Tema 3: Aprendizaje Automático en Máquinas

Universidad Pontificia de Salamanca

Manuel Martín-Merino

Contenido

- Aprendizaje estadístico: Introducción
- Aprendizaje estadístico: Sesgo-Varianza
- Problemas dentro del aprendizaje automático
- Análisis discriminante: Modelos lineales
- Análisis discriminante probabilístico
- Redes Neuronales: Modelos no lineales
- Estimación de errores y comparación modelos
- Tópicos avanzados
- Resúmen: Discusión

Aprendizaje estadístico: Introducción (I)

- Sea $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ una muestra finita de objetos donde $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ es la representación vectorial del objeto i y y_i es la salida óptima.
- El problema de **aprendizaje automático** trata de encontrar una función $f : \mathcal{X} \to \mathbb{R}$ que permita *predecir la salida para nuevas observaciones* con error mínimo.
- $y_i \in \mathbb{R}$ en problemas de predicción mientras que $y_i \in \{\pm 1\}$ en problemas de clasificación.

Aprendizaje estadístico: Introducción (II)

Matemáticamente, dada una muestra de datos $(\boldsymbol{x}_i, y_i)$ independientes e idénticamente distribuidos según $p(\boldsymbol{x}, y)$ desconocida, tratamos de encontrar f.

Esto requiere:

- Proponer una familia de funciones aproximadoras suficientemente flexible $f(\boldsymbol{x}; w)$.
- Proponer una función de error $L(y, f(\boldsymbol{x}; \omega))$ que mida la calidad de la aproximación.
- Minimizar el error promedio $E(L(y, f(\boldsymbol{x}; \omega)))$.

Aprendizaje estadístico: Introducción (III)

Objetivo: Aprender la función óptima $f(\boldsymbol{x};\omega)$ que minimiza:

$$E(w) = \int L(y, f(\boldsymbol{x};\omega))p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy \tag{1}$$

como $p(\boldsymbol{x}, y)$ es desconocida, el riesgo funcional se aproxima por el empírico más un término proporcional a la complejidad del modelo:

$$E(\omega) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i;\omega)) + \lambda\Omega(\omega) \tag{2}$$

Aprendizaje estadístico: Introducción (IV)

Objetivo: Encontrar una función $f(\boldsymbol{x}_i; \omega)$ que obtenga un balance entre error de entrenamiento y complejidad del modelo.

- $f(\boldsymbol{x}_i; \omega)$ muy compleja implica error entrenamiento 0 pero $\Omega(\omega)$ grande.
- $f(\boldsymbol{x}_i; \omega)$ muy sencilla implica minimizar $\Omega(\omega)$ a costa de error de entrenamiento grande.

Aprendizaje estadístico: Sesgo-Varianza (I)

La complejidad del modelo debe buscar un balance entre sesgo-varianza.

Consideramos que existen varios conjuntos de entrenamiento $\mathcal{D}$ con n patrones y obtenidos de la misma distribución de probabilidad $p(\boldsymbol{x}, y)$.

Por ser la muestra finita $f_{\mathcal{D}}(\boldsymbol{x}; \omega)$ dependerá de la muestra particular elegida. Por ello el error se calcula como un promedio para todas las muestras de entrenamiento

$$\mathcal{E}_{\mathcal{D}}\{f_{\mathcal{D}}(\boldsymbol{x}; \omega) - g(\boldsymbol{x}))^2\} \tag{3}$$

Aprendizaje estadístico: Sesgo-Varianza (II)

Este error se puede descomponer como:

$$\mathcal{E}_{\mathcal{D}}[\{f_{\mathcal{D}}(\boldsymbol{x};\omega) - g(\boldsymbol{x})\}^2] =$$
$$\underbrace{\{\mathcal{E}_{\mathcal{D}}[f_{\mathcal{D}}(\boldsymbol{x};\omega)] - g(\boldsymbol{x})\}^2}_{\text{sesgo}^2} + \underbrace{\mathcal{E}_{\mathcal{D}}[\{f_{\mathcal{D}}(\boldsymbol{x};\omega) - \mathcal{E}_{\mathcal{D}}[f_{\mathcal{D}}(\boldsymbol{x};\omega)]\}^2]}_{\text{varianza}} \quad (4)$$

- *Sesgo*: Mide la desviación de la función sobre el valor óptimo del promedio de las estimaciones para diferentes conjuntos de entrenamiento.

- *Varianza*: Mide la sensibilidad del estimador de la muestra a la muestra particular elegida.

Aprendizaje estadístico: Sesgo-Varianza (III)

Ejemplo: *Consideramos el problema de regresión* $y = (\text{sen}\, 2\pi x)^2 + \epsilon$, *donde* ϵ *es una variable aleatoria de media* 0 *y* $\sigma_\epsilon = 0{,}2$. *El tamaño del conjunto de entrenamiento es* $n = 30$

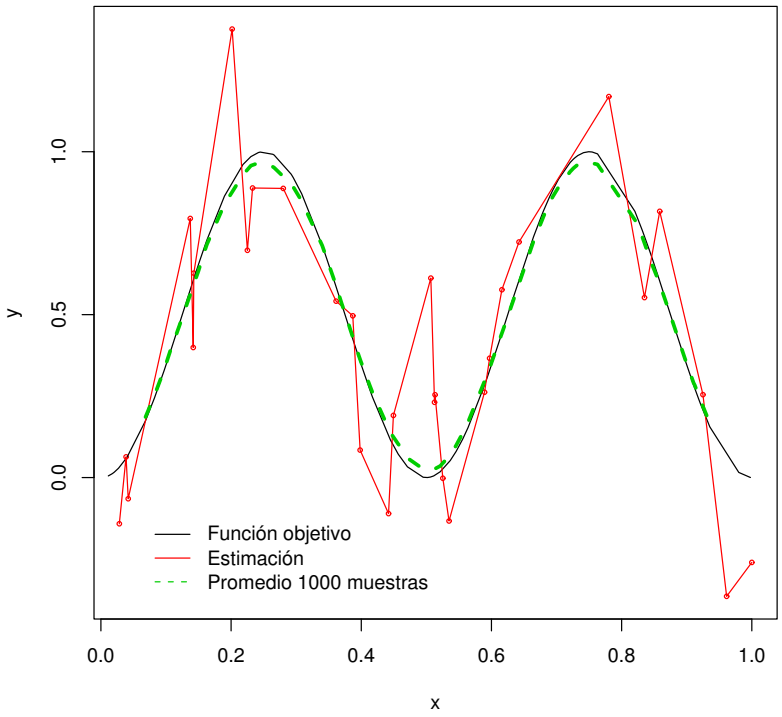


Fig. 1: Grados de libertad 30. Varianza elevada.

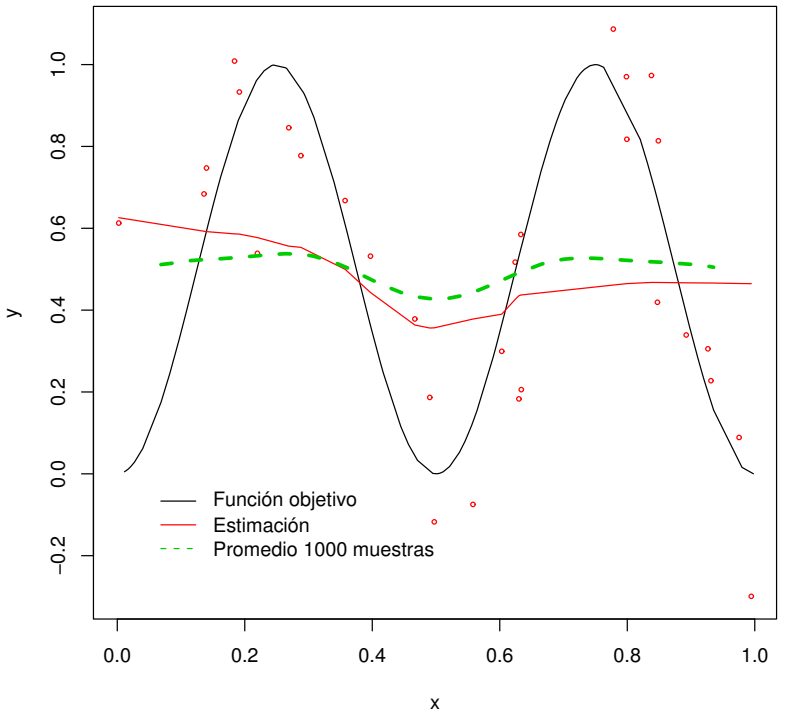
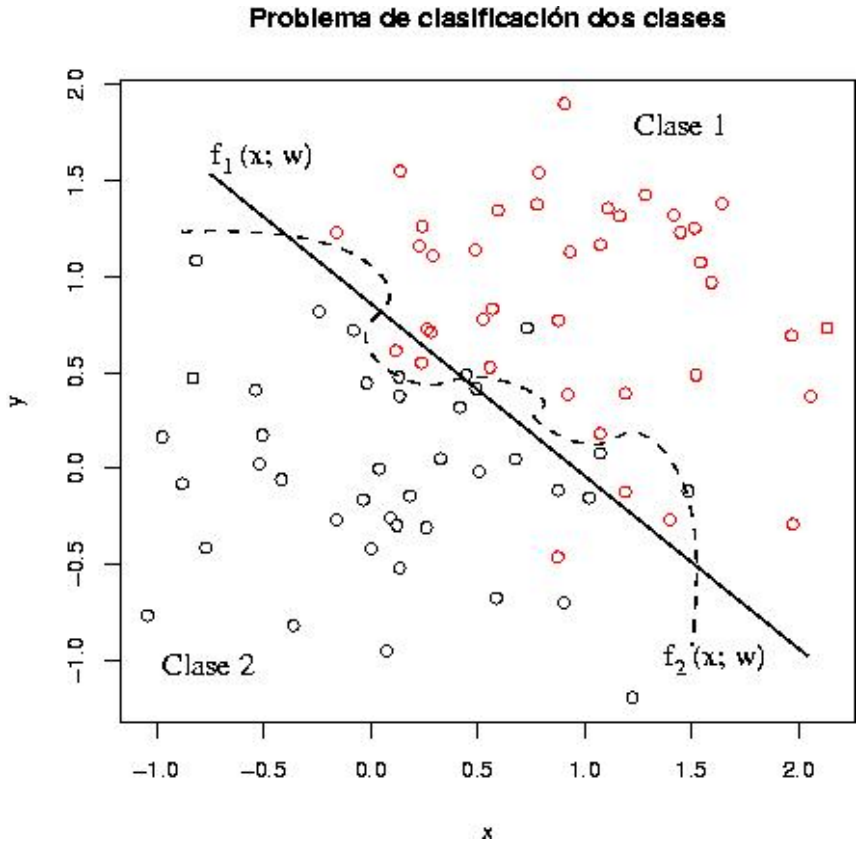


Fig. 2: Grados de libertad $2{,}39$. Sesgo elevado.

Problemas dentro del aprendizaje automático (I)

Clasificación



Ejemplo clasificación

- Proponer una familia de funciones (por ej. polinomios)
- Proponer una función de error (por ej. cuadrático)
- Estimar los parámetros: Optimizar

Clasificación

Objetivo: Estimar la $p(C_j|\boldsymbol{x})$ para nuevos patrones no utilizados en la fase de entrenamiento.

Enfoques:

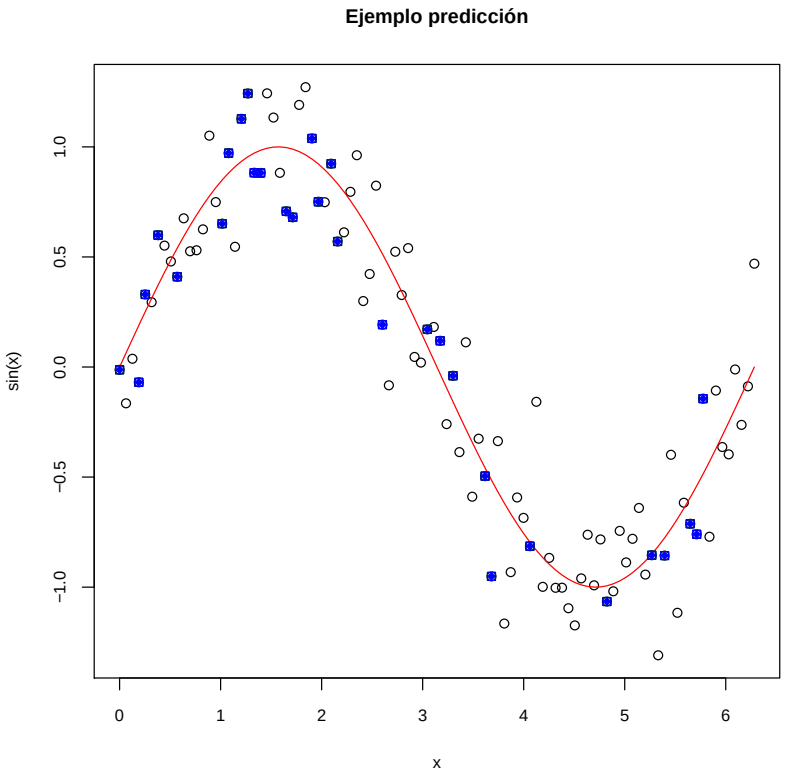
- Probabilístico. Estima $p(\boldsymbol{x}|C_j)$ y aplica la regla de decisión Bayesiana:

$$\boldsymbol{x} \in C_k \quad \text{si} \quad p(\boldsymbol{x}|C_k)p(C_k) > p(\boldsymbol{x}|C_j)p(C_j)\,\forall j \neq k \qquad (5)$$

- Redes Neuronales y regresión logística. Estima directamente $p(C_j|\boldsymbol{x})$
- Análisis discriminante. Estima directamente la ecuación de la frontera $f(\boldsymbol{x};\omega)$

Problemas dentro del aprendizaje automático (II)

Predicción: Identificación funciones

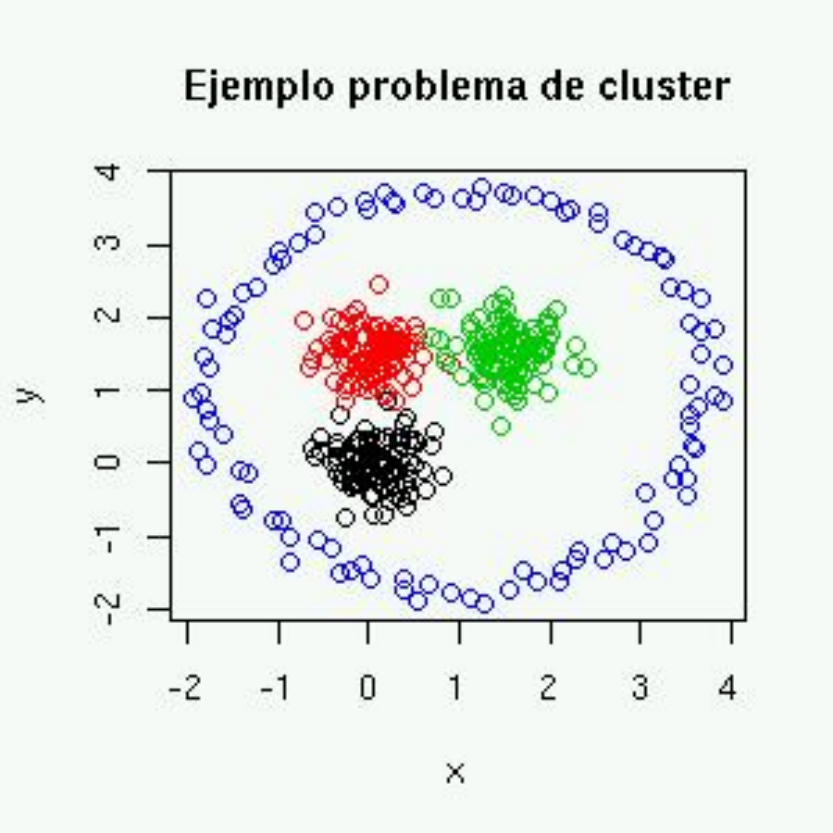


Ejemplo identificación funciones

- Proponer una familia de funciones (por ej. polinomios)
  - Modelos lineales: Regresión lineal, logística.
  - Modelos no lineales: Redes Neuronales.
- Proponer una función de error (por ej. cuadrático)
- Estimar los parámetros: Optimizar

Problemas dentro del aprendizaje automático (III)

Análisis de Cluster

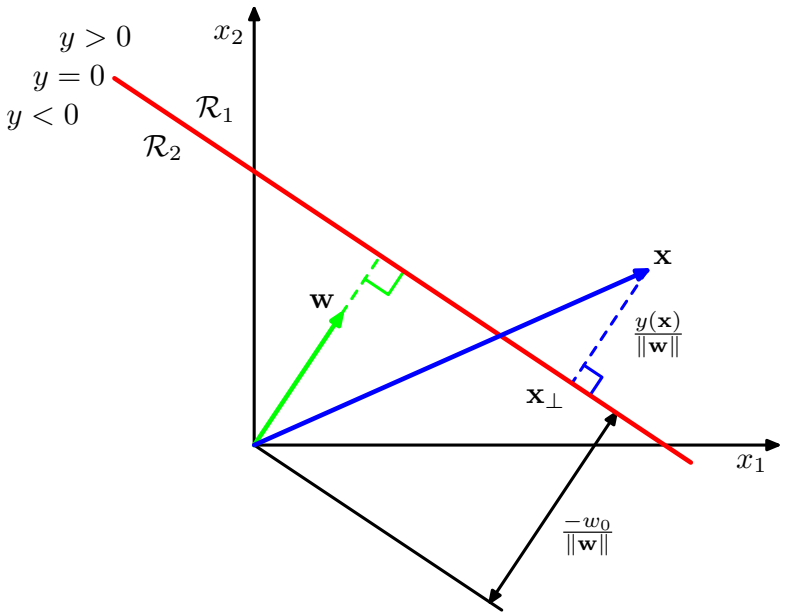


Ejemplo *cluster*

- Proponer un modelo
- Definir función de error que mida calidad cluster
- Estimar los parámetros: Optimizar

Análisis discriminante: Modelos lineales (I)

- Sea $\{\boldsymbol{x}_i, \boldsymbol{t}_i\}_{i=1}^{N}$ una muestra finita de patrones de entrenamiento y $\{C_j\}_{j=1}^{M}$ el conjunto de clases. $t_{ij} = 1$ si $\boldsymbol{x}_i$ pertenece a la clase C_j.

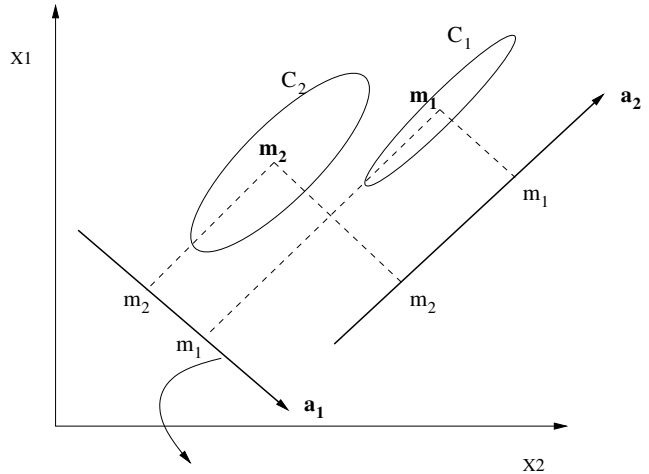


Hiperplano discriminante lineal

- La función discriminante lineal se escribe como: $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$
  $f(\boldsymbol{x}) \geq 0$ entonces $\boldsymbol{x} \in C_1$, en otro caso $\boldsymbol{x} \in C_2$.
- ¿ Función de error a optimizar ?

Análisis discriminante: Modelos lineales (II)

Enfoque Fisher: Proyecta los datos sobre un hiperplano lineal que maximiza la separabilidad entre clases.



Dirección generada por el discriminante de Fisher

Criterio de separabilidad (biclase):

- Maximiza separación entre centros de clases proyectadas $(\bar{m}_2 - \bar{m}_1)^2$.
- Minimiza la dispersión de cada clase en torno a su media $(s_1^2 + s_2^2)$ donde $s_k^2 = \sum_{i \in C_k} (\bar{x}_i - \bar{m}_k)^2$.

Criterio de Fisher

La función de error a optimizar se puede escribir como:

$$J(\boldsymbol{w}) = \frac{(\bar{m}_2 - \bar{m}_1)^2}{s_1^2 + s_2^2} = \tag{6}$$

$$= \frac{\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\boldsymbol{w}}{\sum_{x_i \in C_1} \boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^T\boldsymbol{w} + \sum_{x_i \in C_2} \boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^T\boldsymbol{w}} \tag{7}$$

Definimos las matrices de covarianza interclase e intraclase como:

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \tag{8}$$

$$S_w = \sum_{x_i \in C_1} \boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^T\boldsymbol{w} + \sum_{x_i \in C_2} \boldsymbol{w}^T(\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^T\boldsymbol{w} \tag{9}$$

Análisis discriminante: Modelos lineales (III)

Esto supone optimizar el siguiente criterio:

$$J_F = \frac{\mid \mathbf{w}^T S_b \mathbf{w} \mid}{\mid \mathbf{w}^T S_w \mathbf{w} \mid}, \tag{10}$$

Esto equivale a resolver el siguiente problema de autovalores y autovectores:

$$S_b \mathbf{w} = S_w \mathbf{w} \Lambda \tag{11}$$

que se puede resolver eficientemente mediante operaciones de álgebra lineal.

Proyectados los datos, se asignan a la clase de centroide más cercano.

Análisis discriminante: Modelos lineales (IV)

Propiedades discriminante Fisher:

- Es atractivo para trabajar en alta dimensión y eficiente computacionalmente.
- La frontera inducida es lineal y no está preparado para trabajar con clases multimodales.
- La dimensión máxima del espacio proyección es $M - 1$.
- Es bastante sensible a atípicos y no maximiza la capacidad de generalización.

Análisis discriminante: Modelos lineales (V)

Discriminante lineal puede **optimizar** también el **error cuadrático medio**:

$$E = \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0 - t_i)^2 \tag{12}$$

Definimos la variable respuesta:

$$t_i = \begin{cases} N/N_1 & \text{si} \quad \boldsymbol{x_i} \in C_1 \\ -N/N_2 & \text{si} \quad \boldsymbol{x_i} \in C_2 \end{cases} \tag{13}$$

Calculando las derivadas e igualando a 0 obtenemos:

$$\frac{\partial E}{\partial w_0} = \sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0 - t_i) = 0 \tag{14}$$

$$\frac{\partial E}{\partial \boldsymbol{w}} = \sum_{i=1}^{N}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0 - t_i)\boldsymbol{x}_i = 0 \tag{15}$$

(16)

Operando se obtiene:

$$(\boldsymbol{S}_w + \frac{N_1 N_2}{N}\boldsymbol{S}_B)\boldsymbol{w} = N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{17}$$

Teniendo en cuenta que $\boldsymbol{S}_B\boldsymbol{w}$ es paralelo a $\boldsymbol{m}_2 - \boldsymbol{m}_1$ el vector $\boldsymbol{w}$ de la función discriminante se puede expresar:

$$\boldsymbol{w} \propto \boldsymbol{S}_w^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{18}$$

que es equivalente a la expresión obtenida maximizando la separabilidad.