

» La Minería de Datos y sus algoritmos



Minería de Datos

se define como un conjunto de técnicas que sirven para analizar automáticamente la información.

La Minería de Datos se define como un conjunto de técnicas que sirven para analizar automáticamente la información, permitiendo descubrir conocimientos interesantes o patrones en los datos.

El Data Minig -así conocido en inglés- se entiende que son las técnicas de extracción de datos y están diseñadas para explicar o comprender el pasado; pero también para predecir el futuro. Dentro de sus finalizaciones, es la investigación lo que motiva el inicio de la prospección de datos; su fin será apoyar la toma de decisiones basadas en hechos reales, en lugar de intuiciones.

Se parte de varios algoritmos, los cuales son aplicados en la Minería de Datos que muestran los resultados que se quieren dentro de las técnicas de análisis. Los más utilizados son: **Árboles de decisión:** Los algoritmos de árbol de decisión consisten en organizar los datos en elecciones que compiten, formando ramas de influencia después de una decisión inicial. El tronco del árbol representa la decisión inicial, y empieza con una pregunta de sí o no, como tomar o no el desayuno. Tomar desayuno y no tomar desayuno serían las dos ramas divergentes del árbol, y cada elección posterior tendría sus propias ramas divergentes que llevan a un punto final.

El algoritmo K-means se basa en el análisis de grupos. Trata de dividir los datos recogidos en “bloques” (“clusters” en inglés) separados, agrupados por características comunes.

Máquinas de vectores de soporte: Los algoritmos de máquinas de vectores de soporte toman datos de entrada y predicen cuál de las dos posibles categorías incluyen los datos de entrada. Un ejemplo sería recoger los códigos postales de un grupo de votantes e intentar predecir si uno es demócrata o republicano.

**Técnicas de extracción de datos,
que están diseñadas para**

*explicar o
comprender
el pasado;
pero, también,
para predecir
el futuro.*

El algoritmo a priori: Normalmente controla los datos de transacciones. Por ejemplo, en una tienda de ropa, el algoritmo podría controlar qué camisas suelen comprar juntas los clientes.

El algoritmo EM: Este algoritmo define parámetros analizando los datos y predice la posibilidad de una salida futura o evento aleatorio dentro de los mismos parámetros de datos. Por ejemplo, el algoritmo EM podría intentar predecir el momento de una siguiente erupción de un géiser; según los datos de tiempo de erupciones pasadas.

Algoritmo PageRank: Es un algoritmo base para los motores de búsqueda. Puntúa y estima la relevancia de un trozo determinado de datos dentro de un gran conjunto, como un único sitio web, en un conjunto mayor de todos los sitios web de Internet.

Algoritmo AdaBoost: Su función, dentro de otros algoritmos de aprendizaje, es la de anticipar un



Se parte de varios algoritmos, los cuales son aplicados en la Minería de Datos, que muestran los resultados que se quieren, dentro de las técnicas de análisis.

comportamiento, según los datos observados, para que sea sensibles a extremos estadísticos. Aunque el algoritmo EM puede sesgarse, debido a un géiser que tiene dos erupciones en menos de un minuto, cuando normalmente tiene una erupción una vez al día, el algoritmo AdaBoost modificaría la salida del algoritmo EM, analizando la relevancia del extremo.

Algoritmo del vecino k más cercano: Este algoritmo reconoce patrones en la ubicación de los datos y los asocia a los datos mismos con un identificador mayor. Por ejemplo, si quieres asignar una oficina postal a cada ubicación geográfica del hogar y tienes un conjunto de datos para cada ubicación geográfica del hogar, el algoritmo del vecino k más cercano asignará las casas a la oficina postal más vecina, según su proximidad.

Naive Baye: El algoritmo predice la salida de una identidad, basándose en los datos de observaciones conocidas. Por ejemplo, si una persona tiene una altura de 6 pies y 6 pulgadas (1,97 m) y lleva una talla 14 de zapatos, el algoritmo Naive Baye podría predecir, con una determinada probabilidad, que la persona es un hombre.

Algoritmo CART: “CART” es una sigla en inglés que significa análisis de árbol regresivo y de

clasificación. Al igual que los análisis de árboles de decisión, organiza los datos según opciones que compiten, como si una persona ha sobrevivido a un terremoto. Al contrario que los algoritmos de árboles de decisión, que solo pueden clasificar una salida o una salida numérica basada en la regresión, el algoritmo CART puede usar los dos para predecir la probabilidad de un evento.

Por lo tanto, existen varias técnicas que permiten una Minería de Datos más exacta, acorde con la indicación del mercado de investigación, en la que se redirigen los datos por necesitar.

Referencias

Los mejores 10 algoritmos en minería de datos. Disponible en https://techlandia.com/mejores-10-algoritmos-mineria-datos-info_295108/. 2020



Ing. Daniela Madriz Pereira.
Bachiller Ingeniería en Informática.