Decision Tree Classifier and Ensembles

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python

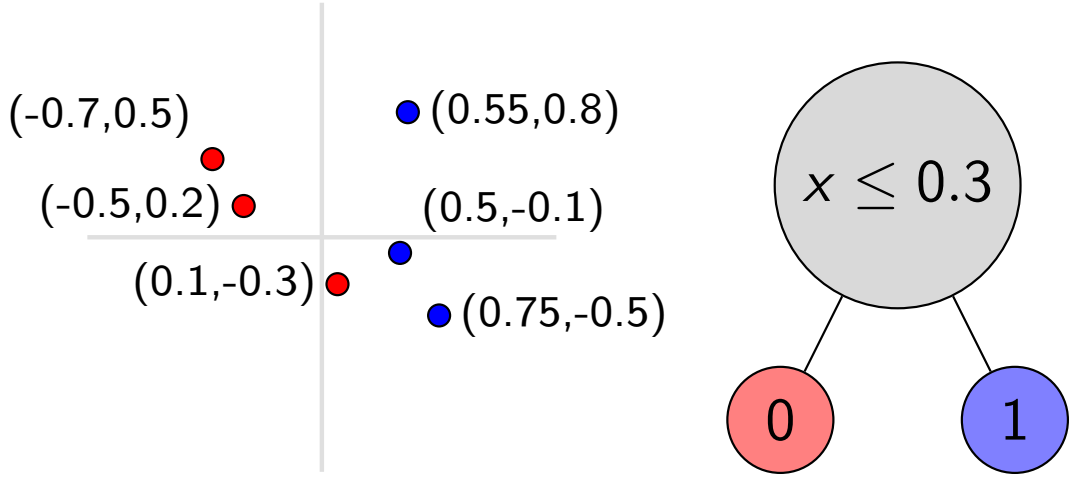
Table of Contents

Árboles de Decisión

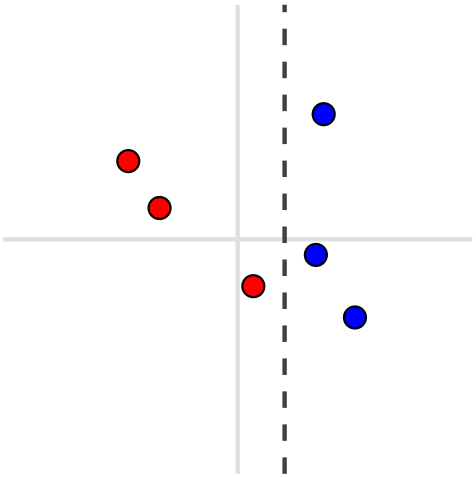
Árboles de Decisión

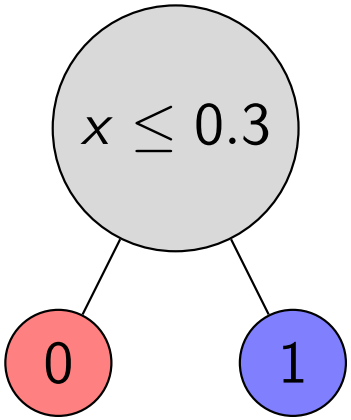
Un árbol de decisión es un módelo predictivo que va de las observaciones de una instancia a conclusiones acerca del valor objetivo de la instancia mediante un árbol. En este árbol, las hojas representan etiquetas de las clases y los nodos de ramificación representan condiciones sobre los valores de las features que llevan a las etiquetas de las clases.

Un Ejemplo Trivial

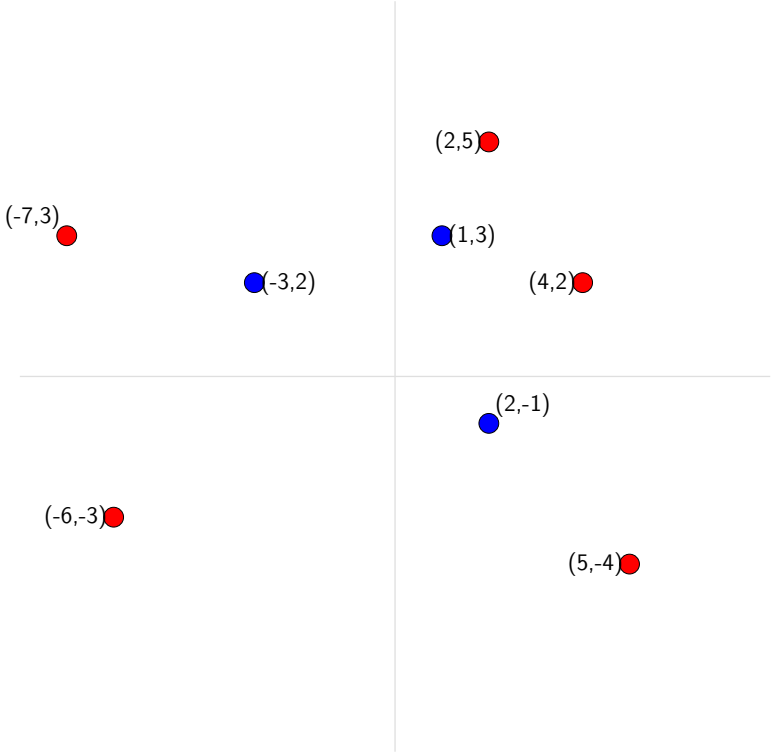


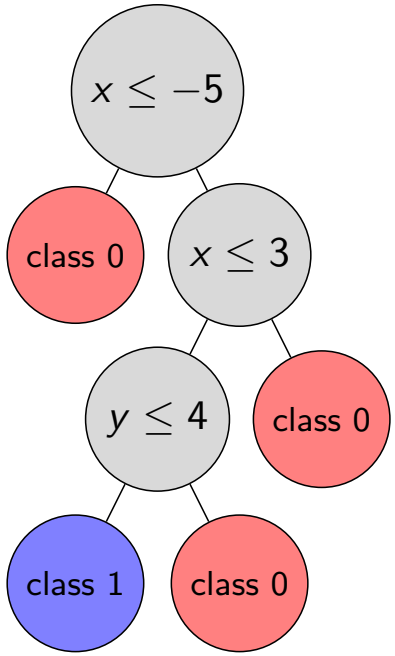
Un Ejemplo Trivial



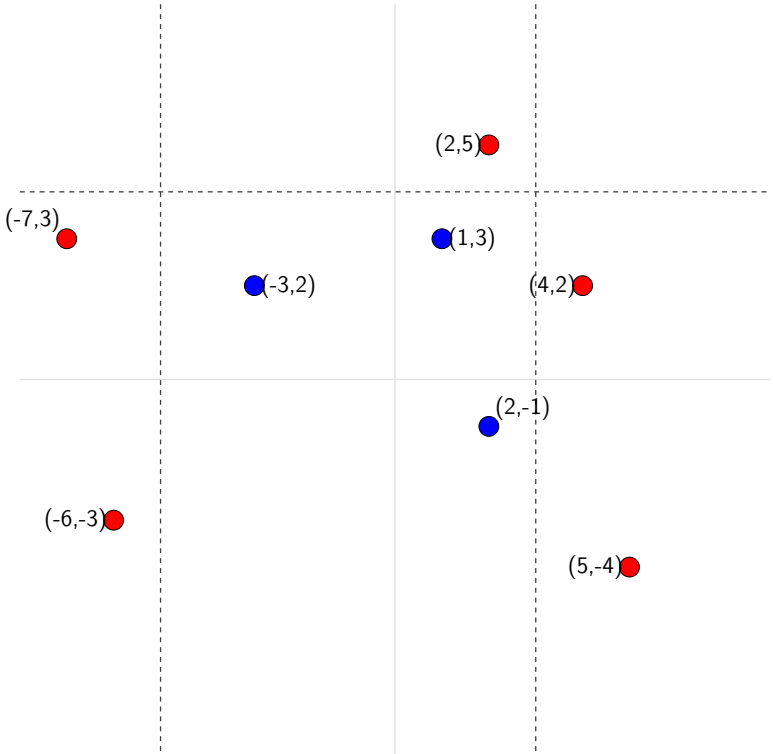


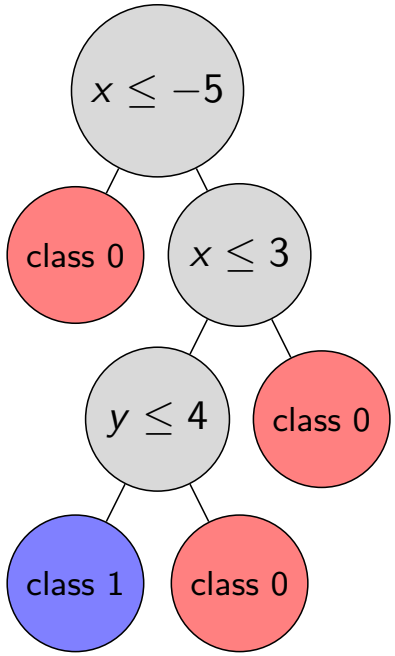
Otro Ejemplo





Otro Ejemplo





Pureza de los nodos

El criterio para elegir la mejor partición en cada nodo interior es la pureza de los nodos.

Pureza de los nodos

El criterio para elegir la mejor partición en cada nodo interior es la pureza de los nodos. Una opción frecuente es la impureza de Gini:

$$I_G(A) = 1 - \sum_{i=1}^{J} p_i^2,$$

para un conjunto A de elementos pertenecientes a J clases $\{1, ..., J\}$ con probabilidades $p_1, ..., p_J$.

Pureza de los nodos

El criterio para elegir la mejor partición en cada nodo interior es la pureza de los nodos. Una opción frecuente es la impureza de Gini:

$$I_G(A) = 1 - \sum_{i=1}^{J} p_i^2,$$

para un conjunto A de elementos pertenecientes a J clases $\{1, ..., J\}$ con probabilidades $p_1, ..., p_J$.

Ejemplo:

Parámetros Importantes

Los parámetros más importantes para un árbol de decisión son:

- Criterio para evaluar las divisiones: gini, entropy, etc.

Parámetros Importantes

Los parámetros más importantes para un árbol de decisión son:

- Criterio para evaluar las divisiones: gini, entropy, etc.
- Profundidad máxima (max_depth) es la máxima profundidad del árbol. De otra forma, los nodos se van expandiendo hasta que son puros o que quedan menos que el siguiente parámetro.

Parámetros Importantes

Los parámetros más importantes para un árbol de decisión son:

- Criterio para evaluar las divisiones: gini, entropy, etc.
- Profundidad máxima (max_depth) es la máxima profundidad del árbol. De otra forma, los nodos se van expandiendo hasta que son puros o que quedan menos que el siguiente parámetro.
- Mínimo número de instancias para dividir.

Ventajas y Desventajas

Advantages:

- DTs requires less effort for data preparation during pre-processing.
- DTs do not require normalization or scaling of data.
- Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
- DTs are very intuitive and easy to explain

Disadvantage:

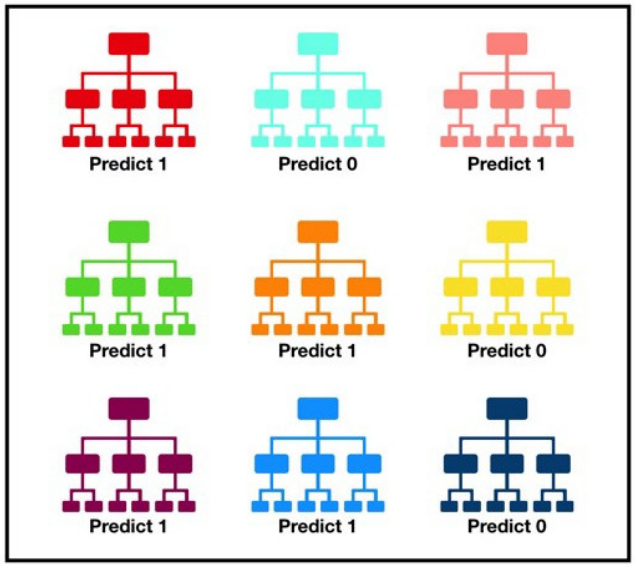
- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- DTs often involve higher time and complexity to train the model.

Table of Contents

Random Forest

Random Forest

Random forests es un método de clasificación que funciona mediante la construcción de varios árboles de decisión en el entrenamiento. La salida de un random forest es la clase seleccionada por la mayoria de árboles.



Bagging: Bootstrap Aggregating

Bagging

Bootstrap aggregating, o bagging, es un meta-algoritmo de ensambles de Machine Learning diseñado para mejorar la estabilidad y precisión de algoritmos de Machine Learning de clasificación y regresión, además de reducir el over-fitting. Usalmente se aplica a árboles de decisión.

El *Bagging* consiste de dos pasos:

- Bootsrapping: Generar varios conjuntos de datos muestreando el conjunto de datos original, con reemplazo. En cada conjunto de datos entrenar un árbol de decisión.

Bagging: Bootstrap Aggregating

Bagging

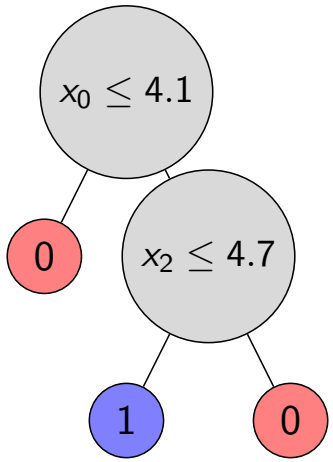
Bootstrap aggregating, o bagging, es un meta-algoritmo de ensambles de Machine Learning diseñado para mejorar la estabilidad y precisión de algoritmos de Machine Learning de clasificación y regresión, además de reducir el over-fitting. Usalmente se aplica a árboles de decisión.

El *Bagging* consiste de dos pasos:

- Bootsrapping: Generar varios conjuntos de datos muestreando el conjunto de datos original, con reemplazo. En cada conjunto de datos entrenar un árbol de decisión.
- Aggregating: Para clasificar una instancia, el ensamble junta las predicciones hechas por cada árbol y toma la clase predicha por la mayoría.

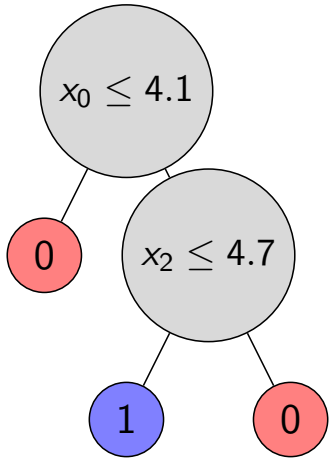
Bagging

	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Bagging

	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



Bootstraping: Se entrenan varios árboles usando muestro aleatorio de instancias, con reemplazo.

	x_0	x_1	y
2			
0			
2			
4			
5			
5			

	x_0	x_1	y
2			
1			
3			
1			
4			
4			

	x_0	x_1	y
4			
1			
3			
0			
0			
2			

	x_0	x_1	y
3			
3			
2			
5			
1			
2			

Bagging

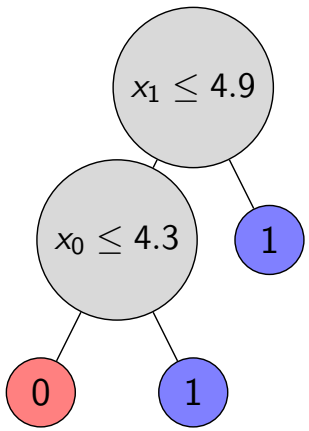
	x_0	x_1	y
2			
0			
2			
4			
5			
5			

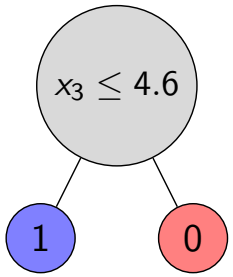
	x_0	x_1	y
2			
1			
3			
1			
4			
4			

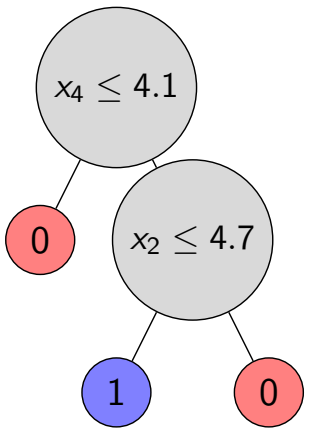
	x_0	x_1	y
4			
1			
3			
0			
0			
2			

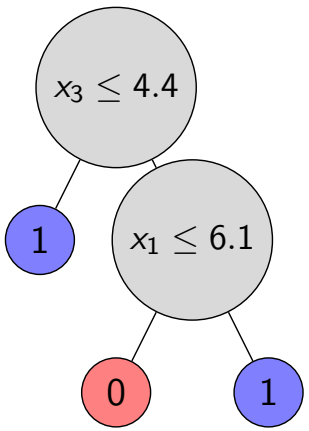
	x_0	x_1	y
3			
3			
2			
5			
1			
2			

Aggregating: La predicción del ensamble es la predicción mayoritaria.









Ventajas y Desventajas

Ventajas:

- Es muy intuitivo, los hiper-parámetros son fáciles de entender.

Desventajas:

Ventajas y Desventajas

Ventajas:

- Es muy intuitivo, los hiper-parámetros son fáciles de entender.
- Es posible conocer la importancia de las features, de acuerdo a los criterios de impureza.

Desventajas:

Ventajas y Desventajas

Ventajas:

- Es muy intuitivo, los hiper-parámetros son fáciles de entender.
- Es posible conocer la importancia de las features, de acuerdo a los criterios de impureza.
- Tienden a evitar el over-fitting.

Desventajas:

Ventajas y Desventajas

Ventajas:

- Es muy intuitivo, los hiper-parámetros son fáciles de entender.
- Es posible conocer la importancia de las features, de acuerdo a los criterios de impureza.
- Tienden a evitar el over-fitting.

Desventajas:

- La gran cantidad de árboles y la profundidad de estos pueden hacer al método lento para usar en tiempo real.