Proyecto: "Análisis Discriminante Lineal para Clasificación y reducción de dimensionalidad"

Análisis Numérico II - 2024

Pietro Palombini

13 de Noviembre de 2024

Resumen

Este informe estudia el Análisis Discriminante Lineal (LDA) y sus aplicaciones en clasificación y reducción de dimensionalidad. Comenzamos presentando los fundamentos de la clasificación, donde los clasificadores asignan vectores de características a etiquetas de clase, y analizamos cómo el clasificador óptimo de Bayes minimiza el error de predicción mediante el uso de probabilidades a posteriori.

Dado que en la práctica raramente se dispone de las probabilidades a posteriori exactas, mostramos cómo LDA aproxima el clasificador de Bayes bajo el supuesto de densidades condicionales gaussianas con matrices de covarianza iguales. Explicamos el proceso de estimación de estos parámetros a partir de datos de entrenamiento para construir funciones discriminantes prácticas.

Posteriormente, examinamos el rol de LDA en la reducción de dimensionalidad, donde se proyectan los datos a un espacio de menor dimensión maximizando la separación entre clases. Esto conlleva resolver un problema de optimización que equilibra la varianza entre clases y dentro de las clases mediante descomposición en autovalores.

El informe incluye tanto las derivaciones teóricas como un algoritmo computacionalmente eficiente para implementar LDA.

1. Introducción

El Análisis Discriminante Lineal (LDA) es una técnica fundamental en el reconocimiento de patrones y aprendizaje automático que cumple un doble propósito: actúa como clasificador y como método de reducción de dimensionalidad. Desarrollado originalmente por R. A. Fisher, LDA encuentra combinaciones lineales de características que optimizan la separación entre clases, lo que lo hace valioso tanto para el análisis como para el preprocesamiento de datos de alta dimensionalidad.

Como clasificador, LDA opera bajo el supuesto de que las clases siguen distribuciones gaussianas con matrices de covarianza iguales pero medias diferentes. Esto conduce a fronteras de decisión lineales, eficientes e interpretables que son efectivas en diversas aplicaciones, desde el reconocimiento de imágenes hasta la bioinformática.

En su rol de reducción de dimensionalidad, LDA proyecta los datos en un espacio de menor dimensión mientras preserva la información discriminatoria de las clases. Al optimizar la razón entre la varianza entre clases y la varianza dentro de las clases, proporciona reducción de dimensionalidad manteniendo el rendimiento de clasificación.

Este informe presenta los fundamentos teóricos de LDA, comenzando con teoría de clasificación y el clasificador óptimo de Bayes. Mostramos cómo LDA aproxima el clasificador de Bayes y derivamos su formulación utilizando datos de entrenamiento. También examinamos sus aspectos de reducción de dimensionalidad, analizando casos de proyección tanto simples como múltiples.

El trabajo se basa en textos clásicos de aprendizaje estadístico, incluyendo *The Elements of Statistical Learning* [1], *Pattern Recognition and Machine Learning* [2], *Introduction to Statistical Pattern Recognition* [4] y *Pattern Classification* [5].

2. LDA como clasificador

La clasificación constituye una de las tareas fundamentales del aprendizaje automático y el análisis estadístico. En esencia, un clasificador es una función que asigna características de entrada a categorías o clases predefinidas. El objetivo es desarrollar modelos capaces de realizar predicciones precisas sobre datos nuevos y no vistos, mediante el aprendizaje de patrones en ejemplos de entrenamiento. Si bien existen diversos enfoques para la clasificación, desde sistemas basados en reglas simples hasta redes neuronales complejas, nos interesa particularmente comprender la eficacia de un clasificador y qué lo hace óptimo. Esto nos conduce a un marco formal para evaluar el rendimiento del clasificador a través del concepto de error de predicción esperado.

Primero estableceremos las bases teóricas mediante Teoría de Probabilidad para entender qué hace que un clasificador sea óptimo. Esto nos llevará al concepto de probabilidades a posteriori y su papel en las decisiones de clasificación. Posteriormente, pasaremos a la práctica mostrando cómo estos conocimientos teóricos pueden implementarse utilizando datos de entrenamiento y técnicas de álgebra lineal, llegando finalmente al Análisis Discriminante Lineal (LDA) como método de clasificación práctico.

2.1. Teoría de Clasificación

Definición 1. *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.*

El objetivo de la clasificación es hallar un clasificador que prediga correctamente la etiqueta de clase verdadera $Y \in \mathcal{Y}$ para nuevas observaciones $X \in \mathcal{X}$. Para evaluar y comparar distintos clasificadores, necesitamos una manera formal de medir su desempeño. Esto nos lleva al concepto de error de predicción esperado.

Definición 2. *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:

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

La elección de la función de pérdida es crucial para medir el rendimiento de un clasificador. Diferentes funciones de pérdida pueden llevar a clasificadores óptimos distintos. En esta sección, nos enfocaremos en la función de pérdida cero-uno, que asigna una penalización de 1 por cada mala clasificación y 0 por una clasificación correcta.

Definición 3. *La función de pérdida cero-uno se define como:*

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

Estamos interesados en encontrar el clasificador que minimice el error de predicción esperado bajo la función de pérdida cero-uno, que, intuitivamente, es el clasificador que asigna cada entrada a la clase con la mayor probabilidad a posteriori.

Teorema 1. *Sea X un vector aleatorio con función de densidad de probabilidad $f_X : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$, y sea Y una variable aleatoria discreta que toma valores en $\{1, \ldots, K\}$. Supongamos que las funciones de densidad de probabilidad condicional $f_{X|Y}(x|y)$ existen para todo $y \in \{1, \ldots, K\}$.*

El clasificador que minimiza el error de predicción esperado (EPE) bajo la función de pérdida cero-uno asigna cada entrada x a la clase con la mayor probabilidad a posteriori $\mathbb{P}(Y = k \mid X = x)$.

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

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

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

donde usamos el teorema de Bayes y el teorema de la probabilidad total. 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:

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

□

Este resultado teórico se conoce como el clasificador óptimo de Bayes, y proporciona la base para desarrollar algoritmos de clasificación prácticos.

Definición 4. *El clasificador óptimo de Bayes es el clasificador que asigna cada entrada x a la clase k que maximiza la probabilidad a posteriori $\mathbb{P}(Y = k \mid X = x)$:*

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

El clasificador óptimo de Bayes minimiza el error de predicción esperado y proporciona el mejor rendimiento de clasificación posible dadas las verdaderas distribuciones condicionales de las clases.

2.2. LDA: Aproximación al clasificador óptimo de Bayes

En la práctica, obtener el clasificador óptimo de Bayes resulta generalmente imposible, ya que requiere conocer las probabilidades a posteriori $\mathbb{P}(Y = k \mid X = x)$. Estas probabilidades típicamente son desconocidas, pues dependen de la distribución subyacente de los datos, que rara vez está disponible en escenarios reales. Por tanto, necesitamos aproximar el clasificador de Bayes mediante suposiciones y estimaciones.

Para aproximar el clasificador de Bayes, emplearemos el teorema de Bayes. Introducimos la siguiente notación:

- Sea $f_k(x)$ la densidad condicional de X condicionada por la clase $Y = k$.
- Sea π_k la probabilidad a priori de la clase k.

Mediante el teorema de Bayes, podemos expresar la probabilidad a posteriori $\mathbb{P}(Y = k \mid X = x)$ como:

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

Esta formulación nos permite aproximar las probabilidades a posteriori utilizando estimaciones de las densidades condicionales $f_k(x)$ y las probabilidades a priori π_k.

LDA realiza las siguientes suposiciones fundamentales:

1. **Distribuciones gaussianas multivariadas**: Las densidades condicionales $f_k(x)$ siguen distribuciones gaussianas multivariadas con medias μ_k y matriz de covarianza común Σ:

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

2. **Matriz de covarianza común**: Todas las clases comparten la misma matriz de covarianza Σ.

Los parámetros de la distribución y las probabilidades a priori pueden estimarse a partir de datos de entrenamiento. Con este propósito, definimos algunas notaciones:

- Sea N_k el número de observaciones en la clase k.
- Sea $N = \sum_{k=1}^{K} N_k$ el número total de observaciones.
- Sea $\{(x_i, g_i)\}_{i=1}^{N}$ los datos de entrenamiento, donde x_i son los vectores de características y g_i son las etiquetas de clase verdaderas.

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,

$$N > p > K$$

Estimaremos los parámetros del modelo mediante:

- **Probabilidades a Priori de Clase**:

$$\hat{\pi}_k = \frac{N_k}{N}$$

- **Medias de Clase**:

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i:g_i=k} x_i$$

- **Matriz de Covarianza**:

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

Asumiremos que la matriz de covarianza muestral es invertible.

2.2.1. Derivación del clasificador

Partiendo del clasificador óptimo de Bayes, derivaremos el clasificador LDA bajo las suposiciones anteriores y utilizando las estimaciones mencionadas.

$$\begin{aligned}
\hat{Y} &= \arg\max_{k=1}^{K} \mathbb{P}(Y = k \mid X = x) \\
&= \arg\max_{k=1}^{K} \frac{f_k(x)\,\hat{\pi}_k}{\sum_{\ell=1}^{K} f_\ell(x)\,\hat{\pi}_\ell} \\
&= \arg\max_{k=1}^{K} f_k(x)\hat{\pi}_k \\
&= \arg\max_{k=1}^{K} \left\{ \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}^{-1}(x-\hat{\mu}_k) \right) \hat{\pi}_k \right\} \\
&= \arg\max_{k=1}^{K} \left\{ -\frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}^{-1}(x-\hat{\mu}_k) + \log\hat{\pi}_k \right\} \qquad (2) \\
&= \arg\max_{k=1}^{K} \left\{ -\frac{1}{2}x^T\hat{\Sigma}^{-1}x + x^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k \right\} \\
&= \arg\max_{k=1}^{K} \left\{ x^T\hat{\Sigma}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\Sigma}^{-1}\hat{\mu}_k + \log\hat{\pi}_k \right\}
\end{aligned}$$

Hemos derivado un conjunto de funciones que, dado un vector de características no visto x, el clasificador lo asigna a la clase k que maximiza la respectiva función. Estas funciones se conocen como funciones discriminantes.

Definición 5. *La función discriminante para la clase k es:*

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

2.2.2. Implementación numérica

Para evaluar las funciones discriminantes eficientemente, calcularemos y almacenaremos los coeficientes c_k y d_k para cada clase k a partir de los datos de entrenamiento:

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

Luego, la función discriminante se simplifica a:

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

Aunque podríamos calcular explícitamente la matriz de covarianza y su inversa, en la práctica esto se vuelve computacionalmente costoso para conjuntos de datos con un gran número de características, así como numéricamente inestable, por lo que se utilizará un algoritmo diferente.

Definición 6. *La matriz de datos centrados por clase de la clase k se define como la matriz $X^{(k)} \in \mathbb{R}^{N_k \times p}$ cuyos filas son los puntos de datos de la clase k con la respectiva media restada:*

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

Definición 7. *La matriz de datos centrados por clase $X_c \in \mathbb{R}^{N \times p}$ es la concatenación de las matrices de datos centrados por clase para todas las clases:*

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

Asumiremos para simplificar que X_c tiene rango completo.

Observemos que ahora podemos escribir la matriz de covarianza muestral definida en la Ecuación 1 como:

$$\hat{\Sigma} = \frac{1}{N-K}X_c^T X_c$$

Sea $X_c = UDV^T$ la descomposición en valores singulares (SVD) de la matriz de datos centrados por clase X_c. Entonces, la matriz de covarianza muestral se puede escribir como:

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

Y la inversa de la matriz de covarianza muestral como:

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

Para reducir el número de multiplicaciones, definimos los coeficientes intermedios α_k y β_k para cada clase k:

$$\begin{aligned}\alpha_k &= V^T \mu_k \\ \beta_k &= D^{-1} \alpha_k\end{aligned}$$

Ahora, los coeficientes c_k se simplifican a

$$\begin{aligned}c_k &= \Sigma^{-1} \mu_k \\ &= V D^{-2} V^T (N-K) \mu_k \\ &= V D^{-2} \alpha_k (N-K)\end{aligned}$$

y los coeficientes d_k se simplifican a

$$\begin{aligned}d_k &= -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \\ &= -\frac{1}{2} \mu_k^T V D^{-2} V^T (N-K) \mu_k + \log \pi_k \\ &= -\frac{1}{2} \alpha_k^T D^{-2} \alpha_k (N-K) + \log \pi_k \\ &= -\frac{1}{2} \beta_k^T \beta_k (N-K) + \log \pi_k \\ &= -\frac{N-K}{2} \|\beta_k\|_2^2 + \log \pi_k\end{aligned}$$

El algoritmo resultante para entrenar el clasificador LDA se resume en el Algoritmo 1.

3. LDA para reducción de dimensionalidad

3.1. Representación $K-1$ dimensional

Observemos que en la ecuación 2, en la derivación del clasificador, tenemos una expresión que se asemeja a una distancia euclidiana al cuadrado, salvo por la presencia de $\hat{\Sigma}^{-1}$. Para simplificar esto, podemos transformar los datos de modo que la matriz de covarianza se convierta en la identidad, lo que nos permitirá usar directamente distancias euclidianas.

En términos del vector aleatorio X, la transformación W que queremos encontrar debe satisfacer $\text{Cov}(WX) = I$. Podemos desarrollar el lado izquierdo de esta ecuación como:

$$\begin{aligned}\text{Cov}(WX) &= \mathbb{E}\left[(WX - \mathbb{E}[WX])(WX - \mathbb{E}[WX])^T\right] \\ &= \mathbb{E}\left[(WX - W\mu)(WX - W\mu)^T\right] \\ &= \mathbb{E}\left[W(X-\mu)(X-\mu)^T W^T\right] \\ &= W\mathbb{E}\left[(X-\mu)(X-\mu)^T\right] W^T \\ &= W\text{Cov}(X)W^T \qquad (3)\end{aligned}$$

En LDA, estimamos $\text{Cov}(X)$ con $\hat{\Sigma}$, por lo que queremos encontrar W tal que $W\hat{\Sigma}W^T = I$. A partir de la computación de la SVD que realizamos anteriormente, sabemos que $\hat{\Sigma} = \frac{1}{N-K} V D^2 V^T$, donde V es ortogonal y D es diagonal. Por lo tanto,

Algorithm 1 Algoritmo de Entrenamiento de LDA

Require: Datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^{N}$, con $x_i \in \mathbb{R}^p$, $y_i \in \{1, \ldots, K\}$
Ensure: Coeficientes $\{(c_k, d_k)\}_{k=1}^{K}$ para las funciones discriminantes

Inicialización:

- $N \leftarrow$ número total de muestras
- Para cada clase k, inicializar $N_k \leftarrow 0$, $\hat{\mu}_k \leftarrow 0$

Calcular Conteos de Clases y Sumas:
for $i = 1$ to N **do**
  $k \leftarrow y_i$
  $N_k \leftarrow N_k + 1$
  $\hat{\mu}_k \leftarrow \hat{\mu}_k + x_i$
end for
Calcular Medias de Clases y Priors:
for $k = 1$ to K **do**
  $\hat{\mu}_k \leftarrow \hat{\mu}_k / N_k$
  $\hat{\pi}_k \leftarrow N_k / N$
end for
Construir la Matriz de Datos Centrados:
Inicializar $X_c \in \mathbb{R}^{N \times p}$
$j \leftarrow 1$
for $k = 1$ to K **do**
  for cada x_i tal que $y_i = k$ **do**
    $X_c(j, :) \leftarrow x_i - \hat{\mu}_k$
    $j \leftarrow j + 1$
  end for
end for
Calcular SVD de X_c:
$[U, D, V^\top] \leftarrow \text{svd}(X_c)$
Calcular Coeficientes para Cada Clase:
for $k = 1$ to K **do**
  $\alpha_k \leftarrow V^\top \hat{\mu}_k$
  $\beta_k \leftarrow D^{-1} \alpha_k$
  $c_k \leftarrow (N - K) V D^{-2} \alpha_k$
  $d_k \leftarrow -\dfrac{N - K}{2} \|\beta_k\|_2^2 + \ln \hat{\pi}_k$
end for
return Coeficientes $\{(c_k, d_k)\}_{k=1}^{K}$

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

Entonces, podemos ver que $W = \sqrt{N-K}D^{-1}V^T$ satisface la ecuación. Efectivamente,

$$
\begin{aligned}
W\hat{\Sigma}W^T &= (\sqrt{N-K}D^{-1}V^T)\frac{1}{N-K}VD^2V^T(\sqrt{N-K}D^{-1}V^T)^T \\
&= (D^{-1}V^T)VD^2V^T(D^{-1}V^T)^T \\
&= D^{-1}V^TVD^2V^TVD^{-1} \\
&= D^{-1}D^2D^{-1} \\
&= D^{-1}D^{1/2}D^{1/2}D^{-1} \\
&= I
\end{aligned}
$$

Esto se llama una transformación de sphering o whitening. Entonces podemos definir:

- La muestra transformada: $x^* = \sqrt{N-K}D^{-1}V^Tx$
- Las medias transformadas: $\mu_k^* = \sqrt{N-K}D^{-1}V^T\hat{\mu}_k$

Bajo esta transformación, el criterio de decisión se simplifica a:

$$
\begin{aligned}
\hat{Y} &= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}^{-1}(x-\hat{\mu}_k)+\log\hat{\pi}_k\right\} \\
&= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}(VDx^*-VD\mu_k^*)^TV^TD^{-2}V(VDx^*-VD\mu_k^*)+\log\hat{\pi}_k\right\} \\
&= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}(Dx^*-D\mu_k^*)^TD^{-2}(Dx^*-D\mu_k^*)+\log\hat{\pi}_k\right\} \\
&= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}(x^*-\mu_k^*)^T(DD^{-2}D)(x^*-\mu_k^*)+\log\hat{\pi}_k\right\} \\
&= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}(x^*-\mu_k^*)^T(x^*-\mu_k^*)+\log\hat{\pi}_k\right\} \\
&= \arg\max_{k=1}^{K}\left\{-\frac{1}{2}\|x^*-\mu_k^*\|_2^2+\log\hat{\pi}_k\right\}
\end{aligned}
$$

Definición 8. *La función discriminante para la clase k en el espacio transformado es:*

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

Esta transformación hace que la regla de decisión sea más intuitiva, reduciendo efectivamente el problema de clasificación a uno basado en distancias euclidianas a K puntos en el espacio transformado, mientras que mantiene las mismas fronteras de decisión. Observemos que solo la proyección de x en el subespacio abarcado por las medias transformadas de clase afecta la decisión. Esto sugiere que LDA puede usarse como una técnica de reducción de dimensionalidad, proyectando los datos en un espacio de menor dimensión mientras se maximiza la separabilidad de las clases.

Teorema 2. *El subespacio generado por las medias transformadas μ_k^* es de dimensión menor o igual a $K-1$.*

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

Demostración. Asumimos que los datos están centrados, por lo tanto:

$$\mu = 0$$
$$\frac{1}{N}\sum_{k=1}^{K} N_k \mu_k = 0$$
$$\frac{1}{N}\sum_{k=1}^{K} N_k W \mu_k = W0$$
$$\frac{1}{N}\sum_{k=1}^{K} N_k \mu_k^* = 0$$

Esto significa que las medias μ_k^* son linealmente dependientes, y por lo tanto, el subespacio que abarcan debe tener una dimensión estrictamente menor que K. □

En bibliotecas modernas de aprendizaje automático como `scikit-learn`, a menudo se aplica una transformación de de-sphering como paso final cuando se usa LDA para reducción de dimensionalidad, de modo que los datos conserven la estructura de covarianza original. A cambio, la regla de decisión es menos intuitiva si se realiza un gráfico en el espacio transformado.

3.2. Representación L-dimensional

Con frecuencia, $K-1$ dimensiones siguen siendo demasiadas para visualización o procesamiento adicional. En estos casos, podemos desear proyectar los datos en un subespacio de L dimensiones, donde $L \leq K-1$, mientras maximizamos la separabilidad entre clases. Para ello, buscamos la matriz $W \in \mathbb{R}^{p\times L}$ que proyecte los datos en un subespacio de L dimensiones, maximizando la varianza entre clases mientras minimiza la varianza dentro de las clases.

Para cuantificar la varianza dentro de clases y entre clases, definimos las siguientes matrices:

Definición 9. *La matriz de dispersión dentro de clases S_W se define como:*

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

Observemos que S_W puede expresarse en términos de la matriz de datos centrados por clase X_c o de la matriz de covarianza muestral $\hat{\Sigma}$ como:

$$S_W = X_c^T X_c = (N-K)\hat{\Sigma}$$

Definición 10. *La matriz de dispersión entre clases S_B se define como:*

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

Bajo nuestra suposición de datos centrados, S_B se simplifica a:

$$S_B = \sum_{k=1}^{K} N_k \mu_k \mu_k^T$$

El objetivo es encontrar la matriz de proyección W que maximice la razón de la dispersión entre clases a la dispersión dentro de clases.

3.2.1. Caso $L = 1$

Para $L = 1$, la matriz de proyección W es un vector $w \in \mathbb{R}^p$, y a partir de la ecuación 3, vemos que maximizar la dispersión entre clases equivale a maximizar $w^T S_B w$, mientras que minimizar la dispersión dentro de clases equivale a minimizar $w^T S_W w$. Unificando estos dos objetivos, definimos el criterio de Fisher.

Definición 11. *El criterio de Fisher se define como la razón de la dispersión entre clases a la dispersión dentro de clases:*

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

Teorema 3. *El criterio de Fisher es invariante bajo un escalado del vector de proyección w.*

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

Demostración.

$$\begin{aligned} J(\alpha w) &= \frac{(\alpha w)^T S_B (\alpha w)}{(\alpha w)^T S_W (\alpha w)} \\ &= \frac{\alpha^2 w^T S_B w}{\alpha^2 w^T S_W w} \\ &= \frac{w^T S_B w}{w^T S_W w} \\ &= J(w) \end{aligned}$$

□

Debido a esta invariancia, podemos imponer la restricción $w^T S_W w = 1$ para simplificar el problema de optimización, que ahora se escribe como:

$$\begin{aligned} \underset{w \in \mathbb{R}^p}{\text{maximize}} \quad & w^T S_B w \\ \text{subject to} \quad & w^T S_W w = 1 \end{aligned} \tag{4}$$

Teorema 4. *La solución al problema de optimización 4 está dada por el autovector correspondiente al mayor autovalor de la matriz $S_W^{-1} S_B$.*

Demostración. El Lagrangiano del problema de optimización es:

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

Para derivar con respecto a w, utilizamos la siguiente identidad del cálculo matricial:

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

Derivando el Lagrangiano con respecto a w e igualando a cero, obtenemos:

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

Por lo tanto, la solución debe ser un autovector de $S_W^{-1}S_B$, asociado al autovalor λ. Reemplazando en la expresión original,

$$\begin{aligned} J(w) &= \frac{w^T S_B w}{w^T S_W w} \\ &= \frac{w^T \lambda S_W w}{w^T S_W w} \\ &= \lambda \end{aligned}$$

Por lo tanto, w debe ser el autovector correspondiente al mayor autovalor de $S_W^{-1}S_B$. □

3.2.2. Caso $L > 1$

Para $L > 1$, 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 $L \times L$. Existen diferentes generalizaciones del criterio de Fisher, la mayoría de las cuales, sin embargo, tienen la misma solución. Una generalización común es la siguiente:

Definición 12. *El criterio de Fisher generalizado se define como:*

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

Teorema 5. *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 $S_W^{-1}S_B$.*

La demostración de este teorema es sustancialmente más compleja que el caso $L = 1$, ya que el uso de la derivada del operador traza respecto a una matriz está involucrado, haciendo las ecuaciones más intrincadas. La estructura de la prueba puede encontrarse en [4].

Referencias

[1] Hastie, T., Tibshirani, R., & Friedman, J. *The elements of statistical learning.* Springer, 2009.

[2] Bishop, C. M. *Pattern recognition and machine learning.* Springer, 2006.

[3] Qi Wei. *Mathematical Foundations of Machine Learning.* Broadview Press, 2023.

[4] Fukunaga, K. *Introduction to statistical pattern recognition.* Academic press, 1990.

[5] Duda, R. O., Hart, P. E., & Stork, D. G. *Pattern classification.* John Wiley & Sons, 2012.

[6] Welling, M. *Fisher Linear Discriminant Analysis.* Department of Computer Science, University of Toronto, 2006.