Feature Extraction
(Extracción de características)

manuel.smontanes@uam.es
Manuel Sánchez-Montañés

Reducción de la Dimensionalidad

Dos estrategias:

Selección de características: se selecciona un subconjunto de los atributos originales.

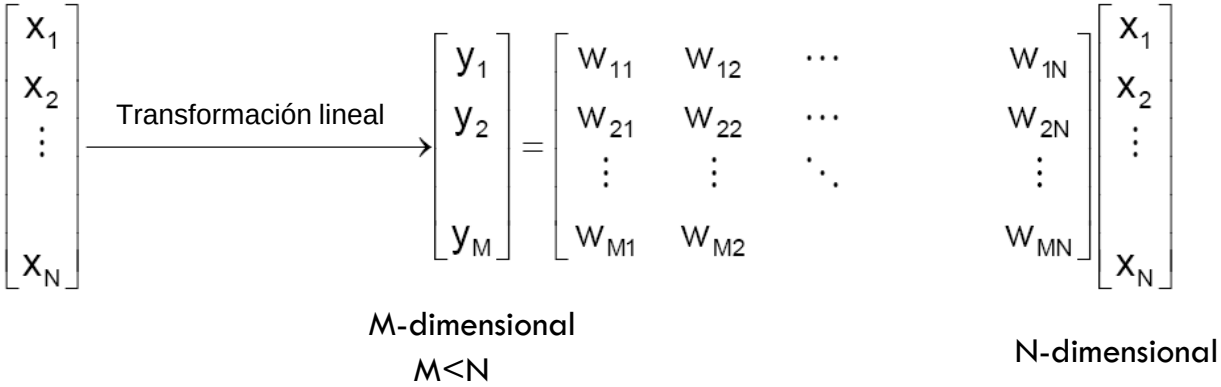
- **Algoritmos de Filtrado:** la calidad de los atributos se mide usando alguna medida estadística general (Chi Cuadrado, Correlación, Información Mutua, etc.)
- **Algoritmos de Wrapping:** la calidad de un conjunto de atributos se evalúa calculando la calidad promedio de un modelo entrenado con esos atributos
- **Algoritmos embebidos:** la selección es parte del algoritmo de aprendizaje

Extracción / construcción de características: la idea es construir nuevas características que condensen / resuman la información relevante de los atributos originales.
Por ejemplo PCA, LDA

Reducción de la dimensionalidad con transformaciones de los atributos (1)

- Idea: encontrar una transformación y=f(x) que conserve la información acerca del problema, minimizando el número de componentes
- En general, la función óptima y=f(x) será no lineal
- Sin embargo, no hay una forma de generar sistemáticamente transformaciones no lineales:
  - La selección de un subconjunto particular de transformaciones depende del problema
  - Por esta razón, la limitación a transformaciones lineales ha sido ampliamente aceptada, $y = W^T x$
    ➔ **y es una proyección lineal de x**

Reducción de la dimensionalidad con transformaciones de los atributos (2)



- Por el momento nos centraremos en transformaciones lineales

Representación de la señal *versus* clasificación (PCA vs. LDA)

- La selección de la transformación extractora de características, **y=f(x)** , está guiada por una función objetivo que buscamos maximizar (o minimizar)
- Dependiendo del criterio usado por la función objetivo, las técnicas de extracción de características se dividen en dos categorías:

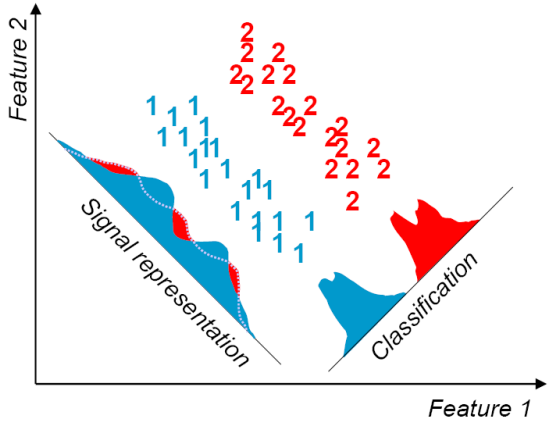
  - Clasificación: El objetivo de la transformación extractora de características es resaltar en un espacio de menos dimensiones la información discriminante de clases

  - Representación de la señal: El objetivo de la transformación extractora de características es representar los vectores de atributos

  de manera precisa en un espacio de menos dimensiones

- Hay dos técnicas principales en la extracción lineal de características:

  - Análisis Discriminante Lineal (LDA), que utiliza el criterio de clasificación
  - Análisis de Componentes Principales (PCA), que usa el criterio de representación de la señal



Análisis Discriminante Lineal (LDA)

- Análisis Discriminante Lineal, dos clases
- Análisis Discriminante Lineal, C clases
- Limitaciones de LDA
- Variantes de LDA
- Otros métodos de reducción de la dimensionalidad basados en LDA

Análisis Discriminante Lineal, dos clases (1)

- El objetivo de **LDA** es realizar una **reducción de la dimensionalidad** preservando el máximo posible de **información sobre la clase**.

Tenemos un conjunto de vectores en **D** dimensiones $\{x_1, x_2, \ldots, x_N\}$, donde N_1 son de clase ω_1, y N_2 de clase ω_2

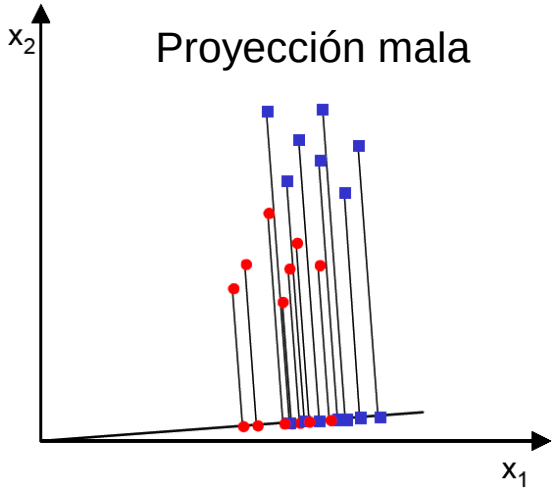
Buscamos obtener un nuevo feature "**y**" proyectando los vectores **x** sobre un vector **w**:

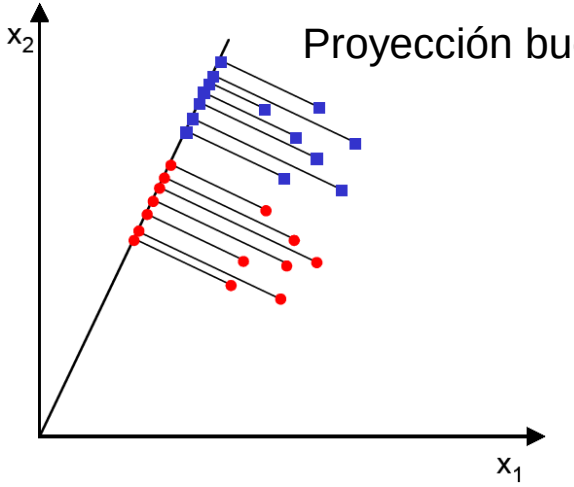
$$y = w^T x$$

Análisis Discriminante Lineal, dos clases (2)

- De todos los posibles **w**, nos gustaría seleccionar el que maximiza la separación de las clases en la proyección $y = w^T x$

- Ilustramos a continuación esta idea para el caso de vectores x con 2 dimensiones:





Análisis Discriminante Lineal, dos clases (3)

- Para poder encontrar un buen vector de proyección, necesitaremos definir una medida de separación entre las proyecciones

- El vector promedio de cada clase en los espacios **x** e **y** es:

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \qquad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T \mu_i$$

Espacio original — Espacio de la proyección

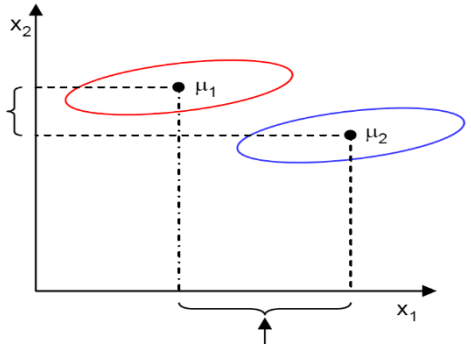
- Podríamos entonces elegir nuestra función objetivo como la distancia entre los promedios proyectados: clasificación por la distancia a las medias

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

Análisis Discriminante Lineal, dos clases (4)

- Sin embargo, la distancia entre los promedios proyectados no es una buena medida ya que no tiene en cuenta la desviación standard dentro de las clases.

este eje proporciona la mejor separación entre clases



este eje maximiza la distancia entre medias

Análisis Discriminante Lineal, dos clases (5)

- La solución propuesta por Fisher es maximizar una función que representa la diferencia entre las medias, normalizada por una medida de la dispersión dentro de las clases

  - Por cada clase definimos la dispersión, un equivalente a la varianza, como:

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} \left(y - \tilde{\mu}_i\right)^2$$

  - donde la cantidad $\left(\tilde{s}_1^2 + \tilde{s}_2^2\right)$ es la dispersión intra clase de los ejemplos proyectados considerando igualdad en la probabilidad a priori de las clases ➔ $\tilde{S}_W$

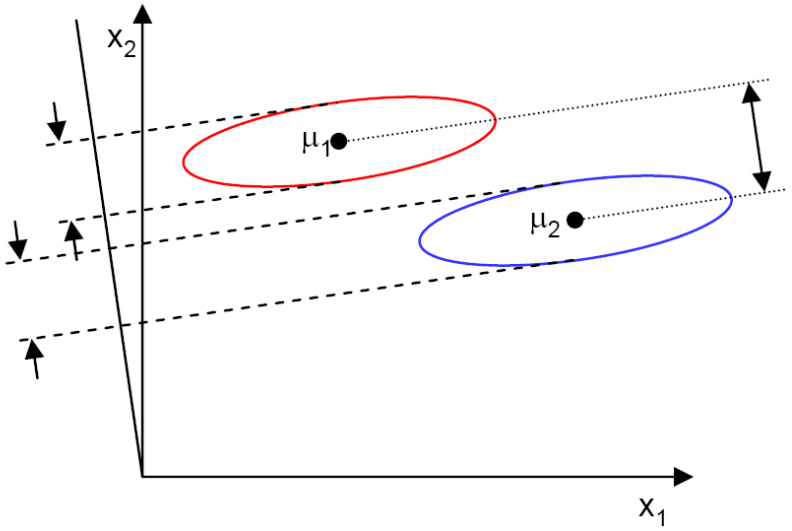
- El discriminante lineal de Fisher se define como la función lineal $w^T x$ que maximiza la función objetivo:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

1. Diferencia entre medias proyectadas aumenta
2. Dispersión intra-clases en la proyección disminuye

Análisis Discriminante Lineal, dos clases (6)

- De esta forma, estaremos buscando una proyección donde los ejemplos de la misma clase son proyectados muy cerca unos de otros (mínima dispersión), y al mismo tiempo, las medias proyectadas están lo más lejos posible.



Análisis Discriminante Lineal, dos clases (7)

- Para poder encontrar la proyección óptima **w***, necesitaremos expresar **J(w)** como una función explícita de **w**
- Primero definiremos las matrices de dispersión en el espacio original:

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_W = \pi_1 S_1 + \pi_2 S_2$$

- donde S_w es la llamada "matriz de dispersión intra clase"

- La dispersión de la proyección y se puede expresar en función de la matriz de dispersión en el espacio original x:

$$\tilde{S}_i = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^T = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)(w^T x - w^T \mu_i)^T = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w$$

$$\tilde{S}_w = \pi_1 \tilde{S}_1 + \pi_2 \tilde{S}_2 = \pi_1 w^T S_1 w + \pi_2 w^T S_2 w = w^T (\pi_1 S_1 + \pi_2 S_2) w = w^T S_w w$$

Análisis Discriminante Lineal, dos clases (8)

- De manera similar, podemos expresar la diferencia entre los promedios proyectados en función de las medias en el espacio original x:

$$\left(\tilde{\mu}_1 - \tilde{\mu}_2\right)^2 = \left(w^T\mu_1 - w^T\mu_2\right)^2 = w^T \underbrace{\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T}_{S_B} w = w^T S_B w$$

S_B es la “matriz de dispersión interclase”.

Como es el producto externo de un vector consigo mismo, tiene rango ≤ 1

Análisis Discriminante Lineal, dos clases (9)

- Con lo que hemos visto, podemos expresar **J(w)** como una función explícita de **w**:

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \longrightarrow J(w) = \frac{w^T S_B w}{w^T S_W w}$$

- Maximizar J(w) respecto a w tiene una solución analítica sencilla:

$$w^* = \underset{w}{\operatorname{argmax}} \left\{ \frac{w^T S_B w}{w^T S_W w} \right\} = S_W^{-1}(\mu_1 - \mu_2)$$

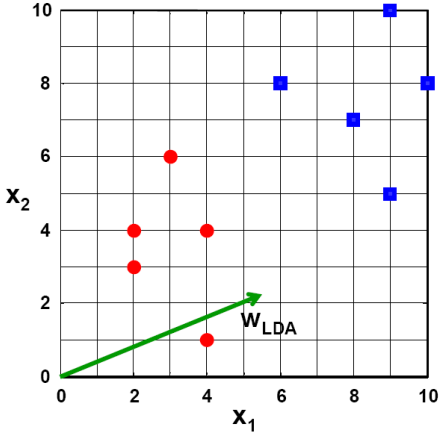
donde el módulo de w* es indiferente

Esta solución es el famoso Discriminante Lineal de Fisher (1936), aunque en realidad no es un discriminante sino la elección de una dirección específica para la proyección de los datos a una dimensión

Ejemplo de LDA

- Calcular la proyección LDA para el siguiente conjunto de datos en dos dimensiones:

$$X1=(x_1,x_2)=\{(4,1),(2,4),(2,3),(3,6),(4,4)\}$$
$$X2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$$



- Solución (a mano):
  - Las estadísticas de las clases son $\mu 1=\begin{pmatrix}3.0\\3.6\end{pmatrix}$, $\mu 2=\begin{pmatrix}8.4\\7.6\end{pmatrix}$

  - Las matrices de dispersión inter- e intra-clase son:

$$S_B=(\mu_1-\mu_2)(\mu_1-\mu_2)^T=\begin{pmatrix}29.16 & 21.60\\21.60 & 16.00\end{pmatrix}$$

$$S_i=\sum_{x\in\omega i}(x-\mu_i)(x-\mu_i)^T \quad S_1=\begin{pmatrix}4 & -2\\-2 & 13.2\end{pmatrix}, S_2=\begin{pmatrix}4 & -2\\-2 & 13.2\end{pmatrix} \quad S_W=\pi_1S_1+\pi_2S_2=\begin{pmatrix}2.67 & 1.33\\1.33 & 8\end{pmatrix}$$

  - La proyección óptima viene entonces dada por:

$$w^*=S_W^{-1}(\mu_1-\mu_2)=\begin{bmatrix}-0.91 & -0.39\end{bmatrix}^T$$

Análisis Discriminante Lineal, C clases (1)

- El Discriminante de Fisher se puede generalizar a problemas con **C** clases (arbitrario)

- En vez de buscar una proyección **y** (escalar), buscamos **(C-1)** proyecciones **[y_1, y_2, …, y_{C-1}]** por medio de **(C-1)** vectores de proyección $\mathbf{w}_i$.

- Definimos por conveniencia la matriz de proyección **W** con **(C-1)** columnas:

  $\mathbf{W} = [\, \mathbf{w}_1 \mid \mathbf{w}_2 \mid \ldots \mid \mathbf{w}_{C-1} \,]$

$$y_i = w_i^T x \Rightarrow y = W^T x$$

Análisis Discriminante Lineal, C clases (2)

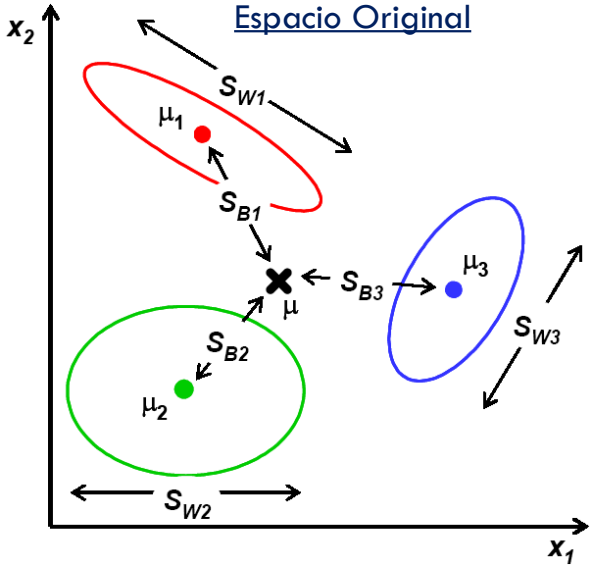
- Matriz de dispersión intra-clase

$$S_W = \sum_{i=1}^{C} \pi_i S_i \qquad S_i = \sum_{x \in \omega i} (x - \mu_i)(x - \mu_i)^T \qquad \mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$$

- Matriz de dispersión inter-clase

$$S_B = \sum_{i=1}^{C} \pi_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mu = \frac{1}{N} \sum_{\forall x} x = \frac{1}{N} \sum_{x \in \omega_i} N_i \mu_i$$



Análisis Discriminante Lineal, C clases (3)

- De manera similar, definimos el vector promedio y las matrices de dispersión de los ejemplos **proyectados** como:

$$\tilde{\mu}_i = E[y \mid y \in \omega_i] \qquad \tilde{S}_W = \sum_{i=1}^{C} \pi_i \tilde{S}_i$$

$$\tilde{\mu} = E[y] = \sum_{i=1}^{C} \pi_i \tilde{\mu}_i \qquad \tilde{S}_B = \sum_{i=1}^{C} \pi_i (\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^T$$

- De manera análoga a cuando teníamos 2 clases, podemos escribir:

$$\tilde{S}_W = W^T S_W W$$

$$\tilde{S}_B = W^T S_B W$$

Análisis Discriminante Lineal, C clases (4)

- Estamos buscando una proyección que maximice la dispersión inter-clase y minimice la dispersión intra-clase.
- Ya que ahora la proyección no es un escalar (tiene C-1 dimensiones), usamos el determinante de las matrices de dispersión para obtener escalares:

$$J(W) = \frac{\left|\tilde{S}_B\right|}{\left|\tilde{S}_W\right|} = \frac{\left|W^T S_B W\right|}{\left|W^T S_W W\right|}$$

Función criterio de Fisher es una función escalar que es grande cuando:

- S_B es grande
- S_W es pequeña

- De esta forma, buscamos la matriz de proyecciones W* que maximiza J (W).

$$\tilde{S}_W = W^T S_W W$$

$$\tilde{S}_B = W^T S_B W$$

Análisis Discriminante Lineal, C clases (5)

- Se puede demostrar analíticamente que la matriz óptima **W*** es la que en sus columnas contiene los **(C-1)** autovectores de la matriz $\mathbf{S_W^{-1}\, S_B}$ correspondientes a los **(C-1)** autovalores más grandes:

.

$$W^* = \left[w_1^* \mid w_2^* \mid \cdots \mid w_{C-1}^* \right] = \operatorname{argmax} \left\{ \frac{\left| W^T S_B W \right|}{\left| W^T S_W W \right|} \right\} \Rightarrow (S_B - \lambda_i S_W) w_i^* = 0$$

$$(S_W^{-1}\, S_B)\, W^* = \lambda\, W^*$$

Análisis Discriminante Lineal, C clases (5)

- Se puede demostrar analíticamente que la matriz óptima **W*** es la que en sus columnas contiene los **(C-1)** autovectores de la matriz $\mathbf{S_W^{-1}\, S_B}$ correspondientes a los **(C-1)** autovalores más grandes:

.

$$W^* = \left[w_1^* \mid w_2^* \mid \cdots \mid w_{C-1}^* \right] = \operatorname{argmax} \left\{ \frac{\left| W^T S_B W \right|}{\left| W^T S_W W \right|} \right\} \Rightarrow (S_B - \lambda_i S_W) w_i^* = 0$$

$$(S_W^{-1}\, S_B)\, W^* = \lambda\, W^*$$

Análisis Discriminante Lineal, C clases (6)

- ¿Por qué (C-1)?

  - S_B es la suma de C matrices de orden 1 o menos

$$S_B = \sum_{i=1}^{C} \pi_i (\mu_i - \mu)(\mu_i - \mu)^T$$

y los vectores media están restringidos por $\frac{1}{C}\sum_{i=1}^{C} \mu_i = \mu$

  - De esta forma, S_B es de rango menor o igual que (C-1)

  - Esto significa que hay como mucho (C-1) autovalores λ_i que no son cero

- LDA se puede también derivar del método de Máxima Verosimilitud para el caso en el que las densidades condicionadas a la clase son gaussianas con las mismas matrices de covarianza.

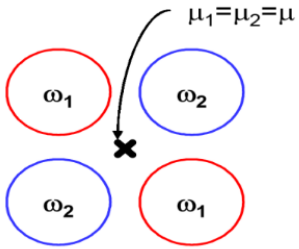
Análisis Discriminante Lineal, C clases (7)

- LDA produce como mucho **C - 1** características proyectadas

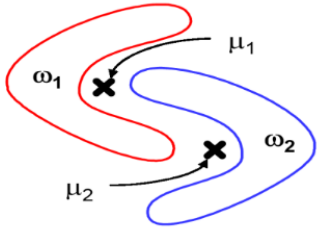
  - Si el error de clasificación estimado es demasiado alto, necesitaremos más características, con lo que deberemos utilizar otro método que proporcione esas características adicionales.

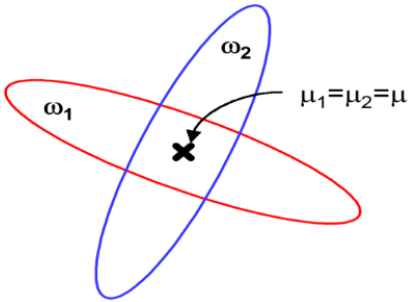
- LDA es un **método paramétrico** ya que asume implícitamente distribuciones **unimodales gaussianas**.

  - Si las distribuciones distan de ser gaussianas, las proyecciones LDA no serán capaces de preservar ninguna estructura compleja en los datos, lo que puede ser necesario para la clasificación.



$S_B = 0$

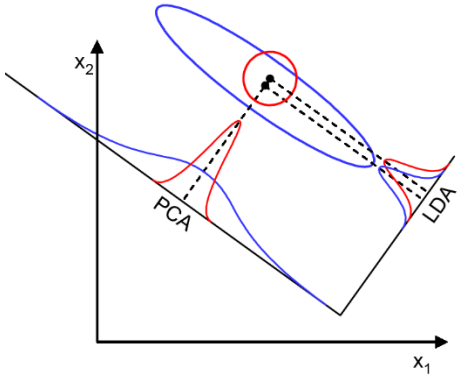




$S_B = 0$

Limitaciones de LDA

- LDA falla cuando la información discriminatoria no está en la media sino en la varianza de los datos:



- Precisa que S_W sea no singular ➔ $(S_W^{-1} S_B) W^* = \lambda W^*$
  - No es aplicable a datos altamente dimensionados donde el número de patrones es menor que el número de características.
- Como discriminante será lineal

Variantes de LDA (1)

- **LDA no paramétrico**, “NPLDA” (Fukunaga)

  - Este método no necesita la suposición de gaussianidad en las distribuciones. Para ello, calcula la matriz de dispersión inter-clase $\mathbf{S_B}$ usando información local y la regla de **K** vecinos más próximos.

  - Como resultado de esto:
    - La matriz $\mathbf{S_B}$ tiene orden máximo, permitiéndonos extraer más de **C-1** características.
    - Las proyecciones son capaces de preservar la estructura de los datos de una manera más precisa.

Variantes de LDA (2)

- **LDA** ortonormal (Okada y Tomita)
  - Se computan proyecciones que maximizan J(w) y a la vez son ortonormales entre sí.
  - Se combina lo obtenido con Fisher con el proceso de ortonormalización de Gram-Schmidt
  - Es capaz de encontrar más de C-1 características.

Variantes de LDA (3)

- **LDA** generalizado (Lowe)

  - Se generaliza lo desarrollado con Fisher incluyendo funciones de costo similares a las usadas al calcular el Riesgo de Bayes.

  - El efecto es una proyección LDA cuya estructura está sesgada por la función de coste.

  - Las clases con costos C_{ij} mayores se separarán más en el espacio de proyecciones.

Variantes de LDA (4)

- Perceptrones multicapa (Webb y Lowe)

  - Estos autores demostraron que las capas ocultas de perceptrones multi-capa (MLP) efectúan un análisis discriminante no lineal maximizando Tr [$S_B S_T^+$], donde las matrices de dispersión se miden a la salida de la última capa oculta. [Nota: $S_T = S_W + S_B$].