Session 11&12

Linear Discriminant Analysis (LDA)

NLP | Machine Learning | Zahra Amini

Telegram: @zahraamini_ai & Instagram:@zahraamini_ai & LinkedIn: @zahraamini-ai

https://zil.ink/zahraamini

معرفی تحلیل افتراقی خطی (LDA)

طبقه‌بندی و کاهش ابعاد دو مفهوم اساسی در یادگیری ماشین هستند و تکنیک تحلیل افتراقی خطی هر دو این اهداف را به طور مؤثر دنبال می‌کند

طبقه‌بندی	طبقه‌بندی یک وظیفه یادگیری تحت نظارت است که هدف آن اختصاص برچسب به داده‌ها بر اساس ویژگی‌های ورودی است به عنوان مثال، در یک مسئله طبقه‌بندی ایمیل، هدف می‌تواند دسته‌بندی ایمیل‌ها به دو دسته "اسپم" یا "غیر اسپم" بر اساس محتوای ایمیل باشد هر نقطه داده به یکی از چندین کلاس تعلق دارد، و ما می‌خواهیم از ویژگی‌ها (مانند فراوانی کلمات برای داده‌های متنی) برای پیش‌بینی کلاس استفاده کنیم
کاهش ابعاد	در بسیاری از مجموعه داده‌ها، به ویژه در تحلیل متن یا پردازش تصویر، با داده‌های با ابعاد بالا کار می‌کنیم، جایی که تعداد ویژگی‌ها (یا ابعاد) بسیار بیشتر از تعداد نقاط داده است. این فضای با ابعاد بالا می‌تواند به مشکلات زیر منجر شود: بیش‌برازش: ابعاد بیشتر می‌تواند منجر به مدلی شود که نویز داده‌ها را به خاطر می‌سپارد هزینه محاسباتی: کار کردن در فضاهای با ابعاد بالا هزینه محاسباتی بالایی دارد مشکل در تجسم: تجسم داده‌ها با صدها یا هزاران ویژگی دشوار است تکنیک‌های کاهش ابعاد هدفشان این است که تعداد ویژگی‌ها را کاهش دهند، در حالی که مهم‌ترین اطلاعات مورد نیاز برای پیش‌بینی را حفظ کنند

معرفی تحلیل افتراقی خطی (LDA)

نیاز به LDA	به‌ویژه زمانی مفید است که داده‌های دارای برچسب (یعنی کلاس‌ها شناخته شده‌اند) داریم و می‌خواهیم ابعاد را کاهش دهیم در حالی که تفکیک‌پذیری کلاس‌ها حفظ شود. به عبارت ساده، این تکنیک به ما کمک می‌کند داده‌ها را به فضایی با ابعاد کمتر پروژکت کنیم، جایی که داده‌های کلاس‌های مختلف تا حد ممکن از یکدیگر متمایز باشند
چرا به LDA در مسائل طبقه‌بندی نیاز داریم؟	تفکیک‌پذیری کلاس‌ها: در بسیاری از موارد، داده‌های با ابعاد بالا ممکن است پیچیده و دشوار برای تجسم باشند. اما در فضاهای با ابعاد کمتر، ممکن است بتوانیم کلاس‌ها را بسیار واضح‌تر جدا کنیم به‌ویژه زمانی مفید است که بخواهیم ترکیب خطی‌ای از ویژگی‌ها را پیدا کنیم که بتواند بهترین تفکیک بین کلاس‌های مختلف را ایجاد کند
LDA چگونه این کار را انجام می‌دهد؟	به واریانس بین کلاس‌ها (یعنی تفاوت کلاس‌ها با یکدیگر) و واریانس درون کلاس‌ها (یعنی میزان پراکندگی نقاط داده هر کلاس در خودشان) نگاه می‌کند هدف این است که واریانس بین کلاس‌ها را به حداکثر و واریانس درون کلاس‌ها را به حداقل برسانیم

مثال

طبقه‌بندی ایمیل‌ها (اسپم در مقابل غیر اسپم)

هر ایمیل می‌تواند با مجموعه‌ای از ویژگی‌ها نشان داده شود. برای مثال:
فراوانی کلمات (مثلاً چند بار کلماتی مانند "رایگان"، "پیشنهاد"، "بخر" ظاهر می‌شوند)
طول ایمیل
حضور برخی کلمات کلیدی خاص

فضای با ابعاد بالا:
این ویژگی‌ها می‌توانند یک فضای با ابعاد بالا تشکیل دهند. تصور کنید که ما ۱۰۰۰ کلمه مختلف در مجموعه داده خود داریم و هر ایمیل با یک بردار ۱۰۰۰ بعدی نشان داده می‌شود (هر بعد نشان‌دهنده فراوانی یک کلمه خاص است). در این حالت، تفکیک ایمیل‌های اسپم از غیر اسپم دشوار می‌شود

کاربرد LDA
به ما کمک می‌کند تا این ابعاد را کاهش دهیم، در حالی که جداسازی بین اسپم و غیر اسپم تا حد ممکن واضح باقی بماند
می‌تواند داده‌ها را به فضایی با ابعاد کمتر (مثلاً ۲ یا ۳ بعد) پروژکت کند، جایی که ایمیل‌های اسپم LDA ،به جای کار با ۱۰۰۰ ویژگی کامل و غیر اسپم متمایزتر و راحت‌تر قابل طبقه‌بندی باشند

معرفی تحلیل افتراقی خطی (LDA)

هدف اصلی تحلیل افتراقی خطی این است که داده‌های با ابعاد بالا را به فضای با ابعاد پایین‌تر پروژه کنیم در حالی که حداکثر جداسازی بین کلاس‌ها را حفظ کنیم. این کار با حداکثر کردن نسبت واریانس بین کلاس‌ها به واریانس درون کلاس‌ها در فضای پروژه‌شده انجام می‌شود

واریانس درون کلاس‌ها ← S_w
این مقدار نشان‌دهنده تغییرات داده‌ها در داخل هر کلاس است، یعنی چگونه داده‌ها در هر کلاس حول میانگین خود پراکنده می‌شوند

واریانس بین کلاس‌ها ← S_B
این مقدار نشان‌دهنده تفاوت میانگین هر کلاس با میانگین کلی داده‌ها است

تابع هدف
LDA یک بردار پروجکشن w پیدا می‌کند که نسبت واریانس بین کلاس‌ها به واریانس درون کلاس‌ها را حداکثر کند

max

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$S_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

$$S_B = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

- x_1: Frequency of "buy".
- x_2: Frequency of "offer".

Data points are as follows:

Class	x_1	x_2
A	1	2
A	2	3
A	3	4
B	5	6
B	6	8
B	7	8

1. **Compute Class Means (μ_k)**

For Class A:

$$\mu_A = \frac{1}{3}\begin{bmatrix}1+2+3\\2+3+4\end{bmatrix} = \begin{bmatrix}2\\3\end{bmatrix}$$

For Class B:

$$\mu_B = \frac{1}{3}\begin{bmatrix}5+6+7\\6+8+8\end{bmatrix} = \begin{bmatrix}6\\7.33\end{bmatrix}$$

Overall mean (μ):

$$\mu = \frac{1}{6}\begin{bmatrix}1+2+3+5+6+7\\2+3+4+6+8+8\end{bmatrix} = \begin{bmatrix}4\\5.33\end{bmatrix}$$

3. Compute the Scatter Matrices

Within-Class Scatter Matrix (S_W):

The within-class scatter matrix measures the variance within each class. For each class:

$$S_{W,k} = \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

For **Class A**:

$$S_{W,A} = (x_1 - \mu_A)(x_1 - \mu_A)^T + (x_2 - \mu_A)(x_2 - \mu_A)^T + (x_3 - \mu_A)(x_3 - \mu_A)^T$$
$$= \begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} -1 & -1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

For **Class B**:

$$S_{W,B} = \begin{bmatrix} -1 & -1.33 \\ -1 & -1.33 \end{bmatrix} \begin{bmatrix} -1 & -1.33 \end{bmatrix} + \cdots$$
$$= \begin{bmatrix} 6 & 8 \\ 8 & 10.66 \end{bmatrix}$$

Combine S_W:

$$S_W = S_{W,A} + S_{W,B} = \begin{bmatrix} 8 & 10 \\ 10 & 12.66 \end{bmatrix}$$

Between-Class Scatter Matrix (S_B):

The between-class scatter matrix measures the variance of the class means relative to the overall mean:

$$S_B = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T$$

For both classes:

$$S_B = 3\begin{bmatrix} -2 & -2.33 \\ -2 & -2.33 \end{bmatrix}\begin{bmatrix} -2 & -2.33 \end{bmatrix} + 3\begin{bmatrix} 2 & 2.33 \\ 2 & 2.33 \end{bmatrix}\begin{bmatrix} 2 & 2.33 \end{bmatrix}$$

$$S_B = \begin{bmatrix} 12 & 14 \\ 14 & 16.66 \end{bmatrix}$$

Solving the Eigenvalue Problem

Step 1: Compute S_W^{-1}

We already have the within-class scatter matrix S_W:

$$S_W = \begin{bmatrix} 8 & 10 \\ 10 & 12.66 \end{bmatrix}$$

We can compute its inverse using the formula for the inverse of a 2x2 matrix:

$$S_W^{-1} = \frac{1}{\det(S_W)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Where for a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, the determinant is:

$$\det(S_W) = ad - bc$$

For the matrix $S_W = \begin{bmatrix} 8 & 10 \\ 10 & 12.66 \end{bmatrix}$, we compute the determinant:

$$\det(S_W) = (8 \times 12.66) - (10 \times 10) = 101.28 - 100 = 1.28$$

Thus, the inverse of S_W is:

$$S_W^{-1} = \frac{1}{1.28} \begin{bmatrix} 12.66 & -10 \\ -10 & 8 \end{bmatrix} = \begin{bmatrix} 9.89 & -7.81 \\ -7.81 & 6.25 \end{bmatrix}$$

Step 2: Form the Matrix Equation

Now, we have the following equation to solve for w:

$$S_W^{-1} S_B w = \lambda w$$

Where $S_B = \begin{bmatrix} 12 & 14 \\ 14 & 16.66 \end{bmatrix}$.

We need to multiply S_W^{-1} with S_B:

$$S_W^{-1} S_B = \begin{bmatrix} 9.89 & -7.81 \\ -7.81 & 6.25 \end{bmatrix} \begin{bmatrix} 12 & 14 \\ 14 & 16.66 \end{bmatrix}$$

Performing the matrix multiplication:

$$S_W^{-1} S_B = \begin{bmatrix} (9.89 \times 12) + (-7.81 \times 14) & (9.89 \times 14) + (-7.81 \times 16.66) \\ (-7.81 \times 12) + (6.25 \times 14) & (-7.81 \times 14) + (6.25 \times 16.66) \end{bmatrix}$$

$$S_W^{-1} S_B = \begin{bmatrix} 118.68 - 109.34 & 138.46 - 129.56 \\ -93.72 + 87.5 & -109.34 + 104.17 \end{bmatrix}$$

$$S_W^{-1} S_B = \begin{bmatrix} 9.34 & 8.90 \\ -6.22 & -5.17 \end{bmatrix}$$

Step 3: Find the Eigenvalues and Eigenvectors

Now, we need to find the eigenvalues and eigenvectors of the matrix $S_W^{-1} S_B$:

$$A = \begin{bmatrix} 9.34 & 8.90 \\ -6.22 & -5.17 \end{bmatrix}$$

To find the eigenvalues, we solve the characteristic equation:

$$\det(A - \lambda I) = 0$$

Where I is the identity matrix, and λ is the eigenvalue.

$$\det\left(\begin{bmatrix} 9.34 - \lambda & 8.90 \\ -6.22 & -5.17 - \lambda \end{bmatrix}\right) = 0$$

This expands to:

$$(9.34 - \lambda)(-5.17 - \lambda) - (-6.22)(8.90) = 0$$

Simplifying the determinant:

$$(9.34 - \lambda)(-5.17 - \lambda) + 55.4 = 0$$

Expanding:

$$\lambda^2 - 4.17\lambda - 48.3 + 55.4 = 0$$

$$\lambda^2 - 4.17\lambda + 7.1 = 0$$

Now we solve this quadratic equation for λ using the quadratic formula:

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

For $a = 1$, $b = -4.17$, and $c = 7.1$, we get:

$$\lambda = \frac{-(-4.17) \pm \sqrt{(-4.17)^2 - 4(1)(7.1)}}{2(1)}$$

$$\lambda = \frac{4.17 \pm \sqrt{17.39 - 28.4}}{2}$$

$$\lambda = \frac{4.17 \pm \sqrt{-11.01}}{2}$$

Thus, the two possible eigenvalues are:

$$\lambda_1 = \frac{4.17 + 6.76}{2} = 5.465$$

$$\lambda_2 = \frac{4.17 - 6.76}{2} = -1.295$$

Since the eigenvalues must be **real and positive**, we choose the eigenvalue:

$$\lambda_1 = 5.465$$

4. Eigenvector Calculation

Now, we find the eigenvector w corresponding to the largest eigenvalue $\lambda_1 = 5.465$. To do this, we substitute the eigenvalue into the equation:

$$(S_W^{-1} S_B - \lambda_1 I)w = 0$$

Substitute $\lambda_1 = 5.465$ into the matrix $S_W^{-1} S_B - \lambda_1 I$:

$$S_W^{-1} S_B - 5.465I = \begin{bmatrix} 9.34 - 5.465 & 8.90 \\ -6.22 & -5.17 - 5.465 \end{bmatrix}$$

This gives:

$$= \begin{bmatrix} 3.875 & 8.90 \\ -6.22 & -10.635 \end{bmatrix}$$

Now, solve for w by solving the system of equations:

$$\begin{bmatrix} 3.875 & 8.90 \\ -6.22 & -10.635 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0$$

This represents two linear equations:

1. $3.875w_1 + 8.90w_2 = 0$
2. $-6.22w_1 - 10.635w_2 = 0$

Step 1: Express w_1 in terms of w_2 (or vice versa)

We can start by solving the first equation for w_1 in terms of w_2:

$$3.875w_1 = -8.90w_2$$

$$w_1 = \frac{-8.90}{3.875}w_2$$

$$w_1 = -2.295w_2$$

Step 2: Substitute into the second equation

Now, substitute $w_1 = -2.295w_2$ into the second equation:

$$-6.22(-2.295w_2) - 10.635w_2 = 0$$

$$(6.22 \times 2.295)w_2 - 10.635w_2 = 0$$

$$14.26w_2 - 10.635w_2 = 0$$

$$3.625w_2 = 0$$

$$w_2 = 0$$

Step 3: Solve for w_1

Now, since $w_2 = 0$, substitute this into the equation $w_1 = -2.295w_2$:

$$w_1 = -2.295(0) = 0$$

Thus, the solution we get is:

$$w_1 = 0, \quad w_2 = 0$$

5. Select the Top k Eigenvectors

In this case, since we want a 1D projection, select the eigenvector corresponding to the largest eigenvalue:

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

6. Projection

Each data point x_i is projected onto the direction w using:

$$y_i = w^T x_i$$

For example: For $x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$:

$$y_1 = w^T x_1$$

Repeat for all points to obtain their projections in 1D space.

Suppose we have the following data point:

$$x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

And the eigenvector w corresponding to the largest eigenvalue is:

$$w = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

To compute y_1, we first calculate the dot product $w^T x_1$:

$$y_1 = w^T x_1 = \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$y_1 = 3 \times 1 + 4 \times 2 = 3 + 8 = 11$$

Therefore, the projection of x_1 onto the eigenvector w is $y_1 = 11$.

Repeat for All Data Points

This process is repeated for each data point x_i to obtain its projection in the 1D space. For example, if we have the following additional data points:

$$x_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$$

We can calculate their projections using the formula $y_i = w^T x_i$.