

Lección 6: Tipos de algoritmos de clasificación y regresión

Dentro del aprendizaje supervisado, más específicamente dentro de los algoritmos de clasificación y regresión, es posible encontrar diversos métodos que llevan a cabo los diferentes procesos que allí se realizan. Si bien son muchos, a continuación, se realizará una breve explicación de algunos de ellos.

- **Árbol de decisión**

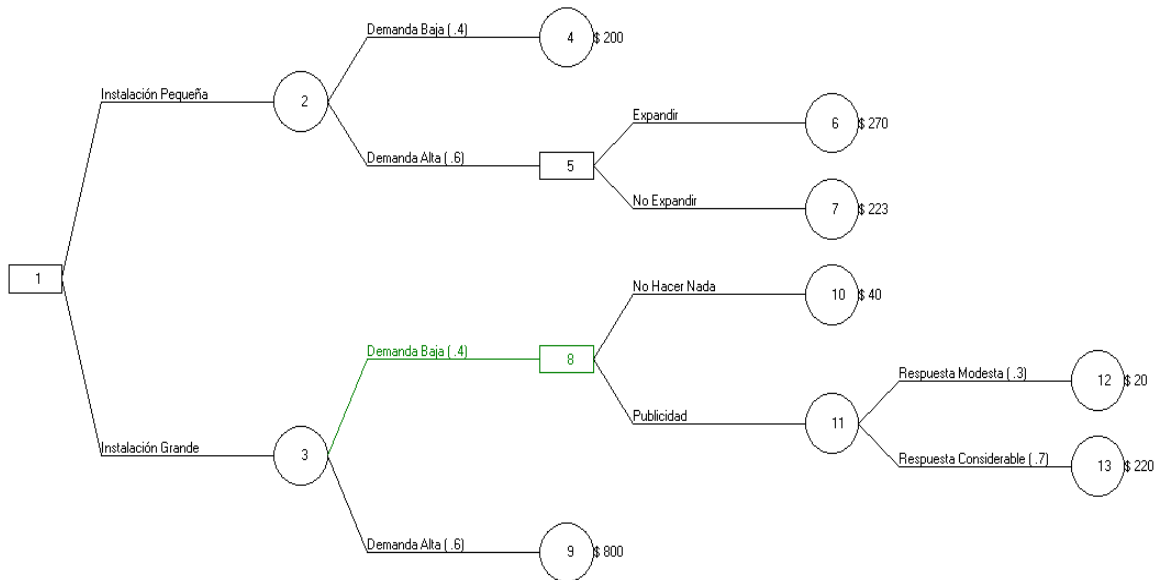
Esta herramienta funciona como un apoyo de decisión. Este utiliza un gráfico que se asemeja a un árbol incluyendo las posibles consecuencias y resultados de eventos aleatorios, el costo de los recursos y su utilidad. Un árbol de decisión es un método analítico que funciona a través de una representación esquemática de alternativas disponibles, que hacen más fácil el proceso de toma de decisiones, especialmente cuando existen riesgos, costos, beneficios o múltiples opciones.

Un árbol de decisión es particularmente útil cuando:

1. Las alternativas o cursos de acción se encuentran bien definidos. Por ejemplo: aceptar o rechazar una propuesta, aumentar o no la capacidad de producción, construir o no una nueva bodega, entre otros.
2. Las incertidumbres tengan la posibilidad de ser cuantificadas. Por ejemplo, probabilidad de éxito de una campaña publicitaria, probable efecto en ventas, probabilidad de pasar de etapas, entre otros.
3. Los objetivos están claros. Por ejemplo, aumentar ventas, maximizar utilizados, minimizar costos, entre otros.

Para los atributos discretos, el algoritmo realiza predicciones en las relaciones comprendidas entre las columnas de entrada de un conjunto de datos. Además, utiliza valores que son conocidos como estados, de estas columnas para predecir los estados de una columna que se denomina como elemento de predicción. Particularmente, el algoritmo logra identificar columnas de entrada que se relacionan con la columna de predicción. Para los atributos continuos, el algoritmo hace uso de la regresión lineal para determinar la división del árbol de decisión.

Si se emplea más de una columna como elemento de predicción, o los datos de entrada contienen una tabla anidada que se haya determinado como unidad de predicción, el algoritmo crea un árbol de decisión independiente para cada columna de predicción.



Fuente: Gestión de operaciones

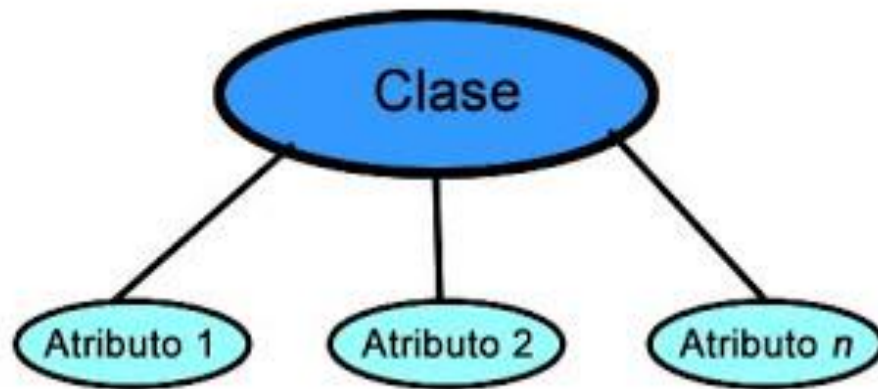
- **Bayes Naïve**

El algoritmo de Bayes naive es un algoritmo de clasificación que se basa en los teoremas de Bayes, y que, además, puede ser usado para el modelo de predicción y exploración. La palabra naive, del termino Bayes naive, proviene del hecho de que el algoritmo utilice técnicas bayesianas sin tener en cuenta las dependencias que puedan existir.

El algoritmo resulta ser menos complejo que otros algoritmos, por lo tanto, es más útil para generar modelos de minería de datos más rápidamente, que puedan detectar relaciones entre columnas de entrada y de predicción. Es posible utilizar el algoritmo para la realización de la exploración inaugural de datos, y posteriormente, emplear los resultados para la creación de modelos de minería de datos adicionales con otros algoritmos más complicados y precisos.

Para emplear el modelo de Bayes naive, son necesarios los siguientes requisitos:

1. Una columna de una sola clave: el modelo debe contener una columna numérica o de texto que identifique cada registro de manera única.
2. Columnas de entrada: todas las columnas deben ser discretas o haber discretizado los valores.
3. Las variables deben ser independientes: los atributos de entrada deben ser independientes unos de otros.
4. Al menos una columna de predicción: el atributo de predicción debe contener valores discretos.



Fuente: Algoritmos de minería de datos

- **Máquinas de vectores de soporte (SVM)**

Una máquina de vectores de soporte es un algoritmo supervisado que puede ser utilizado en clasificación binaria o regresión. Este algoritmo es muy utilizado en aplicaciones como el procesamiento del lenguaje natural, el habla, reconocimiento de imágenes y la visión artificial.

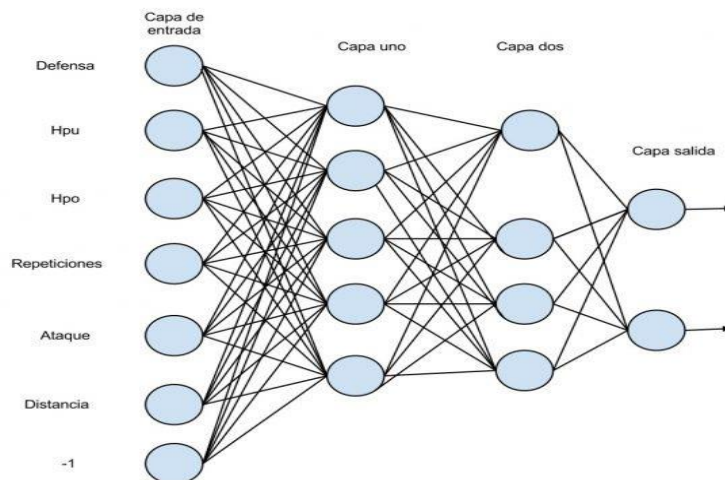
Una SVM crea un hiperplano óptimo en forma de plano de decisión, de tal modo que el borde de separación entre las dos clases de datos se extienda al máximo. Los vectores de soporte se refieren a un conjunto pequeño de las informaciones de entrenamiento que se manejan como soporte para la ubicación óptima de la superficie de decisión.

El entrenamiento de una máquina de vectores de soporte se compone de dos fases:

1. Transformar los predictores o datos de entrada, en un lugar que posea características altamente dimensionales. En esta fase, los datos no se transfiguran explícitamente al espacio de características. El proceso es distinguido como truco de Kernel.
2. Resolver un problema de optimización cuadrática que sea de arreglo a un hiperplano óptimo para catalogar las características transformadas en dos clases.

- **Redes neuronales**

El algoritmo de red neuronal consiste en una implementación de la arquitectura de red neuronal, adaptada al aprendizaje automático. El algoritmo experimenta el funcionamiento de cada posible estado del atributo de entrada, con cada uno de los posibles estados del atributo de predicción. Adicionalmente, calcula las posibilidades de cada composición según los datos de aprendizaje. Se puede utilizar para hacer uso de estas probabilidades en tareas de clasificación o regresión, al igual que para anunciar un resultado en función de algunos atributos de entrada. También se puede usar una red neuronal para el análisis de asociación.

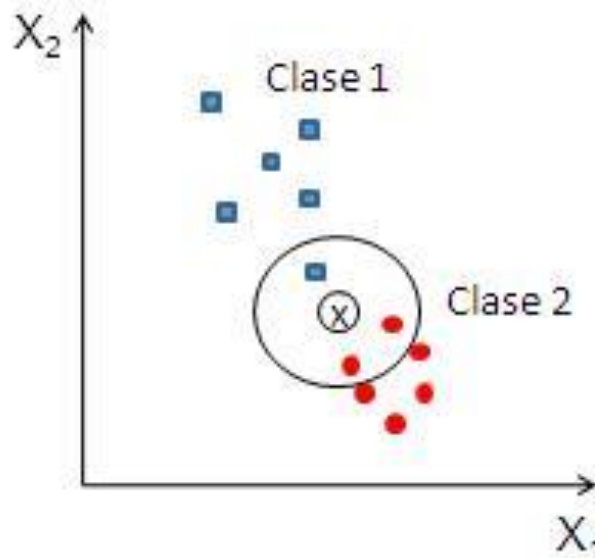


Fuente: Hipertextual

- **Regla de los K vecinos más cercanos**

La regla de los k vecinos más cercanos, consiste en un clasificador supervisado basado en reconocimiento de patrones criterios de vecindad, que se conoce como algoritmo de clasificación "K-NN". Este sistema tiene su punto de partida en la idea de que un nuevo ejemplar será encasillado en la clase a la cual concierna la mayor cuantía de vecinos más cercanos del conjunto de entrenamiento más cercano a esta.

Al aplicar la regla NN, se sumerge en el conocimiento almacenado en el conjunto de entrenamiento, para así determinar cuál será la clase a la que incumbe la nueva muestra, pero únicamente se tiene en cuenta el vecino colindante a ella, por lo que es natural pensar que es viable que no se esté aprovechando de forma eficaz toda la información que se podía extraer del conjunto de entrenamiento.



Fuente: Ecured

Existe una gran cantidad de algoritmos inmersos dentro de la clasificación y regresión. Si bien sus utilidades son distintas, comparten la esencia de generar gran utilidad en los procesos de aprendizaje supervisado.

Referencias:

- GEO tutoriales. (2016). Gestión de operaciones. Recuperado el 7 de octubre de 2018, de <https://www.gestiondeoperaciones.net/procesos/arbol-de-decision/>
- Mathworks. (s.f.). Mathworks. Recuperado el 6 de octubre de 2018, de <https://es.mathworks.com/discovery/svm-maquina-vectores-soporte.html>
- Microsoft. (2018). Microsoft. Recuperado el 7 de octubre de 2018, de <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017>
- Microsoft. (2018). Microsoft. Recuperado el 6 de octubre de 2018, de <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/microsoft-naive-bayes-algorithm?view=sql-server-2017>
- Microsoft. (2018). Microsoft. Recuperado el 6 de octubre de 2018, de <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/microsoft-neural-network-algorithm?view=sql-server-2017>