مختلف یادگیری ماشین دارد.

۱-۱ اهداف

این مطالعه در پنج مرحله‌ی کلیدی سازمان‌دهی شده است:

بررسی مجموعه‌داده‌ها: تحلیل سه مجموعه‌داده‌ی معیار که نماینده‌ی انواع مختلف مسائل هستند: طبقه‌بندی، رگرسیون و خوشه‌بندی.

تحلیل وابستگی‌ها: شناسایی هم‌خطی چندگانه با استفاده از ماتریس‌های همبستگی و نقشه‌های حرارتی [۱].

کاهش بُعد: به‌کارگیری الگوریتم‌هایی مانند تحلیل مؤلفه‌های اصلی [۲]، تحلیل مؤلفه‌های مستقل [۳] و تجزیه مقدار منفرد (SVD) برای فشرده‌سازی فضای ویژگی همراه با حفظ واریانس.

انتخاب ویژگی: استفاده از روش‌های فیلتری [۴] و روش‌های پوششی (حذف بازگشتی ویژگی‌ها RFE) به‌منظور شناسایی مهم‌ترین پیش‌بین‌ها.

ارزیابی مدل: تحلیل مقایسه‌ای عملکرد مدل‌ها (رگرسیون خطی، SGD، K-Means، KNN و جنگل تصادفی) بر روی داده‌های اولیه، استخراج‌شده و انتخاب‌شده.

۱-۲ مروری بر مجموعه‌داده‌ها

برای اطمینان از یک ارزیابی مستحکم، این مطالعه از سه مجموعه‌داده‌ی عمومی و شناخته‌شده استفاده می‌کند:

سرطان پستان ویسکانسین : یک مسئله‌ی طبقه‌بندی دودویی با بُعد بالا که بر تشخیص پزشکی تمرکز دارد.

مسکن بوستون: یک مسئله‌ی رگرسیونی که رابطه‌ی بین عوامل اقتصادی و ارزش املاک را بررسی می‌کند.

---

[۱] Heatmap

[۲] PCA

[۳] ICA

[۴] SelectKBest

آیریس UCI: یک مسئله‌ی پایه‌ای خوشه‌بندی که برای آزمون قابلیت‌های یادگیری بدون‌ناظر به‌کار می‌رود.

با به‌کارگیری نظام‌مند این تکنیک‌ها، این گزارش می‌کوشد راهبرد پیش‌پردازش بهینه را برای انواع مختلف مسائل یادگیری ماشین تعیین کند.

فصل ۲

مروری بر ادبیات

۲-۱ تعریف مفاهیم پایه

۲-۱-۱ استخراج ویژگی

استخراج ویژگی فرآیندی است که طی آن از داده‌های اولیه، مجموعه‌ای از ویژگی‌های جدید، فشرده و معنادار تولید می‌شود تا ساختار درونی داده‌ها به‌صورت بهینه‌تری نمایش داده شود. در بسیاری از مسائل داده‌کاوی، داده‌های خام شامل ویژگی‌های فراوان، هم‌پوشان یا نویزی هستند که استفاده‌ی مستقیم از آن‌ها در مدل‌های یادگیری ماشین مناسب نیست. ازاین‌رو، با به‌کارگیری روش‌هایی مانند PCA، SVD و Autoencoder، فضای ویژگی جدیدی ایجاد می‌شود که روابط پنهان میان متغیرها را آشکار کرده و منجر به کاهش بُعد، افزایش پایداری و بهبود دقت مدل می‌گردد. برخلاف انتخاب ویژگی، استخراج ویژگی داده‌های اولیه را به فضایی جدید که ترکیبی خطی یا غیرخطی از ویژگی‌های اصلی است، نگاشت می‌کند. [۱] [۲]

۲-۱-۲ استقلال خطی

استقلال خطی به وضعیتی اطلاق می‌شود که هیچ‌یک از ویژگی‌ها را نتوان به‌صورت ترکیب خطی از سایر ویژگی‌ها بیان کرد. ویژگی‌های مستقل خطی، اطلاعات منحصربه‌فردی از فضای داده ارائه می‌دهند و از تکرار و افزونگی اطلاعات جلوگیری می‌کنند. در مدل‌های یادگیری ماشین، وجود وابستگی خطی شدید میان ویژگی‌ها (هم‌خطی) می‌تواند باعث ناپایداری در تخمین ضرایب، افزایش واریانس خطا و کاهش

قابلیت تعمیم مدل شود. ازاین‌رو، حفظ استقلال خطی ویژگی‌ها از منظر عددی و آماری اهمیت بالایی دارد.

۲-۱-۳ تحلیل مؤلفه‌های اصلی

تحلیل مؤلفه‌های اصلی یا (PCA)[1] روشی آماری و هندسی برای شناسایی جهت‌هایی در فضای داده است که بیشترین واریانس را توضیح می‌دهند. در این روش، محورهای جدید موسوم به مؤلفه‌های اصلی به‌گونه‌ای انتخاب می‌شوند که متعامد (Orthogonal) و در نتیجه مستقل خطی باشند. هدف اصلی PCA کاهش بُعد داده‌ها، حذف هم‌خطی و تمرکز بر مؤلفه‌هایی است که بیشترین اطلاعات را در خود جای داده‌اند. این روش مبتنی بر تجزیه‌ی مقادیر ویژه یا تجزیه‌ی مقادیر منفرد بوده و در کاربردهایی نظیر فشرده‌سازی داده، حذف نویز و مصورسازی داده‌ها استفاده می‌شود. [۳] [۴]

۲-۱-۴ تحلیل مؤلفه‌های مستقل

تحلیل مؤلفه‌های مستقل یا (ICA) [۲] یک روش آماری پیشرفته برای استخراج ویژگی‌هایی است که نه‌تنها از نظر خطی، بلکه از نظر آماری نیز مستقل از یکدیگر هستند. در حالی که PCA بر بیشینه‌سازی واریانس و حذف همبستگی خطی تمرکز دارد، ICA به دنبال شناسایی منابع آماری مستقل در داده‌ها است. خروجی این روش معمولاً شامل ویژگی‌هایی با قدرت تفکیک بالا و حساسیت کمتر به نویز می‌باشد. [۵]

۲-۱-۵ تجزیه مقادیر منفرد

تجزیه‌ی مقادیر منفرد یا Singular Value Decomposition[۳] یکی از مهم‌ترین ابزارهای جبر خطی عددی است که هر ماتریس حقیقی یا مختلط $\mathbf{X} \in \mathbb{R}^{m \times n}$ را به‌صورت حاصل‌ضرب سه ماتریس به شکل زیر تجزیه می‌کند:

$$\mathbf{X} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^{\mathsf{T}}$$

[1] Principal Component Analysis
[۲] Independent Component Analysis
[۳] (SVD)

که در آن، $\mathbf{U} \in \mathbb{R}^{m \times m}$ و $\mathbf{V} \in \mathbb{R}^{n \times n}$ ماتریس‌های متعامد هستند (به‌طوری که $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}$ و $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$) و ماتریس $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ یک ماتریس قطری غیرمنفی است که مقادیر منفرد را به‌ترتیب نزولی روی قطر اصلی خود قرار می‌دهد.

مقادیر منفرد بیانگر میزان سهم هر مؤلفه در بازنمایی ساختار داده بوده و بردارهای متناظر در ماتریس‌های $\mathbf{U}$ و $\mathbf{V}$ به‌ترتیب پایه‌های متعامد فضای ورودی و فضای ویژگی را تشکیل می‌دهند. یکی از ویژگی‌های مهم SVD این است که با نگه‌داشتن تنها چند مقدار منفرد بزرگ، می‌توان بهترین تقریب کم‌رتبه‌ی ممکن از ماتریس اولیه را به‌دست آورد.

به‌دلیل پایداری عددی و قابلیت اعمال بر ماتریس‌های مربعی و غیرمربعی، SVD به‌طور گسترده در کاربردهایی نظیر کاهش بُعد داده‌ها، فشرده‌سازی تصویر، حذف نویز، بازیابی اطلاعات و به‌عنوان هسته‌ی محاسباتی روش‌هایی مانند PCA مورد استفاده قرار می‌گیرد. [۶]

۲-۱-۶ کاهش بُعد

کاهش بُعد فرآیندی است که طی آن داده‌های با ابعاد بالا به نمایش کم‌بعدتری تبدیل می‌شوند، به‌گونه‌ای که اطلاعات کلیدی و ساختار اصلی آن‌ها حفظ شود. این فرآیند علاوه بر کاهش هزینه‌های محاسباتی، موجب حذف ویژگی‌های غیرمؤثر و بهبود عملکرد مدل‌های یادگیری ماشین می‌گردد. روش‌های کاهش بُعد به‌طور کلی به دو دسته تقسیم می‌شوند:

- روش‌های مبتنی بر نگاشت، مانند PCA و t-SNE
- روش‌های مبتنی بر انتخاب ویژگی

[۷] [۸]

۲-۱-۷ انتخاب ویژگی

انتخاب ویژگی فرآیندی است که در آن زیرمجموعه‌ای از ویژگی‌های اصلی که بیشترین تأثیر را بر خروجی مدل دارند، انتخاب می‌شود. برخلاف استخراج ویژگی، در این روش فضای داده تغییر نمی‌کند، بلکه ویژگی‌های غیرضروری، نویزی یا همبسته حذف می‌گردند. هدف اصلی انتخاب ویژگی افزایش دقت مدل، کاهش پیچیدگی و جلوگیری از بیش‌برازش است. روش‌های انتخاب ویژگی معمولاً به سه دسته‌ی

فیلتر، پوششی [۴] و تعبیه‌شده [۵] تقسیم می‌شوند. [۹]

۲-۱-۸ ماتریس کوواریانس

ماتریس کوواریانس ماتریسی مربعی است که میزان هم‌تغییری بین هر دو ویژگی را نشان می‌دهد. هر درایه‌ی این ماتریس بیانگر آن است که تغییرات یک ویژگی تا چه اندازه با تغییرات ویژگی دیگر همراه است. مقادیر نزدیک به صفر نشان‌دهنده‌ی استقلال تقریبی و مقادیر بزرگ مثبت یا منفی نشان‌دهنده‌ی تغییرات هم‌جهت یا خلاف‌جهت هستند. تحلیل این ماتریس مبنای روش‌هایی مانند PCA می‌باشد.

۲-۱-۹ هم‌خطی

هم‌خطی به وضعیتی گفته می‌شود که در آن چند ویژگی دارای وابستگی خطی قوی با یکدیگر هستند. وجود هم‌خطی می‌تواند باعث تکین شدن ماتریس کوواریانس یا ماتریس طراحی و در نتیجه ناپایداری تخمین ضرایب در مدل‌هایی مانند رگرسیون خطی شود. این پدیده موجب کاهش قابلیت تفسیر مدل و افزایش عدم‌قطعیت ضرایب می‌گردد. برای کاهش هم‌خطی، از روش‌هایی مانند حذف ویژگی‌های وابسته، نرمال‌سازی داده‌ها و الگوریتم‌هایی نظیر PCA و Ridge Regression استفاده می‌شود.

---

[۴] Wrapper

[۵] Embedded

فصل ۳

توصیف مجموعه داده‌ها و تحلیل هم‌خطی چندگانه

۳-۱ توصیف مجموعه‌داده‌ها و تحلیل هم‌خطی چندگانه

در این بخش، سه مجموعه‌داده‌ی مرجع که به‌ترتیب نماینده‌ی مسائل طبقه‌بندی، رگرسیون و خوشه‌بندی هستند، معرفی و بررسی می‌شوند. هدف اصلی این تحلیل، ارزیابی میزان وابستگی خطی میان ویژگی‌ها و شناسایی الگوهای هم‌خطی چندگانه با استفاده از نقشه‌های حرارتی همبستگی است. نتایج این بررسی مبنای انتخاب روش‌های مناسب کاهش بُعد یا انتخاب ویژگی در بخش‌های بعدی خواهد بود.

۳-۱-۱ مجموعه‌داده‌ی سرطان پستان ویسکانسین (طبقه‌بندی)

مرور مسئله این مجموعه‌داده یک مسئله‌ی طبقه‌بندی دودویی را مدل می‌کند که در آن هدف، پیش‌بینی بدخیم یا خوش‌خیم بودن یک توده‌ی پستانی بر اساس ویژگی‌های استخراج‌شده از تصویر است.

ساختار داده داده شامل ۵۶۹ نمونه با ۳۰ ویژگی عددی است که از تصاویر دیجیتالی حاصل از آسپیراسیون با سوزن نازک (FNA)[۱] یک توده‌ی پستانی محاسبه شده‌اند.

[۱] Fine Needle Aspiration

ویژگی‌های کلیدی ویژگی‌ها شامل شاخص‌های هندسی و مورفولوژیکی نظیر شعاع، بافت، محیط، مساحت، صافی و فشردگی هستند. برای هر ویژگی، سه معیار آماری شامل مقدار میانگین، انحراف معیار و مقدار بیشینه (بدترین حالت) ثبت شده است.

تحلیل هم‌خطی چندگانه بررسی ماتریس همبستگی نشان‌دهنده‌ی وجود هم‌خطی چندگانه‌ی شدید میان بسیاری از ویژگی‌های هندسی است.

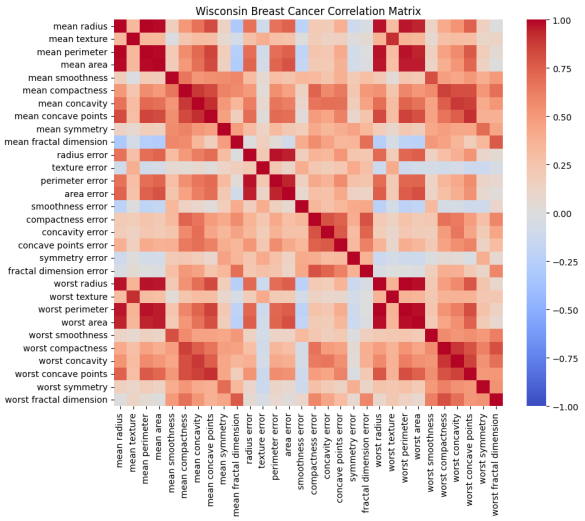
افزونگی هندسی یک بلوک مشخص از همبستگی مثبت بسیار بالا میان میانگین شعاع، میانگین محیط و میانگین مساحت مشاهده می‌شود. این همبستگی از دیدگاه هندسی کاملاً قابل انتظار است، زیرا:

$$\text{مساحت} = \pi r^{۲} \qquad \text{و} \qquad \text{محیط} = ۲\pi r$$

که نشان می‌دهد این ویژگی‌ها در اصل بازنمایی‌های متفاوتی از یک متغیر پنهان مشترک هستند.

وابستگی‌های مورفولوژیک ویژگی‌هایی نظیر تقعر، نقاط تقعر و فشردگی نیز همبستگی‌های خطی قوی با یکدیگر دارند که ناشی از ارتباطات ساختاری بین این شاخص‌های مورفولوژیک است.

نتیجه‌گیری وجود این خوشه‌های به‌شدت همبسته نشان می‌دهد که بُعد مؤثر داده‌ها به‌مراتب کمتر از تعداد اسمی ویژگی‌هاست. استفاده‌ی مستقیم از تمامی ویژگی‌ها در مدل‌های خطی بدون منظم‌سازی می‌تواند منجر به ناپایداری ضرایب و کاهش قابلیت تعمیم مدل شود.



شکل ۳-۱: ماتریس کواریانس داده های پستان ویسکانسین

۳-۱-۲ مجموعه‌داده‌ی مسکن بوستون (رگرسیون)

مرور مسئله این مجموعه‌داده یک مسئله‌ی کلاسیک رگرسیون را نمایش می‌دهد که هدف آن پیش‌بینی ارزش میانه‌ی خانه‌های مالک‌نشین (بر حسب هزار دلار) در ۵۰۶ ناحیه‌ی شهری بوستون است.

ساختار داده داده شامل ۱۳ ویژگی عددی است که جنبه‌های اجتماعی‌اقتصادی، محیطی و ساختاری مناطق مختلف شهری را توصیف می‌کنند.

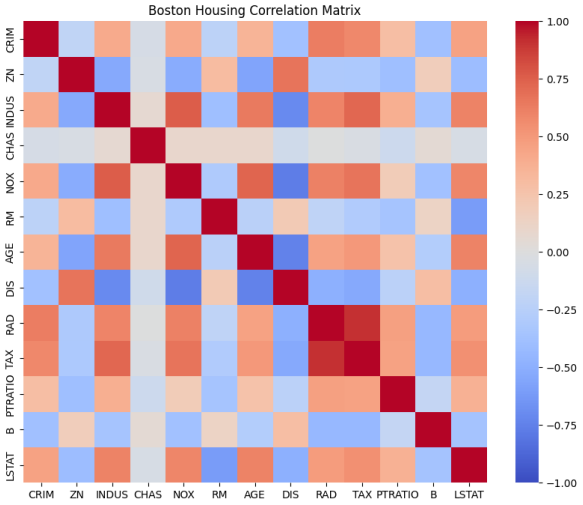
ویژگی‌های کلیدی از جمله ویژگی‌های مهم می‌توان به CRIM (نرخ جرم)، RM (میانگین تعداد اتاق‌ها)، NOX (غلظت اکسید نیتریک) و LSTAT (درصد جمعیت با وضعیت اجتماعی‌اقتصادی پایین) اشاره کرد.

تحلیل هم‌خطی چندگانه نقشه‌ی حرارتی همبستگی وجود چند خوشه‌ی متمایز از ویژگی‌های وابسته به یکدیگر را نشان می‌دهد.

خوشه‌های اجتماعی اقتصادی همبستگی مثبت بسیار قوی میان RAD (دسترسی به بزرگراه‌ها) و TAX (نرخ مالیات بر املاک) مشاهده می‌شود که اغلب از مقدار ۰٫۹۰ فراتر می‌رود. این موضوع نشان‌دهنده‌ی ارتباط ساختاری میان زیرساخت‌های حمل‌ونقل و سیاست‌های مالیاتی مناطق شهری است.

تأثیر صنعتی ویژگی NOX (آلودگی هوا) همبستگی مثبت قابل توجهی با INDUS (درصد مناطق صنعتی) و AGE (قدمت ساختمان‌ها) دارد که بیانگر تمرکز فعالیت‌های صنعتی در مناطق قدیمی‌تر شهری است.

نتیجه‌گیری اگرچه هم‌خطی چندگانه در این مجموعه‌داده وجود دارد، شدت و یکنواختی آن به اندازه‌ی مجموعه‌داده‌ی سرطان پستان نیست. وجود خوشه‌های متمایز نشان می‌دهد که روش‌های انتخاب ویژگی نظیر RFE می‌توانند در این مسئله مؤثرتر از روش‌های کاهش بُعد سراسری مانند PCA باشند.



شکل ۳-۲: ماتریس کواریانس داده های بوستون

۳-۱-۳ مجموعه‌داده‌ی آیریس (UCI) (خوشه‌بندی)

مرور مسئله هدف این مجموعه‌داده، گروه‌بندی ۱۵۰ نمونه گل زنبق به سه گونه‌ی Setosa، Versicolor و Virginica بر اساس ویژگی‌های فیزیکی آن‌هاست.

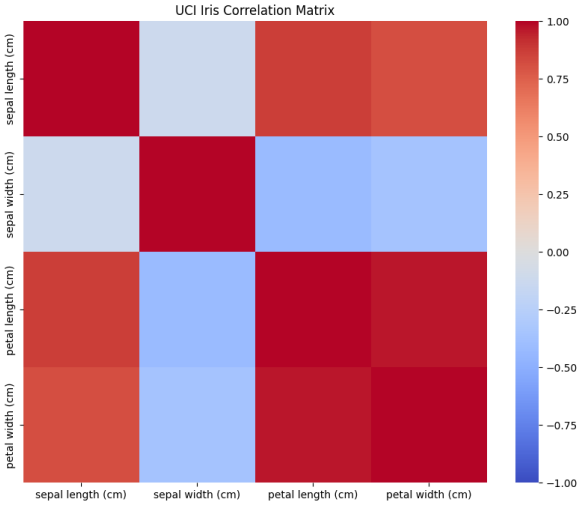
ساختار داده این داده شامل ۱۵۰ نمونه‌ی متعادل (۵۰ نمونه برای هر گونه) و ۴ ویژگی عددی شامل طول و عرض کاسبرگ و طول و عرض گلبرگ است.

تحلیل هم‌خطی چندگانه ساختار همبستگی در این مجموعه‌داده ساده اما بسیار معنادار است.

ویژگی‌های غالب همبستگی مثبت بسیار بالا ($r > ۰/۹۶$) میان طول گلبرگ و عرض گلبرگ مشاهده می‌شود که نشان‌دهنده‌ی وابستگی تقریباً کامل این دو ویژگی است.

استقلال نسبی عرض کاسبرگ کمترین میزان همبستگی را با سایر ویژگی‌ها دارد و در برخی موارد حتی همبستگی منفی نشان می‌دهد، که آن را به یکی از مستقل‌ترین ویژگی‌ها تبدیل می‌کند.

نتیجه‌گیری همبستگی شدید میان ابعاد گلبرگ بیانگر افزونگی اطلاعاتی یکی از این ویژگی‌هاست. کاهش بُعد با استفاده از PCA به دو بُعد احتمالاً بخش عمده‌ای از واریانس موردنیاز برای تفکیک خوشه‌ها را حفظ خواهد کرد.



شکل ۳-۳: ماتریس کواریانس داده های آیریس

۳-۲ استخراج ویژگی و کاهش ابعاد

در این مرحله، سه روش متداول کاهش بُعد شامل (PCA)[۲]، (SVD)[۳] و (ICA)[۴] بر روی مجموعه‌داده‌های نرمال‌سازی‌شده اعمال شدند. انتخاب تعداد مؤلفه‌ها (k) به‌صورت دلخواه انجام نشد، بلکه به‌گونه‌ای تعیین گردید که حداقل ۹۵ درصد از واریانس (اطلاعات) اولیه‌ی داده‌ها حفظ شود. این رویکرد امکان مقایسه‌ی منصفانه‌ی میزان افزونگی ساختاری در مجموعه‌داده‌های مختلف را فراهم می‌کند. در ادامه، نتایج کاهش بُعد برای هر مجموعه‌داده به‌صورت جداگانه تحلیل شده است.

۳-۲-۱ سرطان پستان ویسکانسین (فشرده‌سازی قابل‌توجه)

ابعاد داده ابعاد اولیه برابر با ۳۰ ویژگی بوده که پس از کاهش بُعد به ۱۰ مؤلفه‌ی اصلی تقلیل یافته است.

تحلیل الگوریتم کاهش بُعد موفق شده است حدود ۶۶ درصد از فضای ویژگی اولیه را حذف کند، در حالی که تنها ۵ درصد از واریانس کل داده‌ها از دست رفته است. این میزان فشرده‌سازی نشان‌دهنده‌ی افزونگی ساختاری شدید در داده‌هاست.

[۲] Principal Component Analysis
[۳] Singular Value Decomposition
[۴] Independent Component Analysis

تبیین همان‌گونه که در تحلیل ماتریس همبستگی مشاهده شد، بسیاری از ویژگی‌های هندسی نظیر شعاع، محیط و مساحت به‌شدت همبسته‌اند و در عمل بازنمایی‌های متفاوتی از یک مفهوم مشترک هستند. الگوریتم PCA این هم‌پوشانی اطلاعاتی را شناسایی کرده و آن‌ها را در قالب تعداد محدودی مؤلفه‌ی اصلی فشرده می‌کند. این مؤلفه‌ها نمایانگر ویژگی‌های کلی‌تری نظیر «اندازه» و «شکل» توده بوده و هم‌زمان اثر نویز ناشی از اندازه‌گیری‌های منفرد را کاهش می‌دهند.

۳-۲-۲ مسکن بوستون (فشرده‌سازی متوسط)

ابعاد داده ابعاد اولیه شامل ۱۳ ویژگی بوده که پس از کاهش بُعد به ۹ مؤلفه کاهش یافته است.

تحلیل کاهش بُعد در این مجموعه‌داده محدود بوده و تنها ۴ ویژگی حذف شده‌اند. این نتیجه نشان می‌دهد که بخش عمده‌ای از اطلاعات در ویژگی‌های متمایز و نسبتاً مستقل توزیع شده است.

تبیین برخلاف داده‌های سرطان پستان، مجموعه‌داده‌ی بوستون شامل متغیرهایی با ماهیت‌های اجتماعی، اقتصادی و محیطی متفاوت است؛ برای مثال، نرخ جرم، آلودگی هوا و ساختار مالیاتی نمایانگر پدیده‌های مستقل هستند. از آن‌جا که واریانس داده‌ها در خوشه‌های افزونه‌ی مشخص متمرکز نشده است، PCA توانایی فشرده‌سازی شدید داده‌ها را ندارد. این ویژگی‌ها نشان می‌دهد که در این مسئله، روش‌های انتخاب ویژگی می‌توانند در برخی کاربردها مؤثرتر از کاهش بُعد سراسری باشند.

۳-۲-۳ آیریس UCI (مصورسازی کامل)

ابعاد داده داده‌های اولیه شامل ۴ ویژگی عددی بوده که به ۲ مؤلفه‌ی اصلی کاهش یافته‌اند.

تحلیل کاهش بُعد به دو مؤلفه امکان نمایش کامل داده‌ها را بر روی یک صفحه‌ی دوبعدی فراهم کرده است، بدون آن‌که اطلاعات معنادار برای تفکیک گونه‌ها از دست برود.

تبیین همبستگی بسیار بالای میان طول و عرض گلبرگ باعث شده است که بخش عمده‌ای از واریانس داده‌ها در دو مؤلفه‌ی اول متمرکز شود. این نتیجه نشان می‌دهد که مسئله‌ی آیریس ذاتاً کم‌بعد بوده و ساختار خوشه‌ای آن را می‌توان با تعداد کمی ویژگی به‌خوبی نمایش داد.

۳-۲-۴ خلاصه‌ی نتایج کاهش بُعد

جدول ۳-۱: خلاصه‌ی نتایج کاهش بُعد برای مجموعه‌داده‌های مختلف

مجموعه‌داده	ویژگی‌های اولیه	ویژگی‌های لازم برای ۹۵٪ واریانس	نسبت کاهش	تفسیر
سرطان پستان ویسکانسین	۳۰	۱۰	≈ ۶۶٪	افزونگی بالا (ویژگی‌های هندسی)
مسکن بوستون	۱۳	۹	≈ ۳۰٪	افزونگی کم (عوامل اقتصادی متمایز)
آیریس UCI	۴	۲	۵۰٪	افزونگی بالا (نسبت‌های زیستی)

۳-۳ انتخاب ویژگی

در این مرحله، رویکرد تحلیل از کاهش بُعد مبتنی بر تبدیل خطی فضا (مانند PCA) به سمت انتخاب هدفمند مؤثرترین ویژگی‌های اصلی تغییر می‌یابد. برخلاف استخراج ویژگی، در انتخاب ویژگی فضای داده ثابت باقی می‌ماند و تنها زیرمجموعه‌ای از ویژگی‌های اولیه که بیشترین نقش را در پیش‌بینی متغیر هدف دارند، حفظ می‌شود. در این مطالعه، دو رویکرد متمایز انتخاب ویژگی مورد استفاده قرار گرفته است:

- **روش فیلتری** [۵]: انتخاب ویژگی‌ها بر اساس قدرت آماری فردی آن‌ها، مستقل از مدل یادگیری، مانند همبستگی یا معیارهای آماری مشابه با متغیر هدف.
- **روش پوششی** [۶]: حذف بازگشتی ضعیف‌ترین ویژگی‌ها بر اساس عملکرد یک مدل پیش‌بینی، با هدف یافتن ترکیب بهینه‌ی ویژگی‌ها.

۱-۳-۳ سرطان پستان ویسکانسین (قدرت فردی در برابر اثر ترکیبی)

ویژگی‌های منتخب توسط SelectKBest میانگین محیط، میانگین نقاط تقعر، بدترین شعاع، بدترین محیط و بدترین نقاط تقعر.

تحلیل روش فیلتری به‌طور کامل بر ویژگی‌های هندسی مرتبط با اندازه و شکل توده تمرکز کرده است. این نتیجه نشان می‌دهد که بزرگی فیزیکی تومور و میزان ناهنجاری‌های هندسی آن قوی‌ترین شاخص‌های فردی برای تشخیص بدخیمی هستند. از منظر آماری، این ویژگی‌ها بالاترین همبستگی خطی را با برچسب کلاس دارند.

ویژگی‌های منتخب توسط RFE میانگین شعاع، خطای بافت (Texture Error)، بدترین شعاع، بدترین فشردگی و بدترین تقعر.

تحلیل برخلاف روش فیلتری، RFE مجموعه‌ای متنوع‌تر از ویژگی‌ها را انتخاب کرده است. حضور «خطای بافت» که در انتخاب فیلتری نادیده گرفته شده بود، نشان می‌دهد که اگرچه اندازه‌ی توده مهم‌ترین عامل تشخیصی است، اما تغییرپذیری بافت در ترکیب با ویژگی‌های اندازه‌محور نقش مکمل و

[۵] SelectKBest

[۶] Recursive Feature Elimination - RFE

تعیین‌کننده‌ای در بهبود عملکرد مدل ایفا می‌کند. این نتیجه برتری روش‌های پوششی را در شناسایی تعاملات غیرآشکار بین ویژگی‌ها نشان می‌دهد.

۳-۳-۲ مسکن بوستون (ویژگی‌های اجتماعی اقتصادی در برابر عوامل زمینه‌ای)

ویژگی‌های منتخب توسط SelectKBest INDUS (درصد مناطق صنعتی)، RM (میانگین تعداد اتاق‌ها)، TAX (نرخ مالیات املاک)، PTRATIO (نسبت دانش‌آموز به معلم)، و LSTAT (درصد جمعیت با وضعیت اجتماعی اقتصادی پایین).

تحلیل روش فیلتری ویژگی‌هایی را انتخاب کرده است که به‌طور مستقیم و خطی با قیمت مسکن همبستگی دارند. به‌ویژه، LSTAT و RM که قوی‌ترین همبستگی خام را با متغیر هدف نشان می‌دهند، به‌طور طبیعی در صدر انتخاب‌ها قرار گرفته‌اند. این نتیجه بازتاب‌دهنده‌ی تمرکز روش فیلتری بر اثرات فردی و مستقل متغیرهاست.

ویژگی‌های منتخب توسط RFE CHAS (مجاورت با رودخانه چارلز)، NOX (آلودگی هوا)، RM، DIS (فاصله تا مراکز اشتغال)، و PTRATIO.

تحلیل روش پوششی ویژگی‌هایی را شناسایی کرده است که ماهیتی زمینه‌ای و تعاملی دارند و در روش فیلتری نادیده گرفته شده بودند. به‌طور خاص:

- CHAS یک متغیر دودویی است که در سطح کل داده همبستگی خطی قوی با قیمت ندارد، اما برای زیرمجموعه‌ای از خانه‌های با ارزش بالا نقش تعیین‌کننده‌ای ایفا می‌کند.
- NOX و DIS نشان می‌دهند که کیفیت محیط زیست و دسترسی مکانی، تعدیل‌کننده‌های مهم ارزش مسکن هستند؛ حتی اگر به‌تنهایی قوی‌ترین پیش‌بین‌ها نباشند.

این نتایج نشان می‌دهد که RFE قادر به شناسایی اثرات غیرمستقیم و ترکیبی ویژگی‌هاست.

۳-۳-۳ آیریس UCI (اجماع کامل روش‌ها)

ویژگی‌های منتخب توسط SelectKBest طول گلبرگ (سانتی‌متر)، عرض گلبرگ (سانتی‌متر).

ویژگی‌های منتخب توسط RFE طول گلبرگ (سانتی‌متر)، عرض گلبرگ (سانتی‌متر).

تحلیل هر دو روش انتخاب ویژگی به اجماع کامل رسیده‌اند. این نتیجه به‌طور قطعی نشان می‌دهد که ابعاد کاسبرگ (طول و عرض) نقش معناداری در تفکیک گونه‌های زنبق ندارند و عمده‌ی تفاوت زیستی میان گونه‌ها در هندسه‌ی گلبرگ‌ها نهفته است. چنین اجماعی بیانگر ساختار ساده، کم‌بعد و خوش‌تعریف این مسئله‌ی طبقه‌بندی است.

۳-۴ تحلیل مقایسه‌ای مدل‌ها

در مرحله‌ی نهایی، تأثیر راهبردهای مختلف مهندسی ویژگی بر عملکرد واقعی مدل‌های یادگیری ماشین ارزیابی شد. برای این منظور، مدل‌های پایه‌ی استاندارد شامل رگرسیون لجستیک برای مسائل طبقه‌بندی و رگرسیون خطی برای مسائل رگرسیون، بر روی سه نسخه‌ی متفاوت از هر مجموعه‌داده آموزش داده شدند:

- **داده‌های اصلی**: استفاده از تمامی ویژگی‌ها به‌عنوان حالت پایه،
- **داده‌های کاهش‌یافته با PCA**: نمایش فشرده‌ی داده‌ها در فضای کم‌بعد،
- **ویژگی‌های انتخاب‌شده**: زیرمجموعه‌ای از ویژگی‌های منتخب با استفاده از روش‌های RFE یا SelectKBest.

برای ارزیابی عملکرد، از معیار دقت [۷] در مسائل طبقه‌بندی و از ضریب تعیین R^2 به‌همراه خطای میانگین مربعات [۸] در مسائل رگرسیون استفاده شده است.

۱-۴-۳ سرطان پستان ویسکانسین (اثر نویززدایی)

نتایج عملکرد

- داده‌های اصلی: دقت ۹۵٫۶۱٪
- ویژگی‌های انتخاب‌شده: دقت ۹۵٫۶۱٪
- داده‌های کاهش‌یافته با PCA: دقت ۹۸٫۲۵٪

[۷] Accuracy
[۸] MSE

تحلیل نتایج نشان می‌دهد که داده‌های تبدیل‌شده با PCA بالاترین دقت را در میان سه حالت به دست آورده‌اند و حتی از استفاده‌ی مستقیم از تمامی ویژگی‌های اولیه نیز عملکرد بهتری داشته‌اند. این رفتار بیانگر آن است که مجموعه‌داده‌ی اولیه حاوی نویز و افزونگی قابل توجهی بوده که باعث کاهش توان تعمیم مدل می‌شده است. فشرده‌سازی داده‌ها به ۱۰ مؤلفه‌ی اصلی، امکان استخراج سیگنال زیربنایی مرتبط با تغییرات کلی شکل و اندازه‌ی توده‌ها را فراهم کرده و اثر نویز ناشی از ویژگی‌های همبسته را کاهش داده است.

در مقابل، مدل مبتنی بر ویژگی‌های انتخاب‌شده با وجود استفاده از تنها ۵ ویژگی، به همان دقت حالت پایه دست یافته است. این نتیجه نشان می‌دهد که بخش عمده‌ای از ویژگی‌های جمع‌آوری‌شده برای این مسئله‌ی تشخیصی خاص نقشی اساسی در بهبود عملکرد مدل نداشته‌اند.

۳-۴-۲ مسکن بوستون (محدودیت فشرده‌سازی خطی)

نتایج عملکرد

- داده‌های اصلی: $R^2 = ۰/۶۷$ (MSE: ۲۹.۲۴)
- ویژگی‌های انتخاب‌شده: $R^2 = ۰/۶۶$ (MSE: ۱۹.۲۵)
- داده‌های کاهش‌یافته با PCA: $R^2 = ۰/۶۰$ (MSE: ۵۱.۲۹)

تحلیل برخلاف مجموعه‌داده‌ی سرطان پستان، اعمال PCA بر داده‌های مسکن بوستون منجر به افت محسوس عملکرد مدل شده است. این نتیجه تأیید می‌کند که بسیاری از ویژگی‌های این مجموعه‌داده ماهیت متمایز و نسبتاً مستقلی دارند و هر یک حامل اطلاعات مشخصی درباره‌ی قیمت مسکن هستند. تبدیل این متغیرها به مؤلفه‌های انتزاعی، تفسیرپذیری و اثرگذاری مستقیم آن‌ها را کاهش داده و در نتیجه دقت پیش‌بینی افت کرده است.

در مقابل، استفاده از ویژگی‌های انتخاب‌شده توسط RFE منجر به عملکردی تقریباً هم‌تراز با داده‌های اصلی شده است. این امر نشان می‌دهد که متغیرهایی نظیر LSTAT و RM محرک‌های اصلی قیمت مسکن هستند و افزودن ویژگی‌های کم‌اثرتر تنها به افزایش پیچیدگی مدل بدون بهبود معنادار عملکرد منجر می‌شود.

۳-۴-۳ آیریس UCI (از دست رفتن اطلاعات تفکیکی)

نتایج عملکرد

- داده‌های اصلی: دقت ٪۱۰۰
- ویژگی‌های انتخاب‌شده: دقت ٪۱۰۰
- داده‌های کاهش‌یافته با PCA: دقت ٪۹۰٫۰۰

تحلیل با وجود آن‌که PCA حدود ۹۵ درصد از واریانس کل داده‌ها را حفظ می‌کند، افت دقت به ۹۰ درصد نشان می‌دهد که بخش کوچکی از واریانس حذف‌شده حاوی اطلاعات تفکیکی حیاتی برای مرز میان گونه‌های Versicolor و Virginica بوده است. این گونه‌ها به‌طور ذاتی دارای هم‌پوشانی هستند و تفکیک آن‌ها به اطلاعات ظریف و محلی وابسته است که لزوماً با بیشترین واریانس هم‌راستا نیست.

در مقابل، انتخاب مستقیم ویژگی‌های کلیدی شامل طول و عرض گلبرگ توانسته است دقت کامل مدل را حفظ کند. این نتیجه نشان می‌دهد که در مسائل دارای مجموعه‌ای محدود از ویژگی‌های بسیار قدرتمند، انتخاب ویژگی می‌تواند رویکردی ایمن‌تر و مؤثرتر از استخراج ویژگی مبتنی بر تبدیل خطی باشد.

فصل ۴

نگاهی عمیق تر

۴-۱ بررسی رگرسیون

۴-۱-۱ بررسی عمیق: پایداری رگرسیون و همگرایی الگوریتم‌ها

به‌منظور درک عمیق‌تر رفتار مدل‌های رگرسیونی در حضور هم‌خطی چندگانه، یک آزمایش دوبخشی بر روی مجموعه‌داده‌ی مسکن بوستون انجام شد. در این آزمایش، روش‌های تحلیلی مبتنی بر حل بسته‌ی رگرسیون خطی کلاسیک با روش‌های بهینه‌سازی مبتنی بر مشتق، به‌ویژه رگرسیون نزول گرادیان تصادفی [۱]، مقایسه شده‌اند. تمرکز اصلی این تحلیل بر پایداری ضرایب و سرعت همگرایی الگوریتم‌ها است.

بهینه‌سازی تحلیلی: پایداری ضرایب

در رگرسیون حداقل مربعات معمولی [۲]، وجود هم‌خطی چندگانه میان ویژگی‌ها منجر به افزایش واریانس برآورد ضرایب می‌شود. این پدیده اغلب به تخصیص ضرایب بزرگ و متضاد به ویژگی‌های همبسته می‌انجامد که از آن به‌عنوان «ناپایداری ضرایب» یا انفجار ضرایب یاد می‌شود.

برای ارزیابی این اثر، بزرگی بردار ضرایب مدل $\mathbf{w}$ برای دو حالت داده‌های اصلی و داده‌های تبدیل‌شده با PCA مورد مقایسه قرار گرفت. به‌عنوان معیار، مجموع قدر مطلق ضرایب (norm-ℓ_1) در نظر گرفته شد:

---

[۱] SGD Regressor

[۲] (Ordinary Least Squares - OLS)

- داده‌های اصلی: $\sum |w_i| = ۲۱٫۸۰$
- داده‌های کاهش‌یافته با PCA: $\sum |w_i| = ۱۳٫۹۶$

تحلیل کاهش قابل‌توجه مجموع بزرگی ضرایب در مدل مبتنی بر PCA نشان‌دهنده‌ی پایداری عددی بالاتر این مدل است. در داده‌های اصلی، ویژگی‌هایی نظیر LSTAT و DIS به‌دلیل همبستگی متقابل، برای تبیین تغییرات متغیر هدف با یکدیگر رقابت کرده و ضرایب بزرگی با حساسیت بالا ایجاد می‌کنند. در مقابل، مؤلفه‌های اصلی به‌دلیل متعامد بودن، وابستگی خطی را حذف کرده و امکان توزیع یکنواخت‌تر وزن‌ها را فراهم می‌سازند. این رفتار معادل اعمال یک منظم‌سازی ضمنی بر مدل بوده و خطر بیش‌برازش نسبت به ویژگی‌های نویزی را کاهش می‌دهد.

بهینه‌سازی مبتنی بر مشتق: همگرایی SGD

برای بررسی تأثیر استقلال ویژگی‌ها بر دینامیک یادگیری، یک مدل رگرسیون مبتنی بر نزول گرادیان تصادفی بر روی هر دو نسخه‌ی داده‌ها آموزش داده شد. روند تغییرات خطای آموزش در طول تکرارها برای دو حالت مورد بررسی قرار گرفت.

تفسیر رفتار همگرایی

- **داده‌های کاهش‌یافته با PCA**: مدل در طی یک تا دو تکرار اولیه به ناحیه‌ی کمینه همگرا می‌شود و منحنی خطا تقریباً تخت است. این رفتار نشان می‌دهد که الگوریتم مسیر مستقیمی به سمت کمینه‌ی تابع هزینه در اختیار دارد.
- **داده‌های اصلی**: مدل یک منحنی یادگیری تدریجی را نمایش می‌دهد و حدود ۵ تا ۱۰ تکرار زمان نیاز دارد تا به ناحیه‌ی پایدار خطا برسد.

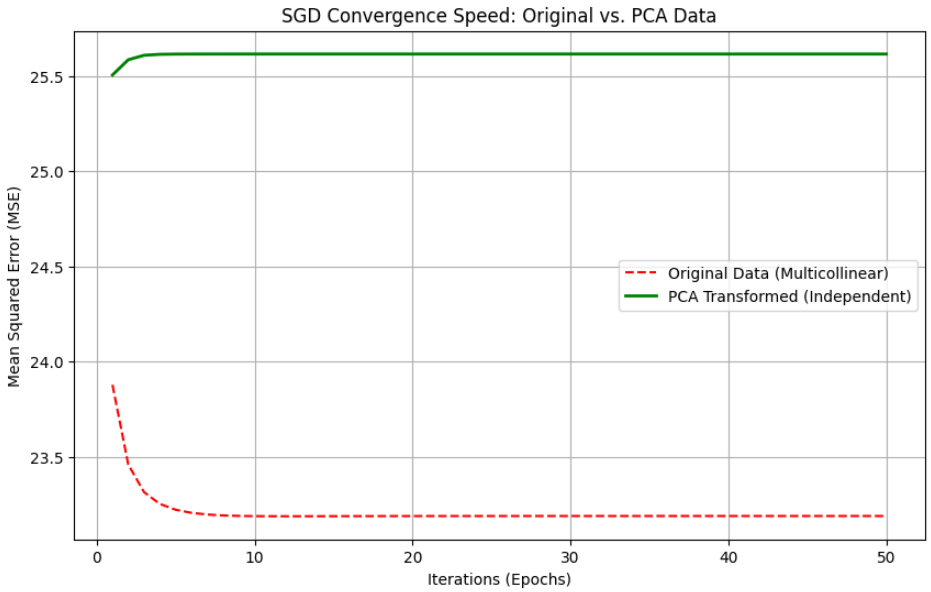
تبیین هندسی تابع هزینه رفتار مشاهده‌شده را می‌توان با شکل سطح تابع هزینه توضیح داد:

- **داده‌های اصلی**: به‌دلیل هم‌خطی چندگانه، سطح تابع هزینه شکلی شبیه به یک دره‌ی بلند و باریک دارد. در این حالت، گرادیان‌ها عمدتاً عمود بر مسیر بهینه هستند و الگوریتم با حرکت‌های زیگزاگی کوچک، به‌آهستگی به کمینه همگرا می‌شود.

- **داده‌های PCA**: استقلال و نرمال‌سازی مؤلفه‌ها باعث می‌شود سطح تابع هزینه شکلی نزدیک به یک کاسه‌ی متقارن داشته باشد. در نتیجه، بردار گرادیان مستقیماً به سمت کمینه‌ی سراسری اشاره کرده و همگرایی سریع و پایدار حاصل می‌شود.

جمع‌بندی

اگرچه مدل مبتنی بر داده‌های اصلی در نهایت به مقدار خطای MSE اندکی کمتر دست می‌یابد، داده‌های کاهش‌یافته با PCA از منظر پایداری عددی، یکنواختی ضرایب و سرعت همگرایی الگوریتم‌های بهینه‌سازی برتری محسوسی دارند. این ویژگی‌ها، PCA را به گزینه‌ای مناسب برای سامانه‌های مقیاس‌پذیر و داده‌محور تبدیل می‌کند؛ به‌ویژه در کاربردهایی که کارایی محاسباتی و سرعت آموزش مدل نقشی حیاتی ایفا می‌کنند.



شکل ۴-۱: مقایسه عملکرد

۴-۲ مقایسه کلاستر

۴-۲-۱ بررسی عمیق: تحلیل خوشه‌بندی (تأثیر بُعد)

الگوریتم‌های خوشه‌بندی مبتنی بر فاصله، نظیر K-Means، به‌طور مستقیم به معیار فاصله‌ی اقلیدسی برای سنجش میزان شباهت میان نمونه‌ها متکی هستند. در فضاهای با بُعد بالا، وجود ویژگی‌های نامرتبط

یا نویزی می‌تواند منجر به اعوجاج در مقیاس فاصله‌ها شده و در نتیجه کیفیت جداسازی خوشه‌ها را کاهش دهد. در این بخش، عملکرد الگوریتم K-Means بر روی مجموعه‌داده‌ی آیریس در دو حالت داده‌های اصلی چهاربُعدی و داده‌های کاهش‌یافته با PCA در فضای دوبُعدی مورد مقایسه قرار گرفته است.

کیفیت خوشه‌ها (امتیاز سیلوئت)

نتایج

- داده‌های اصلی: امتیاز سیلوئت ۰٫۴۵۹۹
- داده‌های کاهش‌یافته با PCA: امتیاز سیلوئت ۰٫۵۰۹۲

تحلیل افزایش امتیاز سیلوئت در داده‌های مبتنی بر PCA نشان‌دهنده‌ی بهبود هم‌زمان دو مؤلفه‌ی اصلی این معیار است: افزایش تراکم درون‌خوشه‌ای و افزایش فاصله‌ی بین خوشه‌ها. این نتیجه بیانگر آن است که خوشه‌ها در فضای کم‌بعد، تعریف واضح‌تر و مرزهای مشخص‌تری پیدا کرده‌اند.

تبیین همان‌گونه که در مرحله‌ی انتخاب ویژگی مشاهده شد، ابعاد مرتبط با کاسبرگ،به‌ویژه عرض کاسبرگ،همبستگی ضعیفی با گونه‌های زیستی دارند. حضور این ویژگی‌ها در فضای چهاربُعدی داده‌های اصلی باعث تضعیف فاصله‌های معنادار میان نمونه‌ها شده و نمونه‌های متعلق به خوشه‌های متفاوت را به‌طور مصنوعی به یکدیگر نزدیک می‌کند. فرافکنی داده‌ها بر روی دو مؤلفه‌ی اصلی، که عمدتاً تغییرات گلبرگ را بازنمایی می‌کنند، این نویز ساختاری را حذف کرده و فضای خوشه‌بندی را شفاف‌تر ساخته است.

تابع زیان خوشه‌بندی (اینرسی)

نتایج

- داده‌های اصلی: اینرسی ۱۳۹٫۸۲
- داده‌های کاهش‌یافته با PCA: اینرسی ۱۱۵٫۰۲

تحلیل کاهش مقدار اینرسی در داده‌های کاهش‌یافته نشان می‌دهد که نمونه‌ها به‌طور میانگین فاصله‌ی کمتری از مراکز خوشه‌ی متناظر خود دارند. اگرچه بخشی از این کاهش به‌صورت ذاتی ناشی از محاسبه‌ی فاصله در فضای کم‌بعدتر است، اما هم‌زمانی آن با بهبود امتیاز سیلوئت نشان می‌دهد که این کاهش عمدتاً حاصل حذف ابعاد نویزی بوده است، نه از دست رفتن اطلاعات ساختاری مرتبط با خوشه‌ها.

کارایی محاسباتی

نتایج

- داده‌های اصلی: ۰٫۰۷۶۶ ثانیه
- داده‌های کاهش‌یافته با PCA: ۰٫۰۲۱۳ ثانیه

تحلیل کاهش بُعد منجر به افزایش قابل توجه سرعت اجرای الگوریتم شده است؛ به‌طوری که نسخه‌ی مبتنی بر PCA بیش از ۳ برابر سریع‌تر از نسخه‌ی مبتنی بر داده‌های خام اجرا شده است. با وجود آن‌که تعداد تکرارهای همگرایی در هر دو حالت یکسان بوده است، هزینه‌ی محاسباتی هر تکرار به‌طور چشمگیری کاهش یافته، زیرا محاسبات فاصله در فضای دوبُعدی به‌مراتب کم‌هزینه‌تر از فضای چهاربُعدی است. این مزیت در مقیاس‌های بزرگ داده، نقشی تعیین‌کننده در عملی‌بودن الگوریتم‌های خوشه‌بندی ایفا می‌کند.

جمع‌بندی

نتایج این تحلیل نشان می‌دهد که در مسائل یادگیری بدون‌ناظر مبتنی بر فاصله، استفاده از کاهش بُعد با PCA به‌طور معناداری بر به‌کارگیری داده‌های خام برتری دارد. این رویکرد نه‌تنها کیفیت خوشه‌بندی را بهبود می‌بخشد (امتیاز سیلوئت بالاتر)، بلکه با کاهش هزینه‌ی محاسباتی، سرعت اجرا و همگرایی الگوریتم را نیز افزایش می‌دهد. ازاین‌رو، کاهش بُعد را می‌توان به‌عنوان یک گام کلیدی پیش‌پردازشی در تحلیل‌های خوشه‌بندی بزرگ‌مقیاس در نظر گرفت.

۴-۳ طبقه بندی

۴-۳-۱ بررسی عمیق: تحلیل طبقه‌بندی

الگوریتم‌های مبتنی بر نمونه در برابر الگوریتم‌های مبتنی بر درخت

الگوریتم‌های یادگیری ماشین بسته به سازوکار درونی خود، واکنش‌های متفاوتی به ساختار فضای ویژگی نشان می‌دهند. در این بخش، عملکرد الگوریتم k-Nearest Neighbors (KNN) به‌عنوان یک روش مبتنی بر فاصله و الگوریتم Random Forest به‌عنوان نماینده‌ای از مدل‌های مبتنی بر درخت تصمیم، با استفاده از مجموعه‌داده‌ی سرطان پستان ویسکانسین مورد مقایسه قرار گرفته است. هدف این تحلیل، بررسی تأثیر کاهش بُعد و انتخاب ویژگی بر این دو خانواده‌ی الگوریتمی با ماهیت‌های کاملاً متفاوت است.

مدل اول: k-Nearest Neighbors (رفع اثر نفرین بُعد)

نتایج عملکرد

- داده‌های اصلی (۳۰ ویژگی): دقت ۰٫۹۴۷۴
- داده‌های کاهش‌یافته با PCA (۱۰ مؤلفه): دقت ۰٫۹۵۶۱

تحلیل افزایش دقت مدل KNN پس از کاهش بُعد نشان می‌دهد که این الگوریتم به‌شدت تحت تأثیر پدیده‌ی «نفرین بُعد» قرار دارد. در فضای با بُعد بالا، فاصله‌ی اقلیدسی بین نمونه‌ها به‌دلیل پراکندگی داده‌ها و حضور ویژگی‌های نویزی، قدرت تمایز خود را از دست می‌دهد. فرافکنی داده‌ها به فضای ۱۰بُعدی مؤلفه‌های اصلی، سیگنال‌های معنادار را متمرکز کرده و امکان شناسایی همسایگان واقعی‌تر را برای الگوریتم فراهم می‌سازد. در نتیجه، دقت طبقه‌بندی به‌طور محسوسی بهبود یافته است.

نکته‌ی محاسباتی اگرچه از منظر تئوریک انتظار می‌رود کاهش بُعد منجر به کاهش زمان پیش‌بینی شود، در این آزمایش خاص زمان اجرا برای داده‌های مبتنی بر PCA افزایش یافته است. این پدیده احتمالاً ناشی از اندازه‌ی نسبتاً کوچک مجموعه‌داده ($N = ۵۶۹$) است؛ به‌گونه‌ای که سربار محاسباتی فرافکنی داده‌ها به فضای مؤلفه‌های اصلی از صرفه‌جویی ناشی از کاهش محاسبات فاصله بیشتر بوده است. در مسائل بزرگ‌مقیاس، مزیت محاسبه‌ی فاصله در فضای کم‌بعد معمولاً غالب خواهد شد.

مدل دوم: جنگل تصادفی (برتری ویژگی‌های خام)

نتایج عملکرد

- داده‌های اصلی: دقت ۰٫۹۶۴۹
- داده‌های کاهش‌یافته با PCA: دقت ۰٫۹۵۶۱
- ویژگی‌های انتخاب‌شده: دقت ۰٫۹۵۶۱

تحلیل جنگل تصادفی بالاترین دقت خود را در حالت استفاده از داده‌های خام به دست آورده است. این رفتار با ماهیت درخت‌های تصمیم سازگار است، زیرا این مدل‌ها به‌صورت درونی فرآیند انتخاب ویژگی را انجام می‌دهند و در هر گره، بهترین تقسیم را بر اساس معیارهای اطلاعاتی انتخاب می‌کنند. در نتیجه، ویژگی‌های کم‌اهمیت یا نویزی به‌طور طبیعی نادیده گرفته می‌شوند و نیازی به پیش‌پردازش صریح برای کاهش بُعد وجود ندارد.

محدودیت PCA برای مدل‌های درختی اعمال PCA منجر به تولید ویژگی‌های ترکیبی و انتزاعی می‌شود که فاقد تفسیر فیزیکی مستقیم هستند (برای مثال ترکیب خطی چند ویژگی هندسی). درخت‌های تصمیم که بر اساس آستانه‌های مشخص و قابل‌تفسیر عمل می‌کنند (مانند «اگر شعاع بزرگ‌تر از مقدار معینی باشد»)، در تقسیم‌بندی چنین ویژگی‌هایی با چالش مواجه می‌شوند. در نتیجه، کاهش شفافیت مرزهای تصمیم می‌تواند به افت جزئی عملکرد منجر شود.

انتخاب ویژگی مدل آموزش‌دیده بر زیرمجموعه‌ی ویژگی‌های انتخاب‌شده عملکردی مشابه مدل مبتنی بر PCA نشان داده است، اما نتوانسته از مدل آموزش‌دیده بر داده‌های کامل پیشی بگیرد. این نتیجه نشان می‌دهد که جنگل تصادفی از پایداری و انعطاف‌پذیری کافی برخوردار است تا مجموعه‌ی کامل ویژگی‌ها را بدون نیاز به کاهش دستی بُعد به‌طور مؤثر مدیریت کند.

جمع‌بندی

نتایج این تحلیل نشان می‌دهد که تأثیر مهندسی ویژگی به‌شدت به نوع الگوریتم وابسته است. الگوریتم‌های مبتنی بر فاصله مانند KNN از کاهش بُعد و حذف نویز سود قابل‌توجهی می‌برند، در حالی که مدل‌های مبتنی بر درخت نظیر جنگل تصادفی، به‌طور ذاتی نسبت به افزونگی ویژگی‌ها مقاوم بوده و در بسیاری

از موارد با داده‌های خام بهترین عملکرد را ارائه می‌دهند. این تفاوت، اهمیت انتخاب راهبرد مهندسی ویژگی متناسب با خانواده‌ی الگوریتمی را برجسته می‌سازد.

فصل ۵

جمع‌بندی و نتیجه‌گیری و پیشنهادات

۵-۱ نتیجه‌گیری

در این مطالعه، تأثیر راهبردهای مختلف مهندسی ویژگی،به‌ویژه کاهش بُعد مبتنی بر (PCA)[۱] و انتخاب ویژگی با استفاده از (RFE)[۲]، به‌صورت نظام‌مند بر عملکرد الگوریتم‌های یادگیری ماشین در سه دسته‌ی اصلی مسائل طبقه‌بندی، رگرسیون و خوشه‌بندی مورد بررسی قرار گرفت. تحلیل تجربی بر روی مجموعه‌داده‌های سرطان پستان ویسکانسین، مسکن بوستون و آیریس UCI نشان داد که هیچ راهبرد واحدی به‌طور مطلق برتر نیست و انتخاب روش پیش‌پردازش باید به ماهیت داده و الگوریتم وابسته باشد. نتایج اصلی این پژوهش را می‌توان در چهار محور کلیدی خلاصه کرد.

۱. واقعیت پدیده‌ی «نفرین بُعد» در مدل‌های مبتنی بر فاصله

نتایج تجربی به‌روشنی نشان داد که الگوریتم‌هایی که به معیار فاصله‌ی اقلیدسی متکی هستند، مانند K-Means و KNN، به‌شدت از کاهش بُعد سود می‌برند. در مجموعه‌داده‌ی آیریس UCI، اعمال PCA منجر به افزایش کیفیت خوشه‌بندی (افزایش امتیاز سیلوئت از حدود ۰٫۴۶ به ۰٫۵۱) و کاهش قابل‌توجه هزینه‌ی محاسباتی (بیش از ۳٫۵ برابر) شد. به‌طور مشابه، در مجموعه‌داده‌ی سرطان پستان، کاهش بُعد باعث بهبود دقت الگوریتم KNN از ٪۹۴٫۷ به ٪۹۵٫۶ گردید. این نتایج تأیید می‌کنند که حذف ابعاد نویزی و همبسته، فاصله‌های معنادار را برجسته کرده و امکان شناسایی «همسایگان واقعی‌تر» را برای

[۱] Principal Component Analysis
[۲] Recursive Feature Elimination

مدل فراهم می‌سازد.

۲. تفسیرپذیری در برابر انتزاع در مسائل رگرسیونی

در مسئله‌ی رگرسیون مسکن بوستون، کاهش بُعد کورکورانه با PCA اثر منفی بر عملکرد مدل داشت و ضریب تعیین R^2 را از ۰٫۶۷ به ۰٫۶۰ کاهش داد. ویژگی‌های این مجموعه‌داده،نظیر نرخ جرم، تعداد اتاق‌ها و ساختار مالیات نمایانگر عوامل اجتماعی اقتصادی متمایز و نسبتاً مستقلی هستند. فشرده‌سازی این متغیرها در مؤلفه‌های انتزاعی، اثرات مشخص و قابل‌تفسیر آن‌ها بر قیمت مسکن را تضعیف می‌کند. در مقابل، انتخاب ویژگی با RFE ضمن ساده‌سازی مدل و حذف متغیرهای کم‌اثر، تفسیرپذیری عوامل کلیدی مانند LSTAT و RM را حفظ کرده و عملکردی هم‌تراز با مدل پایه ارائه داد. این نتیجه نشان می‌دهد که در مسائل رگرسیونی تفسیرمحور، انتخاب ویژگی اغلب راهبرد مناسب‌تری نسبت به کاهش بُعد است.

۳. برتری داده‌ی خام برای مدل‌های مبتنی بر درخت

طبقه‌بند جنگل تصادفی بالاترین دقت خود (۹۶٫۵٪) را در حالت استفاده از داده‌های اصلی با بُعد بالا به دست آورد. درخت‌های تصمیم و مدل‌های تجمیعی مبتنی بر آن‌ها، به‌صورت ذاتی فرآیند انتخاب ویژگی را در ساختار درونی خود انجام می‌دهند و نسبت به نویز و افزونگی ویژگی‌ها مقاوم هستند. اعمال PCA با چرخاندن فضای ویژگی و تولید متغیرهای ترکیبی، تعیین مرزهای تصمیم‌گیری صریح و قابل‌تفسیر را برای این مدل‌ها دشوارتر کرده و منجر به افت جزئی عملکرد شد. این مشاهده تأکید می‌کند که برای الگوریتم‌های درخت‌محور، داده‌های خام اغلب بهترین ورودی را فراهم می‌کنند.

۴. نقش کاهش بُعد در پایداری و همگرایی بهینه‌سازی

تحلیل رفتار الگوریتم‌های مبتنی بر مشتق، به‌ویژه SGDRegressor، نشان داد که PCA ابزاری قدرتمند برای افزایش پایداری عددی است. با حذف هم‌خطی چندگانه و متعامدسازی فضای ویژگی، PCA سطح تابع هزینه را به شکلی محدب و متقارن نزدیک می‌کند و در نتیجه، الگوریتم نزول گرادیان می‌تواند با مسیر مستقیم‌تر و سرعت بالاتر به کمینه همگرا شود. این ویژگی، PCA را به گزینه‌ای جذاب برای سامانه‌های بزرگ‌مقیاس و سناریوهایی که سرعت آموزش اهمیت بالایی دارد، تبدیل می‌کند.

جمع‌بندی نهایی

یافته‌های این پژوهش نشان می‌دهد که هیچ راهبرد پیش‌پردازشی «بهترینِ واحد» وجود ندارد. کاهش بُعد با PCA انتخابی مناسب برای مسائل هندسی، خوشه‌بندی و الگوریتم‌های مبتنی بر فاصله، یا در شرایطی است که کارایی محاسباتی اولویت دارد. در مقابل، برای مسائل تفسیرمحور ەویژه در حوزه‌های اقتصادی و همچنین هنگام استفاده از مدل‌های مقاوم و درخت‌محور، انتخاب ویژگی همچنان به‌عنوان معیار طلایی مهندسی ویژگی مطرح است. بنابراین، طراحی خط لوله‌ی یادگیری ماشین باید همواره با در نظر گرفتن ماهیت داده، الگوریتم هدف و محدودیت‌های محاسباتی انجام شود.

منابع و مراجع

[1] Bengio, Yoshua, Courville, Aaron, and Vincent, Pascal. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798–1828, 2013.

[2] Hinton, Geoffrey E. and Salakhutdinov, Ruslan R. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

[3] Hotelling, Harold. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6):417–441, 1933.

[4] Pearson, Karl. On lines and planes of closest fit to systems of points in space. Philosophical Magazine, 2(11):559–572, 1901.

[5] Hyvärinen, Aapo. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks, 10(3):626–634, 1999.

[6] Eckart, Carl and Young, Gale. The approximation of one matrix by another of lower rank. Psychometrika, 1(3):211–218, 1936.

[7] Cunningham, John P. and Ghahramani, Zoubin. Linear dimensionality reduction: Survey, insights, and generalizations. Journal of Machine Learning Research, 16:2859–2900, 2015.

[8] van der Maaten, Laurens and Hinton, Geoffrey. Visualizing data using t-sne. Journal of Machine Learning Research, 9:2579–2605, 2008.

[9] Guyon, Isabelle and Elisseeff, André. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

Abstract

This study investigates the critical role of feature engineering in optimizing the performance of data mining algorithms across classification, regression. The methodology proceeds in distinct stages: first, identifying feature dependencies through correlation analysis; second, applying feature extraction techniques (PCA, ICA, SVD) based on explained variance; and third, implementing feature selection methods (SelectKBest, RFE) to isolate high-value predictors.

Key Words:

Data Mining, Dimensionality Reduction, PCA, Feature Selection, Multicollinearity, Model Optimization, Regression, Clustering, Classification.

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Comptuer Science

M. Sc. Thesis

Seoncd CDM Project

By

Mohammad Sadegh Gholizadeh

Supervisor

Dr. Mahdi Ghatei