Análisis Discriminante Lineal: Clasificación y Reducción de Dimensionalidad

Pietro Palombini

Universidad Nacional de Córdoba

Facultad de Matemática, Astronomía, Física y Computación

20 de Noviembre de 2024

Esquema de la presentación

¿Qué es LDA?

- El Análisis Discriminante Lineal (LDA) es una técnica fundamental en:
  - Reconocimiento de patrones
  - Aprendizaje automático
- Cumple un doble propósito:
  - Clasificador
  - Método de reducción de dimensionalidad
- Desarrollado originalmente por R. A. Fisher
- Encuentra combinaciones lineales de características que optimizan la separación entre clases

Aplicaciones y Ventajas

Aplicaciones

- Reconocimiento de imágenes
- Bioinformática
- Análisis de datos de alta dimensionalidad

Ventajas

- Fronteras de decisión lineales
- Eficiente computacionalmente
- Reduce dimensionalidad preservando información discriminativa
- Óptimo cuando las suposiciones se cumplen

Clasificadores

Definición (Clasificador)

Sea $\mathcal{X} \subseteq \mathbb{R}^p$ el espacio de características y $\mathcal{Y} = \{1, \ldots, K\}$ el conjunto de etiquetas de clase. Un clasificador es una función $\hat{G} : \mathcal{X} \to \mathcal{Y}$ que asigna una etiqueta de clase a cada punto del espacio de características.

Objetivo

Encontrar un clasificador que:

- Prediga correctamente la etiqueta Y para nuevas observaciones X
- Sea computacionalmente eficiente

Error de Predicción Esperado

Definición (Error de Predicción Esperado)

Sea $X \in \mathbb{R}^p$ un vector aleatorio de características, $Y \in \{1, \ldots, K\}$ una variable aleatoria categórica que representa la etiqueta de clase, y $\hat{G} : \mathbb{R}^p \to \{1, \ldots, K\}$ un clasificador. Sea $L : \{1, \ldots, K\} \times \{1, \ldots, K\} \to \mathbb{R}$ una función pérdida que mide el costo de una clasificación incorrecta.

El Error de Predicción Esperado (EPE) del clasificador $\hat{G}(X)$ se define como la esperanza de la función de pérdida $L(Y, \hat{G}(X))$ sobre la distribución conjunta de X e Y:

$$\text{EPE}(\hat{G}) = \mathbb{E}[L(Y, \hat{G}(X))]$$

Clasificador Óptimo de Bayes

Definición (Función de Pérdida)

La función de pérdida cero-uno se define como:

$$L(y,k) = \begin{cases} 0 & \text{si } y = k \\ 1 & \text{si } y \neq k \end{cases}$$

Teorema (Clasificador Óptimo de Bayes)

El clasificador que minimiza el error de predicción esperado bajo la función de pérdida cero-uno asigna cada entrada x a la clase con la mayor probabilidad a posteriori:

$$\hat{G}(x) = \arg \max_{k=1}^{K} \mathbb{P}(Y = k \mid X = x)$$

Demostración

El error de predicción esperado del clasificador $\hat{G}(X)$ se puede escribir como:

$$\begin{aligned}
\mathbb{E}[L(Y, \hat{G}(X))] &= \sum_{y=1}^{K} \int_{\mathbb{R}^p} L(y, \hat{G}(x)) f_{X|Y}(x|y) \mathbb{P}(Y = y)\, dx \\
&= \int_{\mathbb{R}^p} \left(\sum_{y=1}^{K} L(y, \hat{G}(x)) \mathbb{P}(Y = y \mid X = x) \right) f_X(x)\, dx \\
&= \mathbb{E}_X \left[\sum_{y=1}^{K} L(y, \hat{G}(X)) \mathbb{P}(Y = y \mid X) \right]
\end{aligned}$$

Demostración (cont.)

Para minimizar esta esperanza, podemos minimizar punto a punto para cada valor de x:

$$\hat{G}(x) = \arg \min_{k=1}^{K} \sum_{y=1}^{K} L(y,k) \mathbb{P}(Y = y \mid X = x)$$

Usando la función de pérdida cero-uno:

$$\begin{aligned} \hat{G}(x) &= \arg \min_{k=1}^{K} (1 - \mathbb{P}(Y = k \mid X = x)) \\ &= \arg \max_{k=1}^{K} \mathbb{P}(Y = k \mid X = x) \end{aligned}$$

Aproximación al Clasificador de Bayes

Notación

- $f_k(x)$: densidad condicional de X dado $Y = k$
- π_k: probabilidad a priori de la clase k

Teorema de Bayes

La probabilidad a posteriori se expresa como:

$$\mathbb{P}(Y = k \mid X = x) = \frac{f_k(x)\,\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\,\pi_\ell}$$

Suposiciones de LDA

Suposición 1: Distribuciones Gaussianas

Las densidades condicionales $f_k(x)$ siguen distribuciones gaussianas multivariadas:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right)$$

Suposición 2: Matriz de Covarianza Común

Todas las clases comparten la misma matriz de covarianza Σ

Observación

Estas suposiciones conducen a fronteras de decisión lineales.

Notación

- N_k: número de observaciones en la clase k
- $N = \sum_{k=1}^{K} N_k$: número total de observaciones
- $\{(x_i, g_i)\}_{i=1}^{N}$: datos de entrenamiento

Supondremos además que los datos están centrados,

$$\hat{\mu} = \sum_{i=1}^{N} x_i = 0$$

Y que el número de muestras supera al número de características, y este a su vez al número de clases,

$$N > p > K$$

Estimadores

$$\hat{\pi}_k = \frac{N_k}{N} \qquad \text{(probabilidades a priori)}$$

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:g_i=k} x_i \qquad \text{(medias)}$$

$$\hat{\Sigma} = \frac{1}{N-K} \sum_{k=1}^{K} \sum_{i:g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \qquad \text{(covarianza)}$$

Asumiremos que la matriz de covarianza muestral es invertible.

Clasificador de LDA

$$\begin{aligned}\hat{G}(x) &= \arg\max_{k=1}^{K} \mathbb{P}(Y = k \mid X = x) \\ &= \arg\max_{k=1}^{K} \left\{ -\frac{1}{2}(x - \hat{\mu}_k)^T \hat{\Sigma}^{-1}(x - \hat{\mu}_k) + \log \hat{\pi}_k \right\} \\ &= \arg\max_{k=1}^{K} \left\{ x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k \right\}\end{aligned}$$

Definición (Función Discriminante)

La función discriminante para la clase k es:

$$\delta_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log(\hat{\pi}_k)$$

Implementación

La función discriminante se simplifica a

$$\delta_k(x) = x^T c_k + d_k$$

donde:

- $c_k = \hat{\Sigma}^{-1}\hat{\mu}_k$
- $d_k = -\frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k$

Matrices de datos

Matriz de Datos Centrados por Clase de la clase k: Para la clase k:

$$X^{(k)} = \begin{bmatrix} x_1 - \mu_k \\ x_2 - \mu_k \\ \vdots \\ x_{N_k} - \mu_k \end{bmatrix}$$

Matriz de Datos Centrados:

$$X_c = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(K)} \end{bmatrix}$$

Cómputo mediante SVD

Sea $X_c = UDV^T$ la SVD de X_c. Entonces:

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{N-K} X_c^T X_c \\ &= \frac{1}{N-K} V D^2 V^T\end{aligned}$$

Y

$$\begin{aligned}\hat{\Sigma}^{-1} &= \left(\frac{1}{N-K} V D^2 V^T\right)^{-1} \\ &= V D^{-2} V^T (N-K)\end{aligned}$$

Definimos:

$$\alpha_k = V^T \mu_k$$
$$\beta_k = D^{-1} \alpha_k$$

Y tenemos

$$c_k = (N - K) V D^{-2} \alpha_k$$
$$d_k = -\frac{N - K}{2} \|\beta_k\|_2^2 + \log \pi_k$$

Whitening

Objetivo

Transformar los datos para que la matriz de covarianza sea la identidad:

$$\text{Cov}(WX) = I$$

Solución

$$W = \sqrt{N-K}D^{-1}V^T$$

Datos Transformados

- Muestra: $x^* = Wx$
- Medias: $\mu_k^* = W\hat{\mu}_k$

Función Discriminante en el Espacio Transformado

Definición (Función Discriminante Transformada)

En el espacio transformado, la función discriminante se simplifica a:

$$\delta_k^*(x^*) = -\frac{1}{2}\|x^* - \mu_k^*\|_2^2 + \log \hat{\pi}_k$$

Teorema

El subespacio generado por las medias transformadas μ_k^* es de dimensión menor o igual a $K-1$:

$$\dim(\text{span}(\mu_1^*, \ldots, \mu_K^*)) \leq K - 1$$

Matrices de Dispersión

Definición (Matriz de Dispersión Dentro de Clases)

$$S_W = \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T = X_c^T X_c = (N-K)\hat{\Sigma}$$

Definición (Matriz de Dispersión Entre Clases)

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

Objetivo

Encontrar la matriz de proyección W que maximice la razón de la dispersión entre clases a la dispersión dentro de clases

Criterio de Fisher

Definición (Criterio de Fisher)

Para el caso unidimensional ($L = 1$), el criterio de Fisher se define como:

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

Teorema

El criterio de Fisher es invariante bajo un escalado del vector de proyección:

$$\forall \alpha \in \mathbb{R} \setminus \{0\}, \quad J(\alpha w) = J(w)$$

Optimización

Problema de Optimización

$$\begin{aligned} \underset{w \in \mathbb{R}^p}{\text{maximizar}} \quad & w^T S_B w \\ \text{sujeto a} \quad & w^T S_W w = 1 \end{aligned}$$

Teorema

La solución está dada por el autovector correspondiente al mayor autovalor de la matriz $S_W^{-1} S_B$.

Demostración

El Lagrangiano del problema es:

$$\mathcal{L}(w, \lambda) = w^T S_B w - \lambda(w^T S_W w - 1)$$

Para derivar respecto a w, usamos la identidad:

$$\frac{\partial w^T A w}{\partial w} = 2Aw$$

Derivando e igualando a cero:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} &= 2S_B w - 2\lambda S_W w = 0 \\ S_B w &= \lambda S_W w \\ S_W^{-1} S_B w &= \lambda w \end{aligned}$$

Demostración (cont.)

Por lo tanto, w es un autovector de $S_W^{-1} S_B$. Sustituyendo:

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

Entonces w debe ser el autovector correspondiente al mayor autovalor de $S_W^{-1} S_B$.

Caso $L > 1$

La matriz de proyección W es una matriz $W \in \mathbb{R}^{p \times L}$, y $W^T S_B W$ y $W^T S_W W$ son ahora matrices $L \times L$.

Definición

El criterio de Fisher generalizado se define como:

$$J(W) = \text{tr}((W^T S_B W)^{-1}(W^T S_W W))$$

Teorema

El máximo del criterio de Fisher generalizado $J(W)$ es alcanzado por la matriz W cuyas columnas son los autovectores correspondientes a los L mayores autovalores de la matriz $S_W^{-1} S_B$.

Conclusiones

Aspectos Teóricos

- LDA aproxima el clasificador óptimo de Bayes bajo suposiciones gaussianas
- Proporciona una reducción de dimensionalidad supervisada
- Las fronteras de decisión son lineales

Aspectos Prácticos

- Implementación computacionalmente eficiente mediante SVD
- Proyecciones que maximizan la separabilidad entre clases
- Útil tanto para clasificación como para visualización

Referencias

Hastie, T., Tibshirani, R., & Friedman, J.
The Elements of Statistical Learning.
Springer, 2009.

Bishop, C. M.
Pattern Recognition and Machine Learning.
Springer, 2006.

Qi Wei.
Mathematical Foundations of Machine Learning.
Broadview Press, 2023.

Fukunaga, K.
Introduction to Statistical Pattern Recognition.
Academic Press, 1990.

Duda, R. O., Hart, P. E., & Stork, D. G.
Pattern Classification.
Wiley, 2012.

Welling, M.
Fisher Linear Discriminant Analysis.
Department of Computer Science, University of Toronto, 2006.