گزارش بخش دوم: تحلیل تفکیکی خطی فیشر (LDA)

وحید ملکی

۲۹ آذر ۱۴۰۴

۱ تحلیل تفکیکی خطی فیشر (Fisher's LDA)

در این بخش، الگوریتم LDA برای کاهش ابعاد و دسته‌بندی مجموعه داده Seeds پیاده‌سازی شده است. این مجموعه داده شامل ویژگی‌های هندسی هسته‌های گندم بوده و دارای ۳ کلاس مختلف است. هدف انتقال داده‌ها از فضای ۷ بعدی اولیه به یک فضای ۲ بعدی (تعداد کلاس‌ها منهای یک) است که در آن تفکیک‌پذیری کلاس‌ها بیشینه باشد.

۱۰۱ پیاده‌سازی دستی از پایه (سوال ۱۳)

۱۰۱۰۱ تئوری و فرمول‌بندی

هدف LDA یافتن بردارهای تصویری است که نسبت پراکندگی بین کلاسی (S_B) به پراکندگی درون کلاسی (S_W) را بیشینه کنند. ماتریس پراکندگی درون کلاسی به صورت زیر محاسبه شد:

$$S_W = \sum_{c=1}^{C} \sum_{x \in D_c} (x - \mu_c)(x - \mu_c)^T \quad (۱)$$

و ماتریس پراکندگی بین کلاسی:

$$S_B = \sum_{c=1}^{C} N_c(\mu_c - \mu)(\mu_c - \mu)^T \quad (۲)$$

که در آن μ_c میانگین کلاس c و μ میانگین کل داده‌ها است. بردارهای بهینه (اجزای خطی) همان بردارهای ویژه متناظر با بزرگترین مقادیر ویژه ماتریس $S_W^{-1}S_B$ هستند. در پیاده‌سازی دستی، پس از محاسبه این ماتریس‌ها و حل مساله مقدار ویژه، تنها از بخش حقیقی (Real Part) بردارهای ویژه استفاده شد تا از خطاهای محاسباتی ناشی از مقادیر موهومی ناچیز جلوگیری شود.

۱۰۱۰۲ روش دسته‌بندی

پس از انتقال داده‌ها به فضای کاهش‌یافته ۲ بعدی، از معیار **نزدیک‌ترین میانگین (Nearest Mean Classifer)** برای پیش‌بینی استفاده شد. به این صورت که فاصله اقلیدسی نمونه تست تا مرکز هر کلاس در فضای جدید محاسبه شده و کلاس با کمترین فاصله انتخاب گردید.

۳.۱.۱ نتایج پیاده‌سازی دستی

مدل دستی بر روی داده‌های تست به دقت 88.10% دست یافت. جدول زیر گزارش دسته‌بندی را نشان می‌دهد:

جدول ۱: گزارش دسته‌بندی برای مدل دستی LDA

Support	F۱-Score	Recall	Precision	Class
14	0.78	0.64	1.00	1
14	0.97	1.00	0.93	2
14	0.88	1.00	0.78	3
		0.88		**Accuracy**

۲.۱ مقایسه با Scikit-Learn (سوال ۱۴)

در این مرحله، نتایج با پیاده‌سازی استاندارد کتابخانه Scikit-Learn مقایسه شد. دقت مدل آماده برابر با **۸۶%.۹۲** بود.

۱.۲.۱ تحلیل اختلاف دقت

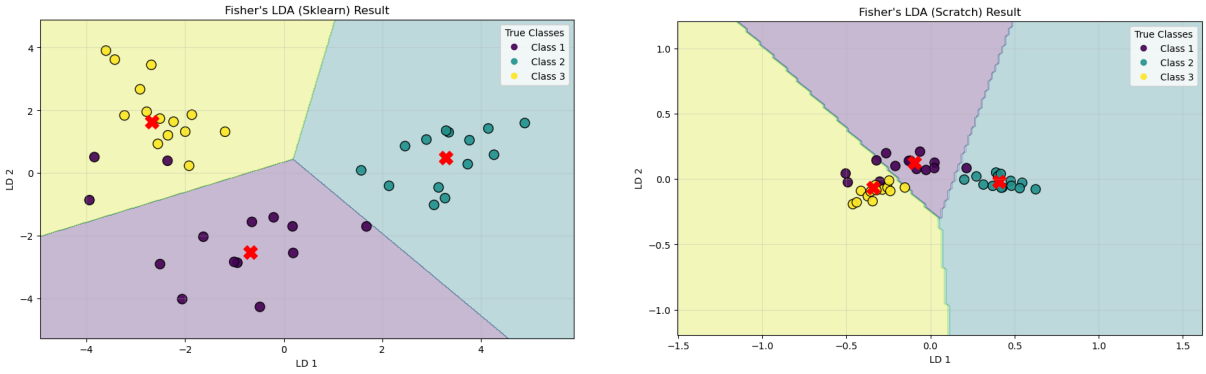
اختلاف دقت بین دو مدل حدود 4.76% است (معادل ۲ نمونه از ۴۲ نمونه تست). دلایل این اختلاف عبارتند از:

۱. **تفاوت در الگوریتم حل:** کتابخانه Sklearn به جای محاسبه مستقیم S_W^{-1} (که ممکن است ناپایدار باشد)، از روش تجزیه مقدار تکین (SVD) استفاده می‌کند که پایداری عددی بیشتری دارد.

۲. **قاعده تصمیم‌گیری:** در پیاده‌سازی دستی، ما صرفاً از "فاصله اقلیدسی تا میانگین" در فضای تصویر شده استفاده کردیم. اما Sklearn به صورت پیش‌فرض یک دسته‌بند بیزی احتمالاتی (Probabilistic) می‌سازد که کواریانس داده‌ها را نیز در نظر می‌گیرد. معیار "نزدیک‌ترین میانگین" تنها زمانی بهینه است که توزیع کلاس‌ها در فضای جدید کروی باشد، در حالی که Sklearn مرزهای دقیق‌تری ترسیم می‌کند.

۲.۲.۱ مقایسه بصری فضای ویژگی

در تصاویر زیر، نحوه توزیع داده‌ها و مرزهای تصمیم‌گیری برای هر دو مدل نمایش داده شده است:



(آ) پیاده‌سازی دستی (Scratch) (ب) پیاده‌سازی Scikit-Learn

شکل ۱: مقایسه فضای تصویر شده و مرزهای تصمیم گیری LDA

همانطور که در شکل ۱ مشاهده می‌شود، هر دو مدل توانسته‌اند ۳ کلاس را به خوبی از هم تفکیک کنند. خوشه‌های داده (نقاط رنگی) در هر دو تصویر ساختار مشابهی دارند، اما جهت محورها یا مقیاس آن‌ها ممکن است متفاوت باشد (که در روش‌های ویژه برداری طبیعی است). همچنین مرزهای تصمیم گیری (نواحی رنگی پس‌زمینه) در مدل Sklearn کمی متفاوت زاویه‌دهی شده‌اند که منجر به اصلاح دسته‌بندی آن ۲ نمونه مرزی شده است.

۲ نتیجه‌گیری کلی

پیاده‌سازی الگوریتم‌های یادگیری ماشین از پایه، درک عمیقی از ریاضیات پشت صحنه (مانند جبر خطی و بهینه‌سازی) فراهم می‌کند. در این تمرین نشان داده شد که پیاده‌سازی دستی LDA عملکرد قابل قبولی (۸۸٪) دارد و بسیار نزدیک به کتابخانه‌های صنعتی عمل می‌کند. تفاوت‌های جزئی موجود ناشی از تکنیک‌های پیشرفته پایداری عددی و تفاوت در جزئیات قاعده تصمیم گیری نهایی است.