Herramientas de aprendizaje supervisado

Clase 1 - Clasificación

Manuel Benjamín

October 7, 2023

Universidad de Buenos Aires

Generalidades

En un problema de clasificación tenemos un conjunto de observaciones $(x_1, g_1), \ldots, (x_n, g_n)$ donde g es una variable categórica que indica la clase que fue observada.

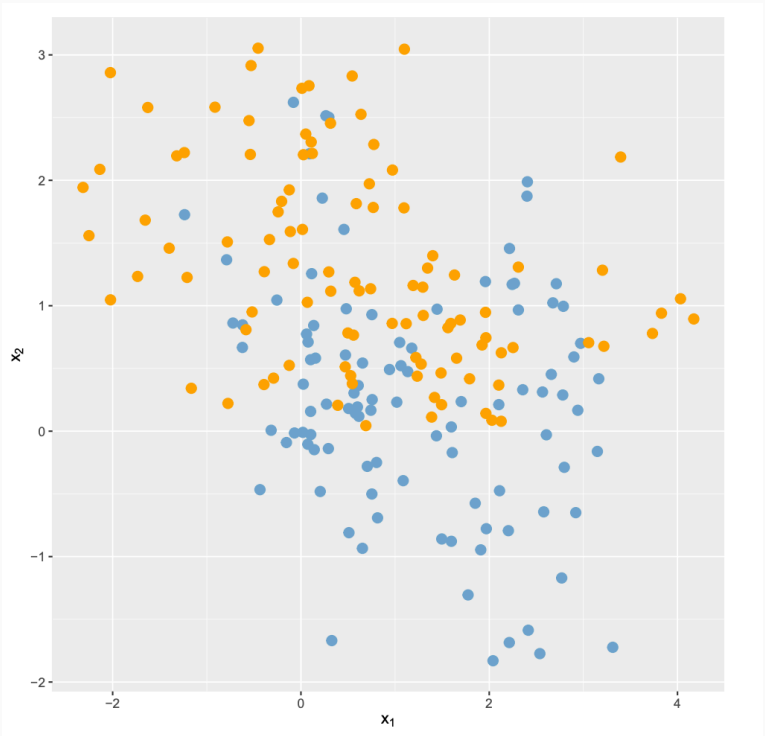


Figure 1: Scatter plot de 200 observaciones (x_{i1}, x_{i2}, g_i). Existen dos clases que esta representados con celeste o naranja.

Un primer intento en construir un clasificador

Supongamos que asignamos el valor 1 a la clase naranja y 0 a la clase celeste.

¿Que sucede si ajustamos un modelo lineal a las observaciones?

$$\hat{\beta} = \textbf{arg min} \sum_{i=1}^{n} (y_i - \beta_0 + \beta x)^2$$

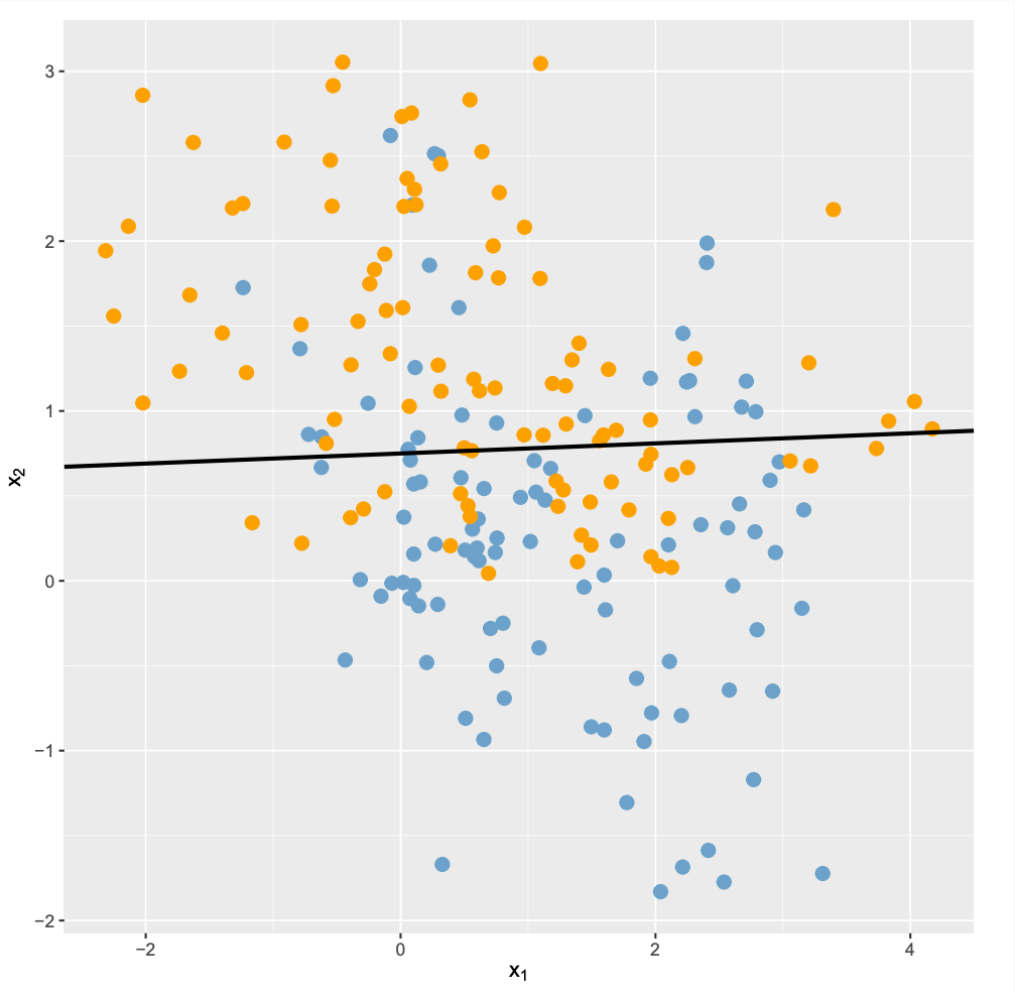


Figure 2: Ajuste lineal

Clasificamos a la clase dependiendo el valor de $\hat{y}$ a la clase predicha según

$$\widehat{G} = \begin{cases} 1 & \text{si } \hat{y} \geq 0.5, \\ 0 & \text{si } \hat{y} < 0.5. \end{cases}$$

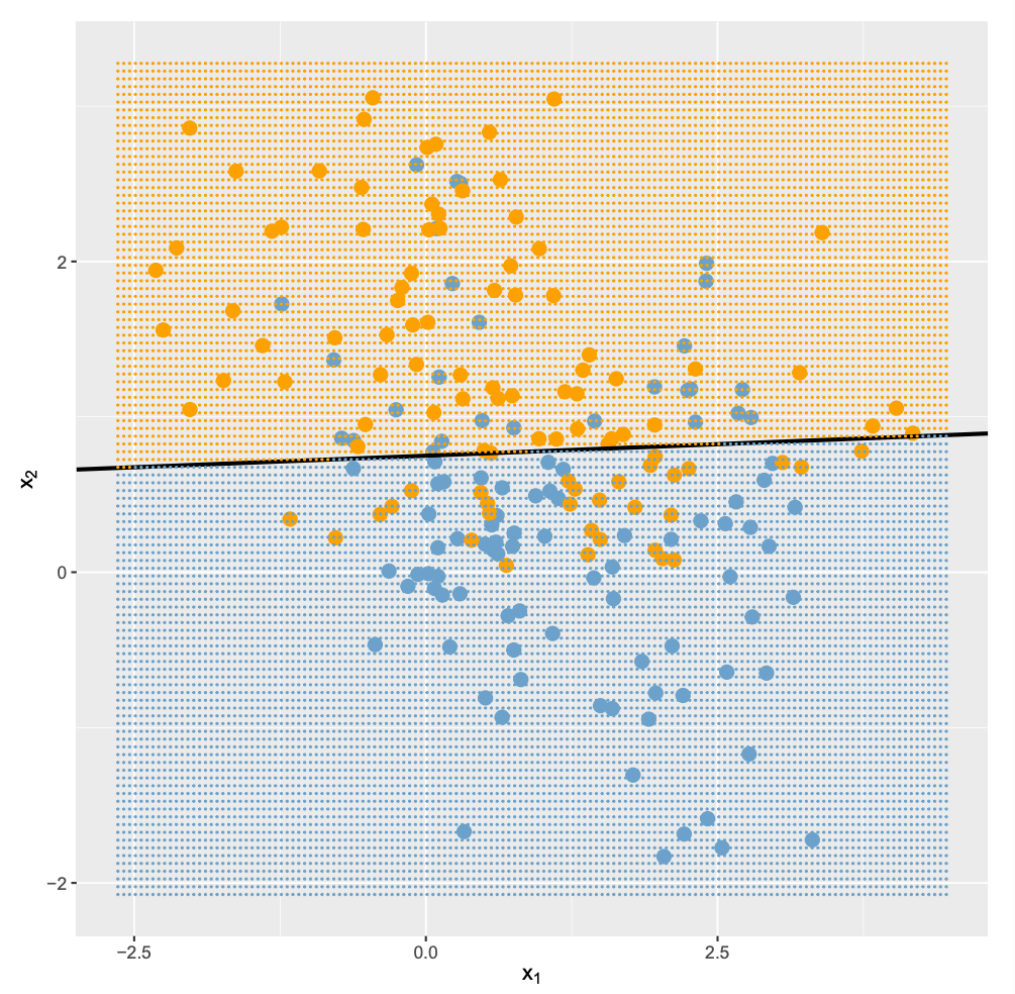


Figure 3: Región de clasificación.

El conjuntos de puntos de $\mathbb{R}^2$ clasificados como NARANJAS corresponde a

$$\{x : x^t\hat{\beta} \geq 0.5\}.$$

La frontera de decisión que separa ambas clases es

$$\{x : x^t\hat{\beta} = 0.5\},$$

que es una frontera lineal en x.

Vecinos cercanos

Se elige la clase con mayor cantidad de representantes entre los K vecinos mas cercanos.

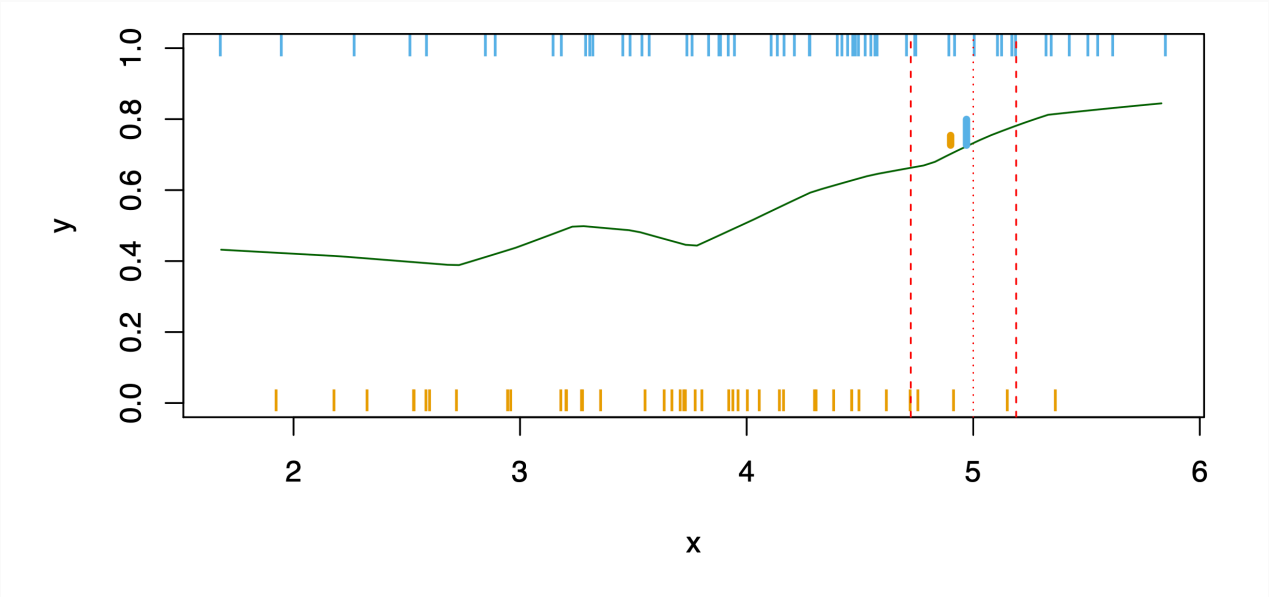


Figure 4: Clasificación por vecinos mas cercanos problema en un problema de clasificación binaria y una única variable predictora.

¿Como medimos la certidumbre de un clasificador?

Un clasificador es una función $C(x) \to \{g_1, \ldots, g_k\}$

La función de perdida que consideramos en un problema de clasificación penaliza las clasificaciones en clases incorrectas

$$L(C(X), y) = \mathbb{I}_{\hat{y} \neq y} = \begin{cases} 1 & \text{si } \hat{y} \neq y, \\ 0 & \text{si } \hat{y} = y. \end{cases}$$

El clasificador óptimo para la teoría de la decisión es el que minimiza

$$E\left(\mathbb{I}_{C(X) \neq Y}\right) = P(\text{Mala clasificación}).$$

Clasificador de Bayes

Este clasificador es el que minimiza el riesgo de una mala clasificación y queda definido por

$$C(x) = \arg\max_{g} P(g|X = x)$$

Es decir, asignar la clase de mayor probabilidad.

- Obviamente este clasificador es incalculable en la práctica.
- El error de clasificación asociado es el **Error de Bayes**
- El clasificador es óptimo en el sentido de que cualquier otro clasificador va a tener un error de clasificación mayor.

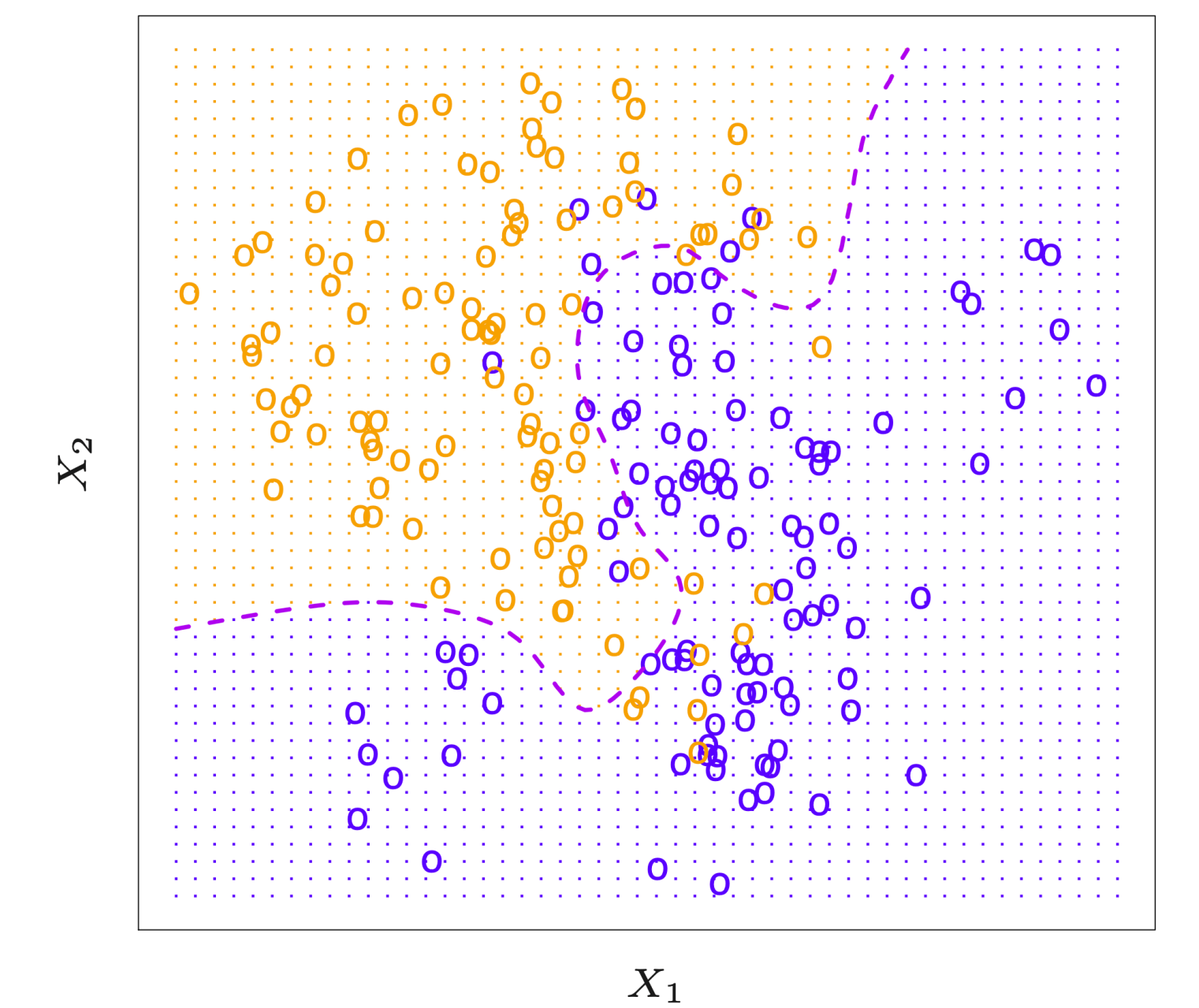
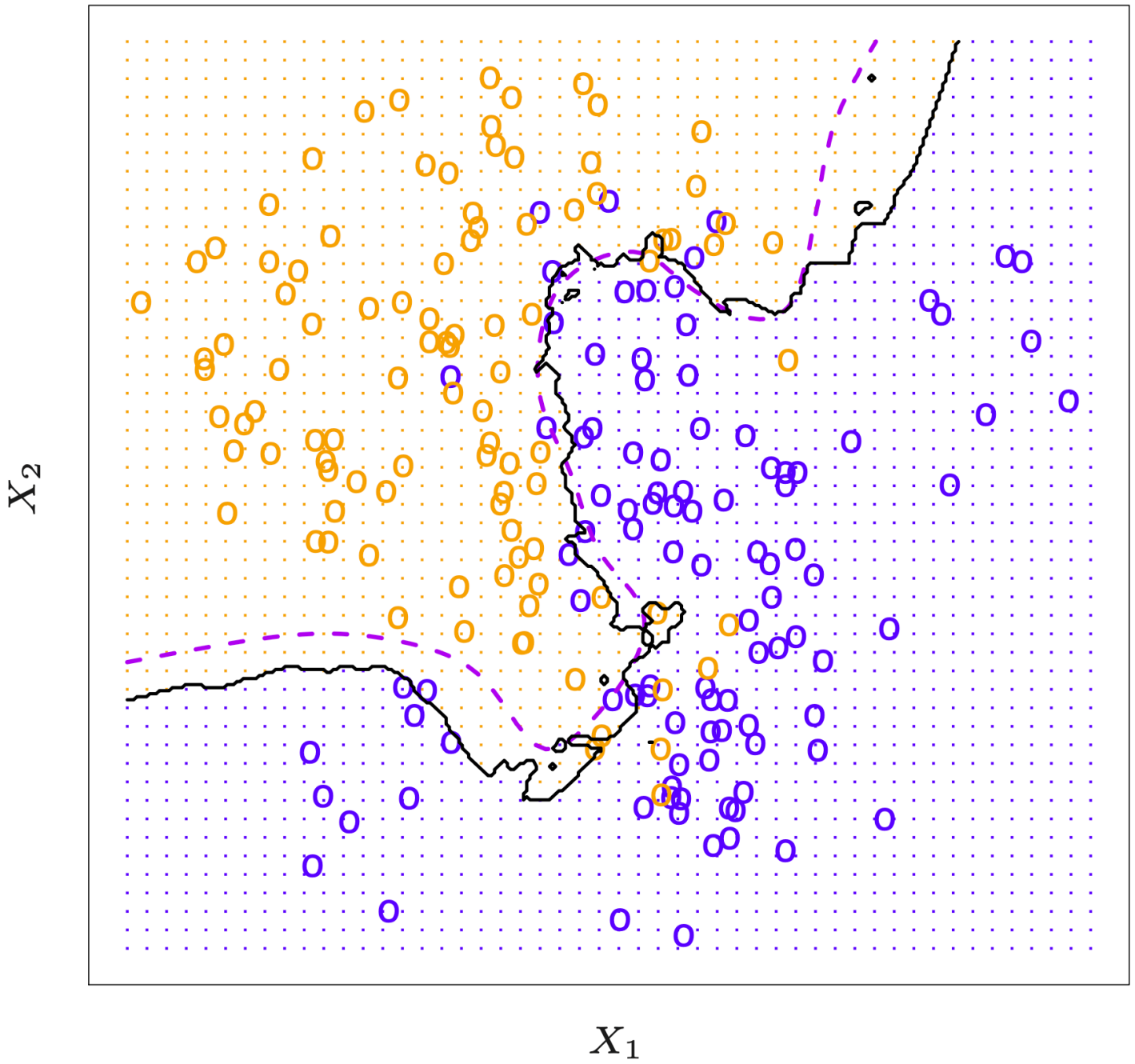
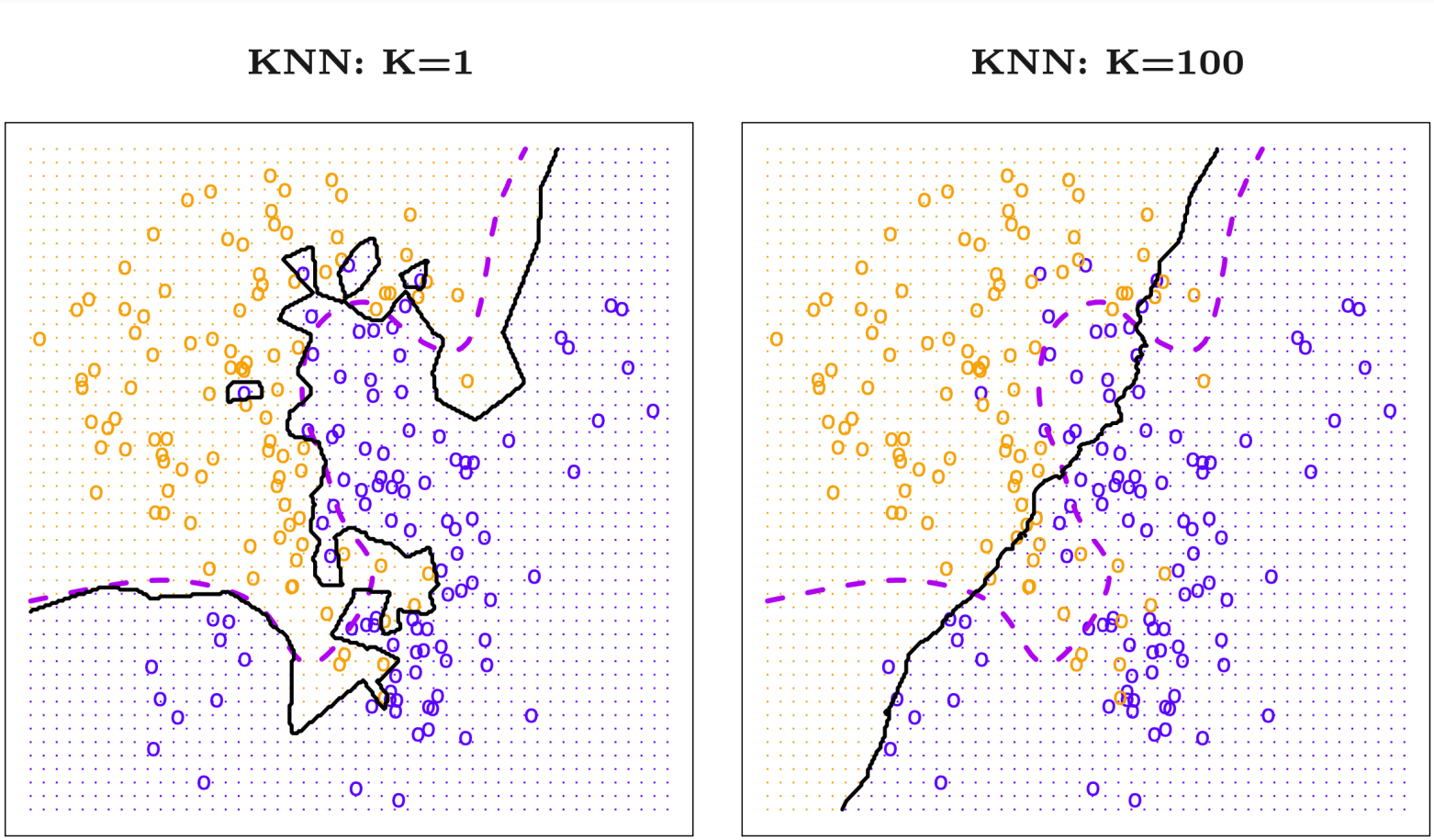


Figure 5: Clasificador de bayes y su frontera de clasificación

KNN: K=10




KNN: K=1
KNN: K=100

Intuición detrás de Vecinos Cercanos

- En cada x del espacio de covariables queda estimada la probabilidad de cada clase

$$\widehat{P}(g_j|X = x) = \frac{\#\text{clases j entre los K vecinos mas cercanos a x}}{K}.$$

- Al elegir la clase mayoritaria estamos asignando

$$C(x) = \arg\max_g \widehat{P}(g|X = x).$$

¿Cómo estimaos el error de clasificación?

Usualmente usamos el error de clasificación

$$\text{Err} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{y_i \neq \hat{y}_i}$$

Esto debemos calcularlo sobre datos de testeo independientes de los datos de entrenamiento.

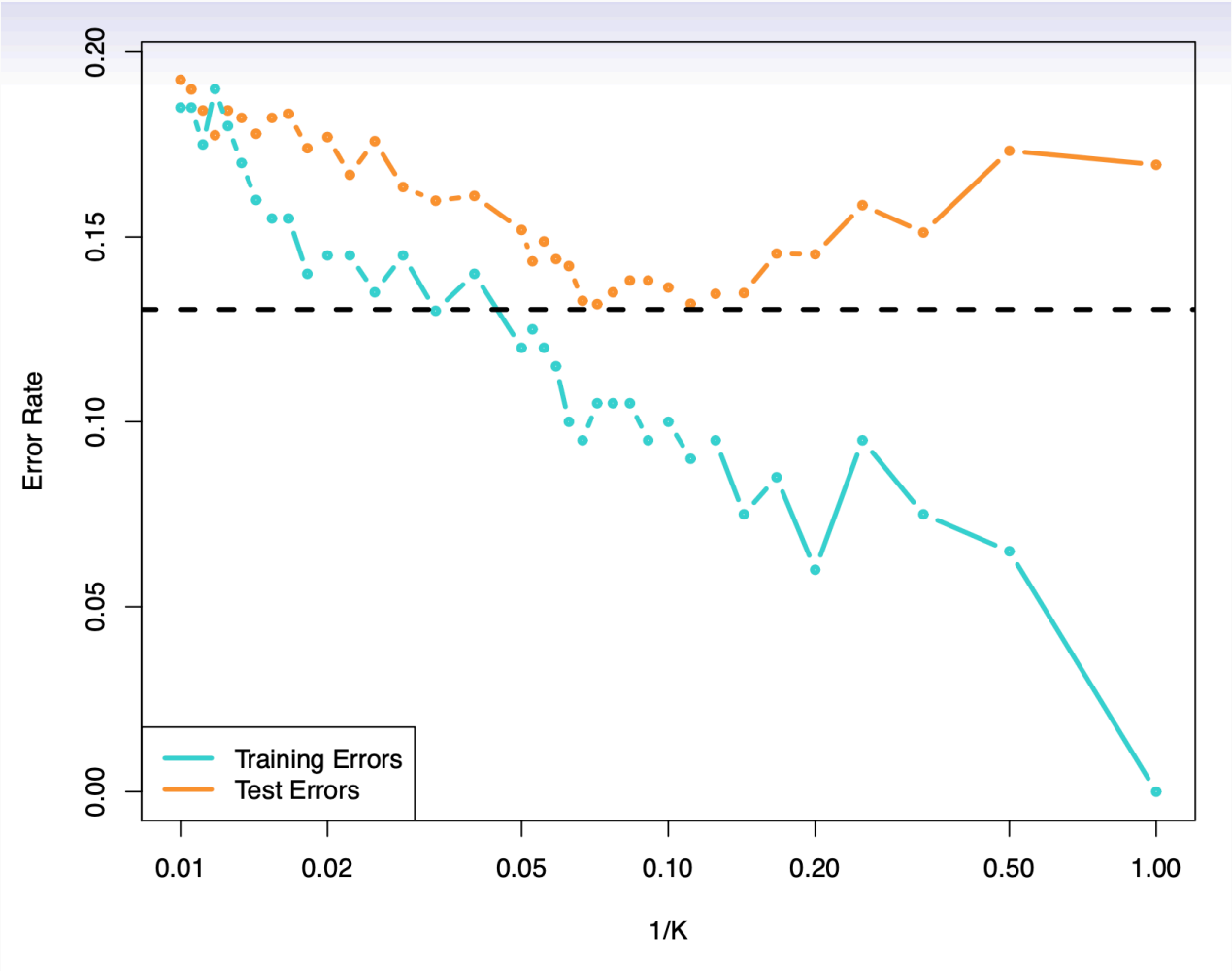


Figure 7: Error de clasificación para KNN. La linea punteada indica el Error de Bayes.

Regresión logística

En un problema de clasificación binaria modelamos la probabilidad de la clase 1

$$p(x) = \frac{exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}.$$

Asumimos

$$Y|X = x \sim \mathcal{B}_e(p(x)).$$

Ajustamos el modelo por máxima verosimilitud

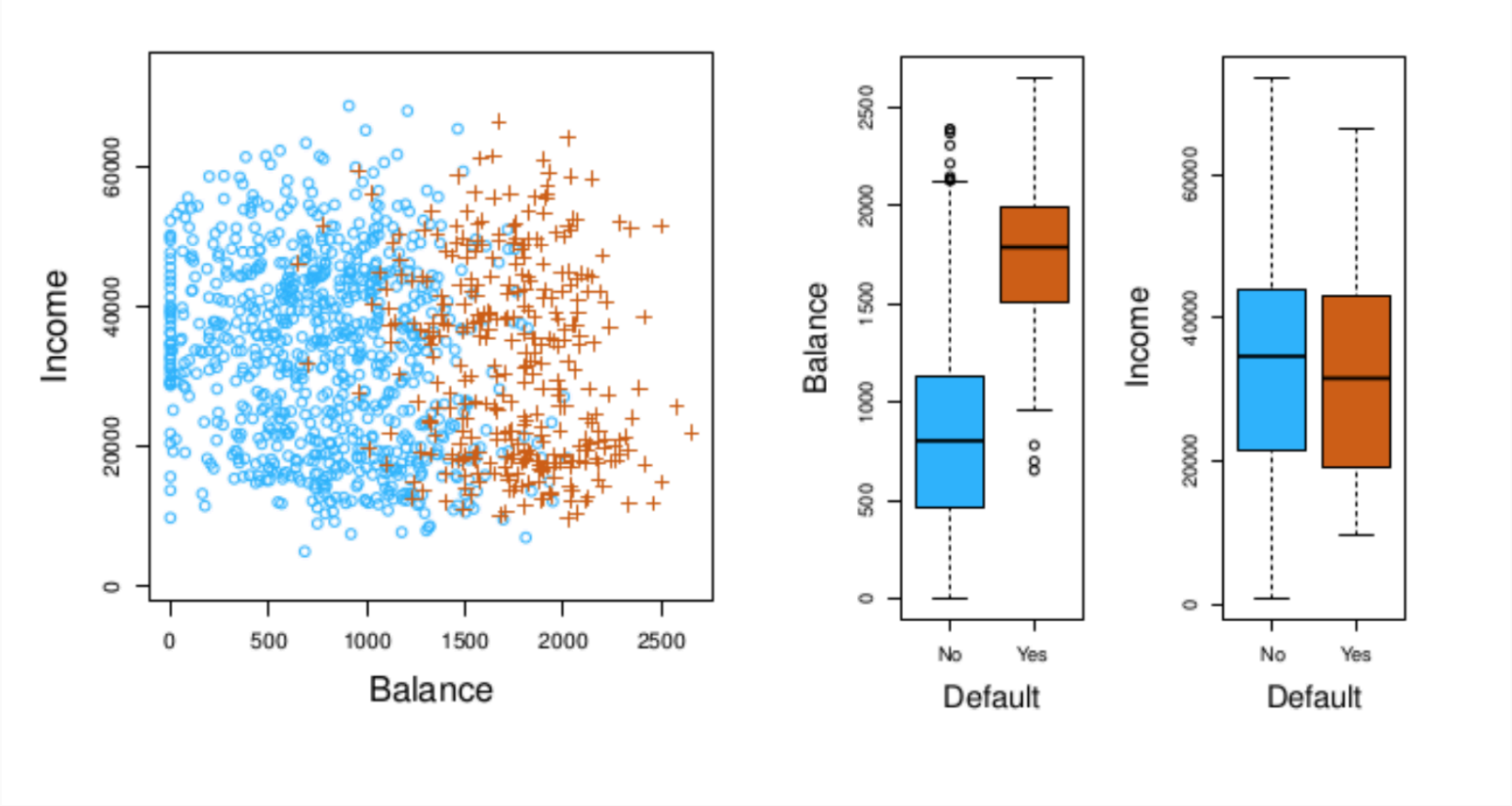


Figure 8: Datos de Default de tarjeta en naranja si cesaron sus pagos, variables explicativas balance e ingreso.

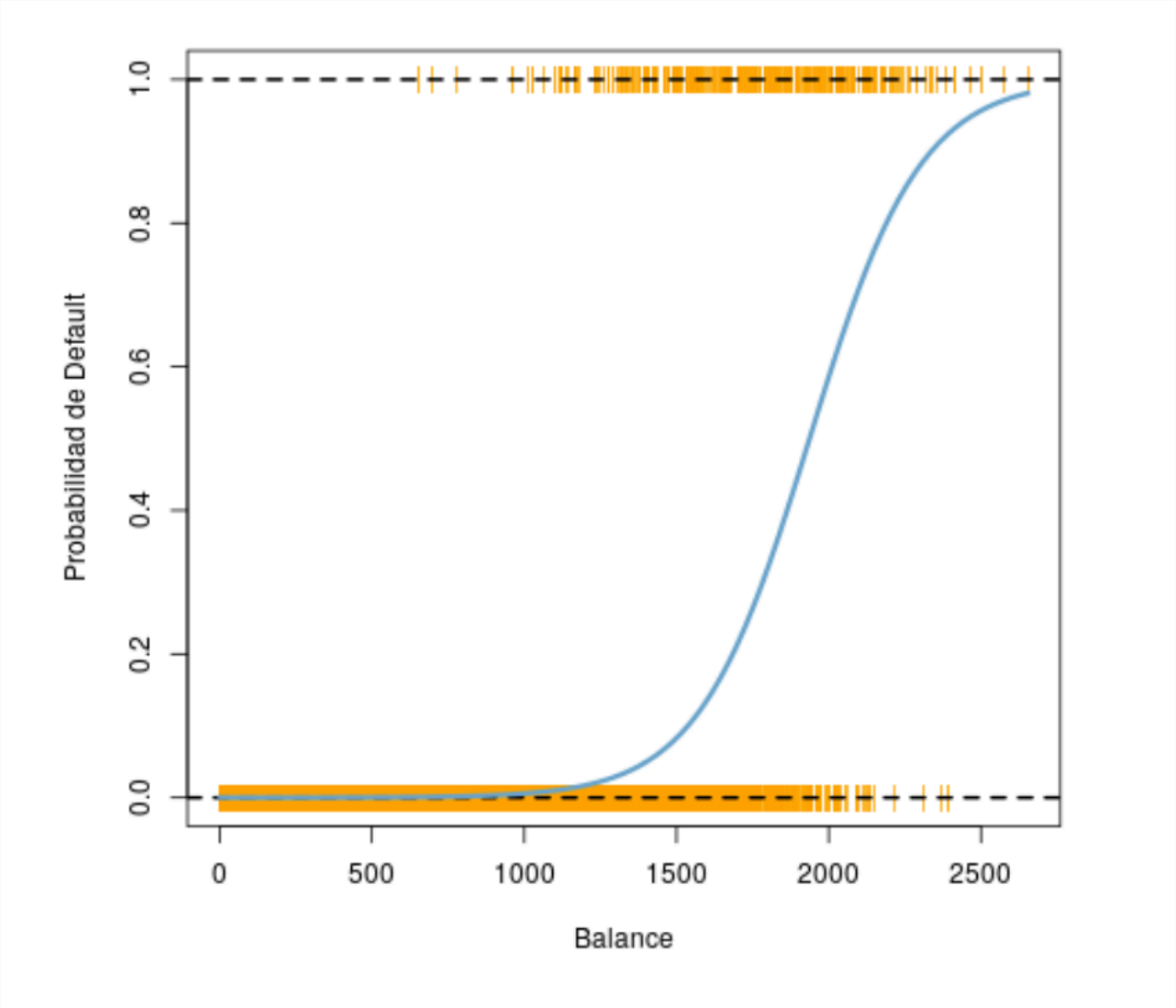


Figure 9: Regresión logistica Default balance