Árboles

Fabián Villena

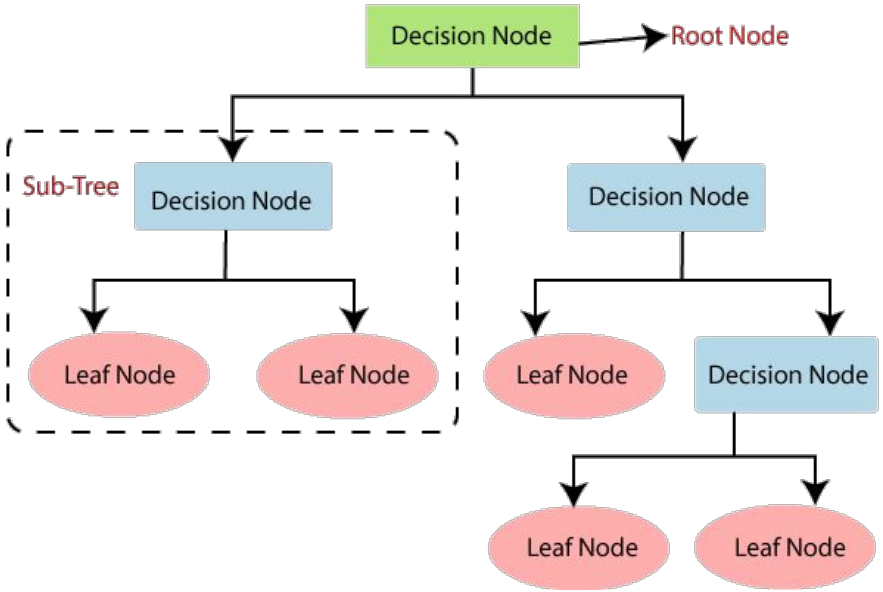
Métodos basados en árboles

Estos métodos son ampliamente utilizados en aprendizaje automático. Estos métodos simulan el proceso de decisión del razonamiento humano, lo que los hace muy intuitivos.
A través de estructuras jerárquicas llamadas árboles de decisión se pueden representar soluciones a problemas de clasificación y de regresión.

Árboles de decisión

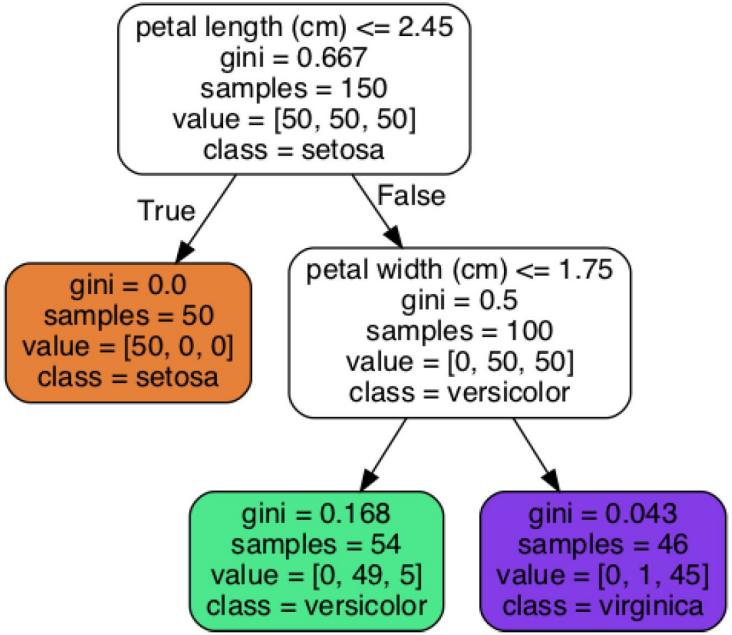
Los árboles de decisión son uno de los algoritmos más utilizados para ajustar modelos de aprendizaje automático. Estos algoritmos pueden ajustar modelos para clasificación y regresión.

Los árboles de decisión son el componente fundamental de uno de los algoritmos más poderosos, Random Forest.



Visualización de un árbol de decisión

Podemos explorar cada una de las decisiones que toma el árbol para poder realizar la predicción final. Esto es muy importante para la explicabilidad del modelo.



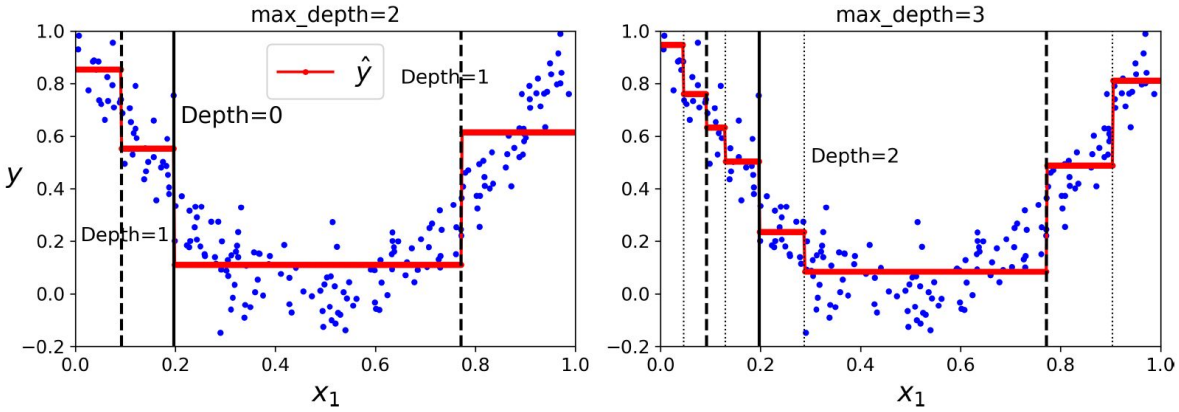
El algoritmo CART

El algoritmo que construye los árboles primero divide el conjunto de datos utilizando una característica k y un umbral de división t_k. El algoritmo busca el par (k, t_k) que produce los subconjuntos más puros (ponderados por su tamaño).

Cuando se dividió el conjunto de datos con el mejor (k, t_k), de manera recursiva se van realizando las divisiones restantes hasta que se ha llegado a la profundidad máxima o la pureza no se puede reducir.

Árboles de decisión para regresión

La diferencia con los árboles para clasificación es que en cada hoja se predice un valor continuo que es el promedio de los valores de la partición. Además el algoritmo CART minimiza el MSE.



Hiperparámetros de los árboles de decisión

Los hiperparámetros más importantes que debemos establecer en los árboles de decisión son:

- La profundidad del árbol: Este hiperparámetro al aumentarlo genera modelos más flexibles.
- Métrica de impureza: Podemos utilizar la impureza de Gino o la entropía de la información
- Mínimo de ejemplos en una hoja: La mínima cantidad de ejemplos que debe tener un nodo final. Al disminuirlo tenemos modelos más complejos.

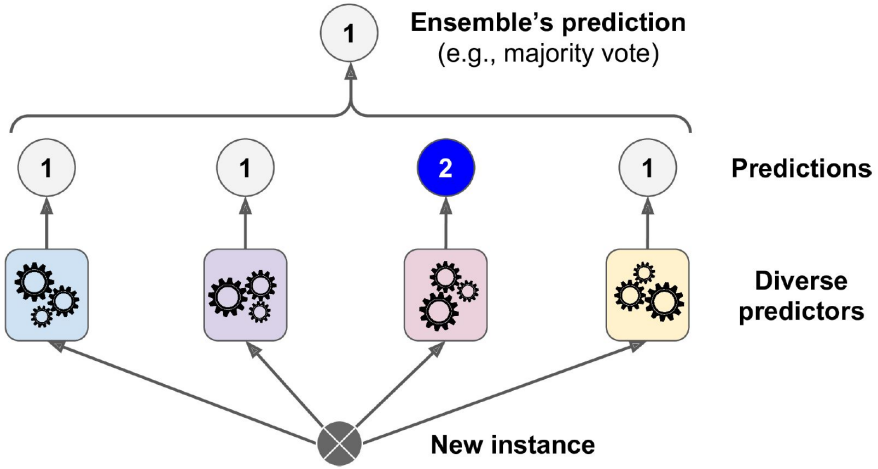
Ensamble de árboles

Si agregamos las decisiones de un grupo de predictores, típicamente se tendrán mejores predicciones que la mejor predicción única.

Podemos ajustar un grupo de árboles de decisión, cada uno ajustado sobre un subconjunto al azar del conjunto de entrenamiento. Para hacer predicciones, simplemente predecimos la clase que tuvo más votos.

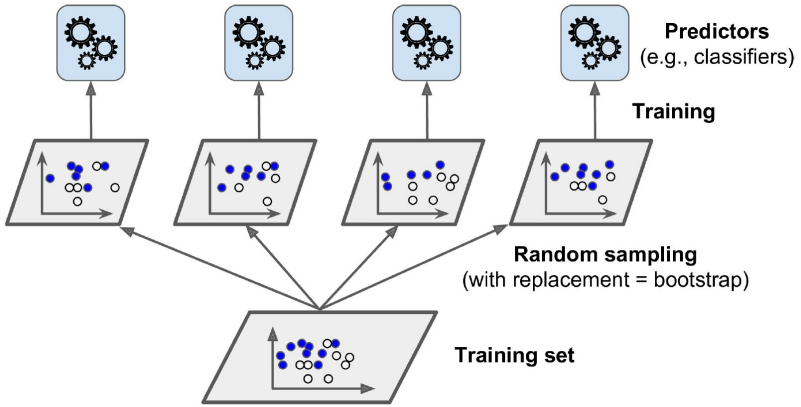
Votación

Si cada clasificador utilizado en un ensamble basado en votación es un predictor débil (que se comporta ligeramente mejor que el azar), el ensamble aún así puede ser un predictor fuerte, sólo con tener una cantidad suficiente de predictores débiles.



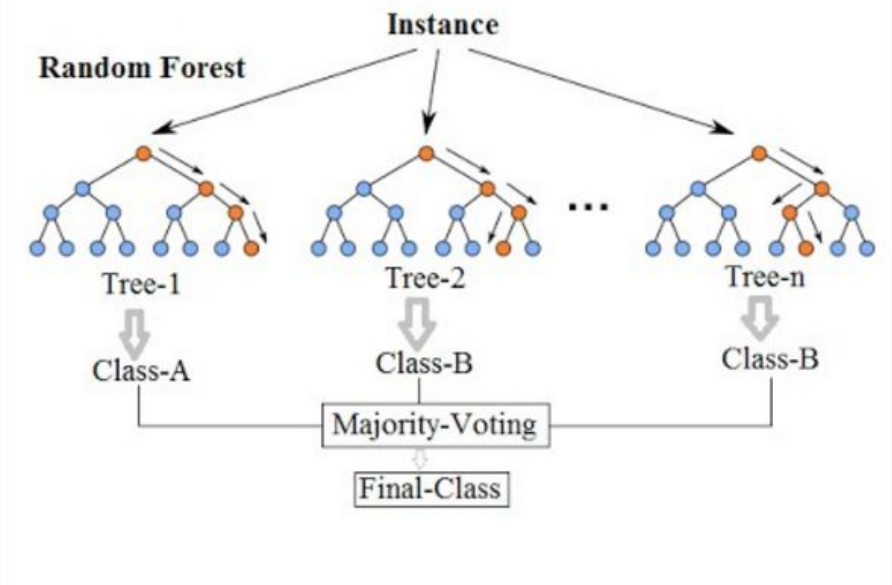
Bagging

Para decorrelacionar cada uno de los predictores, podemos cada uno ajustarlo sobre una muestra sintética generada a través de bootstrapping, generar ejemplos al tomar al azar con reemplazo ejemplos del conjunto de entrenamiento.



Random forests

Para decorrelacionar más cada uno de los árboles, Random Forests selecciona al azar un subconjunto de características para realizar la partición en un nodo del árbol de decisión.



Hiperparámetros de Random Forests

Los hiperparámetros para este algoritmo se heredan de los árboles de decisión y además hay unos específicos para ajustar el ensamble:

- Cantidad de árboles de decisión: Al aumentar la cantidad de árboles complejizamos el modelo.
- Cantidad de características para realizar la partición
- Cantidad de ejemplos que tendrá cada conjunto sintético