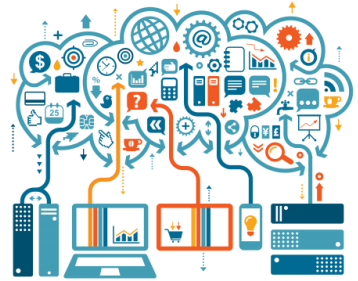


Minería de datos

Ing. Luis Francisco López

Escuela Colombiana de Ingeniería
Julio Garavito

Clasificación (IV)

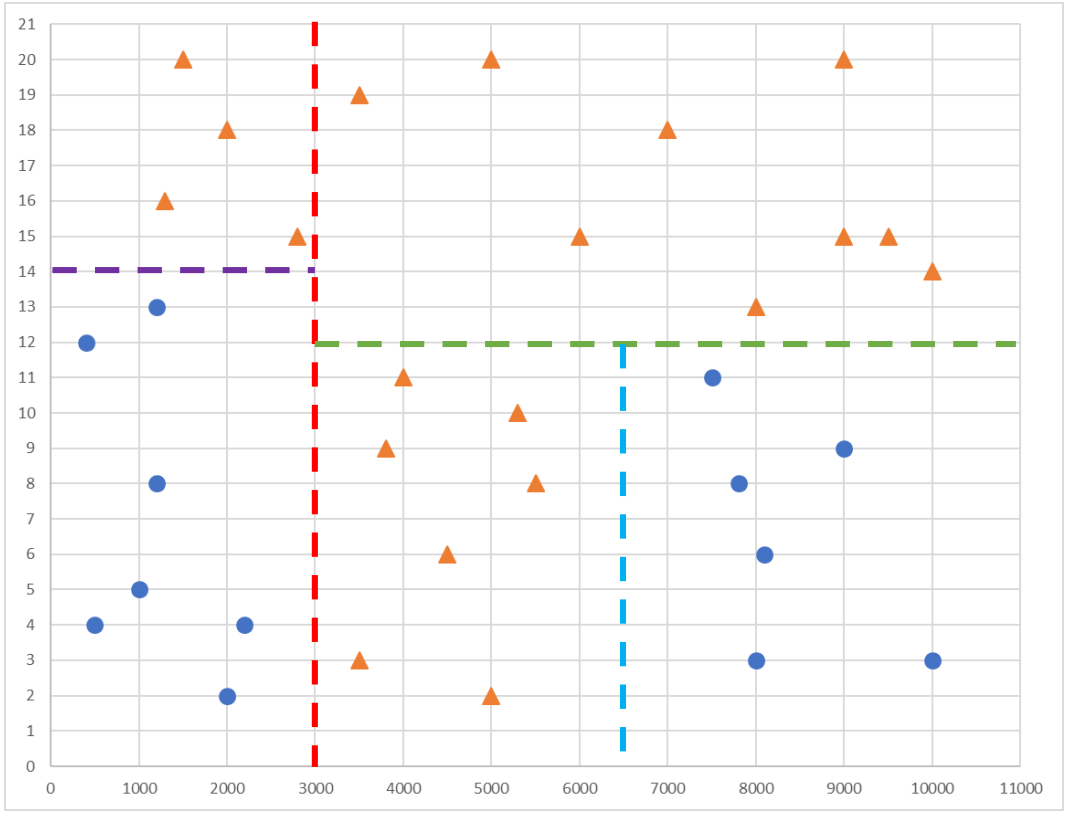
Árboles de decisión

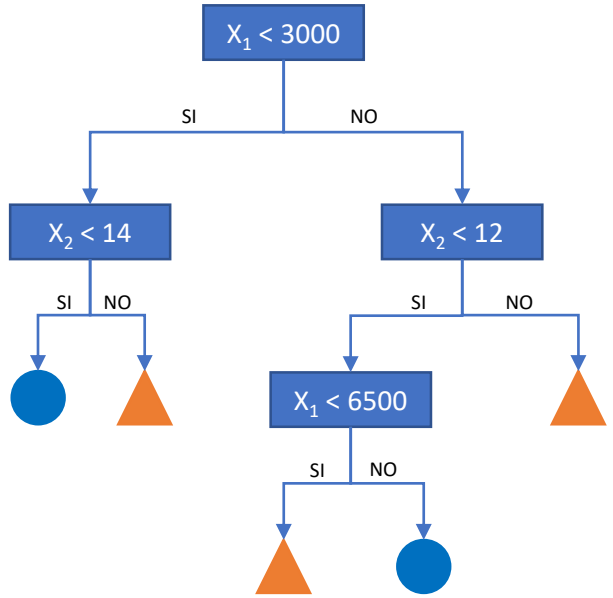
Decision Tree

Concepto

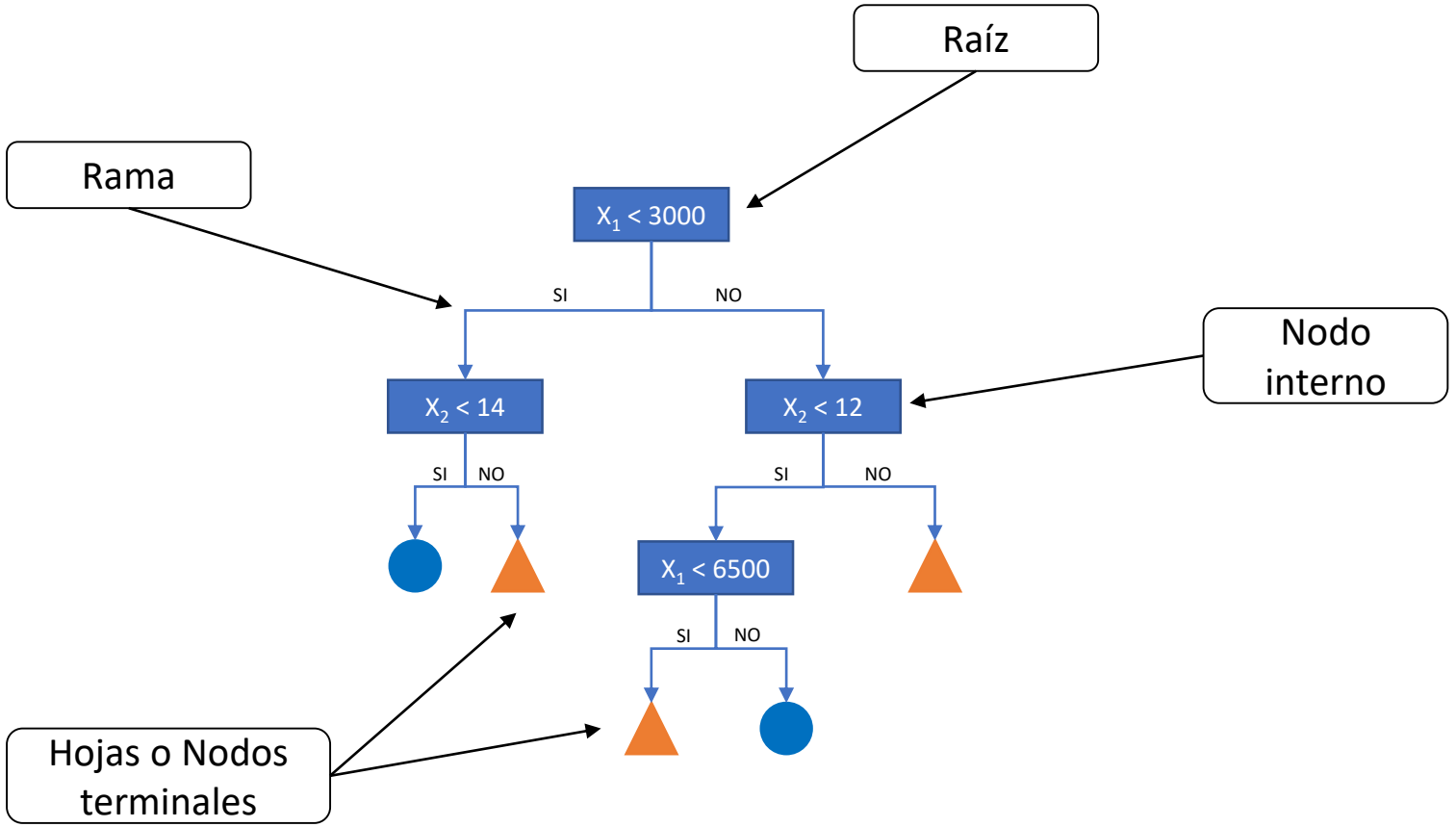
- Utiliza un árbol de decisión como modelo predictivo.
- Utiliza observaciones (o conjunciones) representadas en las ramas para obtener conclusiones sobre el valor objetivo (representado en las hojas).

Ejemplo


21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000


X1 < 3000
SI
NO
X2 < 14
X2 < 12
SI
NO
SI
NO
X1 < 6500
SI
NO

Terminología



Proceso

- El objetivo es encontrar las regiones $R_1, \ldots, R_M$ que minimizan alguna función particular, en este caso la clase predicha será la clase más común en la región (moda).
- $\hat{p}_{mk}$ es la proporción de observaciones de la m-ésima región que son de la k-ésima clase.
- Función de tasa de clasificación errónea:

$$E = \sum_{m=1}^{M} 1 - \max_{k} (\hat{p}_{mk})$$

Proceso

- Índice de Gini definido por:

$$G = \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right)$$

Es una medida de la varianza total de las K clases. Toma valores pequeños si todas las $\hat{p}_{mk}$ tienen valores cercanos a cero o uno.

Proceso

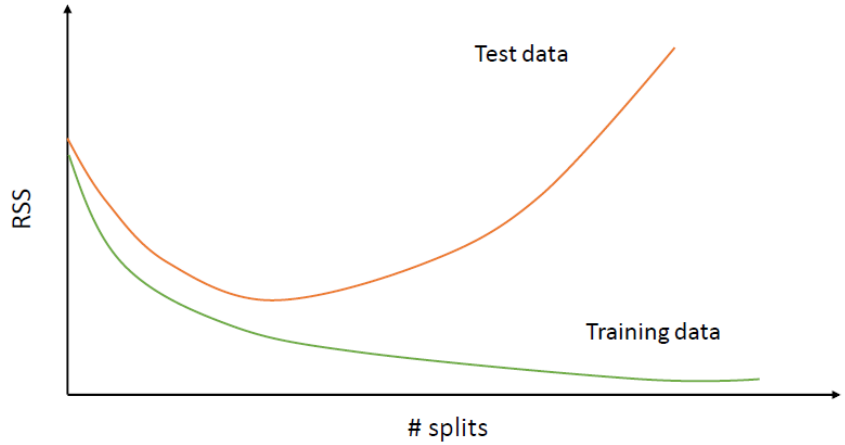
- Ganancia de información:

$$D = -\sum_{m=1}^{M}\sum_{k=1}^{K}\hat{p}_{mk}\log\left(\hat{p}_{mk}\right)$$

Se basa en el concepto de entropía de la teoría de la información.

Sobreajuste (Overfitting)

- Un número muy grande de divisiones puede llevar a sobreajuste que llevan a malas predicciones.



Fuente: ECI - International Summer School/Machine Learning - Dr Ivan Olier (2019)

- Una estrategia es construir un árbol muy grande y luego “podarlo” para obtener un “sub-árbol”

Ventajas y desventajas

- Fáciles de explicar a las personas.
- Son cercanos al proceso de toma de decisiones.
- Pueden ser representados de forma gráfica.
- Pueden incluir variables cualitativas.
- Son modelos de “caja blanca”.
- Desafortunadamente NO tienen el mismo nivel de exactitud que otros métodos de clasificación.

Bosque aleatorio

Random forest

Reducción de la varianza

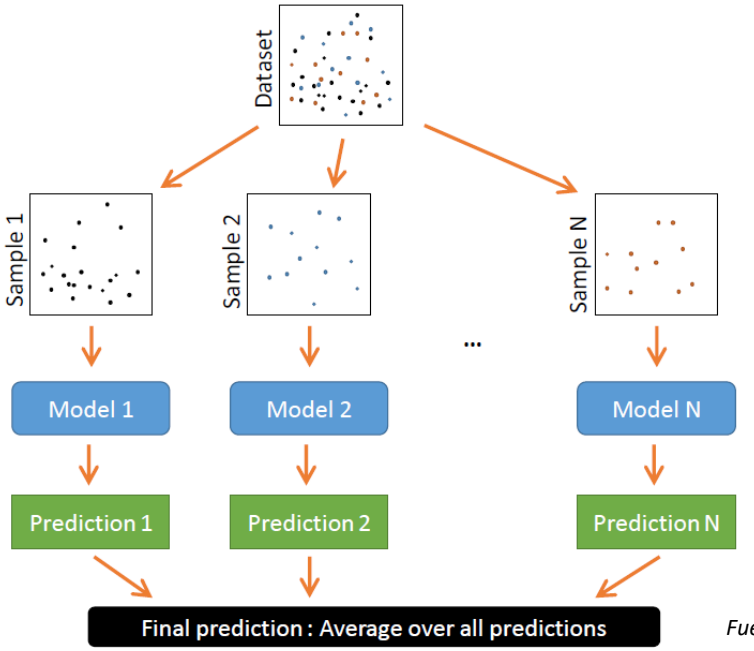
Una forma de reducir la varianza es tener más datos independientes.

Para un conjunto de observaciones independientes $Z_1, Z_2, \ldots, Z_n$ cada uno con varianza σ^2, la varianza de la media $\bar{Z}$ de las observaciones es σ^2/n.

En otras palabras, promediar un conjunto de observaciones reduce la varianza.

Bagging

Utiliza el mismo principio, pero en lugar de usar datos independientes, lo que hace es “re-muestrear” (bootstrap) los datos de entrenamiento.



Fuente: ECI - International Summer School/Machine Learning - Dr Ivan Olier (2019)

Bosques aleatorios

Similares al “bagging” pero utilizan una estrategia para reducir la varianza aún más:

- Se construye un número de árboles de decisión haciendo muestreo (bootstrap) de los datos de entrenamiento.
- Cuando se construyen los árboles, cada vez que se considera una división en el árbol, se selecciona un subconjunto aleatorio de las variables predictoras.
- Reduce el sobreajuste de los árboles de decisión.