Tema 8
Árboles de decisión y métodos de ensamble
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

Objetivos

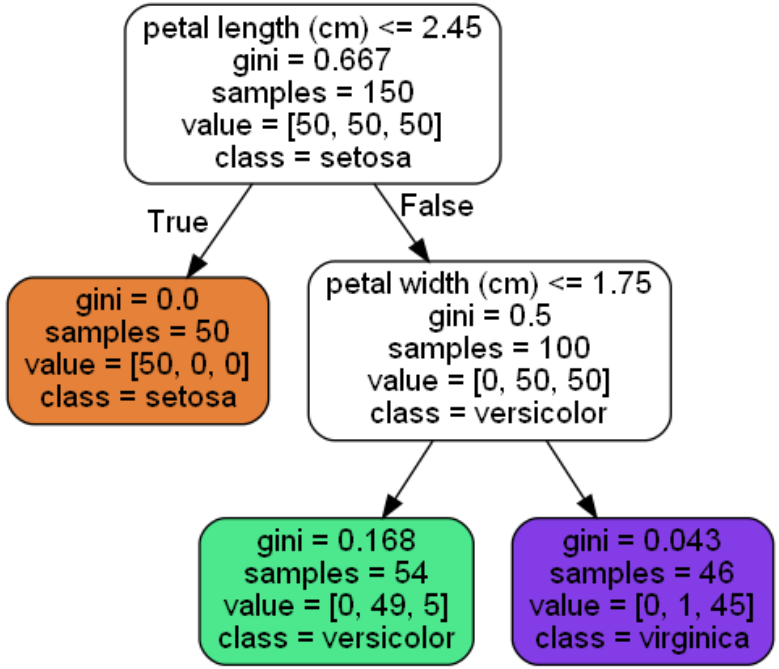
- Árboles de decisión
- Métodos de emsamble
- Pueden usarse en regresión y clasificación
- Ventajas e inconvenientes

- Secuencia de condiciones sobre las variables predictoras y su relación con la variable respuesta
- Condiciones que se suceden en distintos caminos del árbol: ramas de posibilidades $\rightarrow$ regiones
- Predicciones: media/moda de la región a la cual pertenece la observación

Características:

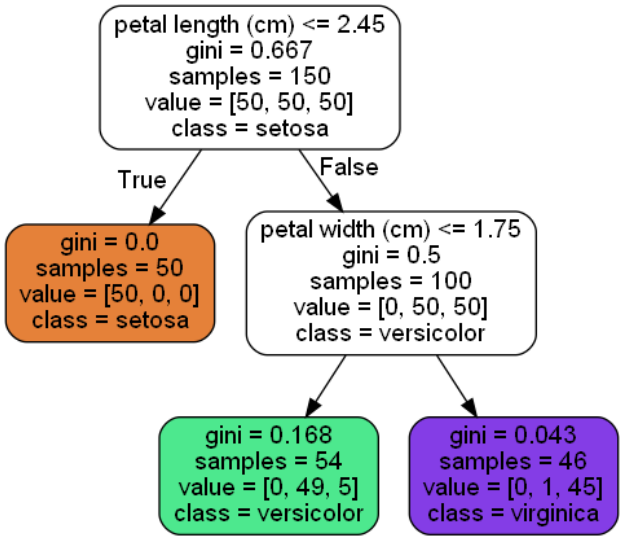
- Simples
- Fáciles de interpretar
- No es necesario escalar las variables predictoras
- Capacidad de ajuste y predicción peor
- Esta capacidad predictiva mejora con los métodos de ensamble

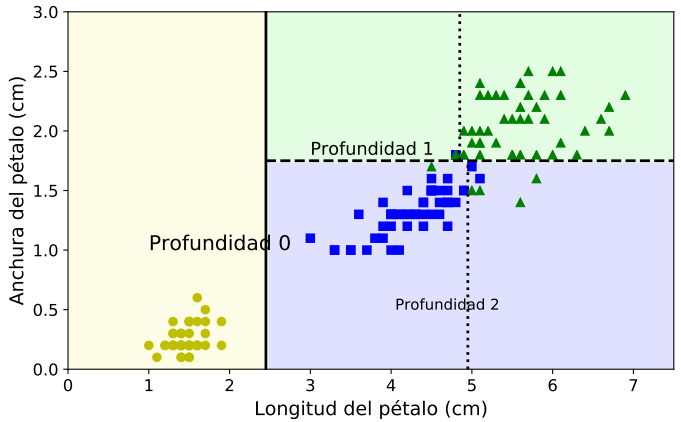
Árboles de decisión: funcionamiento



- Clasificación por profundidad (0,1,2) o posición (Izq.,Der.)
- *sample* cuenta cuántas observaciones de la muestra de entrenamiento pertenecen a un nodo
- *value* indica cuántas observaciones de cada clase
- *gini* es la medida de impureza: si todas las observaciones del nodo son de la misma clase entonces gini$= 0$

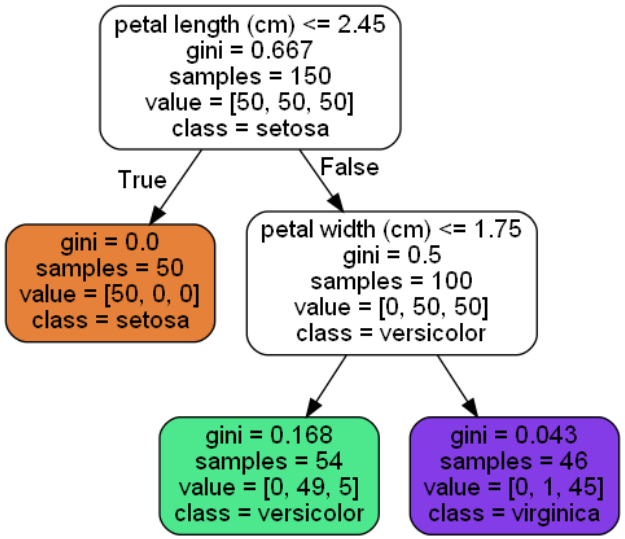
Árboles de decisión: visualización





- La región de la izquierda es pura; no se divide más
- La región de la derecha es impura; se divide con la anchura
- Se fijó la profundidad máxima a 2. Si fuese 3, habría otra división para cada nodo de profundidad 2

Árboles de decisión: estimación



- Mediante el árbol de decisión podemos estimar la probabilidad de que una observación pertenezca a una determinada clase
- Primero, vemos a qué hoja pertenece la observación
- Segundo, obtenemos las proporciones del número de muestras de entrenamiento de cada clase que se encuentran en dicha hoja
- Asignamos a la nueva observación la predicción que obtiene mayor probabilidad de pertenencia según el árbol de decisión

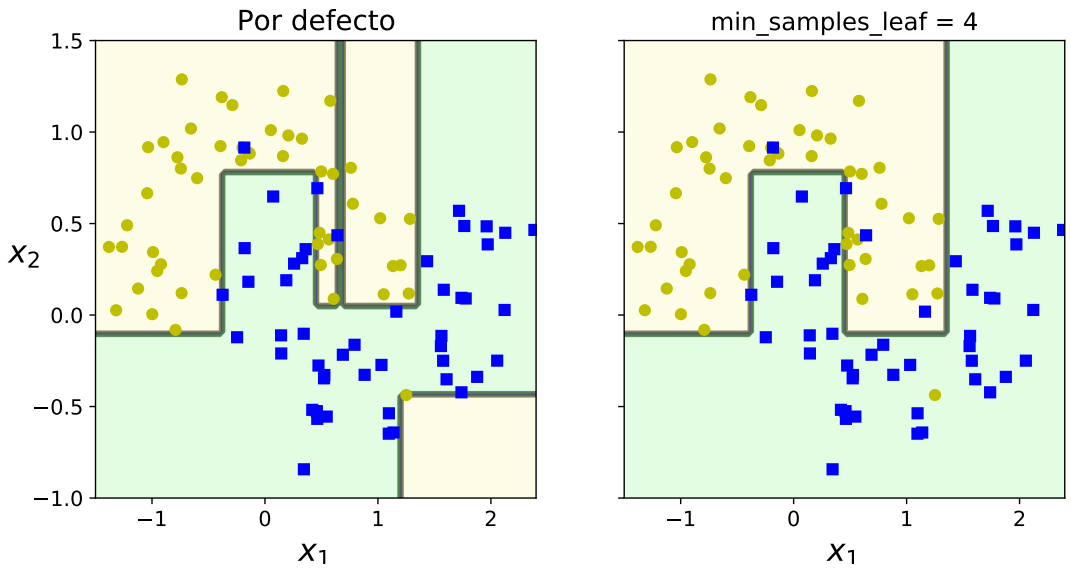
Árboles de decisión: el algoritmo CART

- Se utiliza para generar árboles de decisión
- Utiliza pregunta binaria y genera dos nodos descendientes
- Escoge una de las variables predictoras k
- Fija un umbral t_k, frontera entre las dos regiones creadas
- (k, t_k) se escogen (por minimización) de modo que se consiga obtener una región que sea lo más pura posible según Gini
- Cada una de las dos regiones se divide siguiendo el mismo mecanismo
- Final: profundidad máxima o no se encuentra una división que consiga más pureza o hiperparámetros
- Se comprueba la impureza en cada nivel, pero no se comprueba si la solución es buena en niveles más profundos

Árboles de decisión: hiperparámetros

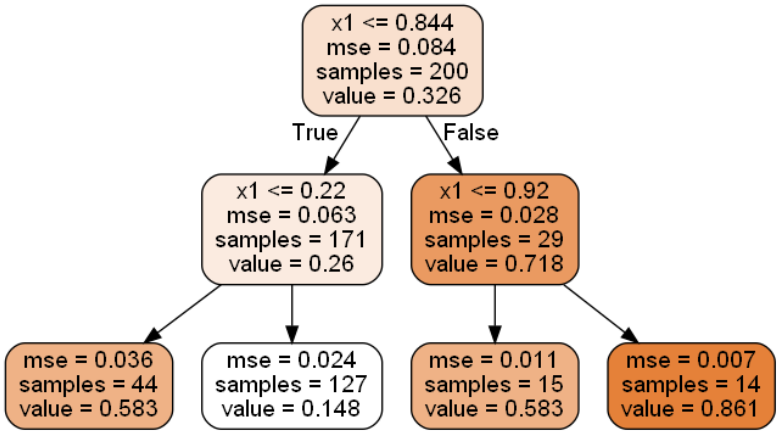
- Utilizados para evitar el sobreajuste que genera por defecto el árbol de decisión
- Restringen la libertad del árbol de decisión durante su creación
- *min_samples_split*: mínimo número de muestras que puede tener un nodo para que se pueda dividir
- *min_samples_leaf*: mínimo número de muestras que puede tener una hoja
- *min_weight_fraction_leaf*: igual pero usando pesos en las observaciones
- *max_leaf_nodes*: máximo número de nodos hoja
- *max_features*: máximo número de variables que son consideradas por el algoritmo para decidir cada división
- También pueden recortarse nodos innecesarios *a posteriori*

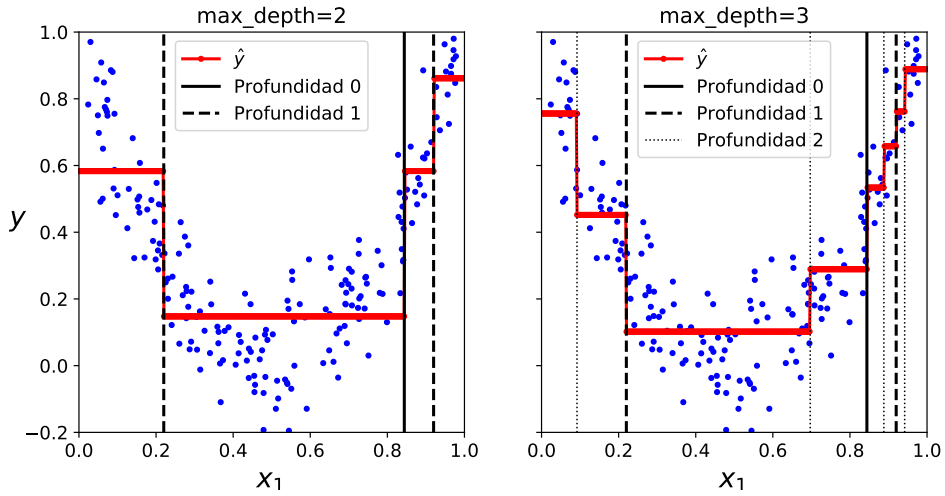
Árboles de decisión: hiperparámetros



- Hiperparámetros por defecto (izq.), es decir sin restricciones vs. argumento min_samples_leaf = 4 (der.)
- el árbol obtenido en el panel de la izquierda presenta sobreajuste

Árboles de decisión: regresión





- En lugar de predecir la clase de cada nodo, el árbol predice un valor
- Este valor es el valor medio de todas las observaciones de entrenamiento que pertenecen a ese nodo
- Se sustituye en la función coste la medida de impureza de Gini por el error cuadrático medio MSE

Árboles de decisión: ventajas e inconvenientes

Ventajas:

- sencillos de implementar
- fáciles de entender e interpretar
- versátiles
- pueden usarse en clasificación y en regresión

Inconvenientes:

- sensibilidad ante rotación de los datos (divisiones perpendiculares a los ejes)
- sensibles ante pequeñas variaciones en las observaciones de la muestra de entrenamiento