Aprendizaje supervisado

Maestría en Estadística

Pablo Vena

Universidad de Buenos Aires

23 de octubre de 2021

Lecturas

ISLR Capítulo 1, 2.

ESL Capítulo 1, 2.

Aprendizaje supervisado

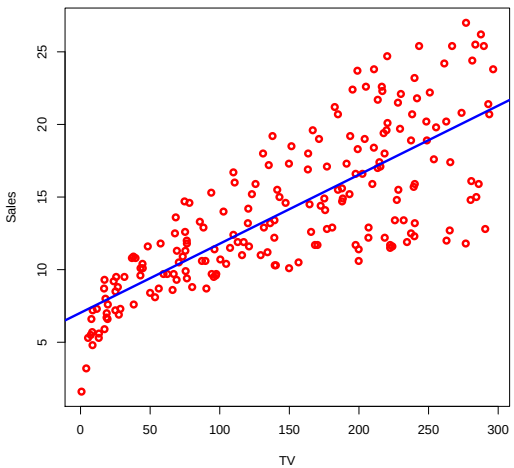
Ejemplo (advertising)

El conjunto de datos `Advertising` consiste en las ventas de un producto en 200 mercados diferentes junto con los presupuestos entre tres medios: `TV`, `radio` y `periódicos`.

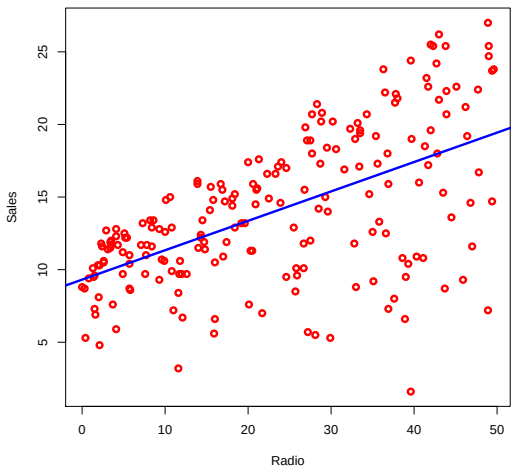
	TV	radio	newspaper	sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
⋮	⋮	⋮	⋮	⋮

El objetivo es predecir las `ventas` en función de los presupuestos.

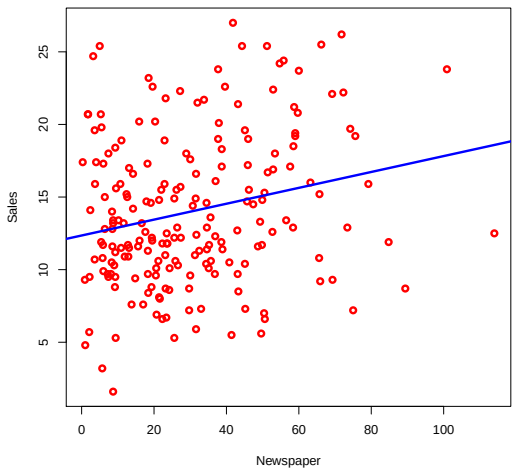
Aprendizaje supervisado



(a) TV



(b) Radio



(c) Periódicos

Figura: Dataset `Advertising`.

Aprendizaje supervisado

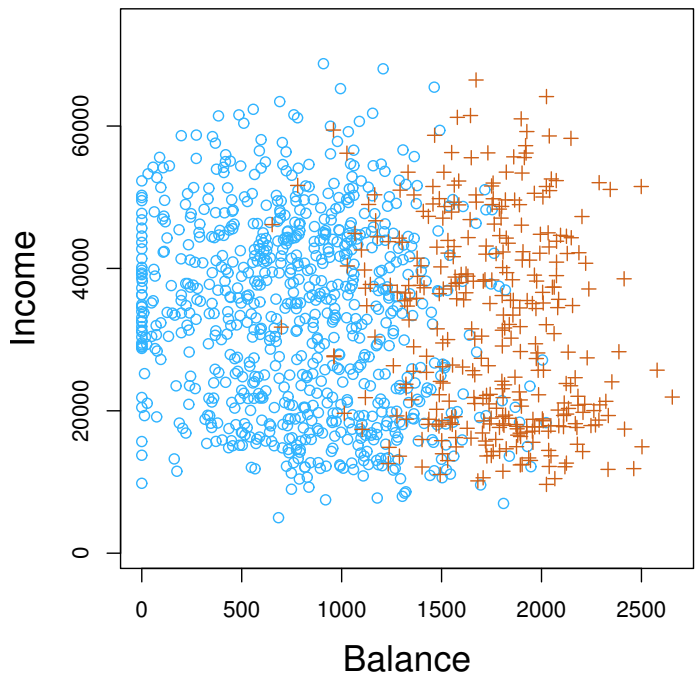
Ejemplo (default)

El conjunto de datos `Default` consiste en los pagos (o no) de tarjetas de crédito. Cada observación cuenta con el `ingreso` anual del inviduo, el `balance` mensual de la tarjeta y si es `estudiante`.

	default	student	balance	income
1	No	No	729.5265	44361.625
2	No	Yes	817.1804	12106.135
3	No	No	1073.5492	31767.139
4	No	No	529.2506	35704.494
5	No	No	785.6559	38463.496
6	No	Yes	919.5885	7491.559
⋮	⋮	⋮	⋮	⋮

El objetivo es predecir si el individuo paga la tarjeta (`default`).

Ajuste de datos



Ajuste de datos

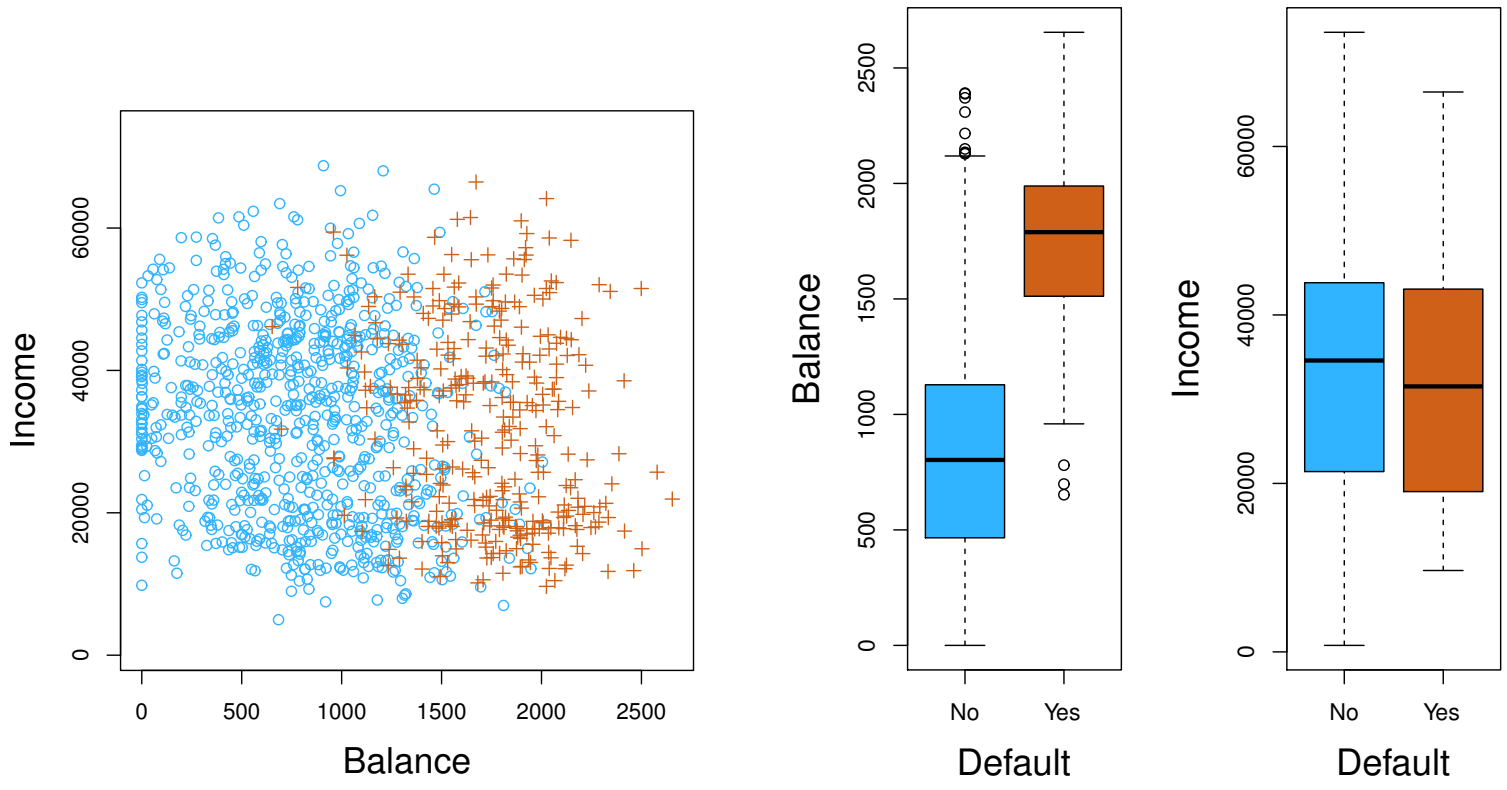


Figura: Dataset `default`.

Ajuste de datos

- pensamos que $y \in \mathbb{R}$ y $x \in \mathbb{R}^p$ se relacionan según

$$y \approx f(x)$$

- x es la **variable independiente**
  - predictor,
  - regresor,
  - covariable,
  - *feature*
- y es la **variabe dependiente** o **respuesta**
- queremos predecir y
- no conocemos la relación verdadera entre x e y

Supervisión

Podemos distinguir dos grandes categorías de problemas de aprendizaje:

supervisado para cada x_i **hay** una respuesta asociada y_i

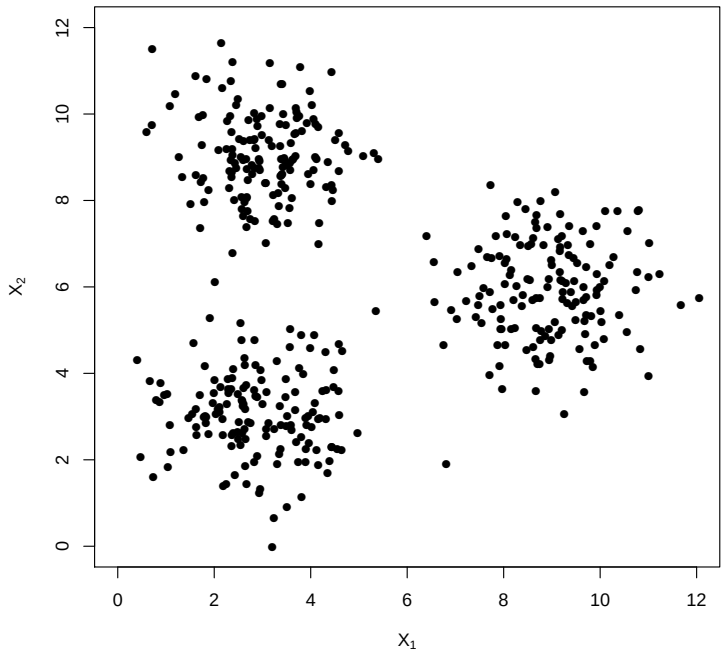
no supervisado para cada x_i **no hay** una respuesta asociada y_i

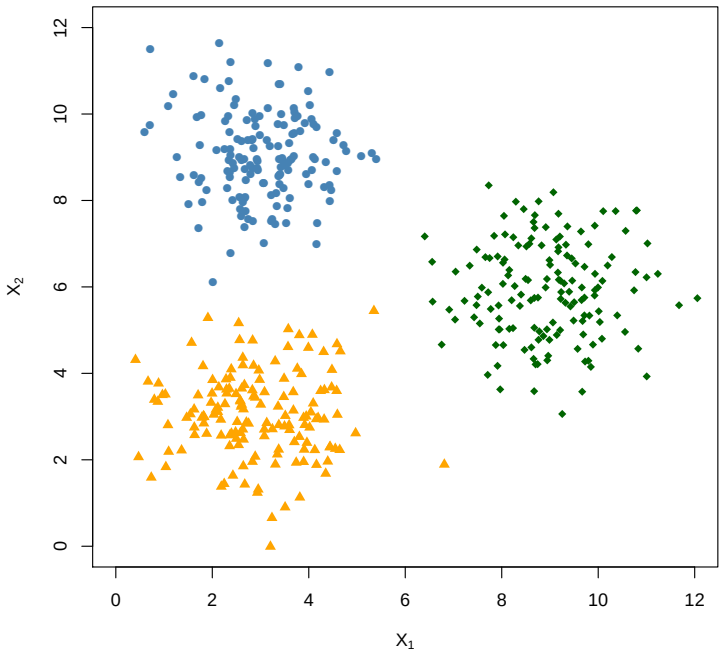
Supervisión

Podemos distinguir dos grandes categorías de problemas de aprendizaje:

supervisado para cada x_i **hay** una respuesta asociada y_i

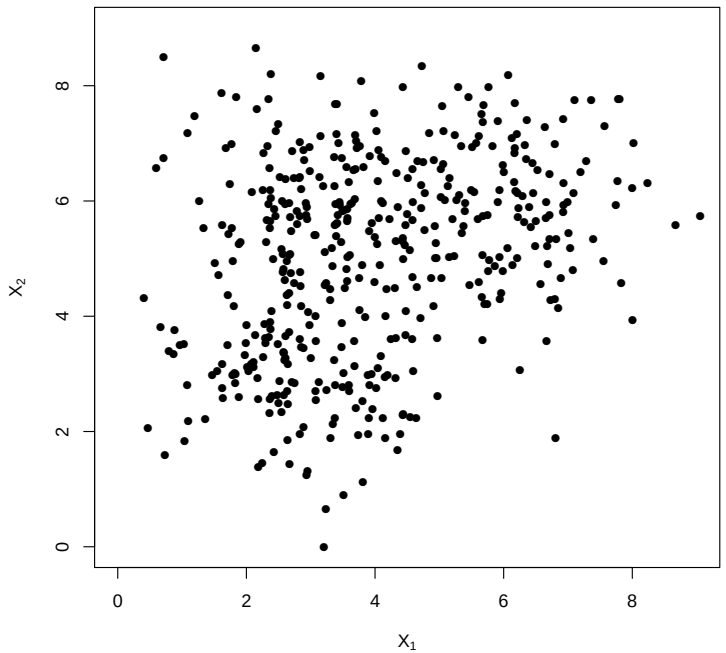
no supervisado para cada x_i **no hay** una respuesta asociada y_i





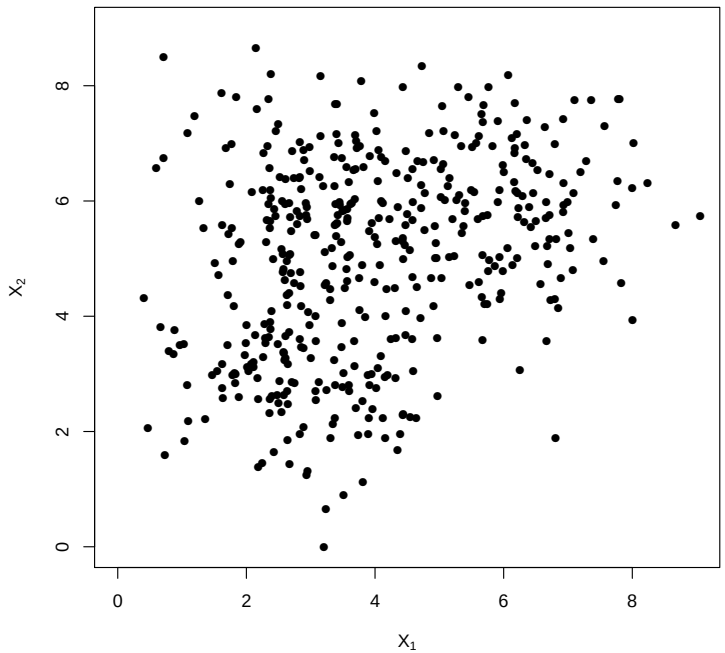
Aprendizaje no supervisado

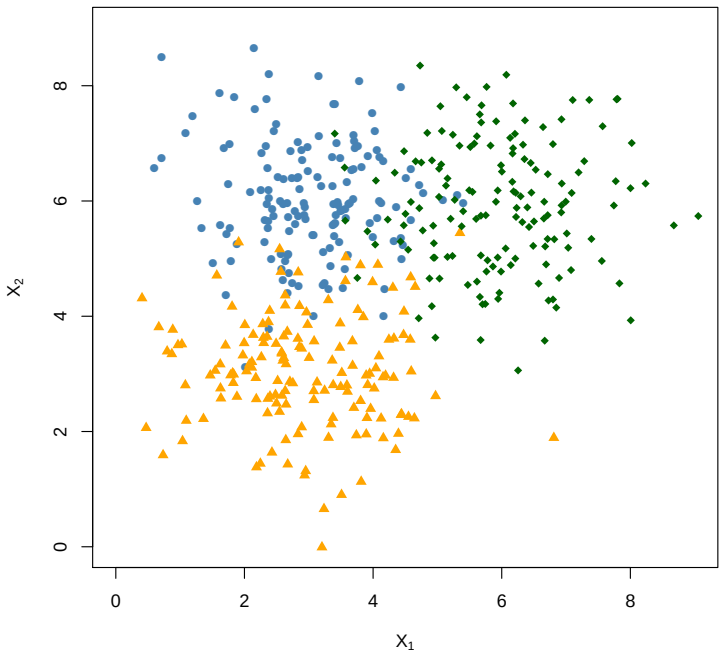
¿Cuántos grupos hay?



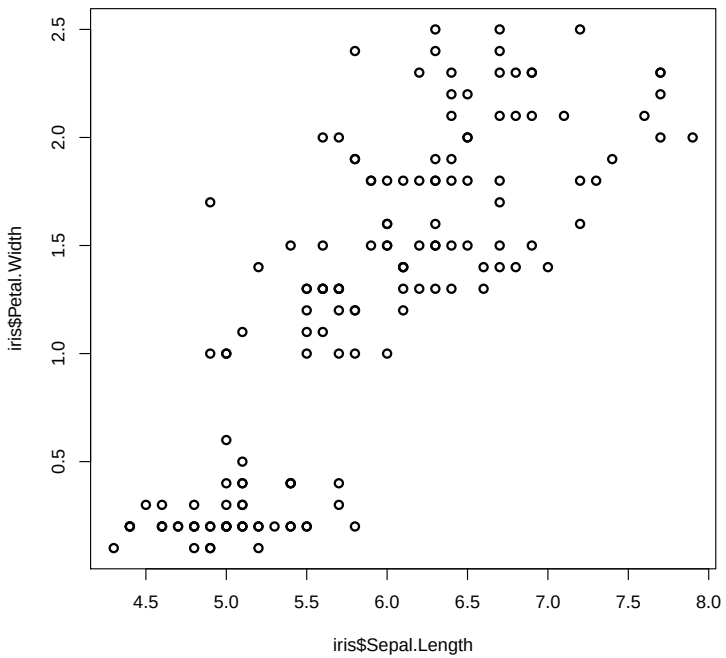
Aprendizaje no supervisado

¿Cuántos grupos hay?

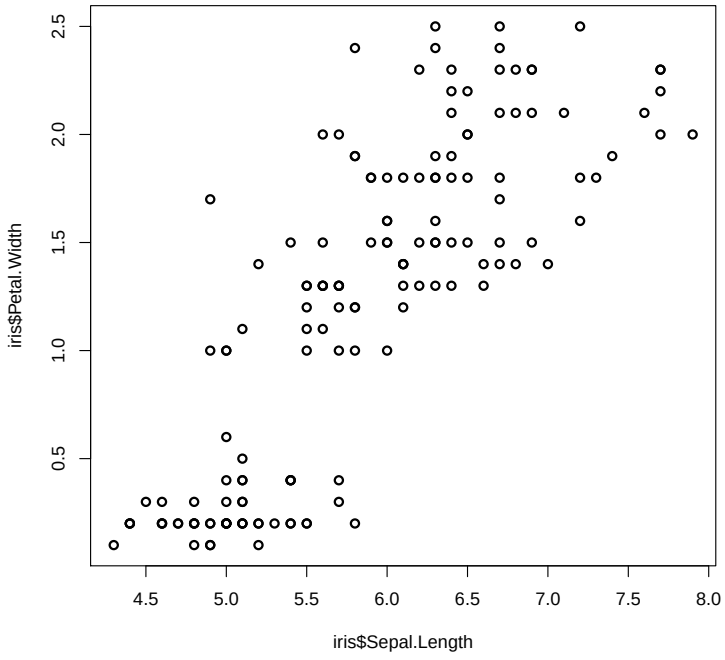


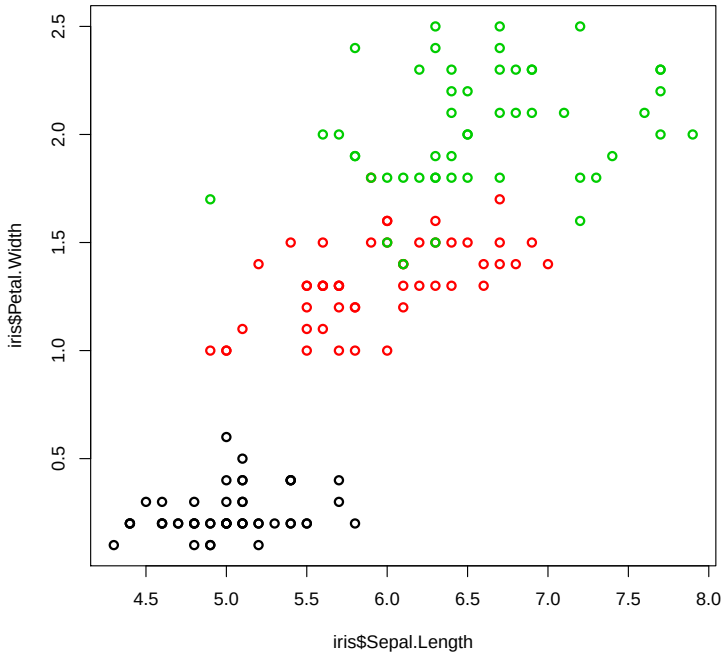


Aprendizaje no supervisado



Aprendizaje no supervisado






```
plot(iris$Sepal.Length,iris$Petal.Width, lwd=2)
```

Aprendizaje no supervisado

Ejemplos

- Clasificación de documentos
- Segmentación de bases de datos de clientes/usuarios
- Segmentación de imágenes (satelitales)
- ...

- k-medias, k-medoids
- clustering jerárquicio
- DBSCAN
- ...

Aprendizaje por refuerzo (reinforcement learning)

Figura: Multiarmed bandit problem.

Queremos asignar un recurso finito (presupuesto) entre varias opciones según recompensas aleatorias (performance, clics, ventas).

Aprendizaje por refuerzo (reinforcement learning)

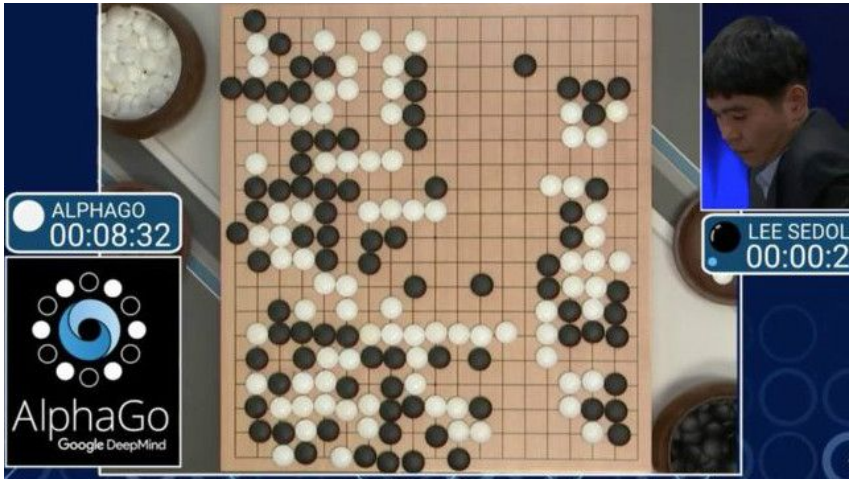


Figura: Lee Sedol vs AlphaGo.

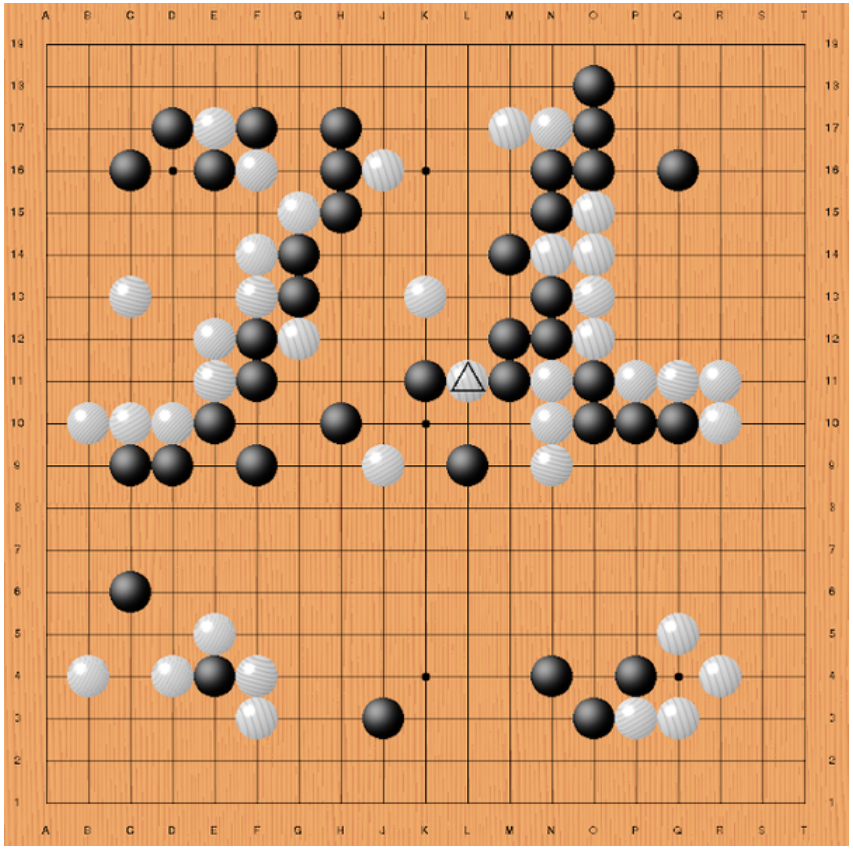


Figura: Move 78.

Clasificación

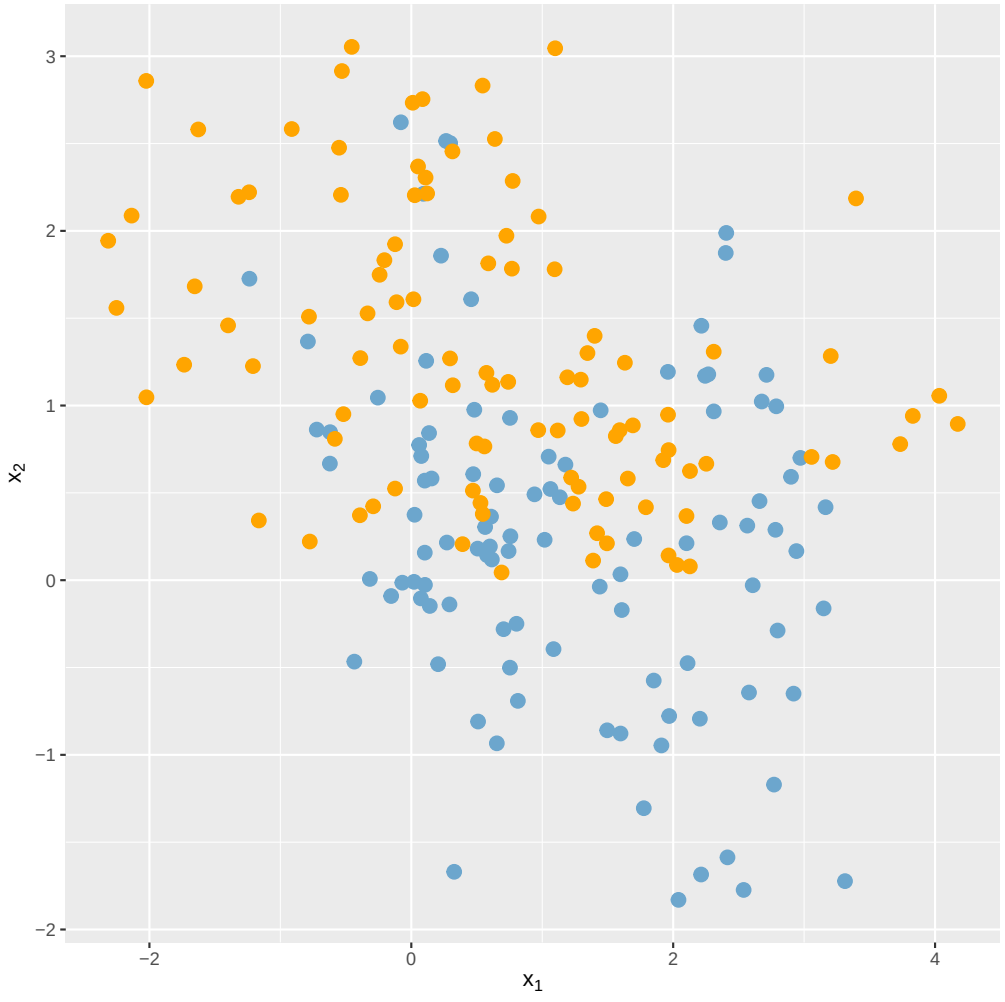


Figura: Scatter plot.

Consideremos un conjunto de datos (sintético) de observaciones (x_1^i, x_2^i, g^i):

- la variable G indica la clase de cada observación,
- hay 200 observaciones, 100 en cada clase (balanceado).

Clasificación: ajuste lineal

Probemos un **ajuste lineal** para construir un clasificador: predecimos la clase de la observación a partir de una combinación lineal de sus coordenadas.

$$\widehat{y} = \widehat{\beta}_0 + x_1\widehat{\beta}_1 + \cdots + x_p\widehat{\beta}_p$$

En forma matricial,

$$\widehat{y} = x^t\widehat{\beta}$$

- El término $\widehat{\beta}_0$ es el **intercept** o **sesgo** (**bias** en *machine learning*).
- Para la escritura matricial agregamos una columna de $(1, \ldots, 1)$ para incluir el *intercept* en el vector de coeficientes.
- La predicción es la proyección del vector del vector de *features* en el plano definido por $\widehat{\beta}$.

Clasificación: ajuste lineal

¿Cómo ajustamos?

Un método muy popular es el **método de cuadrados mínimos**: elegimos los coeficientes β que minimizan la suma de los cuadrados de los residuos.

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - x_i^t \beta\right)^2$$

La función $RSS(\beta)$ es una función cuadrática de β, siempre tiene al menos un mínimo.

Clasificación: ajuste lineal

Derivamos con respecto a β y hallamos la única solución (si X^tX es no singular)

$$\widehat{\beta} = (X^tX)^{-1}X^ty$$

Luego, dada una observación cualquiera x_0, su predicción es

$$\widehat{y}(x_0) = x_0^T\widehat{\beta}$$

- La superficie ajustada queda caracterizada por los p parámetros $\widehat{\beta}$.
- Intuitivamente, no necesitamos demasiados datos para ajustar este modelo.

Clasificación: ajuste lineal

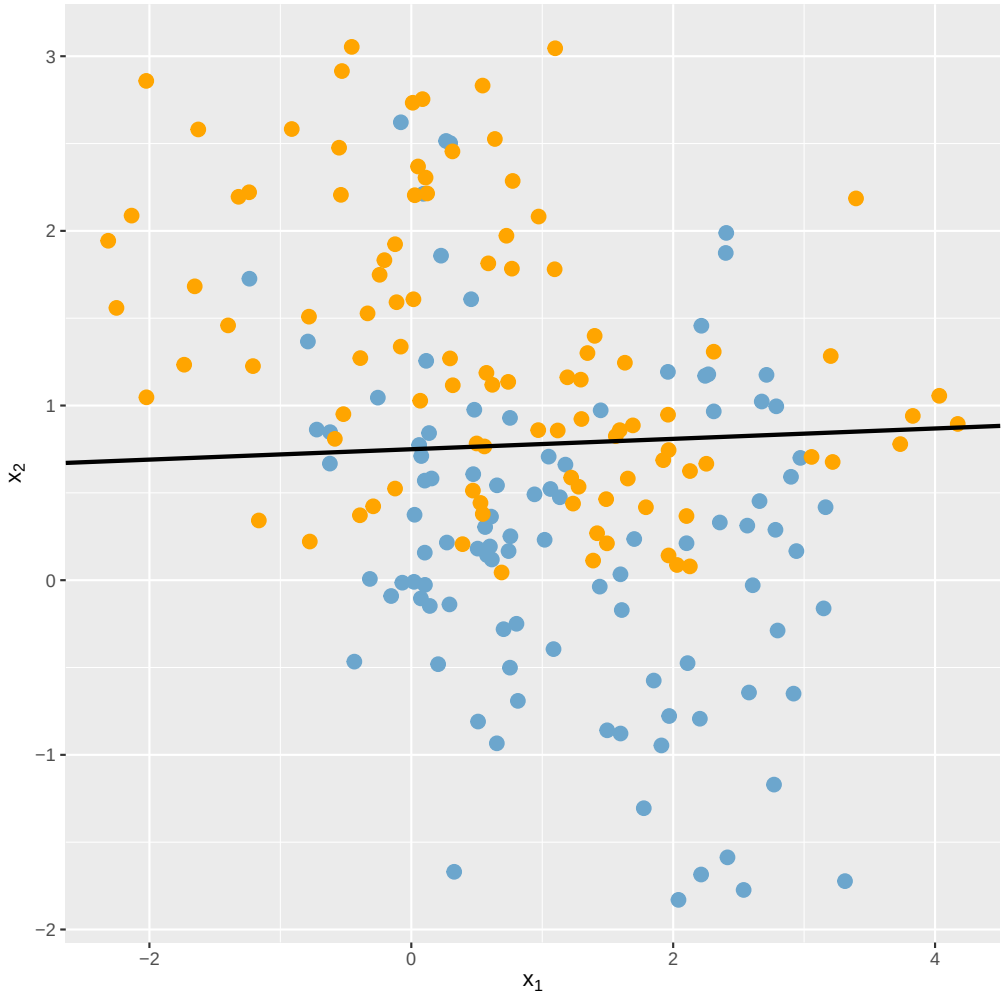


Figura: Ajuste lineal.

En nuestro ejemplo, la variable que queremos predecir es la clase a la que pertenecen las observaciones.

Convertimos nuestras predicciones $\widehat{y}$ a la clase predicha $\widehat{g}$ según

$$\widehat{G} = \begin{cases} 1 & \text{si } \widehat{y} > 0{,}5, \\ 0 & \text{si } \widehat{y} \leq 0{,}5, \end{cases}$$

Clasificación: ajuste lineal

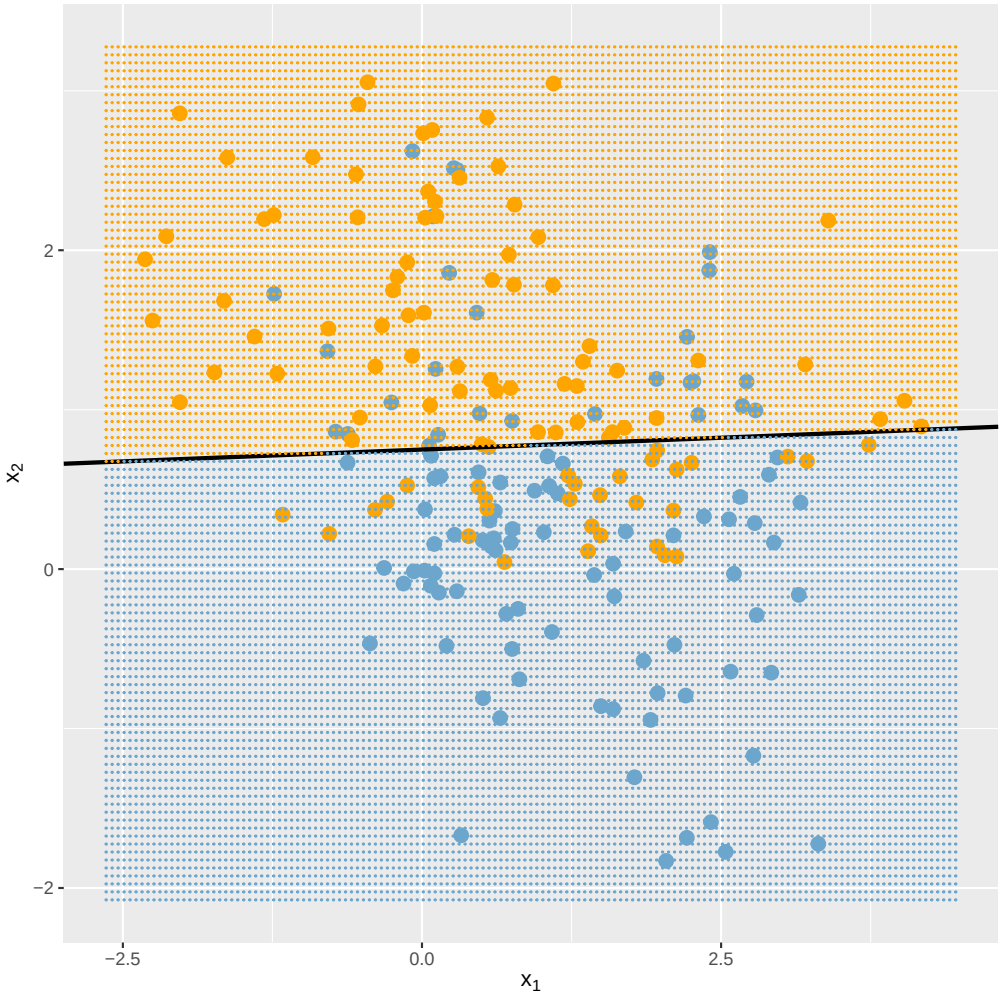


Figura: Regiones de clasificación.

El conjunto de puntos de $\mathbb{R}^2$ clasificado como NARANJA corresponde a

$$\{x : x^t \widehat{\beta} > 0{,}5\}$$

La frontera de decisión que separa ambas clases es

$$\{x : x^t \widehat{\beta} > 0{,}5\}$$

que es lineal.

Clasificación: ajuste lineal

La regla de decisión clasifica de forma errónea varias observaciones. Tal vez nuestro modelo es muy rígido.

Generación de los datos

Consideremos dos posibles escenarios que hayan generado nuestro conjunto de datos:

- escenario 1: los datos de entrenamiento de cada clase fueron generados usando distribuciones normales bivariadas con medias diferentes y coordenadas no correlacionadas.
- escenario 2: cada clase proviene de una mezcla de 10 distribuciones normales de baja varianza, cuyas medias se distribuyen a su vez de forma normal.

Métodos de vecinos más cercanos

K-nearest-neighbours

Los métodos de vecinos más cercanos usan aquellas observaciones del conjunto de entrenamiento $\mathcal{T}$ más cercanas a x para construir su predicción $\widehat{y}$:

$$\widehat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

donde $N_k(x)$ es un entorno de x definido por los k puntos **más cercanos** de la muestra.

Métricas

La cercanía implica una **métrica**, por ejemplo la distancia euclídea. Otras métricas son la distancia Manhattan, Mahalanobis, Coseno, Levenshtein, Hamming, ...

Métodos de vecinos más cercanos

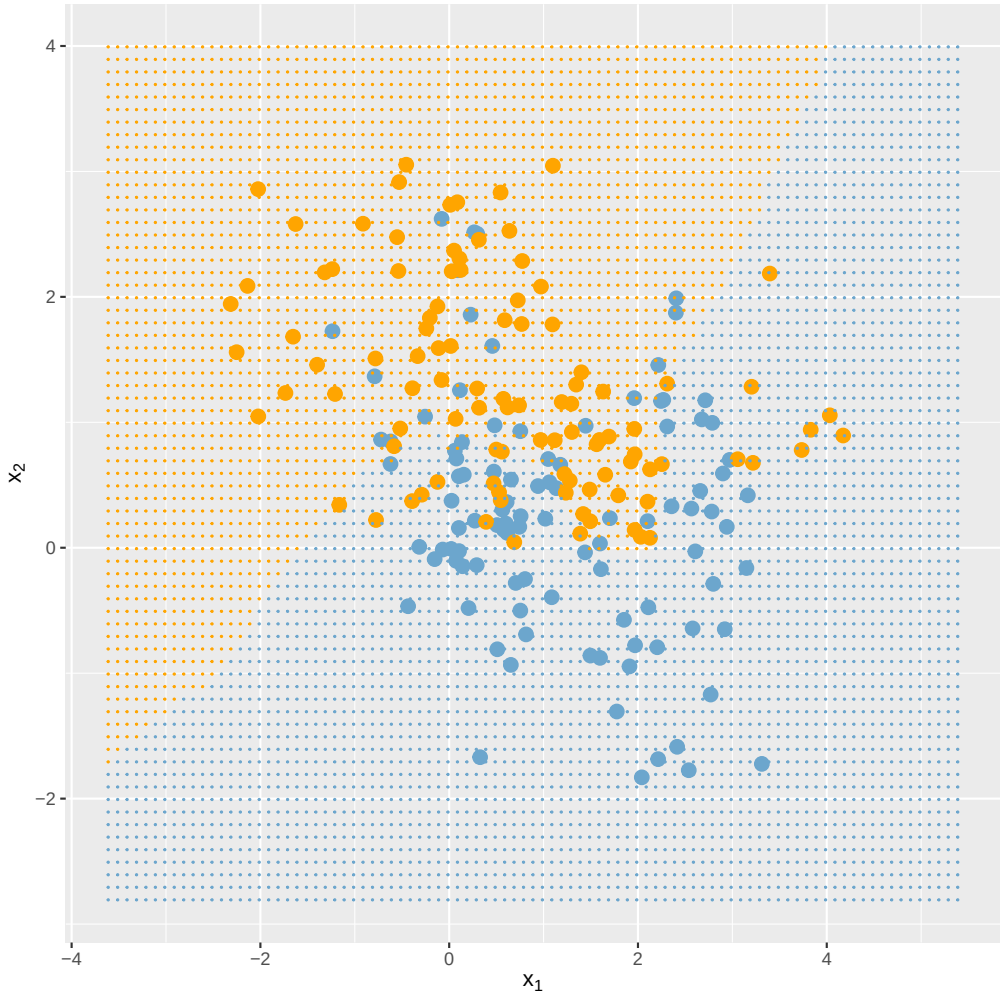


Figura: $k = 25$.

- La predicción $\widehat{y}$ es la proporción de NARANJA en la vecindad. Asignar la clase NARANJA a $\widehat{g}$ si $\widehat{y} > 0{,}5$ es equivalente al voto de la mayoría en el entorno.
- La frontera de decisión es más irregular y responde a grupos locales donde una clase domina.

Métodos de vecinos más cercanos

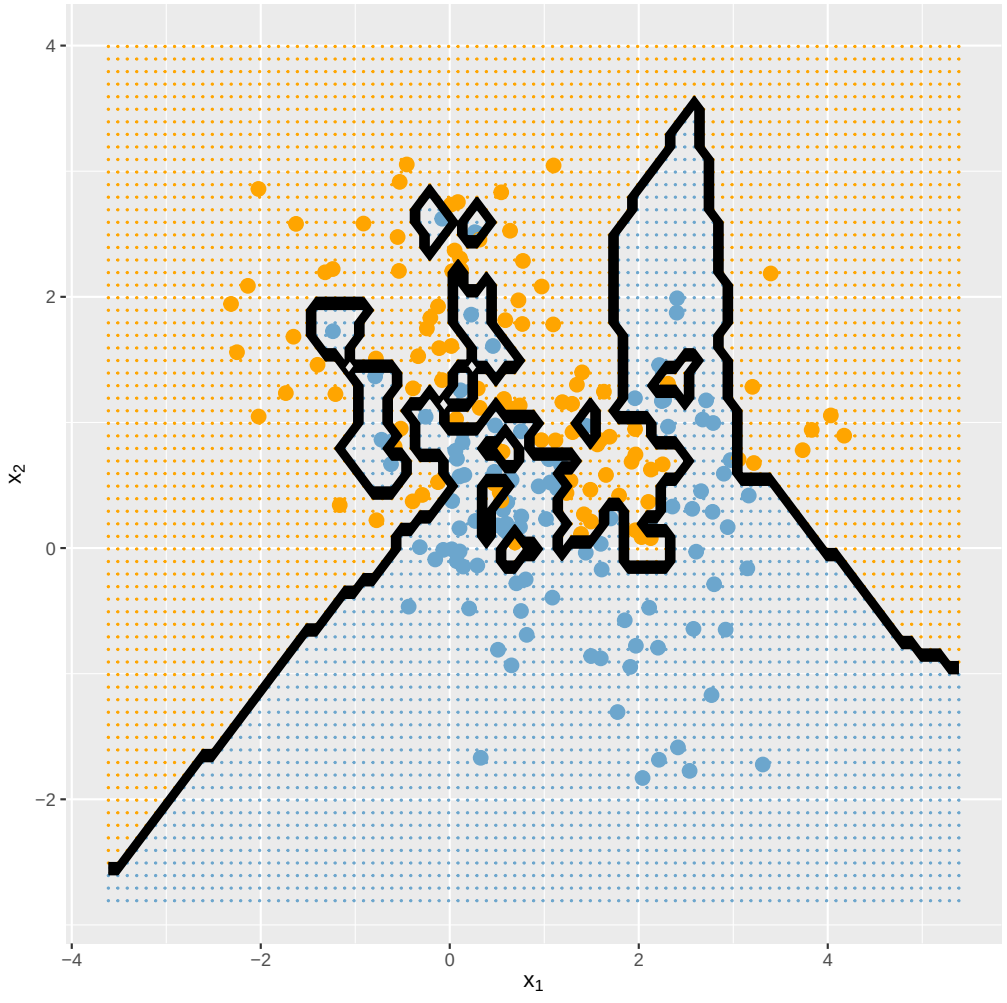


Figura: $k = 1$.

- La predicción $\widehat{y}$ es la clase de la observación más cercana a x en el conjunto de entrenamiento.
- En este caso, las regiones de clasificación corresponden a la *Mosaico (tessellation) de Voronoi* del conjunto de entrenamiento.
- La frontera de decisión es incluso más irregular.

Métodos de vecinos más cercanos

- Para este conjunto de datos, el método de vecinos más cercanos clasifica de manera incorrecta menos observaciones que el ajuste lineal.
- El error en el conjunto de entrenamiento parecería aumentar con k y siempre es 0 cuando $k = 1$.
- Un conjunto de test independiente sería más satisfactorio para comparar entre diferentes métodos.
- El ajuste lineal tiene 3 (p) parámetros.

Métodos de vecinos más cercanos

- Para este conjunto de datos, el método de vecinos más cercanos clasifica de manera incorrecta menos observaciones que el ajuste lineal.
- El error en el conjunto de entrenamiento parecería aumentar con k y siempre es 0 cuando $k = 1$.
- Un conjunto de test independiente sería más satisfactorio para comparar entre diferentes métodos.
- El ajuste lineal tiene 3 (p) parámetros.
- El método de vecinos tiene N/k parámetros *efectivos*, generalmente mayor a p. Si las clases fueran disjuntas, el método de k vecinos ajustaría un parámetro por vecindad (N/k vecindades).
- Parece más apropiado para el *escenario 2*, mientras que en el *escenario 1* la frontera de decisión sería innecesariamente ruidosa.

Los datos

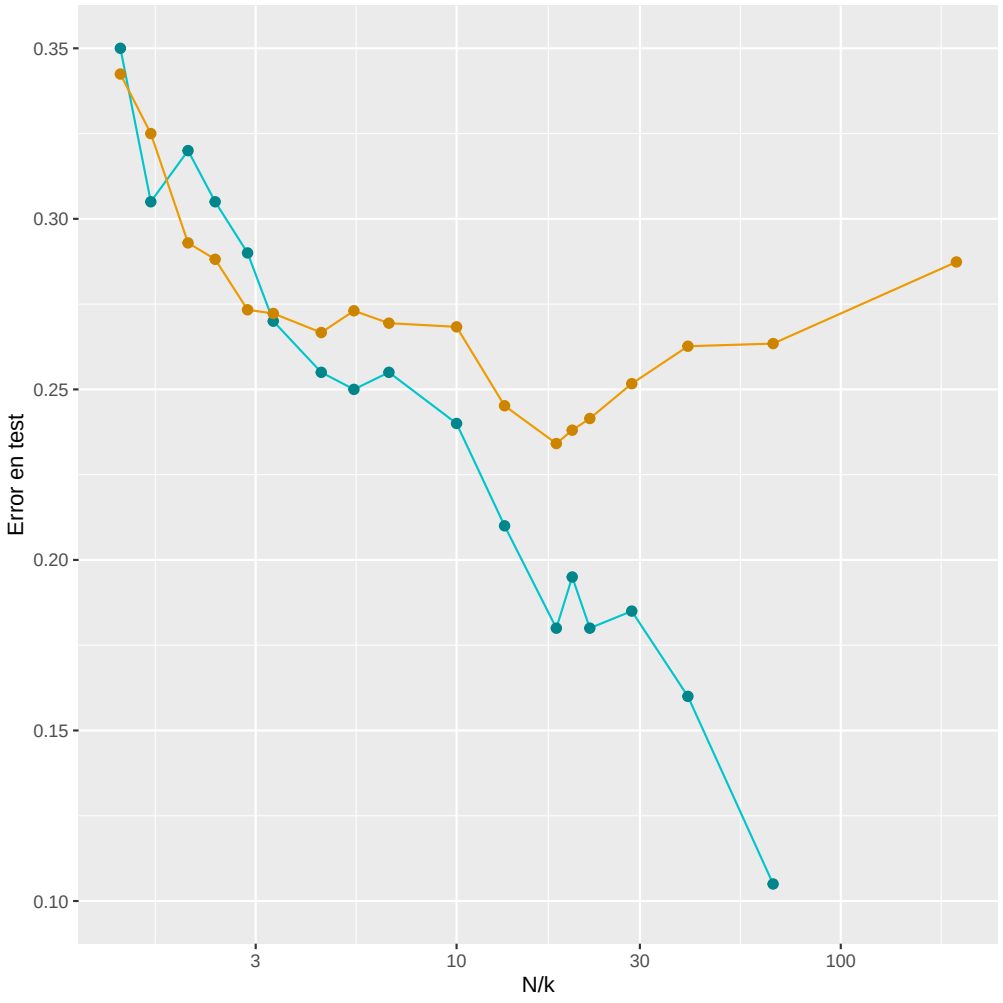
- La frontera de decisión del métodos de cuadrados mínimos es suave y *estable*. Se apoya fuertemente en el supuesto de que una regla de decisión lineal es apropiada. Veremos que tiene poca varianza y mucho sesgo.
- Por otro lado, los procedimientos de k-vecinos más cercanos no siguen fuertes supuestos sobre los datos y pueden adaptarse a diferentes situaciones. Sin embargo, una región en particular de la frontera de decisión depende de un algunas observaciones, es más inestable. Poco sesgo, mucha varianza.

Los datos

Los datos fueron generados a partir de un modelo intermedio más cercano al *escenario 2*:

1. Primero se generaron 10 medias m_k a partir de una distribución normal bivariada $\mathcal{N}((1,0)^t, I)$ y la etiquetamos `AZUL`. Similarmente, otras 10 fueron generadas a partir de $\mathcal{N}((0,1)^t, I)$ y etiquetadas como `NARANJA`.
2. En cada clase generamos 100 observaciones a partir de una mezcla de normales: para cada observación elegimos una media m_k al azar (con probabilidad $1/10$) y generamos $\mathcal{N}(m_k, I/5)$.

Test



- Clasificamos 10000 observaciones nuevas generadas con el procedimiento descripto.
- Comparamos el error de clasificación del método de cuadrados mínimos con vecinos más cercanos para diferentes valores de k.

Funciones de pérdida

- Sean $X \in \mathbb{R}^p$ un vector aleatorio de **covariables** (*features*) e $Y \in \mathbb{R}$ una variable aleatoria real de **respuesta** con distribución conjunta $F_{(X,Y)}$.
- Buscamos una función $f(X)$ para predecir los valores Y dados los valores X. Para penalizar los errores de predicción usamos una **función de pérdida** (*loss function*):

$$L(Y, f(X)) : \mathbb{R} \times \mathbb{R} \longrightarrow [0, \infty)$$

- La función es pérdida es una variable aleatoria porque depende de X. Definimos el **riesgo** mediante la esperanza:

$$R(Y, f(X)) = E(L(Y, f(X))$$

Regresión

Si consideramos la función de pérdida cuadrática

$$L(y, \widehat{y}) = (y - \widehat{y})^2$$

el riesgo es

$$R(f) = E(L(f(y, f(x)))) = E\left((y - f(x))^2\right)$$

Podemos buscar aquella función f que minimice el riesgo (bajo la pérdida cuadrática):

$$R(f) = E\left((y - f(x))^2\right) = E_X E_{Y|X}\left((Y - f(x))^2 \,|X = x\right)$$

Basta con minimizar puntualmente el esperanza condicional:

$$f(x) = argmin_c E_{Y|X}\left((Y - c)^2 \,|X = x\right) = E(Y|X = x)$$

Regresión

La función que minimiza el riesgo es la **función de regresión**.

$$f(x) = E(Y|X = x)$$

Lo mejor que podemos hacer, bajo la pérdida cuadrática, para predecir Y conociendo $X = x$, es calcular la media condicional.

El método de vecinos intenta esta receta directamente:

$$\widehat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

Hay dos aproximaciones,

- esperanzas por promedios y
- condicionar a un punto por una región.

Bajo ciertas condiciones, $\widehat{f}(x) \to E(Y|X = x)$ para $N, k \to \infty$ tal que $k/N \to \infty$.

Regresión

El ajuste lineal **modela** la función de regresión. Es decir, impone una forma funcional, lineal en sus argumentos:

$$f(x) \approx x^t \beta$$

El vector de parámetros β óptimo es

$$\beta = E(XX^t)^{-1} E(XY)$$

- Observemos que no condicionamos en X, la forma funcional nos permite usar la información de todos los valores de X.
- El método de cuadrados mínimos reemplaza la esperanza por promedios sobre el conjunto de entrenamiento.

Regresión

- El ajuste lineal supone que $f(x)$ se aproxima bien **globalmente** por una función lineal.
- kNN supone que $f(x)$ se aproxima bien **localmente** por una función constante.

Clasificación

Para el problema de clasificación cambiamos la función de pérdida para penalizar los errores de predicción (clase incorrecta)

$$L(\widehat{y}, y) = \mathbf{1}_{\{\widehat{y} \neq y\}} = \begin{cases} 1 & \text{si } \widehat{y} \neq y \\ 0 & \text{si } \widehat{y} = y \end{cases}$$

El riesgo resulta

$$\begin{aligned} R(f) &= E(L(f(X), Y)) = E(\mathbf{1}_{\{f(X) \neq Y\}}) \\ &= P(f(X) \neq Y) = P(\text{mala clasificación}) \end{aligned}$$

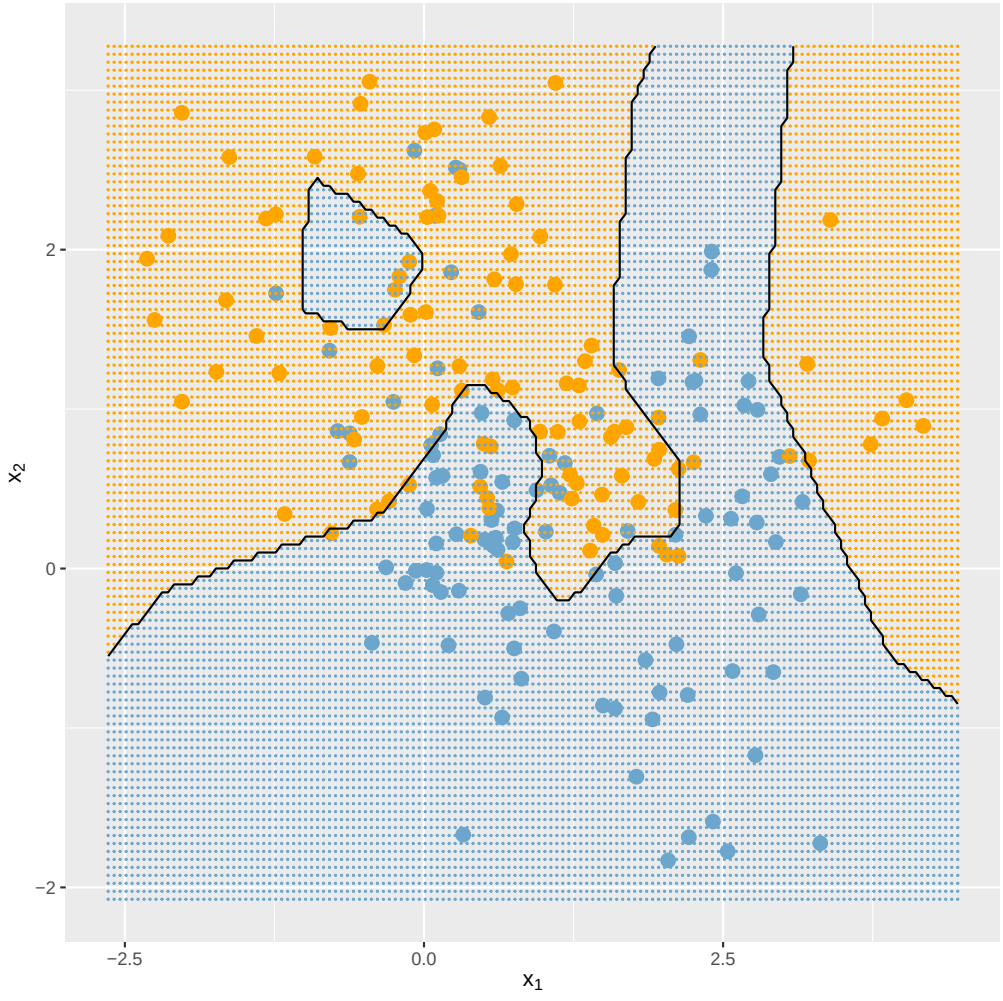
Podemos buscar la regla de clasificación f que minimice la probabilidad de mala clasificación:

$$E(\mathbf{1}_{\{f(X) \neq Y\}}) = E_X \left(\sum_{k=1}^{K} L(y, f(x)) P(y|X) \right)$$

Basta minimizar puntualmente la sumatoria:

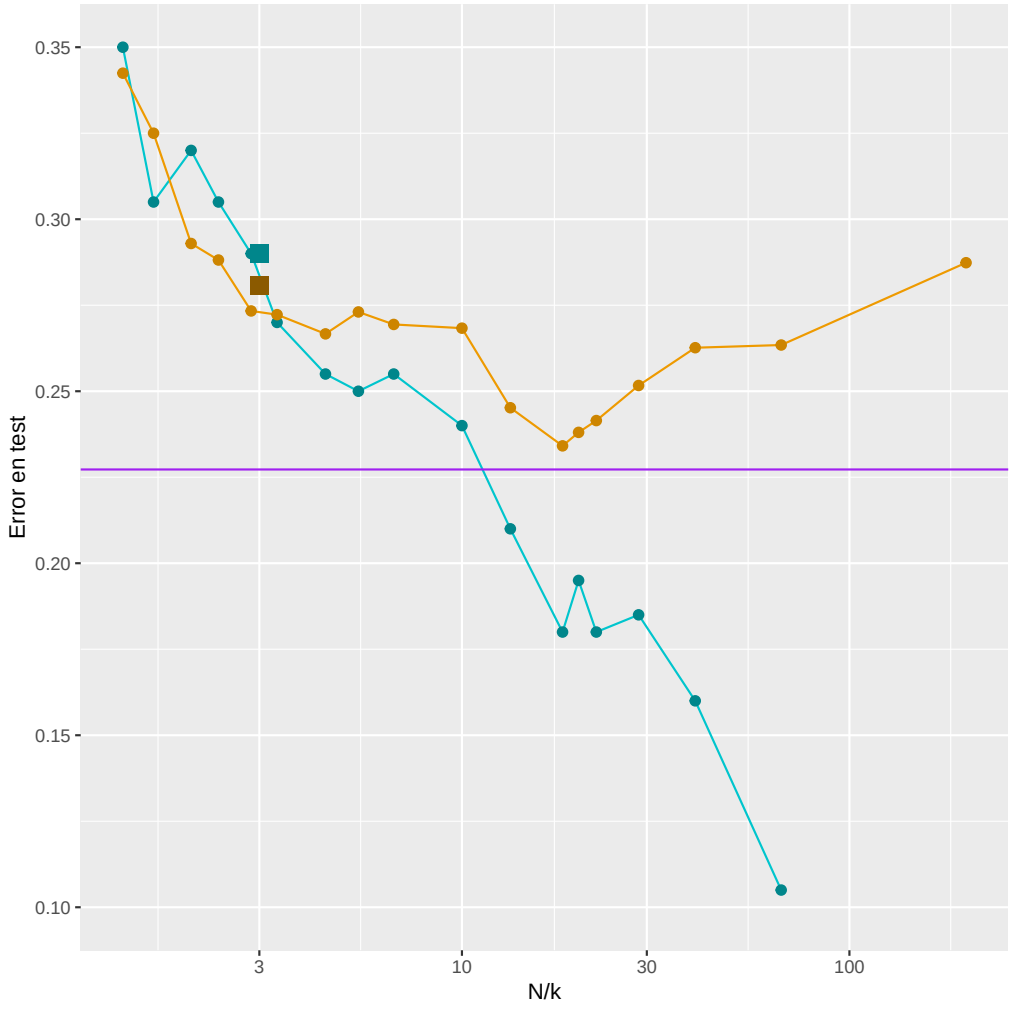
$$f(x) = argmin_g(1 - P(g|X = x)) \Rightarrow f(x) = \max_g P(g|X = x)$$

Clasificación



- Lo mejor que podemos hacer, **el clasificador de Bayes**, es asignar cada punto a la clase más probable según la distribución condicional $P(G|X)$.
- El error de clasificación asociado es el **error Bayes**.

Clasificación



Clasificación

- El método de vecinos más cercanos aproxima esta solución directamente. En un entorno los vecinos votan y gana la mayoría. La probabilidad condicional en un punto se relaja a una probabilidad condicional en un entorno del punto, y las probabilidades se estiman como proporciones de la muestra de entrenamiento.
- Si en un problema de dos clases usamos una variable de respuesta Y binaria entonces

$$\widehat{f} = E(Y|X) = P(G = G_1|X)$$

si G_1 corresponde a $Y = 1$. En la práctica, según el modelo de regresión utilizado, $\widehat{f}(X)$ podría no ser positiva o incluso mayor a 1.