Unidad 3: Análisis utilizando aprendizaje supervisado
3.1: Árboles de decisión

Docente: Pablo Torres Tramón[1,2]

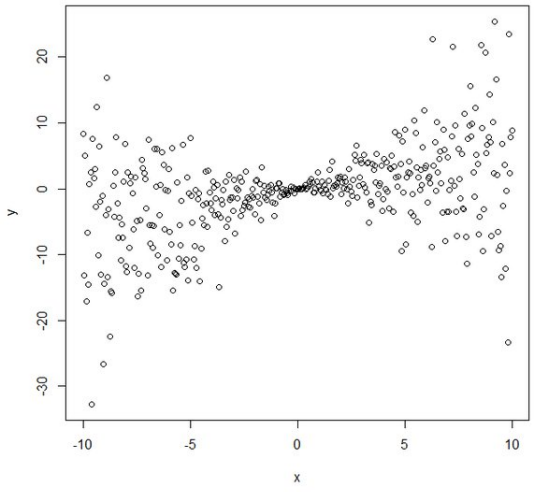
[1]Facultad de Ingeniería
t.pabloandrestorres@uandresbello.edu
[2]Ponencias originales elaboradas por:
Mailiu Díaz Peña y Alejandro Figueroa

Minería de Datos
Otoño 2023

Índice

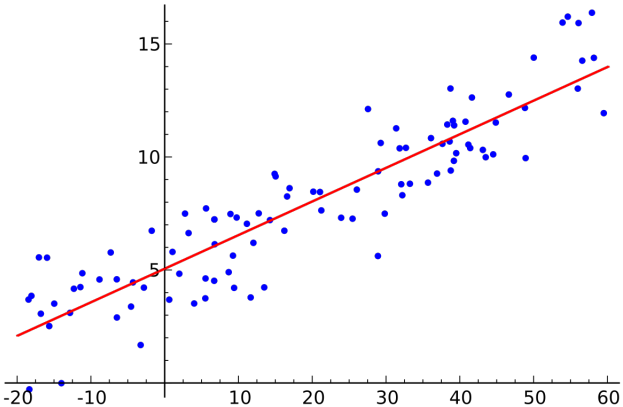
Table of Contents

Motivación: Problemas del modelo lineal



Motivación

Necesitamos normalizar y estándarizar



Motivación

Es posible usar modelo no lineal que capture mejor la data

- ¿Qué es una función no lineal?

Motivación

Es posible usar modelo no lineal que capture mejor la data

- ¿Qué es una función no lineal?
- Modelos lineales pueden aprender usando el gradiente descendiente

Motivación

Es posible usar modelo no lineal que capture mejor la data

- ¿Qué es una función no lineal?
- Modelos lineales pueden aprender usando el gradiente descendiente
- ¿Cómo aprender si usamos una función no lineal?

Motivación

Es posible usar modelo no lineal que capture mejor la data

- ¿Qué es una función no lineal?
- Modelos lineales pueden aprender usando el gradiente descendiente
- ¿Cómo aprender si usamos una función no lineal?
- ¿Podríamos usar variables cualitativas en estos modelos?

Árboles de decisión

¿Se debe realizar la cirugía ocular?

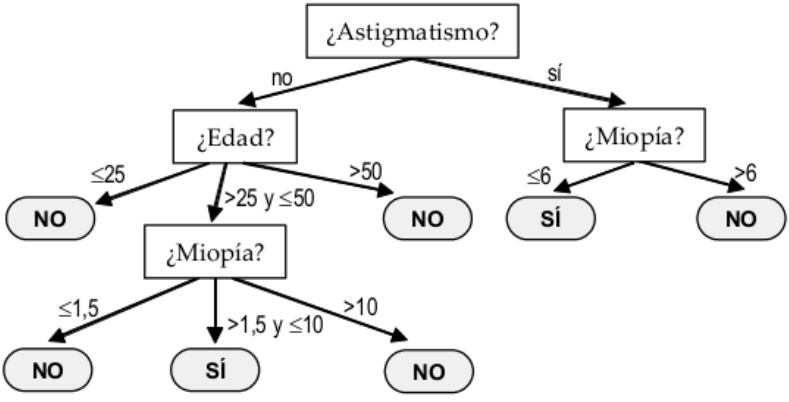


Table of Contents

Terminología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.
- **Hoja**: Nodos sin partición.

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.
- **Hoja**: Nodos sin partición.
- **Poda**: Reducción del árbol de decisión eliminando nodos (opuesto a la partición).

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.
- **Hoja**: Nodos sin partición.
- **Poda**: Reducción del árbol de decisión eliminando nodos (opuesto a la partición).
- **Rama**: Una subsección del árbol de decisión.

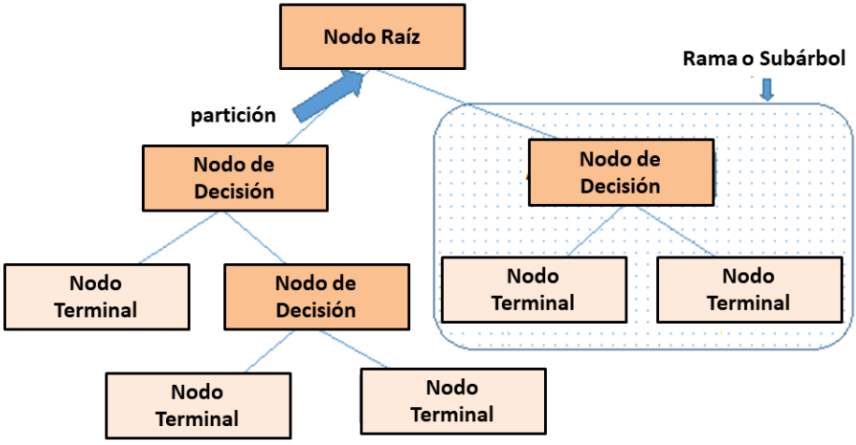
Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.
- **Hoja**: Nodos sin partición.
- **Poda**: Reducción del árbol de decisión eliminando nodos (opuesto a la partición).
- **Rama**: Una subsección del árbol de decisión.
- **Padre**: Nodo que tiene particiones.

Términología de un árbol de decisión

- **Nodo**: Conjunto de atributos sobre el cual dividir la población actual.
- **Raíz**: Representa a toda la población (data).
- **Arista**: División de la población actual de acuerdo a un valor fijo siguiendo el atributo del nodo antecesor.
- **Hoja**: Nodos sin partición.
- **Poda**: Reducción del árbol de decisión eliminando nodos (opuesto a la partición).
- **Rama**: Una subsección del árbol de decisión.
- **Padre**: Nodo que tiene particiones.
- **Hijo**: Nodo que es el resultado de una partición.

Estructura de un árbol de decisión [Ara21]



Cómo crear un árbol de decisión

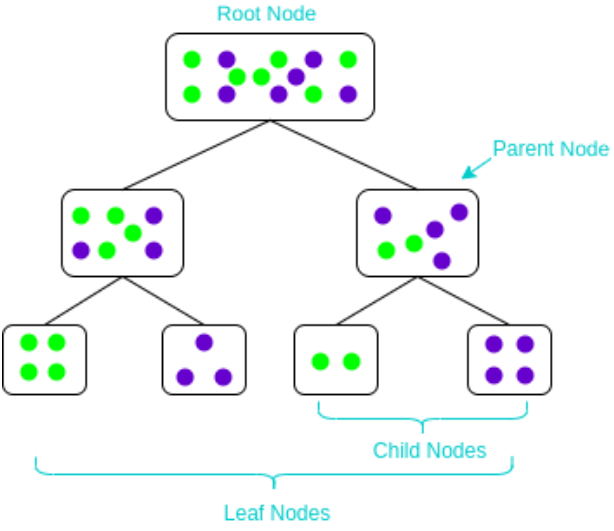
Objetivo

Elegir una partición que divida los datos en ejemplos clasificados correctamente. Esto implica identificar la característica que es más "útil" para la clasificación y luego derivar una regla de decisión utilizando dicha característica.

Algoritmo ID3 [Qui86]

1. Seleccione el **mejor atributo** para dividir.
2. Haga que ese atributo sea un nodo de decisión y divida el conjunto de datos en subconjuntos más pequeños.
3. Repita recursivamente para cada hijo hasta que se cumpla una de las siguientes condiciones:
    1. Todas las tuplas pertenecen al mismo valor de atributo.
    2. No quedan más atributos.
    3. No hay más instancias.
    4. Asigne de una clase específica de acuerdo a la evidencia de la rama.

Ejemplo algoritmo



Decisión para ramificación

- La decisión de hacer divisiones afecta **altamente** la precisión del árbol.
- Los criterios de decisión son diferentes para árboles de clasificación y regresión.
- Existen varios algoritmos para decidir si realizar o no la ramificación.
- La creación de subnodos incrementa la homogeneidad de los subnodos resultantes. Es decir, la pureza del nodo se incrementa respecto a la variable objetivo.
- Se prueba la división con todas las variables y se escoge la que produce sub-nodos más homogéneos.

Criterios de selección

Selección de atributos

Es una heurística (intuición) para seleccionar el criterio de división que divide los datos de la mejor manera posible.

Las medidas de selección más populares son:

- Ganancia de información: es una disminución de la entropía, propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con sus clasificación objetivo.
- Índice de Gini.
- Chi Cuadrado
- Reducción en la varianza

Table of Contents

Entropía

La entropía es una medida de aleatoriedad en un sistema. Para una distribución de probabilidad con $|Y|$ clases, se define como:

$$E(P) = -\sum_{i \in Y} Pr(P_i) \log_2 Pr(P_i) \quad (1)$$

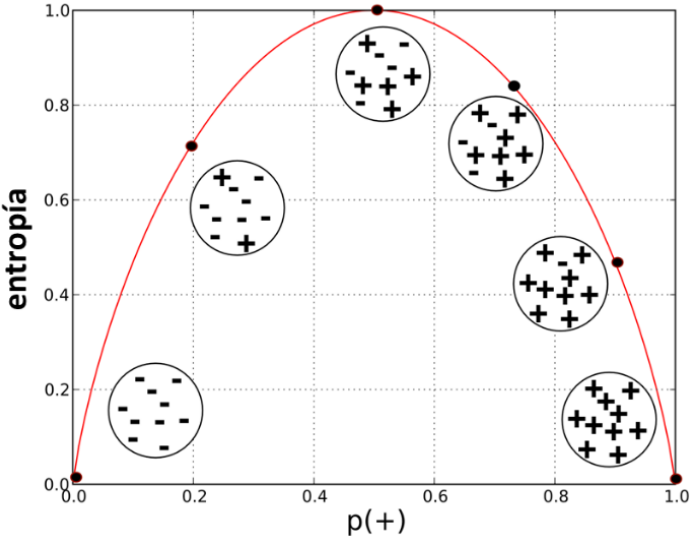
donde $Pr(P_i)$ es la probabilidad de la partición P de pertenecer a la i-ésima clase.

Propiedades de la entropía

Para una **variable aleatoria binaria** x, se tiene que la entropía:

- Alcanza su valor máximo de 1 cuando la probabilidad es 0.5; y hay una probabilidad del 50-50 de que $x = 1$ o $x = 0$.
- La función alcanza un mínimo en 0 cuando $p(x = 1) = 1$ o $p(x = 1) = 0$. En otras palabras, cuando tenemos total certeza.

Curva de entropía [Ara21]



Ganancia de información

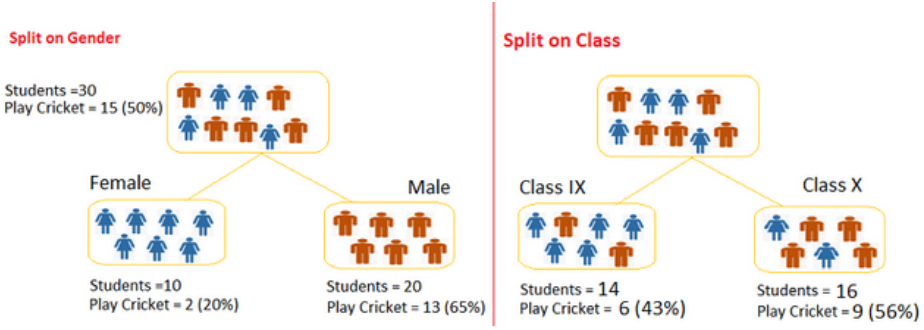
Ganancia de información

Es la diferencia entre la entropia de una partición inicial (padre) y sus posibles sub-particiones basadas en el atributo x_j.

$$S(P, x_j) = E(P) - \sum_{k \in X_j} \left(\frac{|P_{x_j=k}|}{|P|} E(P_{x_j=k}) \right) \tag{2}$$

Donde $k \in X_j$ representa todos los valores posibles del atributo x_j.

Ganancia de información: Ejemplo



$$
\begin{aligned}
E(P^0) &= -Pr(P^0_{y=0}) \log_2 Pr(P^0_{y=0}) - Pr(P^0_{y=1}) \log_2 Pr(P^0_{y=1}) \\
&= -\frac{15}{30} \log_2 \frac{15}{30} - \frac{15}{30} \log_2 \frac{15}{30} \\
&= 1
\end{aligned}
$$

Ganancia de información: Ejemplo

- Para dividir por género (x_0):

$$E(P^0_{x_0=F}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10} = 0{,}72$$
$$E(P^0_{x_0=M}) = -\frac{13}{20}\log_2\frac{13}{20} - \frac{7}{20}\log_2\frac{7}{20} = 0{,}93$$
$$E(P^0_{x_0}) = \frac{10}{30} * 0{,}72 + \frac{20}{30} * 0{,}93 = 0{,}86$$

$$S(P^0_{x_0}) = 1 - 0{,}86 = 0{,}14 \quad (3)$$

Ganancia de información: Ejemplo

- Para dividir por género (x_0):
  $E(P^0_{x_0=F}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10} = 0{,}72$
  $E(P^0_{x_0=M}) = -\frac{13}{20}\log_2\frac{13}{20} - \frac{7}{20}\log_2\frac{7}{20} = 0{,}93$
  $E(P^0_{x_0}) = \frac{10}{30} * 0{,}72 + \frac{20}{30} * 0{,}93 = 0{,}86$

$$S(P^0_{x_0}) = 1 - 0{,}86 = 0{,}14 \quad (3)$$

- Para división por curso (x_1):
  $E(P^0_{x_1=IX}) = -\frac{6}{14}\log_2\frac{6}{14} - \frac{8}{14}\log_2\frac{8}{14} = 0{,}99$
  $E(P^0_{x_1=X}) = -\frac{9}{16}\log_2\frac{9}{16} - \frac{7}{16}\log_2\frac{7}{16} = 0{,}99$
  $E(P^0_{x_1}) = \frac{14}{30} * 0{,}99 + \frac{16}{30} * 0{,}99 = 0{,}99$

$$S(P^0_{x_1}) = 1 - 0{,}99 = 0{,}01 \quad (4)$$

Ganancia de información: Ejemplo

Decisión

Considerando las ecuaciones 3 y 4, entonces:

$$S(P^0_{x_0}) > S(P^0_{x_1}) => G > C \tag{5}$$

Entonces la decisión final es: **el árbol se dividirá por género.**

Table of Contents

Definición

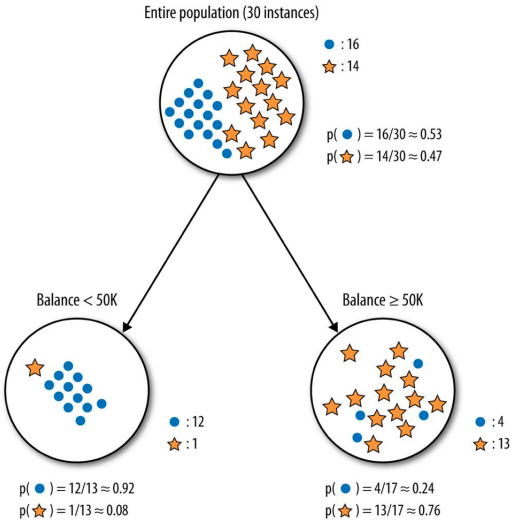
Índice de Gini

El índice de Gini calcula la probabilidad de encontrar aleatoriamente un elemento del conjunto de entrada incorrectamente clasificado, según la siguiente fórmula:

$$Gini(P) = 1 - \sum_{i=1}^{n} Pr_p(Y = i)^2$$

Donde $Pr_p(Y = i)$ es la probabilidad de encontrar un elemento de la clase i en la partición P.

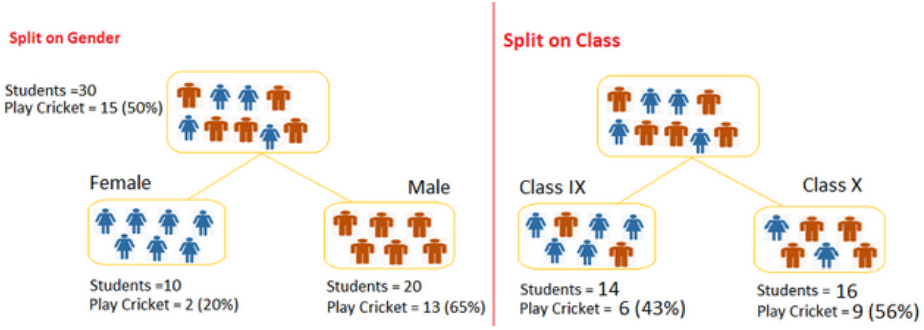
Ejemplo


Entire population (30 instances)
: 16
: 14
p( ) = 16/30 ≈ 0.53
p( ) = 14/30 ≈ 0.47
Balance < 50K
Balance ≥ 50K
: 12
: 1
p( ) = 12/13 ≈ 0.92
p( ) = 1/13 ≈ 0.08
: 4
: 13
p( ) = 4/17 ≈ 0.24
p( ) = 13/17 ≈ 0.76

Propiedades

- Funciona con la variable objetivo categórica (Ej: Éxito o Fracaso).
- Realiza solo divisiones binarias
- Si el valor de Gini $= 0{,}5$, entonces la heterogeneidad es perfecta.
- Si el valor de Gini $= 0{,}0$, entonces la homogeneidad es perfecta.

Índice de Gini: Ejemplo



Ejemplo

- Para dividir en género:
  $Gini(P_F) = 1 - (\frac{2}{10})^2 - \frac{8}{10}^2 = 0{,}32$
  $Gini(P_M) = 1 - (\frac{13}{20})^2 - \frac{7}{20}^2 = 0{,}45$
  $PROM_G = \frac{10}{30} * 0{,}32 + \frac{20}{30} * 0{,}45 = 0{,}41$
- Para división en clase:
  $Gini(P_{IX}) = 1 - (\frac{6}{14})^2 - \frac{8}{14}^2 = 0{,}49$
  $Gini(P_X) = 1 - (\frac{9}{16})^2 - \frac{7}{16}^2 = 0{,}49$
  $PROM_C = \frac{14}{30} * 0{,}51 + 1630 * 0{,}51 = 0{,}51$

Decisión

La división del nodo será por género ya que $PROM_G < PROM_C$.

Medidas de selección de atributos

Para regresión:

- Reducción en la varianza

$$Var(X) = \frac{\sum(X - \bar{X})^2}{n} \tag{6}$$

- Suma de cuadrado residual (RSS)

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

Table of Contents

Modelo

Hipótesis (Varias)

$$H = \{h | h : X \longrightarrow Y\}$$

Función de costo (Entropía)

$$E(P) = -\sum_{i \in Y} Pr(P) \log_2 Pr(P)$$

Función de costo (Ganancia)

$$S(P, x_j) = E(P) - \sum_{k \in X_j} \left(\frac{|P_{x_j=k}|}{|P|} E(P_{x_j=k}) \right)$$

Objetivo

$$\max_{(P,x_j) \in h} S(P, x_j)$$

Árboles de decisión

Ventajas:

- Fácil de entender e interpretar.
- Requiere poca preparación de los datos. No requiere la normalización de datos, no es necesario eliminar ni imputar datos perdidos.
- Capaz de manejar datos numéricos y categóricos.
- Utiliza un modelo de caja blanca. (ej: de un modelo de caja negra es una red neuronal artificial.)
- Es posible validar un modelo utilizando pruebas estadísticas. Eso hace que sea posible tener en cuenta la fiabilidad del modelo.
- Robusto. Se desempeña bien incluso si sus suposiciones son violadas por el verdadero modelo a partir del cual se generaron los datos.
- Funciona bien con grandes conjuntos de datos.

Desventajas

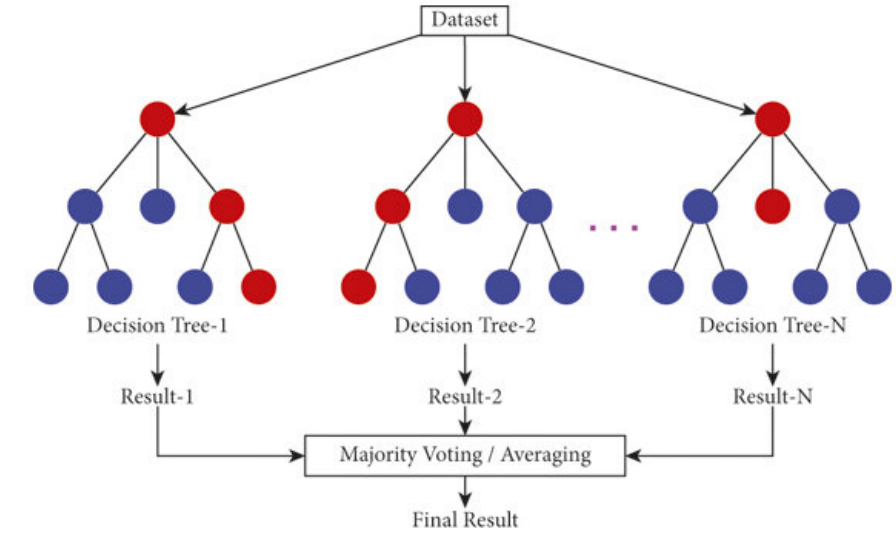
- Tienden al sobreajuste u *overfitting* de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.
- Se ven influenciadas por los *outliers*, creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos outliers.
- No suelen ser muy eficientes con modelos de regresión.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos.
- Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utilizan para categorizar una variable numérica continua.

Formas de evitar el sobreajuste

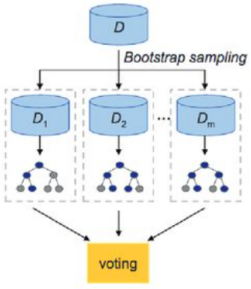
- **Definir restricciones sobre el tamaño del árbol**
  - Mínimo número de muestras u observaciones para dividir un nodo.
  - Mínimo número de observaciones para un nodo terminal.
  - Máxima profundidad del árbol (vertical).
  - Máximo número de nodos hoja.
  - Máximo número de atributos a considerar para la ramificación.
- **Podar el árbol** (cortar, limitar y optimizar el tamaño y la forma de los arboles)
  - Construir el árbol a un profundidad extensa.
  - Remover las hojas que den un valor negativo comparado desde la raíz.

Table of Contents

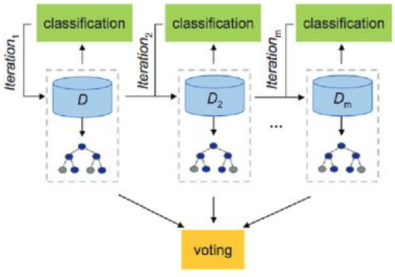
Idea


Dataset
Decision Tree-1
Decision Tree-2
Decision Tree-N
Result-1
Result-2
Result-N
Majority Voting / Averaging
Final Result

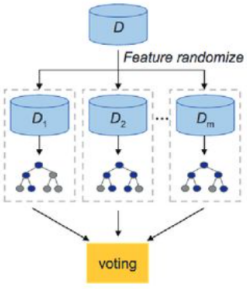
Bagging, Boosting & Random Forest



(a) bagging



(b) boosting



(c) random forests

Ejercicio

Ejercicio 7.1

Utilice el algoritmo de Árboles de Decisión de la biblioteca `Sklearn` en `Python` para clasificar el dataset de semillas[1]. Pruebe utilizando los meta-algoritmos de ensamblaje con Árboles de decisión.

[1] https://archive.ics.uci.edu/ml/datasets/seeds

Table of Contents

Conclusiones

- ¿En qué consiste el método?
- ¿Cuáles son las ventajas y desventajas?
- ¿Cuáles son los pasos o aspectos a considerar para aplicar el método?
- ¿Cuáles son las medidas de selección de atributos?

Table of Contents

Referencias

[Qui86] J. Ross Quinlan. «Induction of decision trees». En: *Machine learning* 1.1 (1986), págs. 81-106.

[Ara21] Carlos Arana. *Modelos de Aprendizaje Automático Mediante Árboles de Decisión*. CEMA Working Papers: Serie Documentos de Trabajo. 778. Universidad del CEMA, feb. de 2021. URL: https://ideas.repec.org/p/cem/doctra/778.html.