Apuntes Machine Learning

Freyman Balaguera

May 24, 2020

1 Modelos Supervisados

Un modelo supervisado es aquél a quien se le entregan datos previamente etiquetados para su entrenamiento.

1.1 Modelos de Clasificación

Los modelos de clasificación pueden ser de **binarios** o **multiclases**.

Los **modelos de dos clases o binarios** son aplicados a eventos como detectar si hay o no fraude, si va o no a llover, etc... elecciones binarias.

Los **modelos de clasificación multiclases** permiten clasificar elementos con una gama amplia de etiquetas. Por ejemplo, la especie de una planta.

1.1.1 kNN - Vecinos Cercanos

El kNN es un algoritmo de clasificación basado en un K que representa la cantidad de vecinos cercanos a tener en cuenta para otorgar una clase determinada.

Etapa de entrenamiento el algoritmo no tiene como tal una definición de la etapa de entrenamiento, basta con tener los puntos con sus etiquetas dispersos sobre el plano.

Etapa de Prueba el algoritmo entonces selecciona un elemento nuevo, y a éste elemento nuevo lo etiquetará según su K vecinos cercanos:

1. Si k = 1 entonces toma la etiqueta que tenga el primer punto mas cercano al nuevo elemento.

2. Si k = 2 entonces verificaría las etiquetas de estos dos puntos, si ambos son de la misma clase, entonces la etiqueta dada al nuevo elemento es dicha clase, pero si son de clases diferentes, el kNN hará una selección aleatoria de la etiqueta de dicho punto.

3. Si k = 3 entonces la etiqueta que tenga la mayoría de puntos será la otorgada a el nuevo elemento.

Es importante aclarar que éste modelo se retroalimenta a partir de recibir nuevos datos y etiquetarlos.

1.2 Decision Tree - Arbol de decisiones

El famosisimo if-else pero optimizado. En el arbol de decisiones lo que se tiene es que a partir de las caracteristicas de un conjunto de datos, vamos a establecer decisiones.

Los nodos de un arbol son decisiones binarias basados en puntos especificos de cada caracteristica determinado por una métrica de pureza como el indice de Gini y la profundidad del arbol.

Etapa de entrenamiento: el árbol de comienza a construir probando cada combinación posible de separación binaria de los datos por característica hasta hallar la que menos impura sea, o hasta que se alcance el máximo nivel de profundidad deseado. Genera un Grafo aciclico dirigido (Arbol de Decisiones).

Etapa de prueba: se comienza a seguir el grafo aciclico dirigido generado por la etapa de entrenamiento hasta llegar a una hoja del arbol que contendrá la etiqueta correspondiente al nuevo dato.

El overfitting en un arbol de decisiones se produce cuando practicamente cada dato se vuelve una etiqueta. Dejando de ser un modelo generalizado.

Hay dos formas de evitar el overfitting de éste modelo, una es colocar un buen nivel de profundidad del cual no se pueda exceder, o una medida de impureza rígida limite, y la otra forma es podar el arbol, es decir, una vez formado el arbol, extraer algunos de sus nodos.

1.3 Naive Bayes

¿Te gustan las probabilidades? Este es el modelo de las probabilidades.

Se basa en el Teorema de Bayes (probabilidad condicionada) para establecer la etiqueta de un elemento.

Etapa de entrenamiento: el modelo comienza a calcular la probabiliad condicionada de cada componente del vector de datos de entrada X si se da la clase C

$$P(X|C) = P(X_1|C) * P(X_2|C) * ... * P(X_n|C)$$

Etapa de prueba: el modelo predice la etiqueta de cada dato en la clase que otorga una mayor probabilidad.