دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

استاد درس: دکتر مهدی قطعی

آذر ۱۴۰۴

تمرین پنجم

درس داده‌کاوی محاسباتی

موضوع پروژه: انتخاب ویژگی با استفاده از مقادیر تکین (SVD) در برابر روش‌های کلاسیک

هدف از انجام پروژه

هدف از این پروژه، عبور از کاربرد «جعبه سیاه» الگوریتم‌ها و درک ریاضیاتی این مسئله است که چگونه **تجزیه مقادیر تکین** (SVD) می‌تواند ساختار پنهان داده‌ها را آشکار کند. در این تمرین، شما به جای استفاده صرف از توابع آماده، با تحلیل مستقیم ماتریس‌های U، Σ و V^T، مهم‌ترین ویژگی‌های یک مجموعه داده صنعتی را شناسایی کرده و آن را با روش‌های کلاسیک (مانند RFE) مقایسه می‌کنید. اهداف اصلی عبارتند از:

- درک نقش **ماتریس بردارهای تکین راست** (V^T) در تعیین سهم هر ویژگی اولیه.
- یادگیری نحوه وزن‌دهی به ویژگی‌ها با استفاده از **مقادیر تکین** (Σ).
- مقایسه **پایداری** (Stability) روش‌های جبری (SVD) در برابر روش‌های تکراری (Wrapper/RFE) در حضور نویز.
- مصورسازی هندسی ویژگی‌ها با استفاده از نمودار **بارگذاری** (Loading Plot).

مفاهیم و مهارت‌هایی که دانشجو خواهد آموخت

- پیاده‌سازی و مقایسه روش‌های کلاسیک انتخاب ویژگی شامل Variance Threshold و RFE.
- استفاده خلاقانه از **تجزیه مقادیر تکین** (SVD) برای **انتخاب** ویژگی (نه صرفاً استخراج).
- تحلیل **ماتریس بارگذاری** (Loadings) برای تفسیر فیزیکی سنسورهای مهم.
- تحلیل حساسیت الگوریتم‌ها نسبت به داده‌های پرت (Outliers) و نویز.
- کار با داده‌های صنعتی واقعی (نامتوازن و ابعاد بالا).

تعاریف مفاهیم پایه

تجزیه مقادیر تکین (SVD)[1] هر ماتریس داده X (با ابعاد $n \times m$) را می‌توان به صورت $X = U\Sigma V^T$ تجزیه کرد. در اینجا:

- ماتریس Σ شامل **مقادیر تکین** (σ_i) است که «انرژی» یا اهمیت هر مؤلفه را نشان می‌دهد.
- ماتریس V^T (با ابعاد $m \times m$) شامل **بردارهای تکین راست** است. هر سطر از V^T متناظر با یک مقدار تکین است و نشان می‌دهد که ویژگی‌های اولیه ($x_1, ..., x_m$) با چه وزن‌هایی ترکیب شده‌اند تا آن مؤلفه را بسازند.

انتخاب ویژگی با SVD (SVD Feature Ranking)[2] برخلاف PCA که ویژگی‌های جدید می‌سازد (استخراج)، ما می‌توانیم از SVD برای **انتخاب** ویژگی‌های قدیمی استفاده کنیم. اگر ویژگی j در بردارهای تکینِ اول (که σ بزرگی دارند) ضریب بزرگی داشته باشد، یعنی آن ویژگی نقش مهمی در ساختار اصلی داده دارد.

[1] Singular Value Decomposition
[2] SVD Ranking

تعریف پروژه و ساختار اجرایی آن

پروژه باید به‌صورت گام‌به‌گام در محیط Google Colab و با زبان Python اجرا شود.

بخش اول: آماده‌سازی داده‌های صنعتی (SECOM)

دیتاست: از دیتاست UCI SECOM (ساخت نیمه‌هادی) استفاده کنید. این داده‌ها دارای صدها سنسور، نویز زیاد و مقادیر گمشده هستند.

- **مرحله ۱.۱:** دیتاست را بارگذاری کنید. ستون‌های ثابت (واریانس صفر) را حذف کنید و NaNها را با روش Median Imputation پر کنید.
- **مرحله ۲.۱:** داده‌ها را نرمال‌سازی کنید (StandardScaler). (Hint: SVD is scale-sensitive. Without standardization, sensors with large units will artificially dominate the singular values.)

بخش دوم: روش‌های کلاسیک (مبنای مقایسه)

- **مرحله ۱.۲ – روش فیلتر:** با استفاده از Mutual Information (تابع `mutual_info_classif`)، تعداد ۲۰ ویژگی برتر را انتخاب کنید.
- **مرحله ۲.۲ – روش پوششی (RFE):** با استفاده از RandomForest و الگوریتم RFE، تعداد ۲۰ ویژگی برتر را بیابید. **گزارش:** زمان اجرای دقیق این مرحله را ثبت کنید.

بخش سوم: روش جبری (رتبه‌بندی با مقادیر تکین)

این بخش قلب ریاضی تمرین است. شما **نباید** از توابع آماده انتخاب ویژگی استفاده کنید، بلکه باید الگوریتم را بسازید.

- **مرحله ۱.۳ – تجزیه:** با استفاده از `numpy.linalg.svd`، ماتریس‌های U, Σ, V^T را محاسبه کنید.

- **مرحله ۲.۳ – تحلیل ریاضی (امتیازدهی):** یک تابع امتیازدهی (Score Function) برای هر ویژگی اولیه j تعریف کنید. ایده: ویژگی j مهم است اگر در k بردار تکین اول، وزن زیادی داشته باشد.

  فرمول پیشنهادی برای پیاده‌سازی:

$$Score_j = \sum_{i=1}^{k} \sigma_i^2 \cdot |V_{ij}|$$

(Hint: Here, V_{ij} is the absolute coefficient of feature j in the i-th singular vector. We weight it by σ_i because vectors with small singular values represent noise.)

- **مرحله ۳.۳ – انتخاب:** بر اساس امتیازهای محاسبه شده، ۲۰ ویژگی برتر را انتخاب کنید.

بخش چهارم: تحلیل هندسی و پایداری

این بخش برای سنجش عمق درک شما از هندسه داده طراحی شده است.

- **مرحله ۱.۴ – نمودار بارگذاری (Loadings Plot):** برای دو مؤلفه اول ($PC1$ و $PC2$)، نمودار پراکندگی ضرایب V را رسم کنید. هر نقطه در این نمودار یک «ویژگی» است. آیا می‌توانید خوشه‌هایی از سنسورها که رفتار مشابهی دارند را تشخیص دهید؟ (ویژگی‌هایی که در یک جهت بردار کشیده شده‌اند).
- **مرحله ۲.۴ – تست پایداری (Stability Test):** ۵٪ نویز تصادفی به داده‌های اصلی اضافه کنید. مجدداً ۲۰ ویژگی برتر را با روش RFE و روش SVD انتخاب کنید. **سوال:** لیست ویژگی‌های منتخب کدام روش تغییر کمتری کرد؟ (روش‌های جبری معمولاً در برابر نویز پایدارترند).

بخش پنجم: مقایسه نهایی

یک مدل Logistic Regression ساده را روی سه زیرمجموعه داده (۲۰ ویژگی MI، ۲۰ ویژگی RFE، ۲۰ ویژگی SVD) آموزش دهید.

۱. **جدول مقایسه:** دقت (Accuracy)، F1-Score و «زمان انتخاب ویژگی» را مقایسه کنید.

۲. **تحلیل همپوشانی:** نمودار ون (Venn Diagram) رسم کنید که نشان دهد چند درصد اشتراک بین روش SVD و RFE وجود دارد.

۳. **نتیجه‌گیری:** برای یک کاربرد صنعتی که نیاز به تفسیرپذیری و سرعت بالا دارد، کدام روش را پیشنهاد می‌کنید؟

نکات پایانی و دستورالعمل ارسال

- نیاز است که گزارش ارسالی حتما در قالب خواسته شده باشد.
- در صورت هرگونه سوالی در گروه بپرسید.
- ذکر منابع اجباری است.
- نیازی به ریز شدن در مطالب نیست، بیان مسئله و مشکل، و چگونگی رفع مسئله با روش خوانده شده کافی است.
- فایل گزارش خود را به صورت PDF بفرستید.
- تمامی کدها، داده‌ها و فایل‌های پروژه را در یک مخزن GitHub Repository بارگذاری کنید، و لینک دسترسی به ریپازیتوری را در انتهای گزارش خود قرار دهید. مخزن باید به‌صورت Public تنظیم شود تا قابل مشاهده باشد.
- به یاد داشته باشید هدف این تمرین، درک اهمیت این درس، و آشنایی با برخی موضوعات است که در آینده بسیار با آن‌ها برخورد خواهید داشت، ملاک نمره حجم گزارش ارسالی نخواهد بود.

موفق باشید.