Material Imprimible

Curso de Data Science

Módulo 4 - Clasificación

Contenidos:

- Qué es un problema de clasificación
- Algoritmos de clasificación
- Árbol de decisión en Scikit Learn
- Métricas de Evaluación

Qué es un problema de clasificación

Un problema supervisado de clasificación es un tipo de problema de Machine Learning en el que se entrena un modelo para asignar una etiqueta o categoría a una observación basándose en un conjunto de datos de entrenamiento previamente etiquetados.

En estos desafíos predictivos, se tienen un conjunto de datos de entrada junto con las etiquetas correspondientes que describen la clase a la que pertenece cada una de las observaciones. El objetivo del modelo de clasificación es aprender la relación entre los datos de entrada y las etiquetas asociadas para poder clasificar de manera precisa nuevas observaciones no vistas previamente.

Se dice que es un problema de clasificación binario cuando la etiqueta tiene dos etiquetas contrapuestas (Verdadero o Falso, “Si” o “No”, “Compra” o “No me compra”). En el caso que las etiquetas sean categorias multiples, se habla de un problema de clasificación multiclase. (Por ej: “Perro”, “Gato” o “Ave”)

Algoritmos de clasificación

Los algoritmos más famosos y utilizados en este tipo de problemas son los siguientes:

- Árboles de decisión: este algoritmo utiliza una estructura de árbol para tomar decisiones basadas en una serie de preguntas o pruebas realizadas en los datos de entrada. Cada nodo del árbol representa una pregunta, y las ramas representan las posibles respuestas a esa pregunta.

- Regresión logística: este algoritmo se utiliza para problemas de clasificación binaria, donde se busca predecir la pertenencia a una de dos posibles categorías. La regresión logística utiliza una función logística para modelar la relación entre los datos de entrada y las etiquetas de salida.

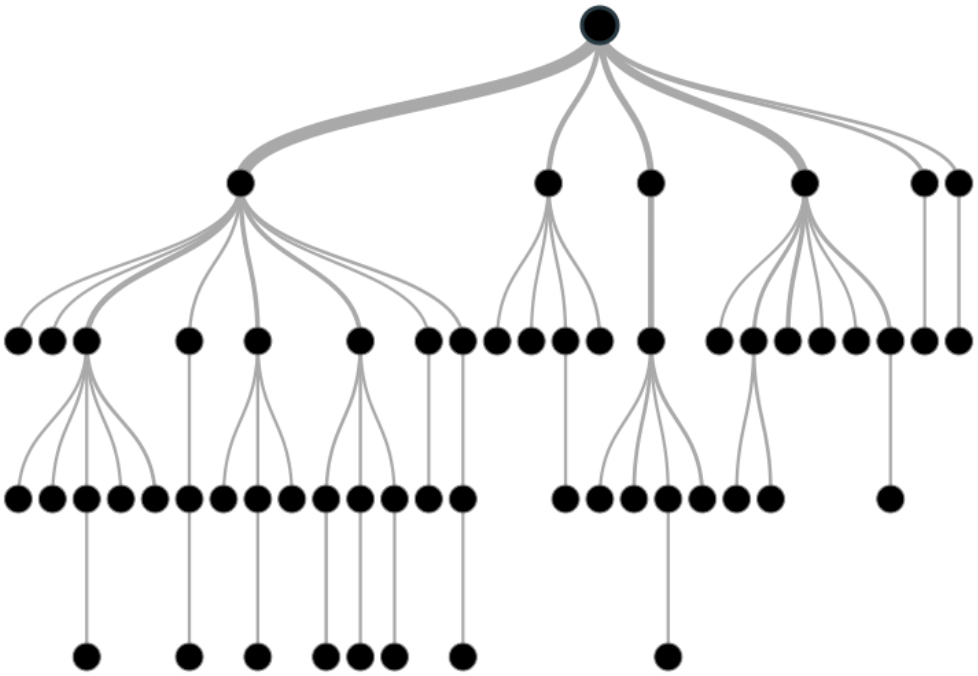
- Support Vector Machines (SVM): este algoritmo busca encontrar el hiperplano que mejor separa las diferentes clases en el espacio de características de los datos de entrada. SVM se utiliza comúnmente en problemas de clasificación binaria y multiclase.

- K-vecinos más cercanos (KNN): este algoritmo clasifica una observación en función de las etiquetas de sus k vecinos más cercanos en el espacio de características. KNN se utiliza comúnmente en problemas de clasificación multiclase.

Existen muchos otros algoritmos y técnicas de ensamble para realizar predicciones de este estilo.

Arbol de decisión en Scikit Learn

Link: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html



Un árbol de decisión es una estructura jerárquica en forma de árbol que se utiliza en Machine Learning para clasificar y predecir la pertenencia a una o varias categorías de un conjunto de datos. Cada nodo del árbol representa una pregunta sobre una característica específica del conjunto de datos, y cada rama del árbol representa una posible respuesta a esa pregunta.

El proceso de construcción de un árbol de decisión comienza con el nodo raíz superior, que incluye todas las observaciones del conjunto de datos de entrenamiento. En cada paso del proceso de construcción, se selecciona una característica que mejor separa las observaciones en diferentes categorías. Para esto, se utiliza una medida de impureza que mide la homogeneidad de las observaciones en cada rama del árbol.

Una vez que se selecciona la variable, se divide el conjunto de datos en dos o más ramas según los valores de esa variable, creando nodos hijo. Luego, el proceso de selección de características y división se repite para cada nodo hijo, y así sucesivamente, hasta que se

llega a un punto en el que no es posible mejorar la precisión del modelo de árbol de decisión.

Finalmente, para hacer una predicción utilizando el árbol de decisión, se comienza en el nodo raíz y se desciende por el árbol hasta llegar a una hoja, que corresponde a una etiqueta de salida o categoría.

Métricas de Evaluación

Luego de entrenar un modelo de clasificación, necesitamos evaluar su rendimiento en un conjunto de datos que no ha sido utilizado para el entrenamiento. Esto se puede hacer utilizando varias métricas, entre ellas el accuracy y la matriz de confusión.

El accuracy (precisión, en español) es una métrica común que se utiliza para medir la precisión global de un modelo de clasificación en un conjunto de datos. Se define como la proporción de observaciones clasificadas correctamente por el modelo, es decir, la cantidad de verdaderos positivos y verdaderos negativos dividida por el total de observaciones. Se puede calcular como:

accuracy = (TP + TN) / (TP + TN + FP + FN)

donde TP (true positive) es el número de casos positivos que fueron correctamente identificados, TN (true negative) es el número de casos negativos que fueron correctamente identificados, FP (false positive) es el número de casos negativos que fueron identificados como positivos y FN (false negative) es el número de casos positivos que fueron identificados como negativos.

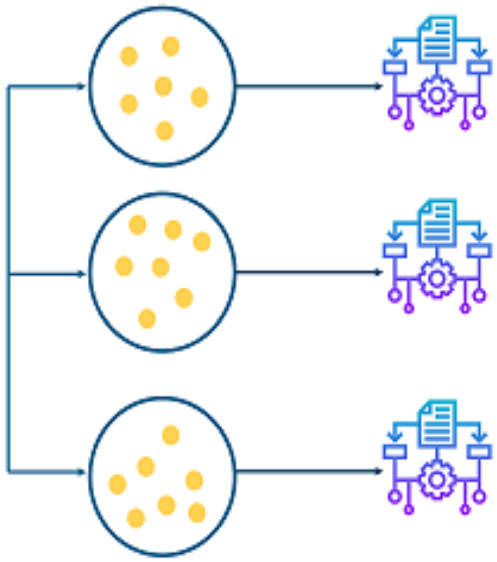
La matriz de confusión, por otro lado, es una tabla que muestra la frecuencia con la que se clasifican correctamente e incorrectamente las observaciones de un modelo de clasificación. La matriz de confusión tiene cuatro celdas, que corresponden a los cuatro posibles resultados de la clasificación: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN).

Bagging y Boosting

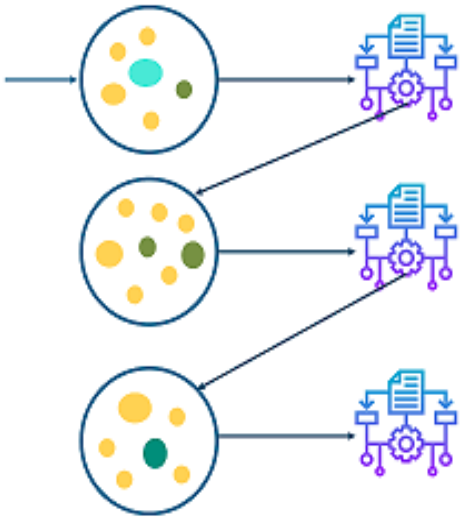
Bagging y Boosting son técnicas de ensamblado de modelos utilizadas en aprendizaje automático para mejorar la precisión de los modelos de predicción. Ambas técnicas se basan en la combinación de múltiples modelos individuales (también llamados modelos base o modelos débiles) para producir un modelo final más preciso.

El Bagging es una técnica de ensamblado que consiste en entrenar en paralelo múltiples modelos individuales con subconjuntos aleatorios del conjunto de datos de entrenamiento, y luego combinar las predicciones de todos los modelos para producir una predicción final. La idea detrás de Bagging es reducir la varianza del modelo final al

agregar múltiples modelos que se ajustan a diferentes partes del conjunto de datos de entrenamiento.



El Boosting es una técnica de ensamblado que construye un modelo fuerte a partir de múltiples modelos débiles, enfocándose en los casos que son más difíciles de predecir. El Boosting se logra mediante la creación de un modelo débil en el conjunto de datos original y luego ajustando su peso para que se enfoque en los casos que fueron mal clasificados. El proceso se repite varias veces para construir múltiples modelos débiles y, finalmente, se combina para producir un modelo fuerte. El modelo final se obtiene agregando los modelos débiles ponderados según su rendimiento.



Fuente de imágenes: machinelearningparatodos.com

Estas técnicas mejoran la posibilidad de hacer un solo algoritmo, apoyándose en las herramientas tecnológicas de avanzada que se ofrecen en la actualidad. Hay que tener en cuenta que cuanto mayor poder tenga nuestro método de ensamble, mayores requerimientos computacionales (procesadores, memoria ram, placas de video si fuese necesario) se tendrán.