IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

REGRESIÓN LOGÍSTICA
DECISION TREES

CLASE 22

CLASIFICACIÓN kNN: CONSIDERACIONES

- ✓ Método intuitivo y simple de entender e implementar.
- ✓ Funciona bien para clasificaciones binarias, o multiclase.
- ✓ Sólo tiene un hiperparámetro (k)

- ✗ El algoritmo es lento para grandes datasets.
- ✗ Funciona bien con pocas variables predictoras, pero falla para problemas de muchas dimensiones.
- ✗ Requiere normalizar los features para evitar problemas de escala.
- ✗ No funciona bien sobre datasets imbalanceados.

Aprendizaje Supervisado
Métodos de Clasificación

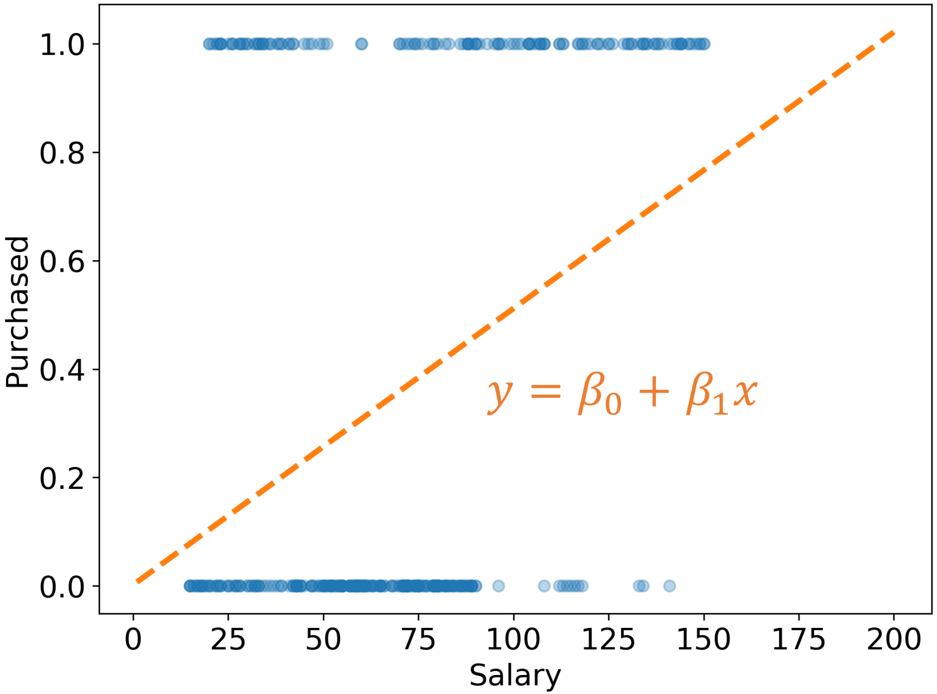
Regresión Logística Simple

CLASIFICACIÓN BINARIA

En el problema de comportamiento de clientes, la clasificación es **binaria:** hay sólo dos outcomes posibles ➔ compra / no compra (1/0).

¿Qué pasaría si tratamos de usar una **regresión lineal**?

- Los resultados para Y no son coherentes con la definición del outcome.
- Para valores extremos del predictor $\boldsymbol{x}$, se obtienen valores de Y fuera del rango [0,1].



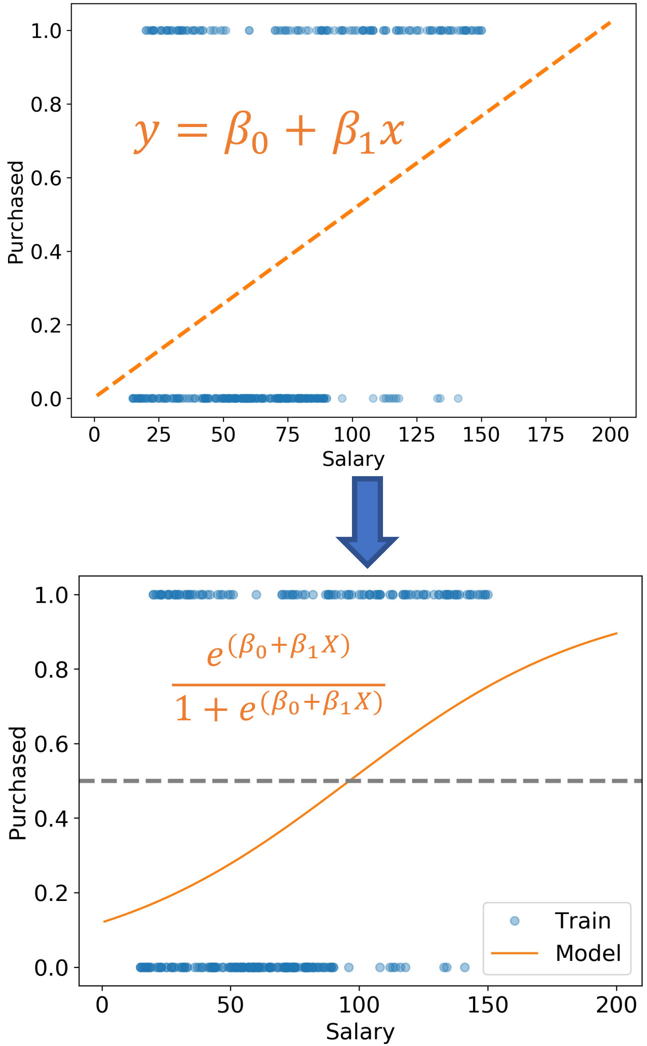
REGRESIÓN LOGÍSTICA SIMPLE

- La situación podría mejorar si tuviéramos una función que transforme la función lineal en algo más apropiado al problema.

- La **Regresión Logística Simple** permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa.

- Para ello, se **transforma el valor devuelto por la regresión lineal** empleando una función cuyo resultado está siempre comprendido **entre 0 y 1.**

- Con esto, resuelve los problemas de rango de predicciones y valores posibles de probabilidad del modelo de regresión lineal.



09:58

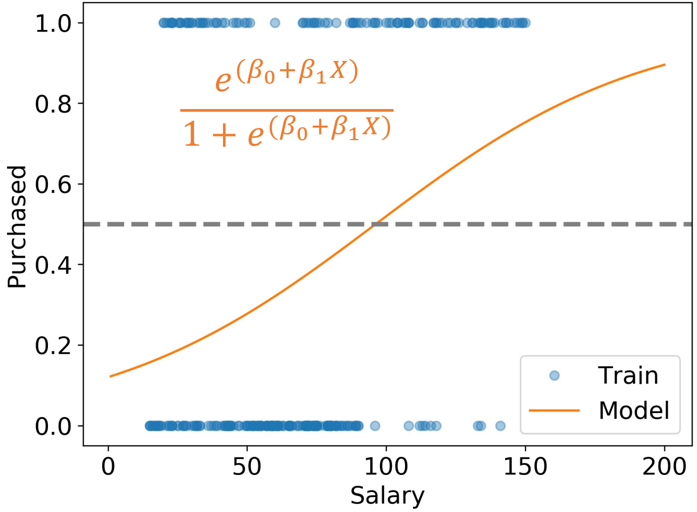
REGRESIÓN LOGÍSTICA SIMPLE

- La función logística más usada es la **sigmoide:**

$$P(Y = 1|X = x) = p = \frac{e^{(\beta_0+\beta_1 X)}}{1 + e^{(\beta_0+\beta_1 X)}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

- **Interpretación:** $P(Y = 1|X = x) = p$ es la probabilidad de que la variable cualitativa Y adquiera valor 1, dado que el predictor X tiene valor x.

- El modelo predice $P(Y = 1|X = x)$ con una curva en forma de S.

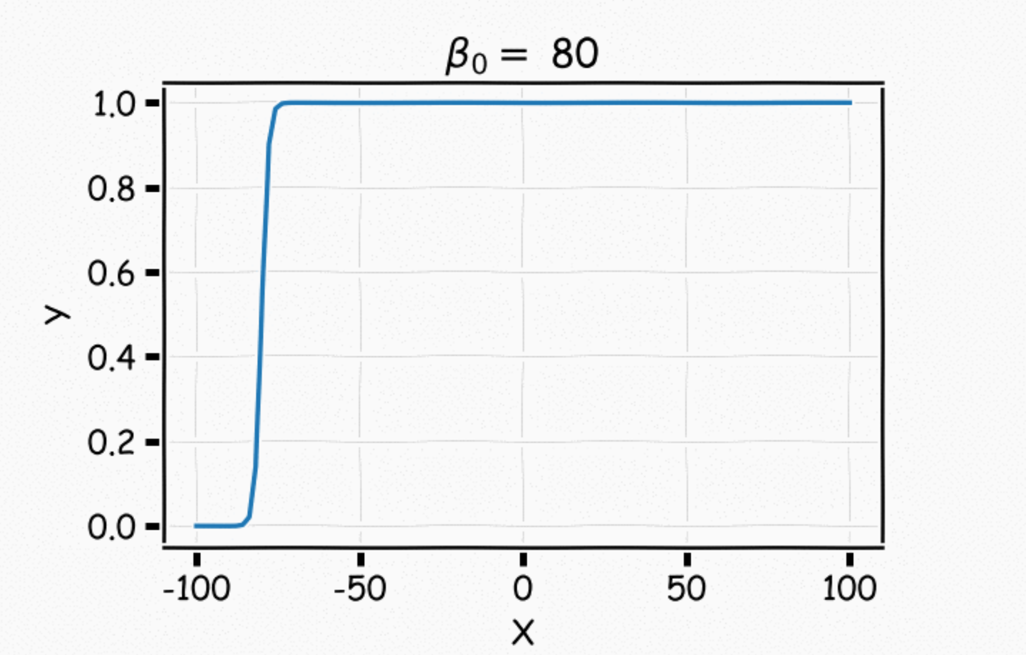


09:58

REGRESIÓN LOGÍSTICA SIMPLE

$$P(Y=1|X=x) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

β_0 mueve la curva hacia la izquierda o derecha en $\mathbf{c} = -\frac{\beta_0}{\beta_1}$.

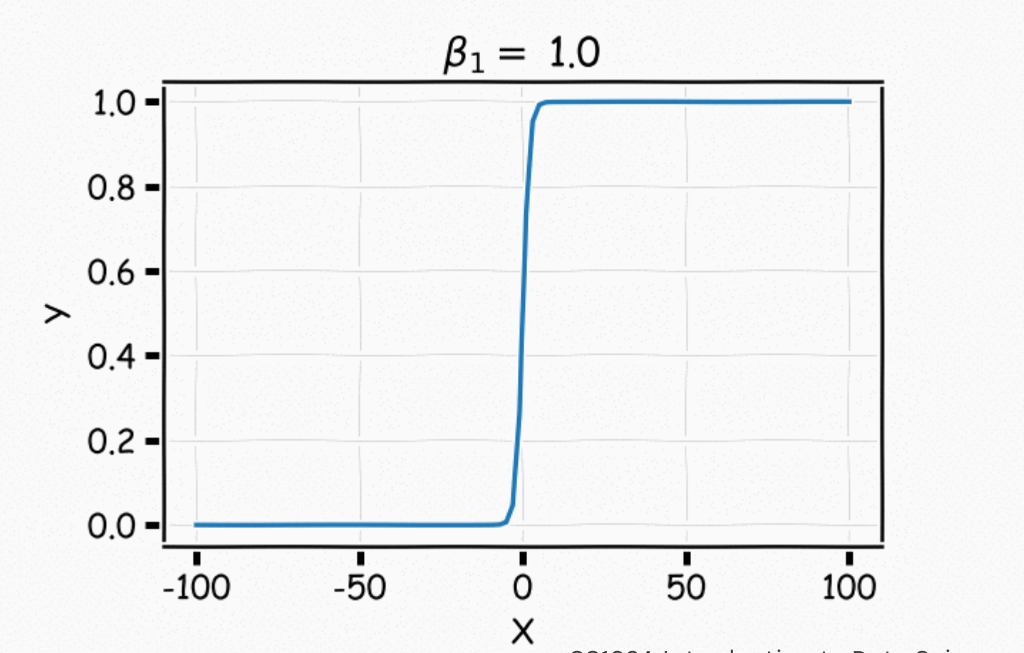


09:58

REGRESIÓN LOGÍSTICA SIMPLE

$$P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

β_1 controla qué tan empinada es la forma S (~pendiente)



09:58

REGRESIÓN LOGÍSTICA SIMPLE

$$P(Y=1|X=x)=p=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

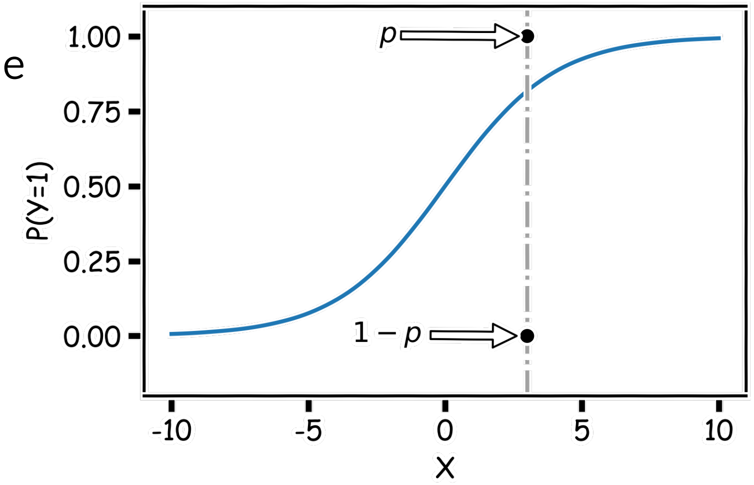
- Con un poco de álgebra, **el modelo logístico se puede reescribir como:**

$$\ln\boxed{\left(\frac{p}{1-p}\right)}=\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1 X$$

Odds: esto es lo que observamos

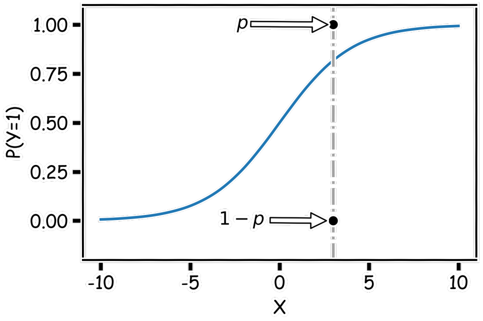
- ¿Cómo ajustamos los parámetros β a partir de un conjunto de observaciones?
  - Método de máxima verosimilitud:

$$L(y,\beta)=\sum_{i=1}^{n} y_i \ln(p_i)+(1-y_i)\ln(1-p_i),$$



REGRESIÓN LOGÍSTICA SIMPLE: MÁXIMA VEROSIMILITUD

- La verosimilitud L es la versión inversa de la probabilidad condicional ➔ $L(b|A) = P(A|B = b)$
- La variable Y es una variable aleatoria discreta que tiene probabilidad p de tomar valor 1 ("éxito"), y probabilidad $(p-1)$ de tomar valor 0 (fracaso) ➔ Distribución de Bernoulli
  - $P(Y = y) = \begin{cases} p & si\ y = 1 \\ 1-p & si\ y = 0 \end{cases}$
  - Esto también se puede escribir como: $P(Y = y_i|p) = p^{y_i}(1-p)^{1-y_i}$



Por lo tanto, la verosimilitud de una observación de p es:

- $L(p|y_i) = P(Y = y_i|p) = p^{y_i}(1-p)^{1-y_i}$

- Dado que las observaciones son independientes, la función de verosimilitud total para p es:
  - $L(p|Y) = \prod_i P(Y = y_i) = \prod_i p^{y_i}(1-p)^{1-y_i}$
  - Y su logaritmo: $l(p|Y) = \ln(L(p|Y) = \sum_i y_i\, ln(p_i) + (1-y_i)\ln(1-p_i)$

- Reemplazando p, obtenemos finalmente para la verosimilitud total:
  - $l(p|Y) = -\sum_i \left[y_i \log \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} + (1-y_i)\log\left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}\right)\right]_i$
- Para encontrar los valores óptimos de los parámetros β, es necesario maximizar esta expresión(derivar, igualar a cero, y resolver el sistema de ecuaciones mediante métodos iterativos).

09:58

REGRESIÓN LOGÍSTICA SIMPLE PARA CLASIFICACIÓN

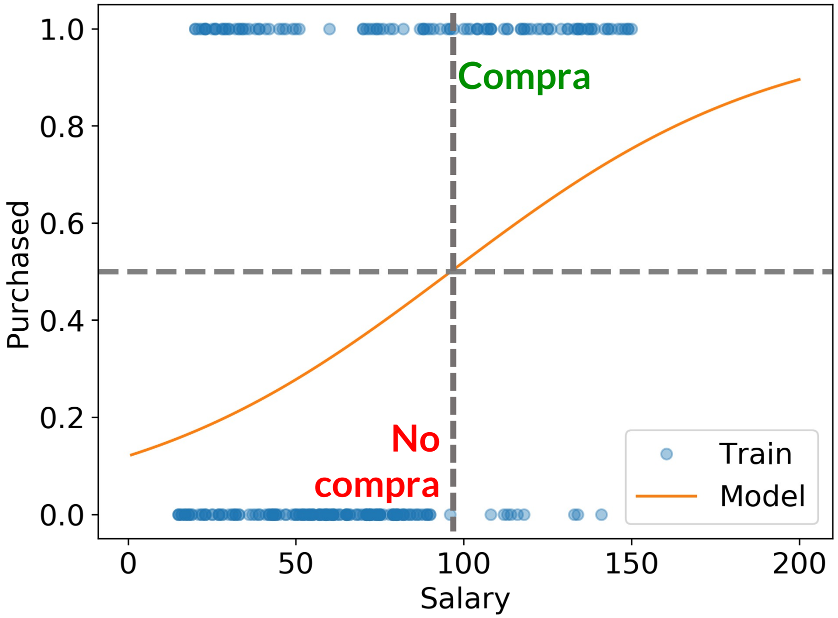
$$P(Y = 1|X = x) = p = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

- Para una nueva observación $X = x$, el modelo de regresión logística entrega la probabilidad $P(Y = 1|X = x)$.

- Para clasificar nuevas observaciones asumimos:

  - Si $P(Y = 1|X = x) \geq 0.5 \rightarrow Y = 1$
  - Si $P(Y = 1|X = x) < 0.5 \rightarrow Y = 1$

- El rendimiento del clasificador se evalúa usando las métricas comunes de clasificación: exactitud, precisión, sensibilidad, F_1



REGRESIÓN LOGÍSTICA MÚLTIPLE

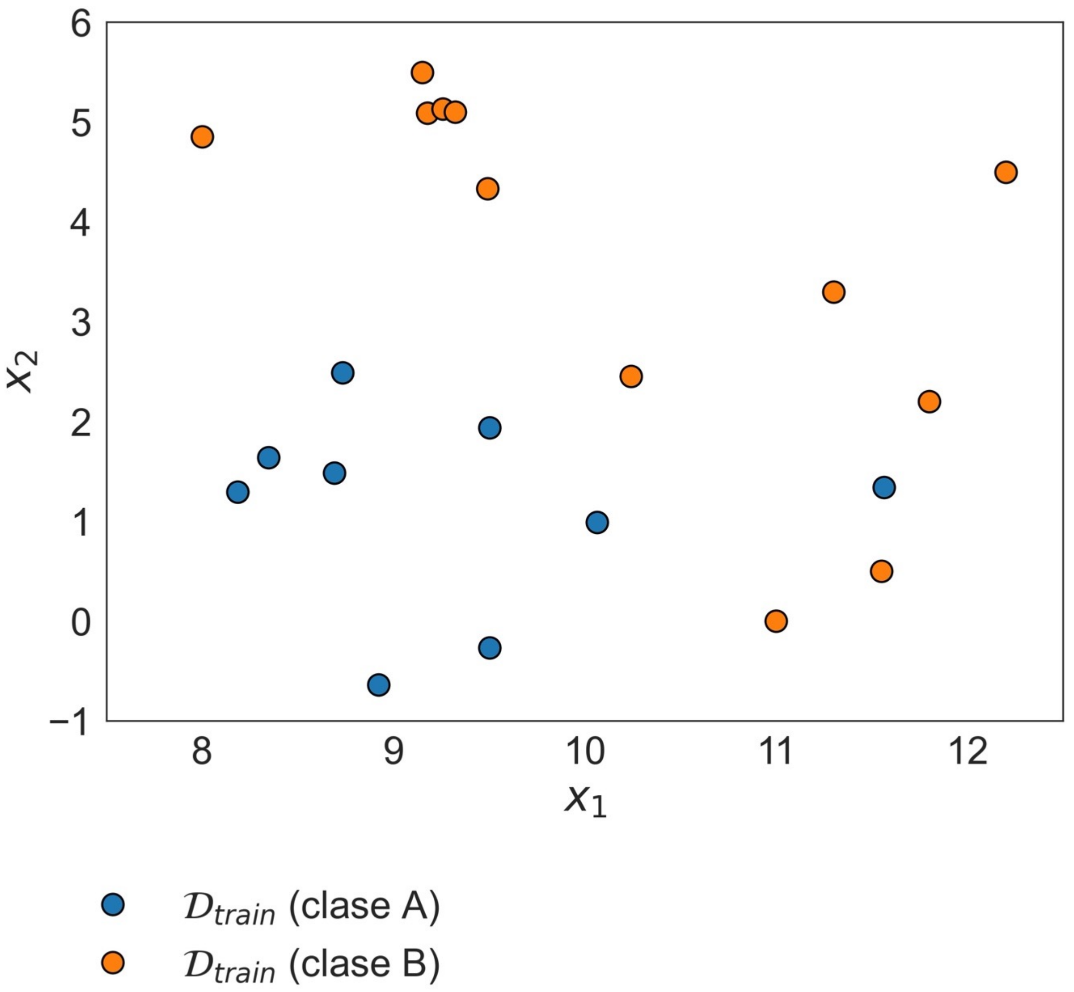
- La regresión logística múltiple es una extensión de la regresión logística simple.
- Se basa en los mismos principios que la regresión logística simple pero ampliando el número de predictores.
- Los predictores pueden ser tanto continuos como categóricos:

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_J X_J$$

Aprendizaje Supervisado
Métodos de Clasificación

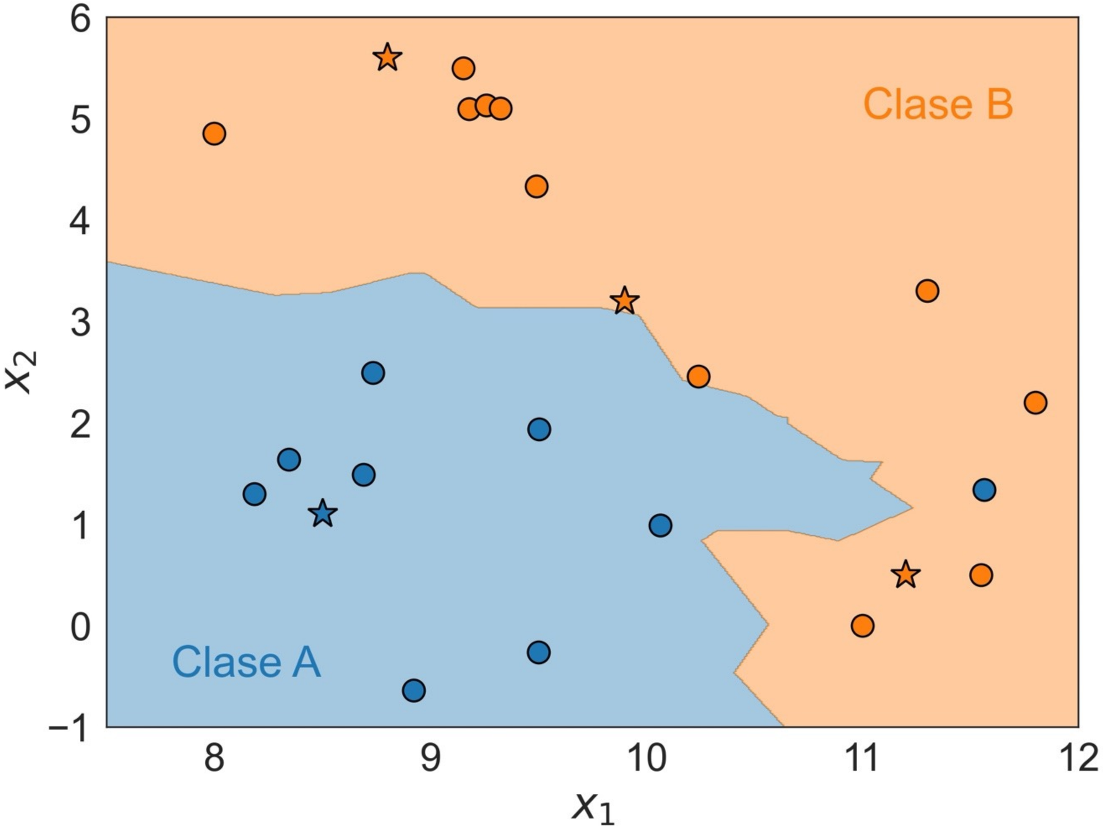
Deceision Trees / Árboles de Decisión

Ejemplo



09:58

Ejemplo



Clasificación kNN

(k=3)

ÁRBOLES DE DECISION / DECISION TREES

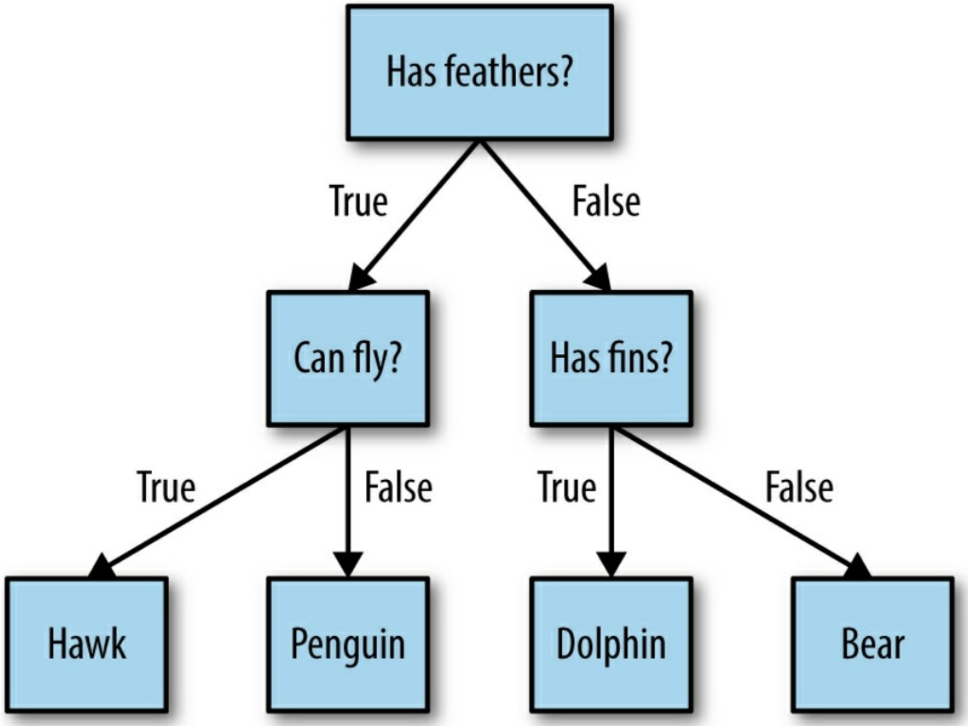
- Los árboles de decisión son modelos de **regresión y clasificación** que se basan en aprender una **jerarquía de preguntas (tests) de tipo if/else** para llegar de la forma más rápida posible a una decisión acertada respecto al valor de la etiqueta de una nueva observación.

ÁRBOLES DE DECISION / DECISION TREES

- Los árboles de decisión son modelos de **regresión y clasificación** que se basan en aprender una **jerarquía de preguntas (tests) de tipo if/else** para llegar de la forma más rápida posible a una decisión acertada respecto al valor de la etiqueta de una nueva observación.

- Las preguntas pueden ser relativas al valor de una variable numérica o categórica.

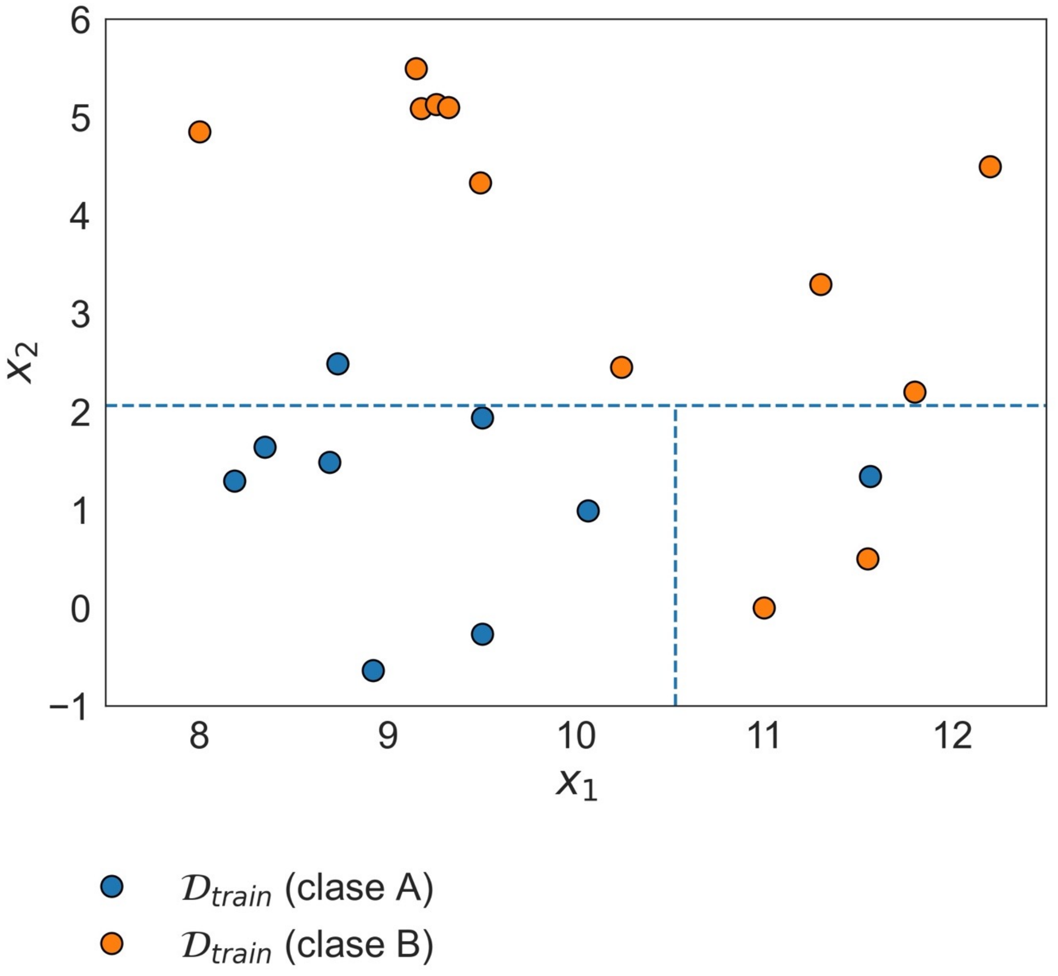
- El algoritmo busca sobre todas las posibles pruebas, y selecciona la que es **más informativa** respecto a la variable objetivo.



Ejemplo

- ¿es x_2 mayor a 2.06? y luego,

- ¿es x_1 menor a 10.5?

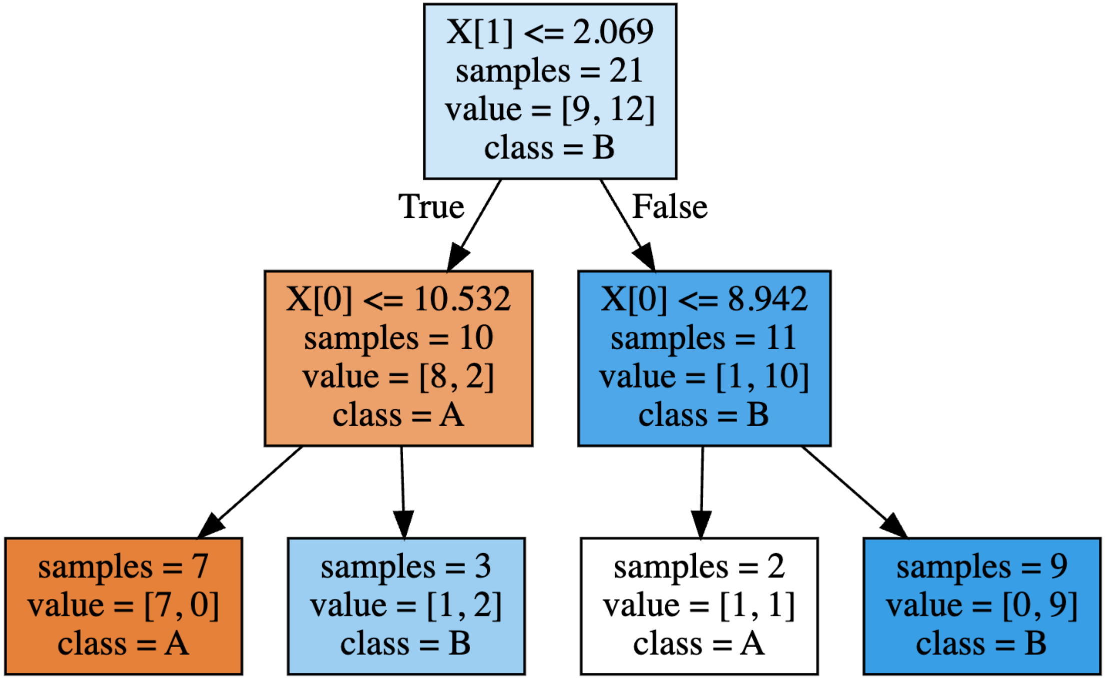


09:58

Ejemplo

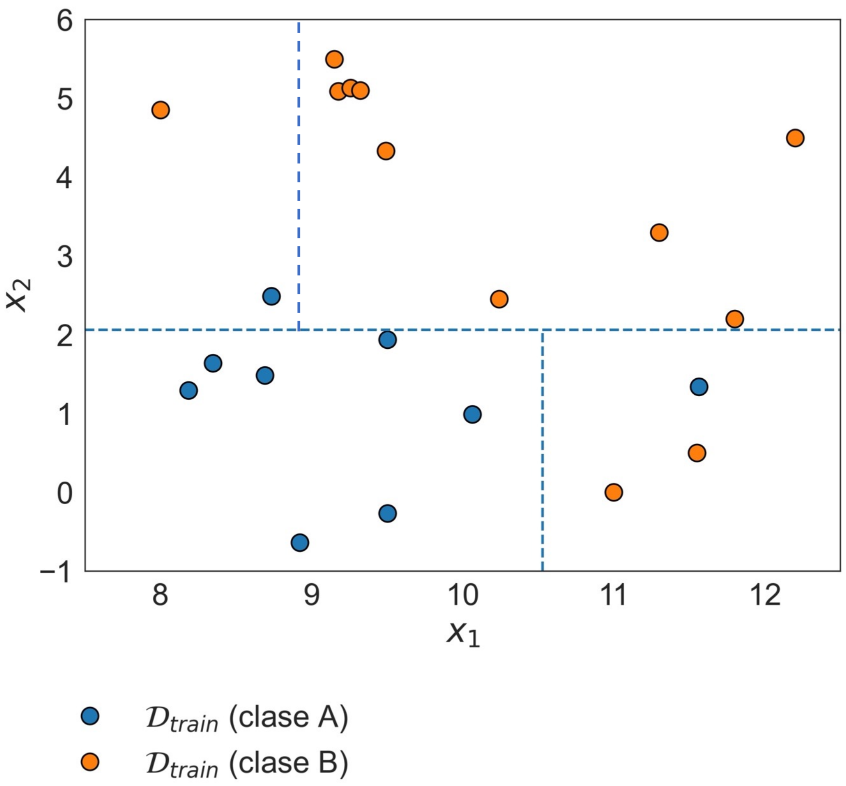
- Para los datos de ejemplo, un árbol de decisión para clasificar las observaciones en clases A o B podría tener la siguiente estructura de tres niveles.

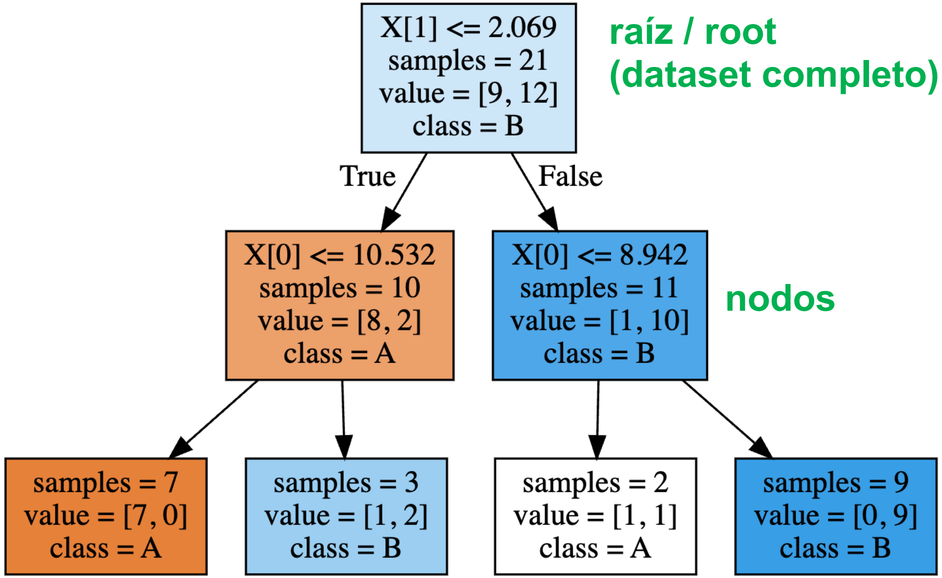
- Cada pregunta concierne **sólo a una variable**, por lo tanto cada pregunta particiona los datos a lo largo de un eje.



09:58

Ejemplo





raíz / root
(dataset completo)

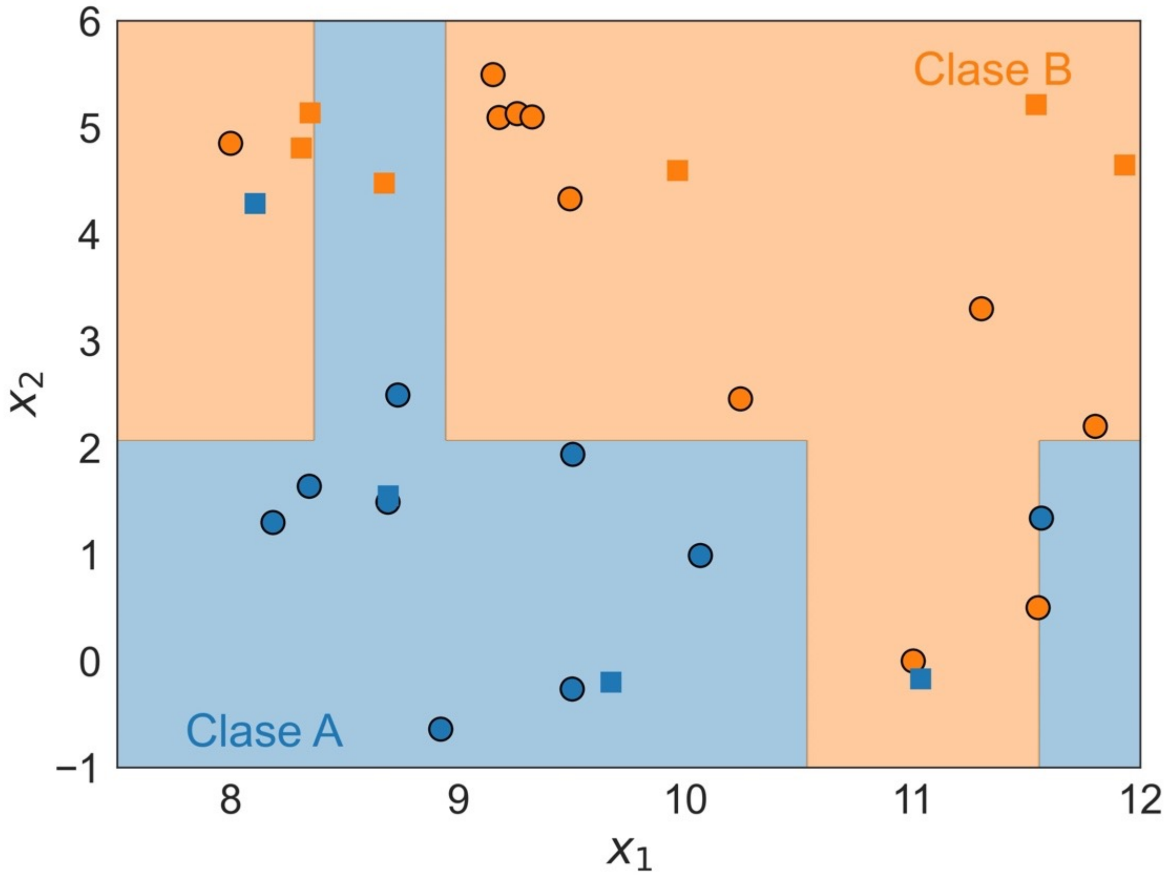
nodos

hoja/leaf

(región en la partición)

09:58

Ejemplo



Frontera de decisión en el plano x_1-x_2

(distinta a la obtenida con el clasificador kNN)

COMPLEJIDAD DEL ÁRBOL DE DECISIÓN

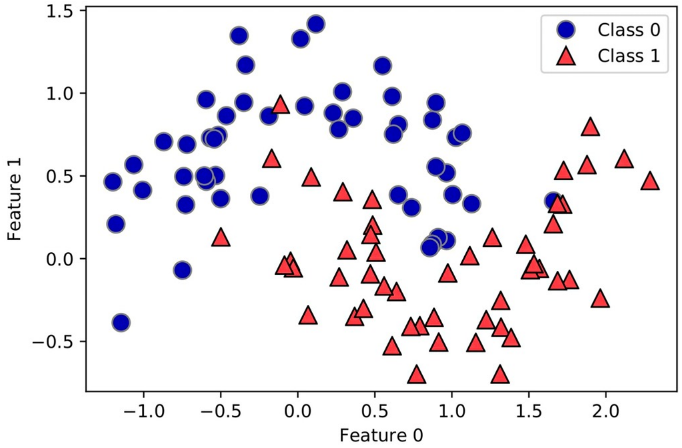
Típicamente, la construcción del árbol de decisión continua hasta que todas las hojas son **"puras": contienen sólo una clase objetivo.**

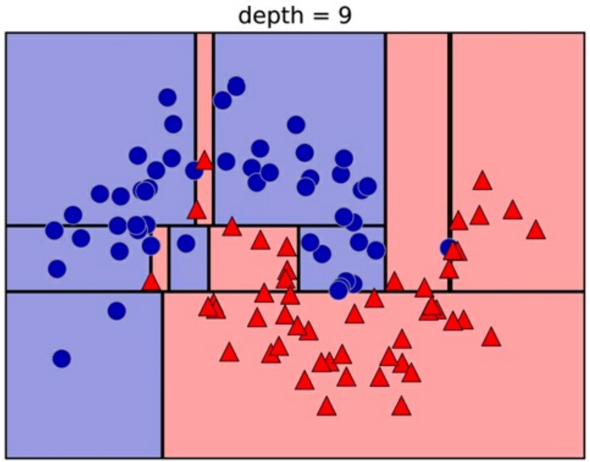
Esto puede llevar a modelos muy complejos ➔ **overfitting**.

- hojas puras ➔ 100% precisión para datos de entrenamiento

Estrategias para prevenir overfitting:

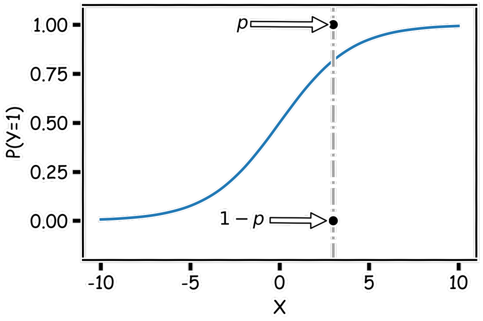
- **Pre-pruning:** cortar tempranamente la creación del árbol
  - Limitar profundidad
  - Limitar nº de hojas
  - Requerir mínimo de puntos en un nodo para dividir
- **Post-pruning:** crear el árbol completo, y colapsar nodos que aportan poca información





REGRESIÓN LOGÍSTICA SIMPLE: MÁXIMA VEROSIMILITUD

- La verosimilitud L es la versión inversa de la probabilidad condicional ➔ $L(b|A) = P(A|B = b)$
- La variable Y es una variable aleatoria discreta que tiene probabilidad p de tomar valor 1 ("éxito"), y probabilidad $(p - 1)$ de tomar valor 0 (fracaso) ➔ Distribución de Bernoulli
  - $P(Y = y) = \begin{cases} p & si\ y = 1 \\ 1 - p & si\ y = 0 \end{cases}$
  - Esto también se puede escribir como: $P(Y = y_i|p) = p^{y_i}(1 - p)^{1-y_i}$



Por lo tanto, la verosimilitud de una observación de p es:

  - $L(p|y_i) = P(Y = y_i|p) = p^{y_i}(1 - p)^{1-y_i}$

- Dado que las observaciones son independientes, la función de verosimilitud total para p es:
  - $L(p|Y) = \prod_i P(Y = y_i) = \prod_i p^{y_i}(1 - p)^{1-y_i}$
  - Y su logaritmo: $l(p|Y) = \ln(L(p|Y) = \sum_i y_i\, ln(p_i) + (1 - y_i)\ln(1 - p_i)$

- Reemplazando p, obtenemos finalmente para la verosimilitud total:
  - $l(p|Y) = -\sum_i \left[y_i \log \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} + (1 - y_i) \log\left(1 - \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}}\right)\right]_i$
- Para encontrar los valores óptimos de los parámetros β, es necesario maximizar esta expresión(derivar, igualar a cero, y resolver el sistema de ecuaciones mediante métodos iterativos).