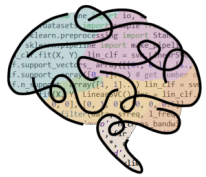
"Machine Learning estadístico para interfaces Cerebro-Computadora"

MODULO III - parte I

Dra. Victoria Peterson

vpeterson@santafe-conicet.gov.ar

Nov 2023



NiCALab

1. Vocabulario

2. Generalidades

3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación

4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

1. Vocabulario
2. Generalidades
3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación
4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

Motivación

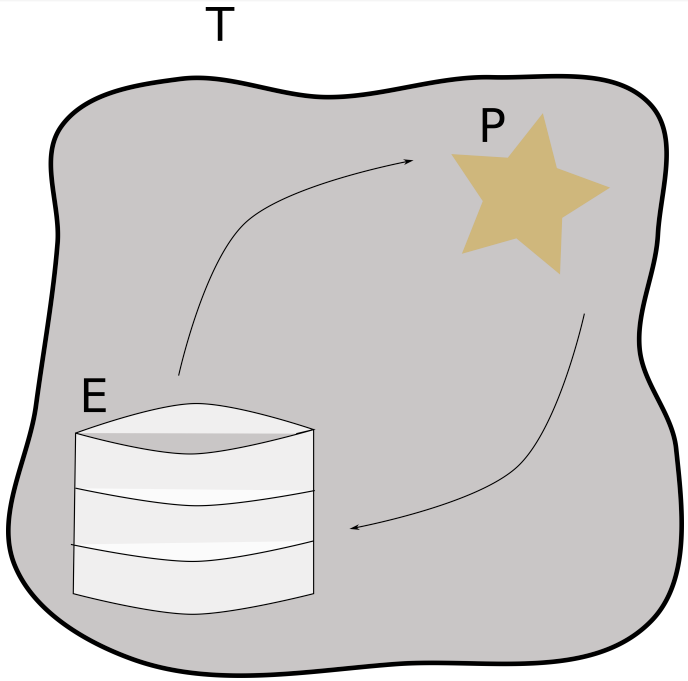
Experiencia con algo nuevo

¿Cómo elegir una papaya rica?

Algoritmos aprendedores

Un programa computacional se dice que aprende de la experiencia ***E*** según sea la tarea ***T*** y la medida de performance ***P***, si su performance para hacer la tarea T, medido mediante P, mejora con la experiencia E" - Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, New York.-

"Aprender es el medio para lograr realizar una tarea"



$$\min \mathscr{L}\{\cdot\}$$

Definiciones

- **Conjunto del dominio**: conjunto arbitrario de datos que definen mi problema a resolver. Se denota con $\mathscr{X}$. Ejemplo papaya: color de la fruta, rigidez.
- **Conjunto variable target**: no siempre disponible. Se denota con $\mathscr{Y}$. Ejemplo papaya: $\mathscr{Y} = \{0, 1\}$, siendo 0 papayas feas y 1 papayas ricas.
- **Datos**: conjunto finito de pares de secuencia en $\mathscr{X} \times \mathscr{Y}$. Se denota $S = \big((\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_N, y_N)\big)$. Ejemplo papaya: conjunto de papayas que han sido saboreadas y se tiene su color, rigidez y gusto.
- **Algoritmo de machine learning**: implementación computacional de un método de ML.
- **Entrenamiento**: proceso de ajuste al cual se somete un algoritmo de machine learning.
- **Modelo**: algoritmo de ML ya entrenado, que es capaz de otorgar una regla de predicción. Se denota $f : \mathscr{X} \to \mathscr{Y}$.
- **Salida**: predicción otorgada por el modelo de ML. Se denota como $\hat{\mathbf{y}}$.

1. Vocabulario

2. Generalidades

3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación

4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

1 Vocabulario

2 Generalidades

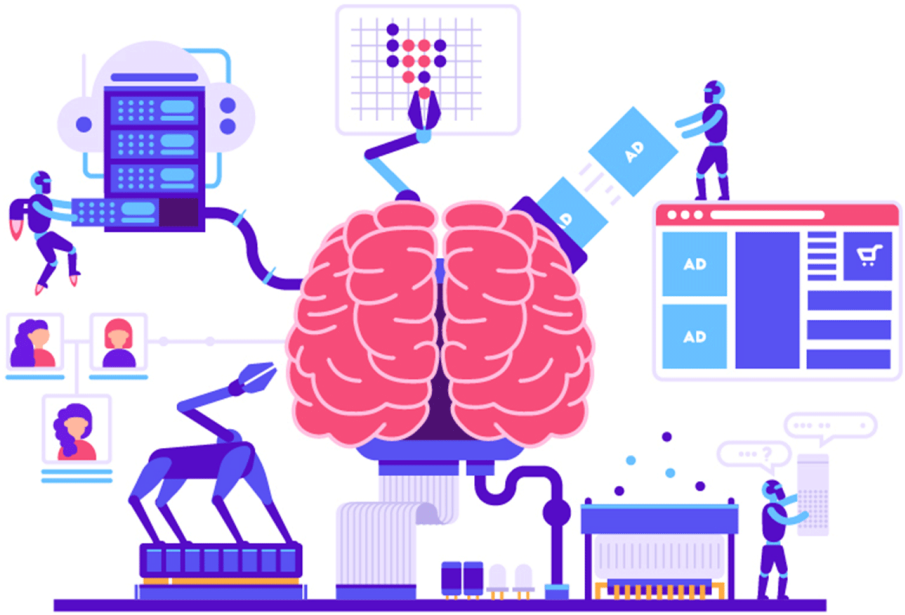
3 Análisis Discriminante Lineal
- Método de transformación lineal
- Hiperplano de separación

4 Evaluación y performance
- Médidas de evaluación
- Estimación del desempeño

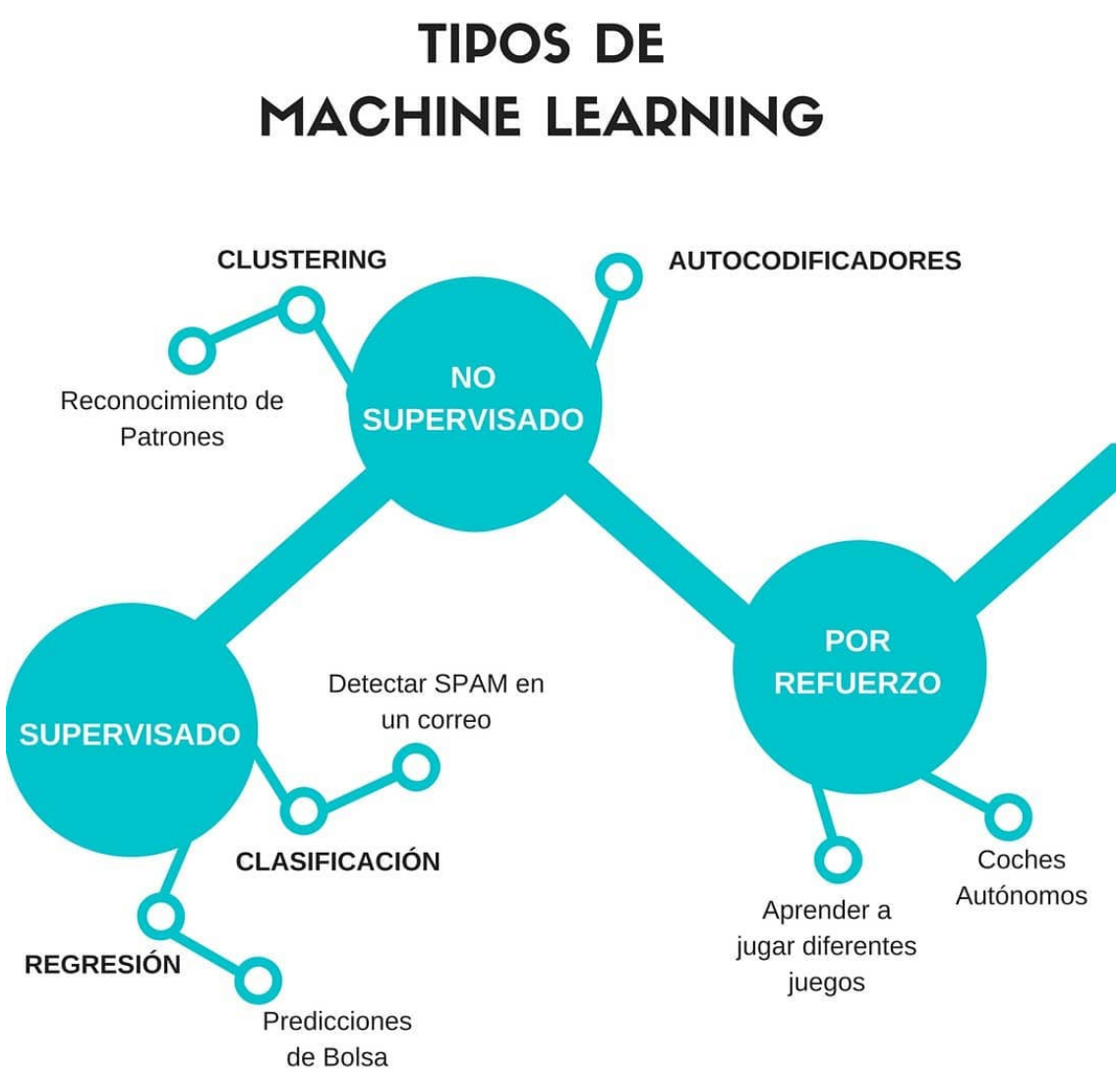
Qué es el aprendizaje maquinal?

Definición-Dr. Bengio

"*Machine learning research is part of research on artificial intelligence, seeking to provide knowledge to computers through data, observations and interacting with the world. That acquired knowledge allows computers to correctly generalize to new settings.*"



Tipos de aprendizaje



Notación

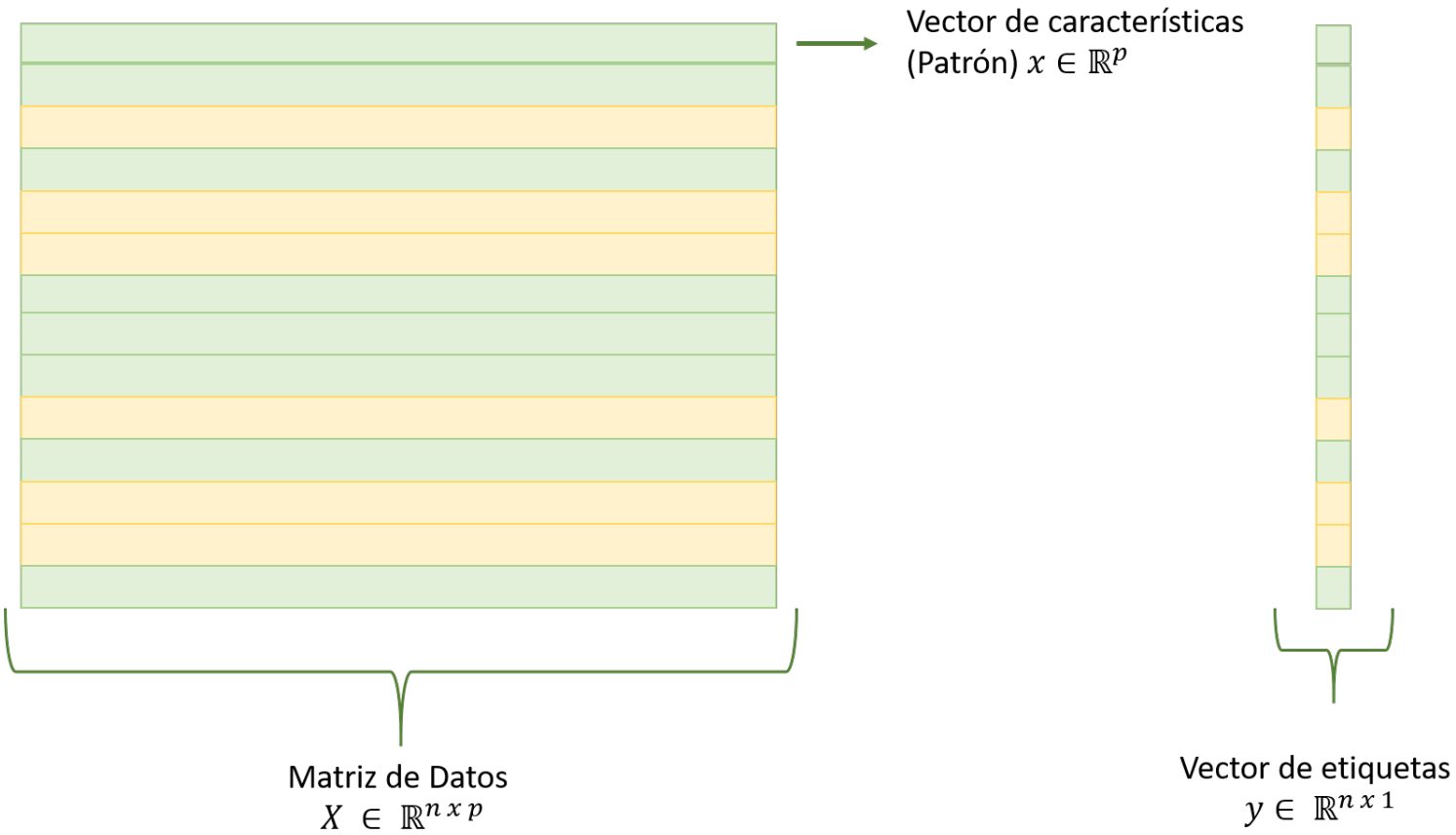
$\mathscr{X}_1, \mathscr{X}_2, \ldots \mathscr{X}_C,$	vectores aleatorios cuyas distribuciones cuyas distribuciones caracterizan a cada una de las clases C de un problema de clasificación
$\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)^T,$	vector de medición, patrón u observación.
$X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T,$	matriz de datos de n patrones de dimensión p.
$\mathbf{y} \in \{1, 2, \ldots, C\}^n$	vector de etiqueta categóricas
$S = \left\{(\mathbf{x}_i, y_i\right\}_{i=0}^n \sim \mathscr{D}^n$	conjunto de datos

si $\mathbf{x}_j$ es una realización de una y sólo una $\mathscr{X}_i \Rightarrow y_j = i,$

Notación

$\mathscr{X}_1, \mathscr{X}_2, \ldots \mathscr{X}_C,$	vectores aleatorios cuyas distribuciones cuyas distribuciones caracterizan a cada una de las clases C de un problema de clasificación
$\mathbf{x} = \left(x_1, x_2, \ldots, x_p\right)^T,$	vector de medición, patrón u observación.
$X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T,$	matriz de datos de n patrones de dimensión p.
$\mathbf{y} \in \{1, 2, \ldots, C\}^n$	vector de etiqueta categóricas
$S = \left\{(\mathbf{x}_i, y_i\right\}_{i=0}^n \sim \mathscr{D}^n$	conjunto de datos
si $\mathbf{x}_j$ es una realización de una y sólo una $\mathscr{X}_i \Rightarrow y_j = i,$	
$p_{\mathbf{x}}(\mathbf{x}),$	densidad de probabilidad (pdf) de $\mathbf{x}$
$\mathbb{E}[\cdot],$	esperanza
$\mu = \mathbb{E}[\mathbf{x}],$	media de un v.a.
$R_{\mathbf{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^T],$	matriz de correlación
$\Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T].$	matriz de covarianza

Aprendizaje supervisado



Aprendizaje supervisado

Objetivo

Estimar la función de predicción $f(x): \mathbb{R}^p \rightarrow \mathcal{Y}$, tal que, dada una observación $\mathbf{x} \in \mathbb{R}^p$, *prediga* la clase $y \in \mathcal{Y}$.

Entrenamiento

En la práctica, la función $f(\cdot)$ debe ser aprendida de un conjunto de *entrenamiento* $S_{ent} = \{\mathbf{x}_j, y_j\}_{j=1}^{n}$ y predecir correctamente futuras observaciones $\{\mathbf{x}_m, y_m\}$

Aprendizaje supervisado

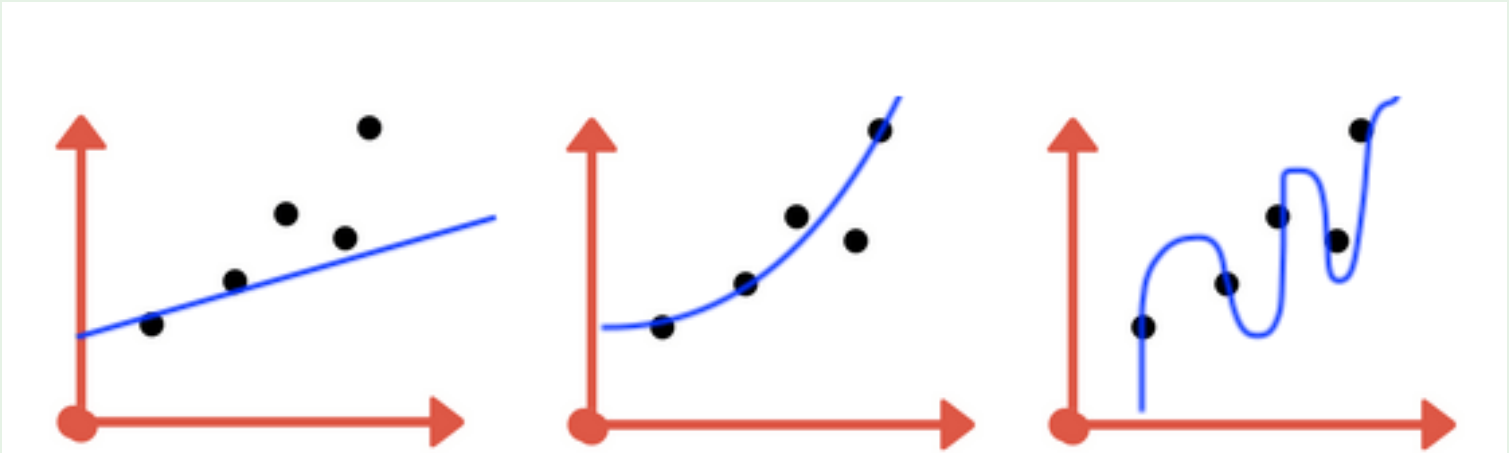
Objetivo

Estimar la función de predicción $f(x): \mathbb{R}^p \rightarrow \mathcal{Y}$, tal que, dada una observación $\mathbf{x} \in \mathbb{R}^p$, *prediga* la clase $y \in \mathcal{Y}$.

Entrenamiento

En la práctica, la función $f(\cdot)$ debe ser aprendida de un conjunto de *entrenamiento* $S_{ent} = \{\mathbf{x}_j, y_j\}_{j=1}^{n}$ y predecir correctamente futuras observaciones $\{\mathbf{x}_m, y_m\}$

Capacidad de generalización



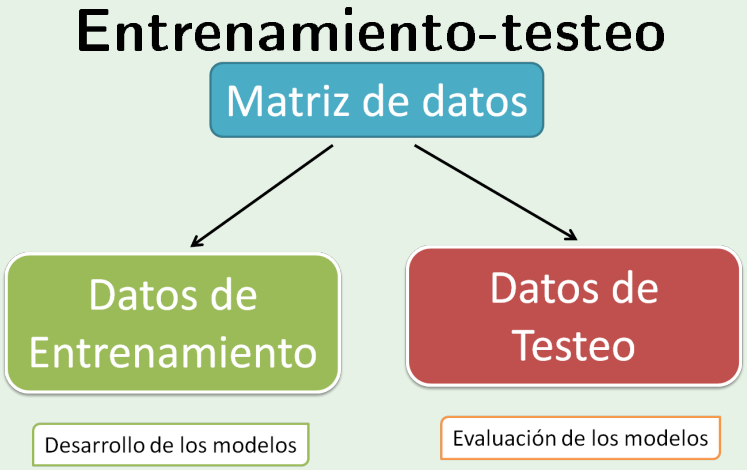
Aprendizaje supervisado

Objetivo

Estimar la función de predicción $f(x) : \mathbb{R}^p \rightarrow \mathcal{Y}$, tal que, dada una observación $\mathbf{x} \in \mathbb{R}^p$, *prediga* la clase $y \in \mathcal{Y}$.

Entrenamiento

En la práctica, la función $f(\cdot)$ debe ser aprendida de un conjunto de *entrenamiento* $S_{ent} = \{\mathbf{x}_j, y_j\}_{j=1}^{n}$ y predecir correctamente futuras observaciones $\{\mathbf{x}_m, y_m\}$



Aprendizaje supervisado

Objetivo

Estimar la función de predicción $f(x): \mathbb{R}^p \rightarrow \mathcal{Y}$, tal que, dada una observación $\mathbf{x} \in \mathbb{R}^p$, *prediga* la clase $y \in \mathcal{Y}$.

Entrenamiento

En la práctica, la función $f(\cdot)$ debe ser aprendida de un conjunto de *entrenamiento* $S_{ent} = \{\mathbf{x}_j, y_j\}_{j=1}^{n}$ y predecir correctamente futuras observaciones $\{\mathbf{x}_m, y_m\}$

Clasificadores lineales

$$f(\mathbf{x}) = \sum_{i=1}^{p} w_i x_i + b = \mathbf{w}^T \mathbf{x} + b,$$

tal que si $sgn(\mathbf{w}^T \mathbf{x} + b) > 0 \Rightarrow \mathbf{x} \in \mathcal{W}_1$

1. Vocabulario

2. Generalidades

3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación

4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

1 Vocabulario

2 Generalidades

3 Análisis Discriminante Lineal
- Método de transformación lineal
- Hiperplano de separación

4 Evaluación y performance
- Médidas de evaluación
- Estimación del desempeño

1 Vocabulario

2 Generalidades

3 Análisis Discriminante Lineal
- Método de transformación lineal
- Hiperplano de separación

4 Evaluación y performance
- Médidas de evaluación
- Estimación del desempeño

Criterio de Fisher (FDA)

Formulación máxima/mínima varianza

Objetivo

Maximizar la separabilidad entre clases y minimizar la disperción intra-clases

Supuestos

- Las características de cada clase tiene distribución normal.
- Todas las clases tienen una matriz de covarianza común $\Sigma_c = \Sigma_t \; \forall c$.
- La verdadera distribución de clases es conocida.

Definimos

Matriz de covarianza intra-clases: $\hat{\mathbf{\Sigma}}_w = \frac{1}{n} \sum_{c=1}^{C} \sum_{i \in I_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T,$

Matriz de covarianza entre-clases: $\hat{\mathbf{\Sigma}}_b = \frac{1}{n} \sum_{c=1}^{C} n_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T,$

Criterio de Fisher (FDA)

Formulación máxima/mínima varianza

Cap. 4**[Webb, 2003]**

Problema

Hallar $\mathbf{w}$ tal que $J_{\mathbf{F}} = \dfrac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \hat{\boldsymbol{\Sigma}}_w \mathbf{w}}$ se maximice.

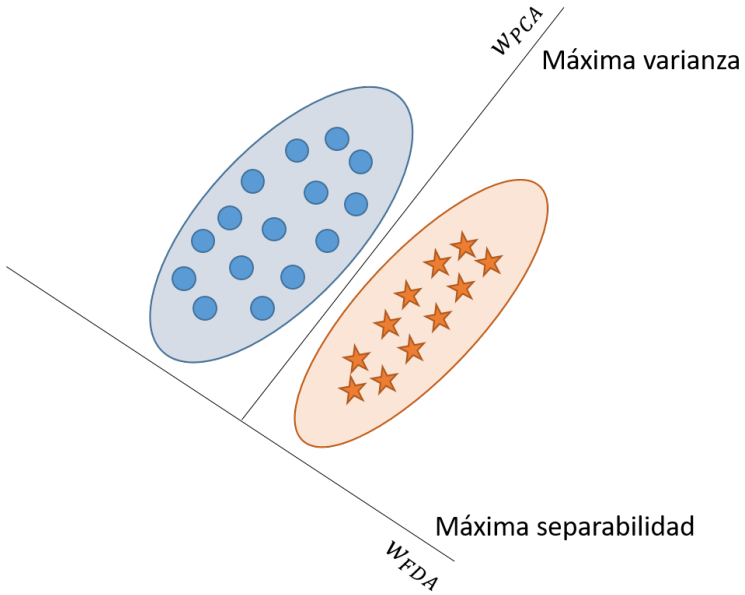
Criterio de Fisher (FDA)

Formulación máxima/mínima varianza

Cap. 4[**Webb, 2003**]

Problema

Hallar $\mathbf{w}$ tal que $J_{\mathbf{F}} = \dfrac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \hat{\boldsymbol{\Sigma}}_w \mathbf{w}}$ se maximice.



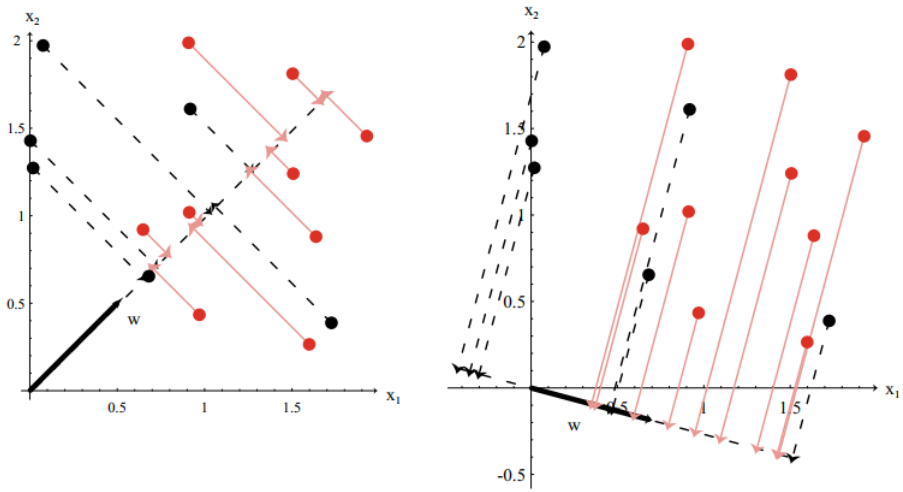
Criterio de Fisher (FDA)

Formulación máxima/mínima varianza

Cap. 4**[Webb, 2003]**

Problema

Hallar $\mathbf{w}$ tal que $J_{\mathbf{F}} = \dfrac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \hat{\boldsymbol{\Sigma}}_w \mathbf{w}}$ se maximice.



Fuente: [Duda et al., 2012]

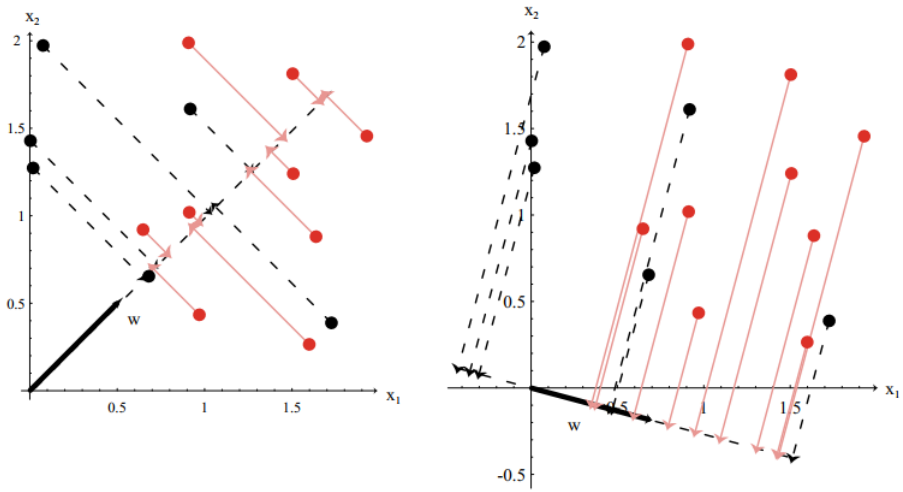
Criterio de Fisher (FDA)

Formulación máxima/mínima varianza

Cap. 4**[Webb, 2003]**

Problema

Hallar $\mathbf{w}$ tal que $J_{\mathbf{F}} = \dfrac{\mathbf{w}^T \boldsymbol{\Sigma}_b \mathbf{w}}{\mathbf{w}^T \hat{\boldsymbol{\Sigma}}_w \mathbf{w}}$ se maximice.



Fuente: [Duda et al., 2012]

COLAB TIME

1 Vocabulario

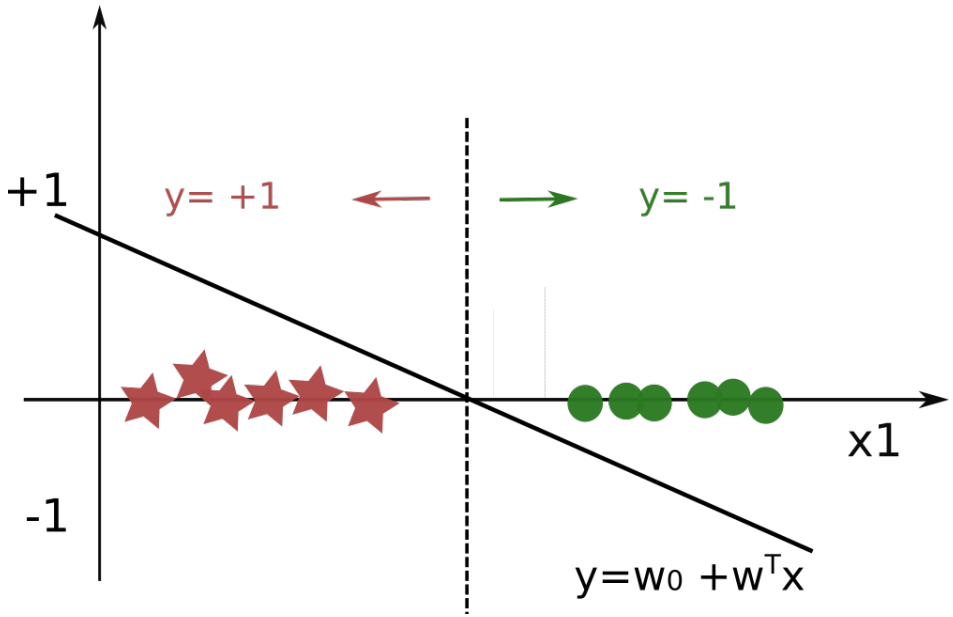
2 Generalidades

3 Análisis Discriminante Lineal
- Método de transformación lineal
- Hiperplano de separación

4 Evaluación y performance
- Médidas de evaluación
- Estimación del desempeño

LDA y regla de decisión

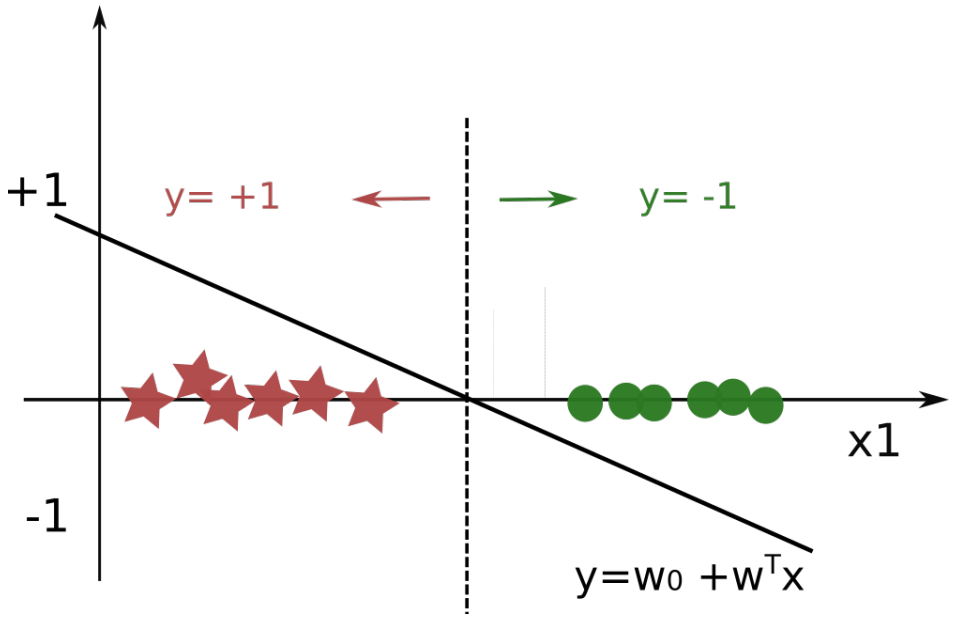
Regla de decisión



$$\hat{y} = \begin{cases} +1, & \text{if } w_0 + \mathbf{w}^T\mathbf{x} \geq 0 \\ -1, & \text{if o.c} \end{cases}$$

LDA y regla de decisión

Regla de decisión

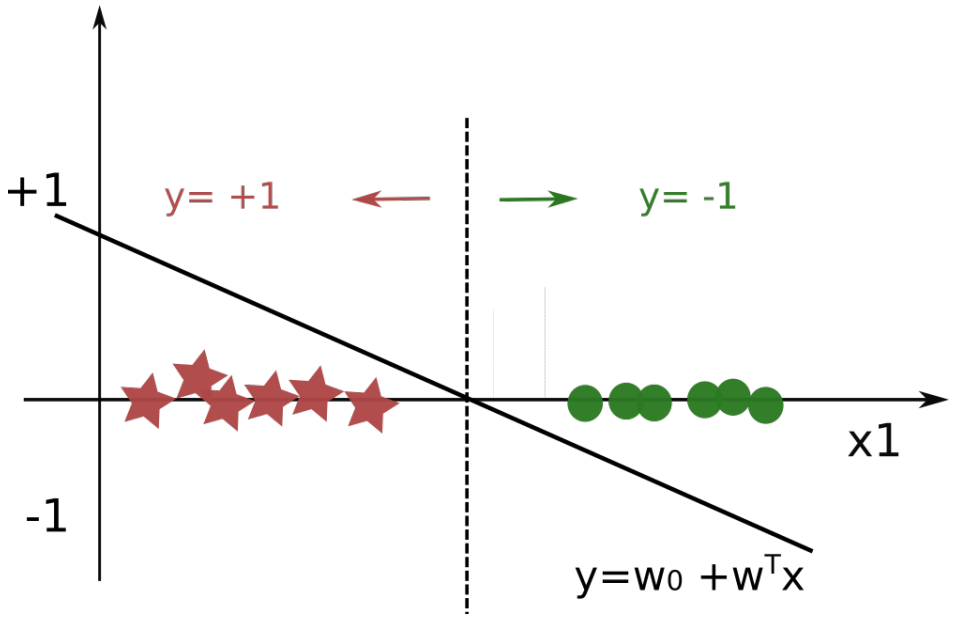


$$\hat{y} = \begin{cases} +1, & \text{if } w_0 + \mathbf{w}^T\mathbf{x} \geq 0 \\ -1, & \text{if o.c} \end{cases}$$

$$\hat{y}(\mathbf{x}) = \text{sign}(w_0 + \mathbf{w}^T\mathbf{x})$$

LDA y regla de decisión

Regla de decisión



$$\hat{y} = \begin{cases} +1, & \text{if } w_0 + \mathbf{w}^T\mathbf{x} \geq 0 \\ -1, & \text{if o.c} \end{cases}$$

$$\hat{y}(\mathbf{x}) = \text{sign}(w_0 + \mathbf{w}^T\mathbf{x})$$

Región de decisión dada por un hiperplano

- en 1D es simplemente un valor de umbral
- en 2D es una recta
- en 3D es una plano

1. Vocabulario

2. Generalidades

3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación

4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

1. Vocabulario

2. Generalidades

3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación

4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

1. Vocabulario
2. Generalidades
3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación
4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

Salidas y probabilidades

Formalización

Sea $\mathscr{D}$ una distribución de probabilidad sobre $\mathscr{X} \times \mathscr{Y}$, donde como antes $\mathscr{X}$ es el conjunto de dominio y $\mathscr{Y}$ el conjunto de etiquetas, podemos pensar dicha distribución como compuesta por dos partes:

- Distribución marginal $\mathscr{D}_x$ de los datos
- Distribución condicional sobre las etiquetas para cada dato del dominio $\mathscr{D}\left((\mathbf{x}, y)|\mathbf{x}\right)$

Salidas y probabilidades

Formalización

Sea $\mathscr{D}$ una distribución de probabilidad sobre $\mathscr{X} \times \mathscr{Y}$, donde como antes $\mathscr{X}$ es el conjunto de dominio y $\mathscr{Y}$ el conjunto de etiquetas, podemos pensar dicha distribución como compuesta por dos partes:

- Distribución marginal $\mathscr{D}_x$ de los datos
- Distribución condicional sobre las etiquetas para cada dato del dominio $\mathscr{D}\left((\mathbf{x}, y)|\mathbf{x}\right)$

$\Rightarrow \quad \hat{y} = \mathbb{P}\left[Y = y | X = \mathbf{x}\right] = f(\mathbf{x})$

Casos de aciertos y errores

Problema binario

Sea $\mathbf{y} \in \{0,1\}$ al vector de etiquetas verdadero de un conjunto de datos en un problema de clasificación (papaya rica $y=1$, papaya fea $y=0$), y sea $\hat{\mathbf{y}}$ el vector de etiquetas predicho por un modelo de clasificación.
Existen cuatro posibles casos:

- **Verdadero Positivo (TP)**: $y_n = 1$ y $\hat{y}_n = 1$
- **Verdadero Negativo (TN)**: $y_n = 0$ y $\hat{y}_n = 0$
- **Falso Positivo (FP)**: $y_n = 0$ y $\hat{y}_n = 1$
- **Falso Negativo (FN)**: $y_n = 1$ y $\hat{y}_n = 0$

Casos de aciertos y errores

Problema binario

Sea $\mathbf{y} \in \{0,1\}$ al vector de etiquetas verdadero de un conjunto de datos en un problema de clasificación (papaya rica $y=1$, papaya fea $y=0$), y sea $\hat{\mathbf{y}}$ el vector de etiquetas predicho por un modelo de clasificación.

Existen cuatro posibles casos:

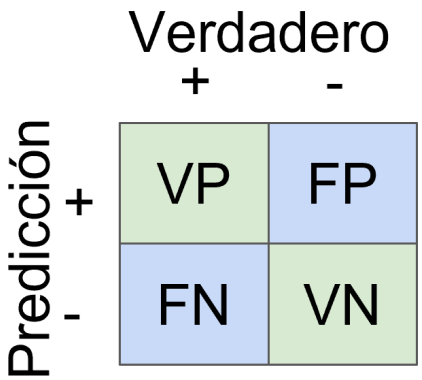
- **Verdadero Positivo (TP)**: $y_n = 1$ y $\hat{y}_n = 1$
- **Verdadero Negativo (TN)**: $y_n = 0$ y $\hat{y}_n = 0$
- **Falso Positivo (FP)**: $y_n = 0$ y $\hat{y}_n = 1$
- **Falso Negativo (FN)**: $y_n = 1$ y $\hat{y}_n = 0$

Matriz de confusión

Predicción \ Verdadero	+	-
+	VP	FP
-	FN	VN

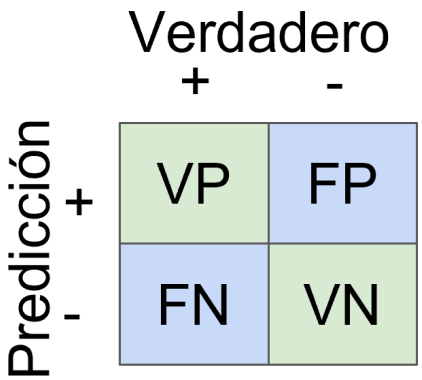
Basadas en la etiqueta

Problema binario



Basadas en la etiqueta

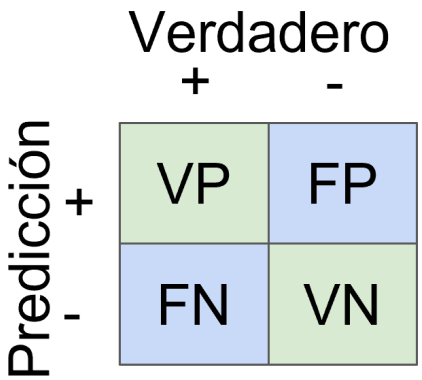
Problema binario



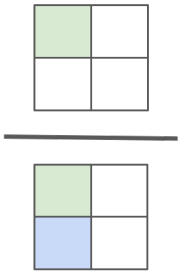
Medidas de evaluación

Basadas en la etiqueta

Problema binario

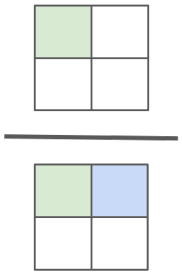


Sensibilidad



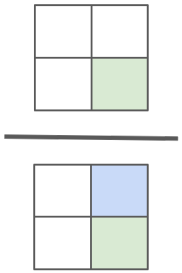
Qué cantidad de los casos verdaderos realmente identifiqué como verdaderos?

Precisión



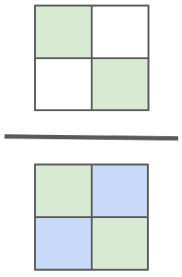
Qué proporción de identificaciones verdaderas eran realmente verdaderas?

Especificidad



Qué cantidad de casos negativos identifiqué?

Exactitud

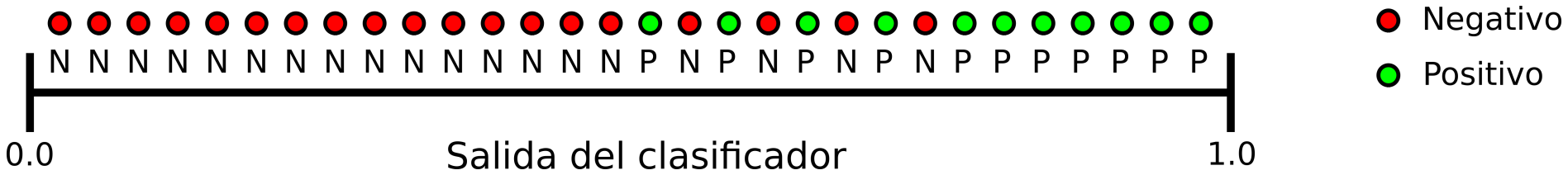


Qué cantidad de casos identifiqué correctamente?

¿Qué pasa si las clases están desbalanceadas?

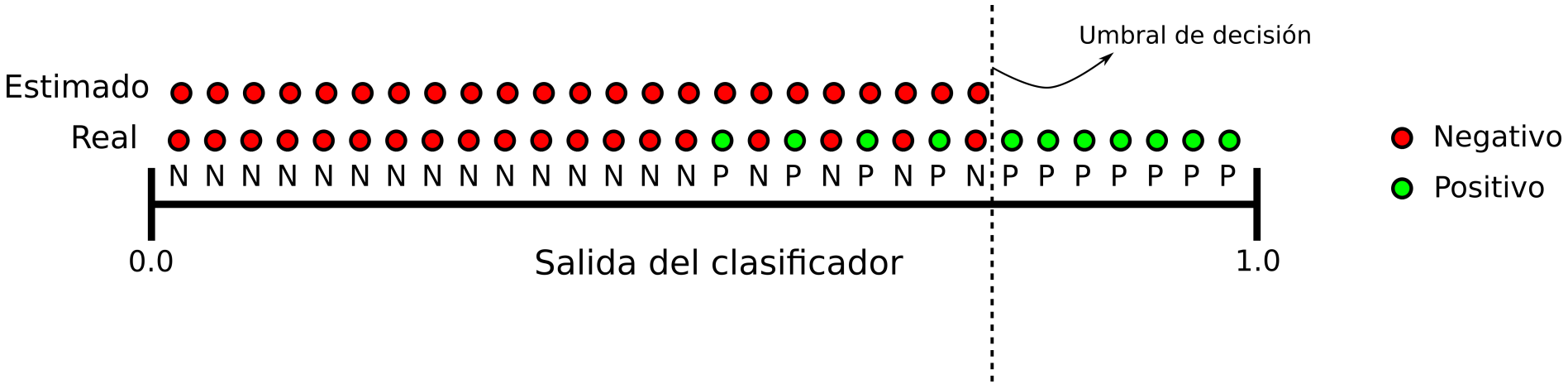
Medidas de evaluación

Basadas en las probabilidades



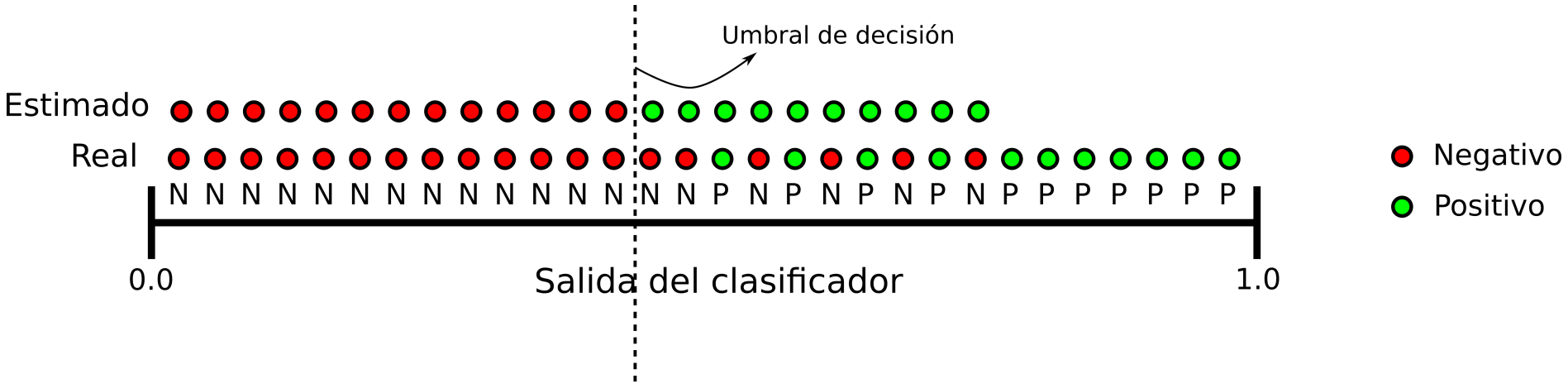
Medidas de evaluación

Basadas en las probabilidades



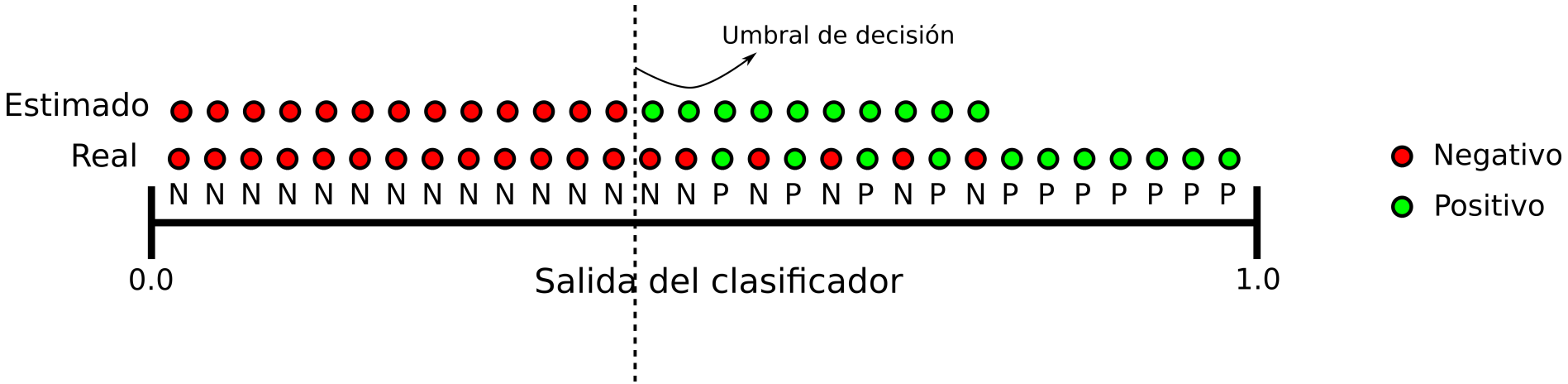
Medidas de evaluación

Basadas en las probabilidades



Medidas de evaluación

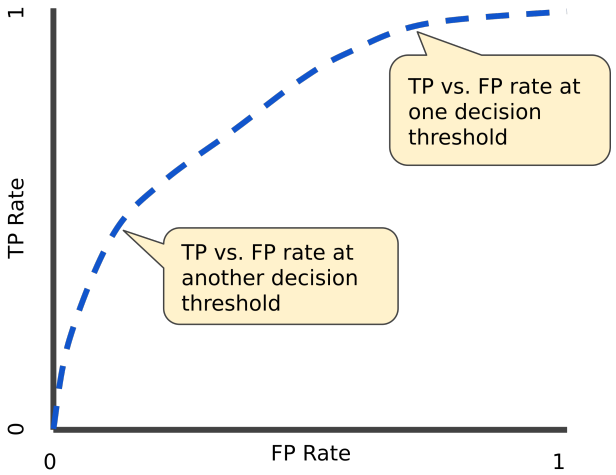
Basadas en las probabilidades



¿Cómo evaluar el desempeño para todo umbral de decisión?

Basadas en las probabilidades

La curva ROC (Receiver operating characteristic curve)

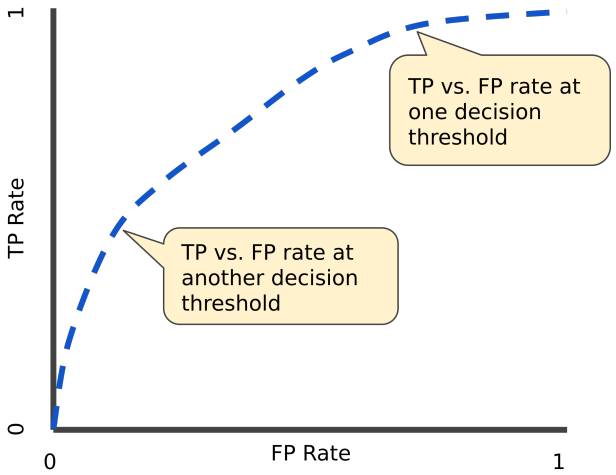


$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}:$$

Basadas en las probabilidades

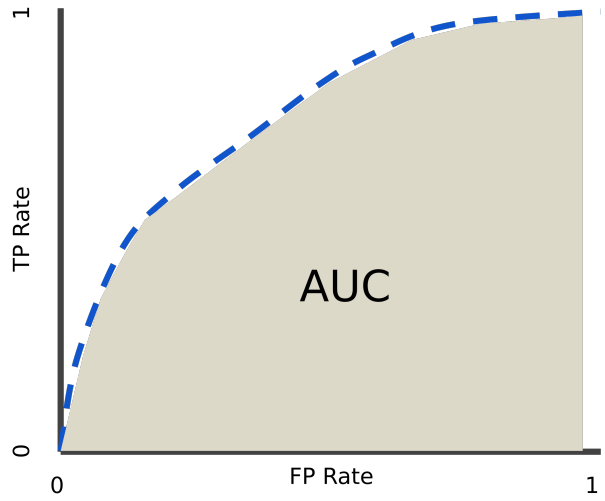
La curva ROC (Receiver operating characteristic curve)



$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}:$$

El área bajo curva ROC (AUC)

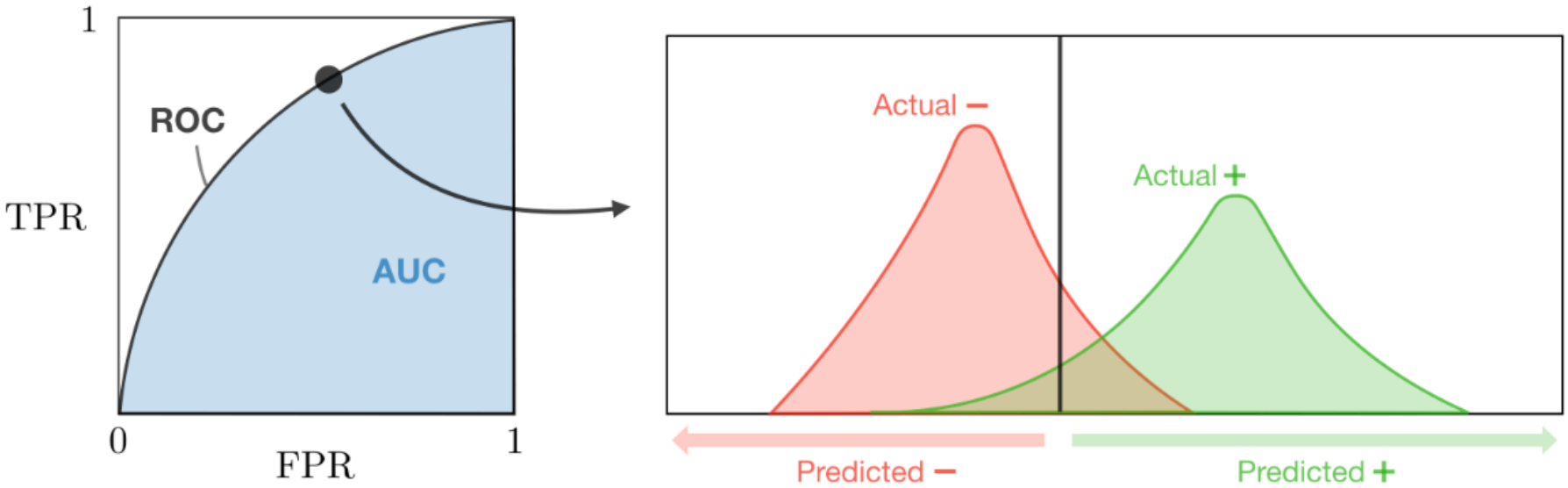


Medidas de evaluación

Área bajo las curvas ROC

[Fawcett, 2006]

ROC-AUC

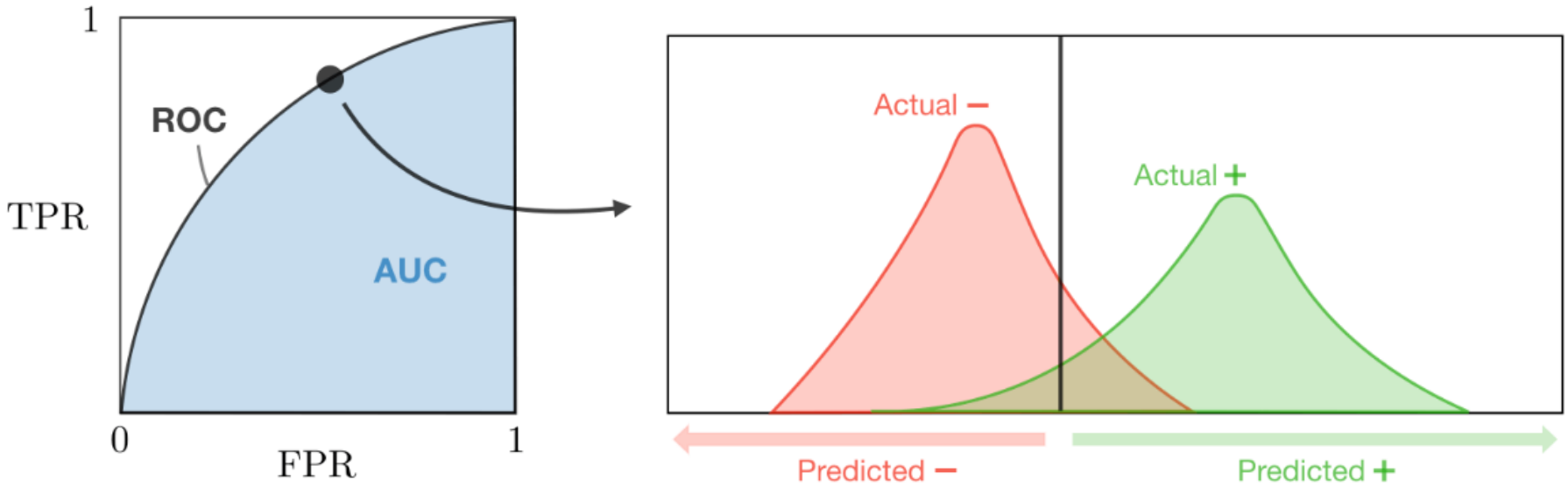


Medidas de evaluación

Área bajo las curvas ROC

[Fawcett, 2006]

ROC-AUC

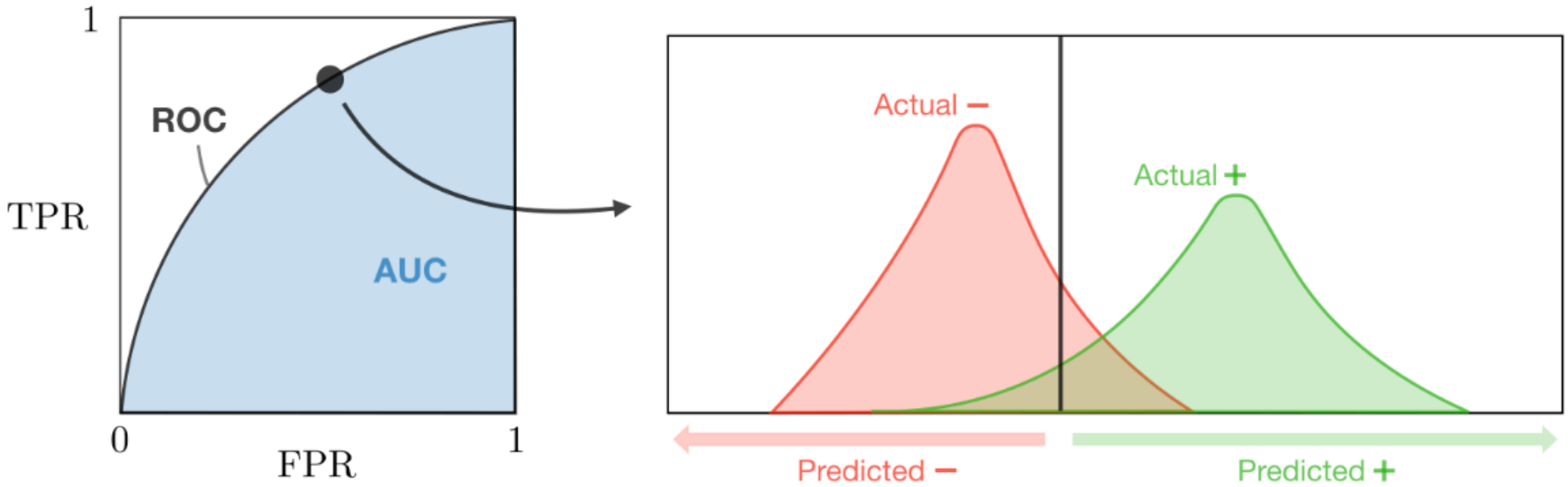


!

Son **insensibles** a los cambios en las distribuciones de clases.

[Fawcett, 2006]

ROC-AUC



!

Son **insensibles** a los cambios en las distribuciones de clases.

⚠

Para problemas de clasificación muy desbalanceados, son incapaces de ver aumentos de FP (error de tipo I).

Medidas de evaluación

Basadas en las probabilidades

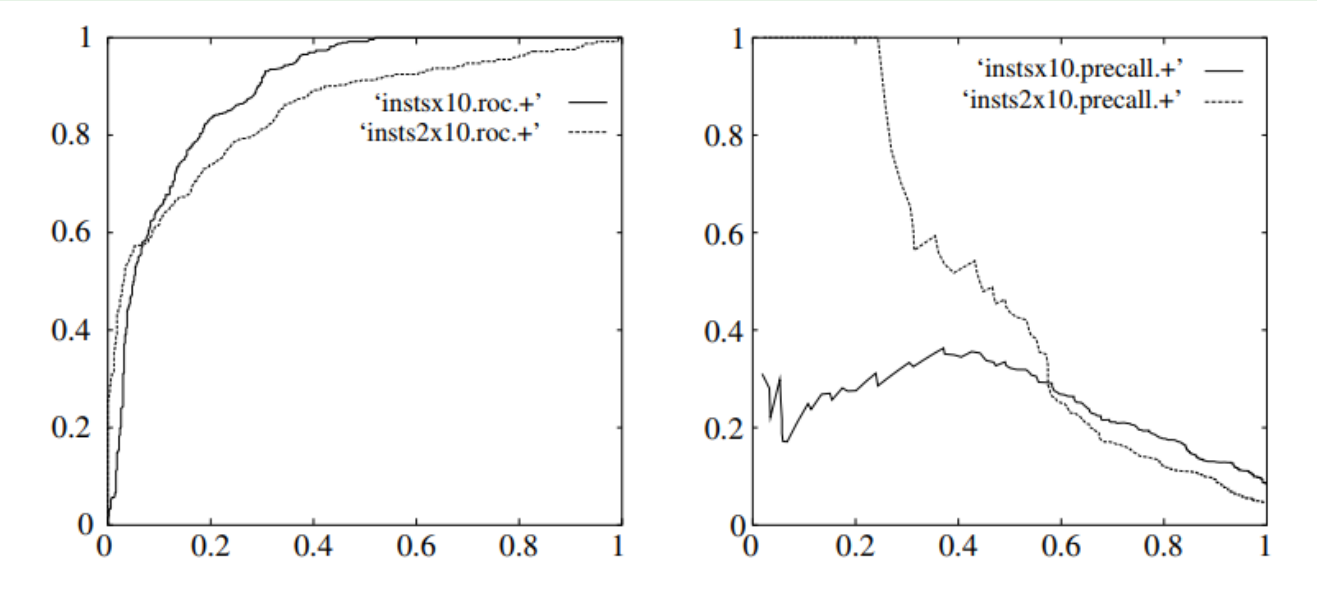
¿Qué pasa si las clases están desbalanceadas?

Medidas de evaluación

Basadas en las probabilidades

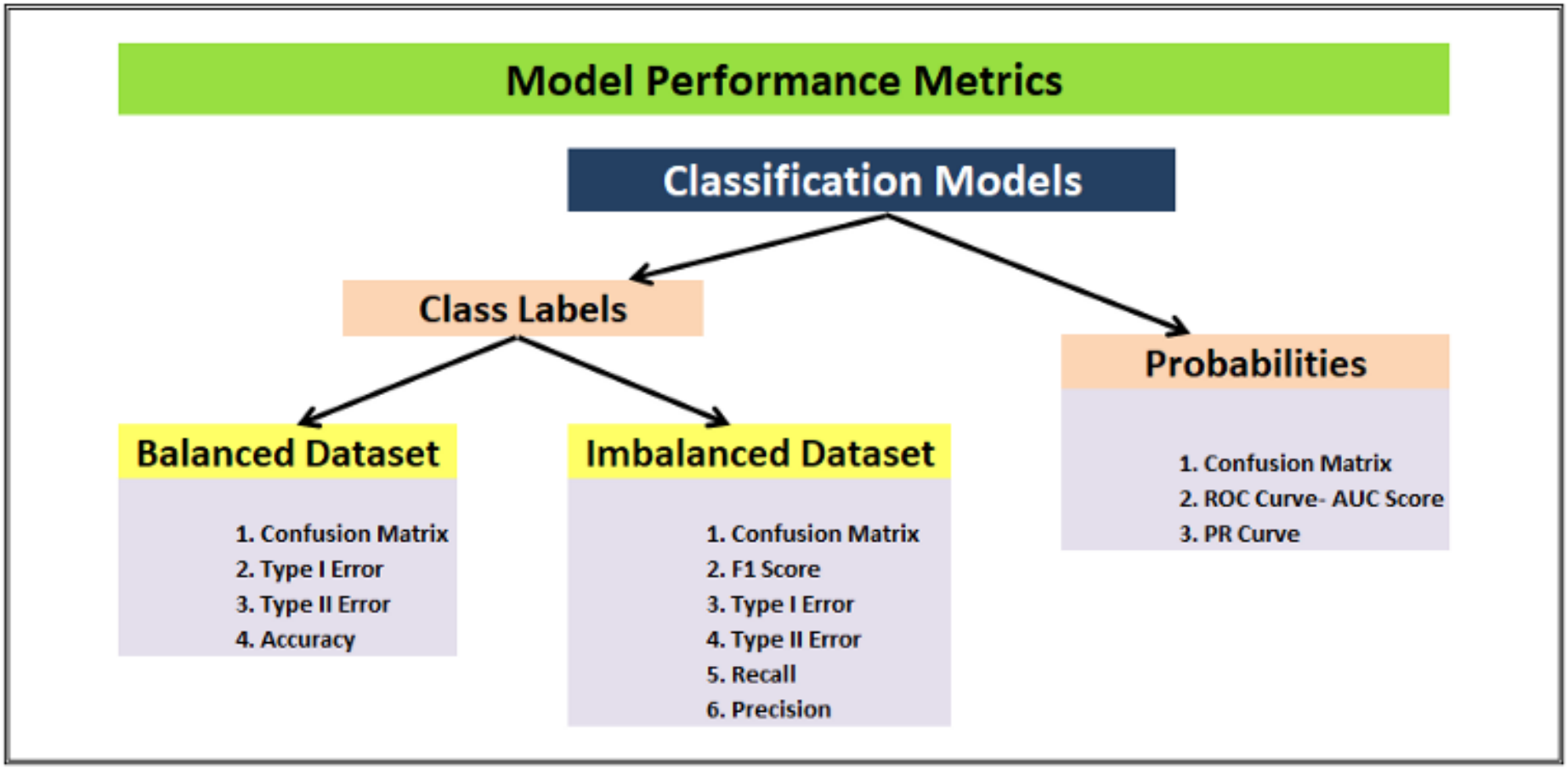
¿Qué pasa si las clases están desbalanceadas?

Curvas Precision-Recall



Fuente: Fawcett, Tom. "An introduction to ROC analysis". Pattern recognition letters 27.8 (2006): 861-874.

Medidas de evaluación



Fuente: Towardsdatascience

Chapter 17
Is It Significant? Guidelines for Reporting BCI Performance

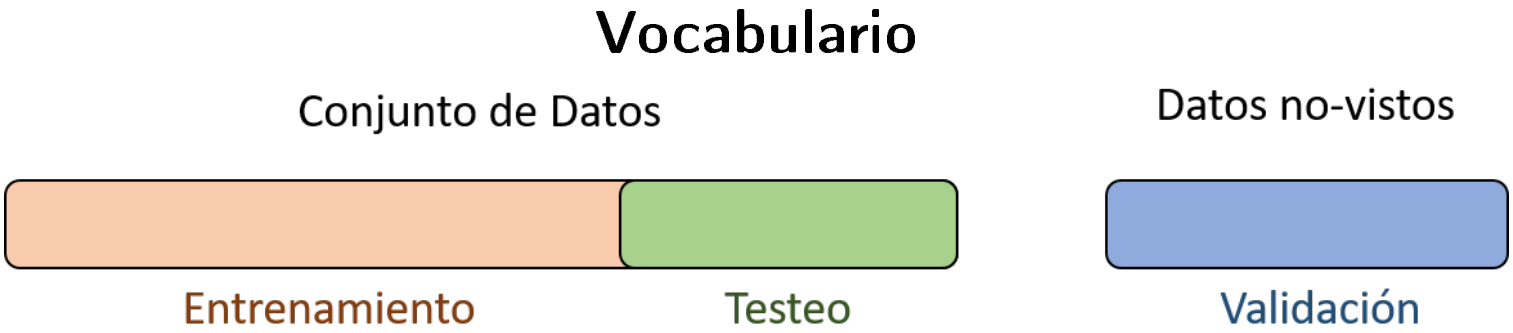
Martin Billinger, Ian Daly, Vera Kaiser, Jing Jin, Brendan Z. Allison, Gernot R. Müller-Putz, and Clemens Brunner

Abstract Recent growth in brain-computer interface (BCI) research has increased pressure to report improved performance. However, different research groups report performance in different ways. Hence, it is essential that evaluation procedures are valid and reported in sufficient detail.

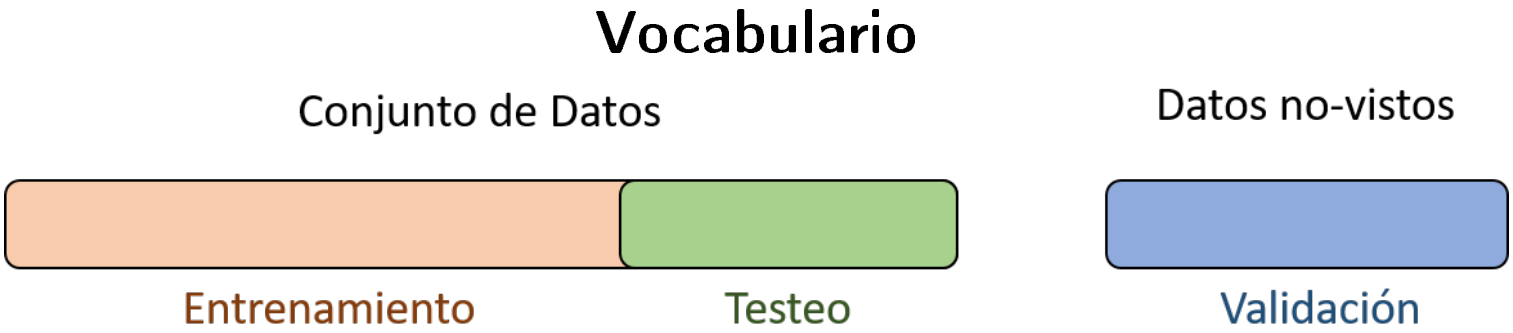
In this chapter we give an overview of available performance measures such as classification accuracy, cohen's kappa, information transfer rate (ITR), and written symbol rate (WSR). We show how to distinguish results from chance level using confidence intervals for accuracy or kappa. Furthermore, we point out common

1. Vocabulario
2. Generalidades
3. Análisis Discriminante Lineal
   - Método de transformación lineal
   - Hiperplano de separación
4. Evaluación y performance
   - Médidas de evaluación
   - Estimación del desempeño

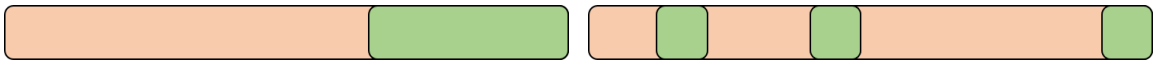
Técnicas de validación



Técnicas de validación

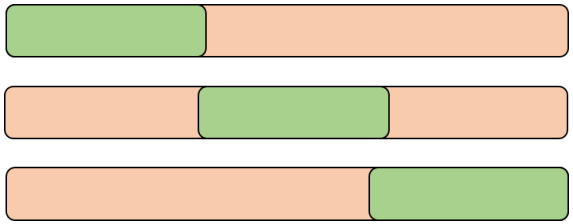


Validación M-fuera(*M-hold out*)



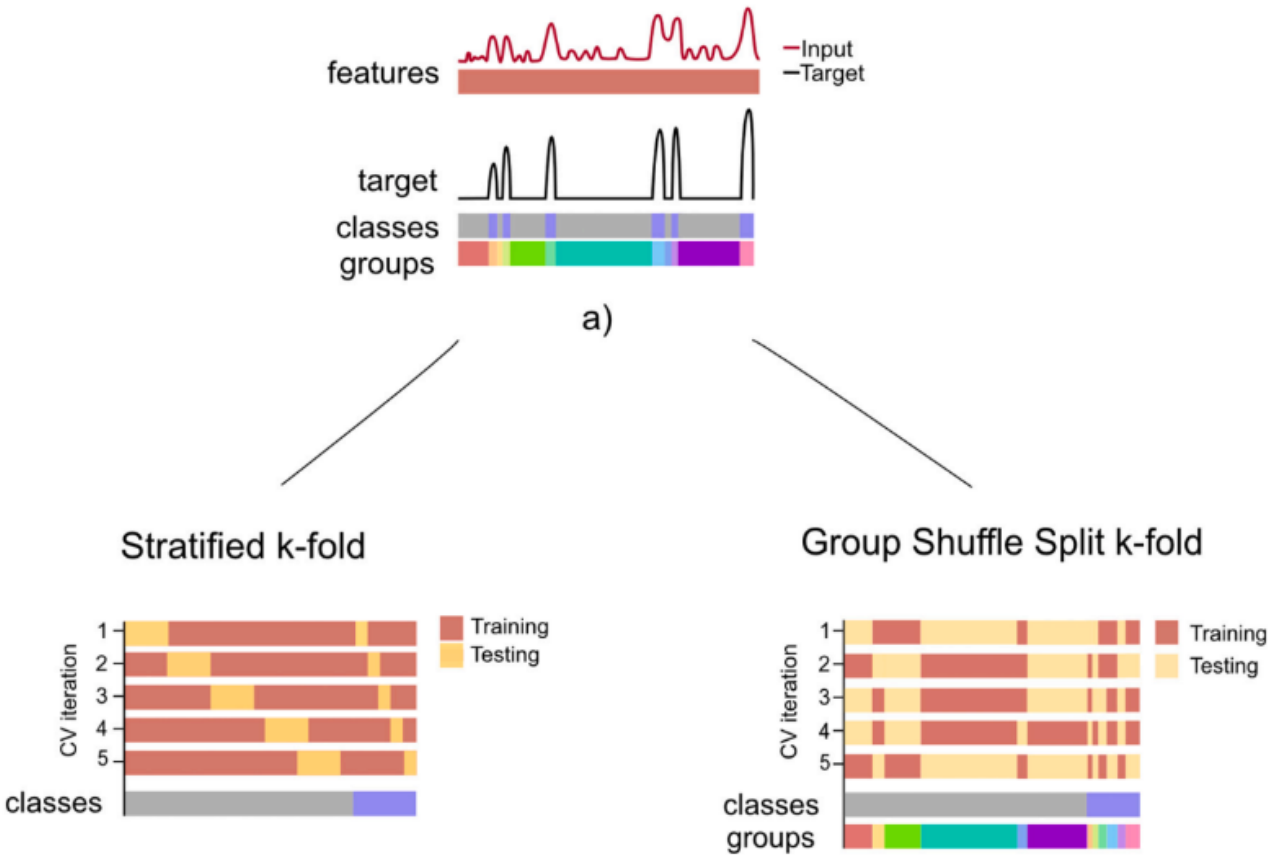
Validación 1-fuera (*Leave one-out*)

Validación cruzada (*Cross-validation*)



Considerando la variable target

Validación cruzada *estratificada*



Fuente: [Merk et al., 2022]

Bibliografía utilizada I

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Merk, T., Peterson, V., Köhler, R., Haufe, S., Richardson, R. M., and Neumann, W.-J. (2022). Machine learning based brain signal decoding for intelligent adaptive deep brain stimulation. *Experimental Neurology*, 351:113993.

Webb, A. R. (2003). *Statistical pattern recognition*. John Wiley & Sons.