Universidad mariano Gálvez de Guatemala

Sede Retalhuleu

Curso: Base de datos II

Catedrático: Jorge santos.



Algoritmo: Arboles de decisión

Nombre: Luis Eduardo castillo Dávila

No carne: 2890-13-19453

**Que es un árbol de decisión.**

Un árbol de decisión es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Los árboles de decisión son muy utilizados a la hora de analizar la toma de decisiones de individuos, son muy diversas las materias donde son utilizados, desde la economía hasta la informática se sirven de su uso. Estos árboles son muy útiles para visualizar las diversas opciones que se tienen y como llegar a ellas, además se les pueden aplicar métodos de inducción, como por ejemplo la inducción hacia atrás, gracias a los cuales mediante sencillos razonamientos se ve cómo se llegaría a la solución final.

**Elementos.**

Elementos de un Árbol de decisión

Los árboles de decisión están formados por nodos, vectores de números, flechas y etiquetas.

Cada nodo se puede definir como el momento en el que se ha de tomar una decisión de entre varias posibles, lo que va haciendo que a medida que aumenta el número de nodos aumente el número de posibles finales a los que puede llegar el individuo. Esto hace que un árbol con muchos nodos sea complicado de dibujar a mano y de analizar debido a la existencia de numerosos caminos que se pueden seguir.

Los vectores de números serían la solución final a la que se llega en función de las diversas posibilidades que se tienen, dan las utilidades en esa solución.

Las flechas son las uniones entre un nodo y otro y representan cada acción distinta.

Las etiquetas se encuentran en cada nodo y cada flecha y dan nombre a cada acción

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión para el modelado de predicción de atributos discretos y continuos.

Para los atributos Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por ejemplo, en un escenario para predecir qué clientes van a adquirir probablemente una bicicleta, si nueve de diez clientes jóvenes compran una bicicleta, pero solo lo hacen dos de diez clientes de edad mayor, el algoritmo infiere que la edad es un buen elemento de predicción en la compra de bicicletas. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto.

Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

Discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción.

**Cómo funciona el algoritmo**

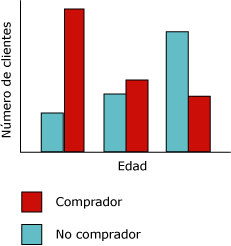
El algoritmo de árboles de decisión de Microsoft genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de SQL Server Data Minan algorithms to improve performance and the quality of analysis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si utiliza demasiados atributos de predicción o de entrada al diseñar un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Algunos métodos para determinar si hay que dividir el árbol son las métricas estándar del sector para la entropía y las redes Bayesianas. Para obtener más información sobre los métodos que se usan para seleccionar los atributos significativos y, después, puntuarlos y clasificarlos,

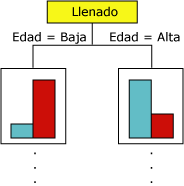
Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobre ajustado o sobreentrenado. Un modelo sobre ajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobre ajustar un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol.

**Predecir columnas discretas**

La forma en que el algoritmo de árboles de decisión de Microsoft genera un árbol para una columna de predicción discreta puede mostrarse mediante un histograma. El siguiente diagrama muestra un histograma que seguimiento una columna de predicción, Bike Buyers, con una columna de entrada, Age. El histograma muestra que la edad de una persona ayuda a distinguir si esa persona comprará una bicicleta.



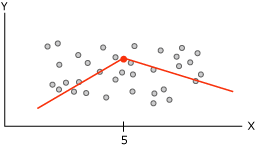
La correlación que aparece en el diagrama hará que el algoritmo de árboles de decisión de Microsoft cree un nuevo nodo en el modelo.



A medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

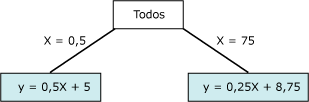
**Predecir columnas continuas**

Cuando el algoritmo de árboles de decisión de Microsoft genera un árbol basándose en una columna de predicción continua, cada nodo contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión. Por ejemplo, considere el siguiente diagrama.



En un modelo de regresión estándar, intentaría derivar una fórmula única que represente la tendencia y las relaciones de los datos como un todo. En cambio, una fórmula única puede hacer un trabajo insuficiente al capturar la discontinuidad en los datos complejos. En su lugar, el algoritmo de árboles de decisión de Microsoft busca los segmentos del árbol que son principalmente lineales y crea fórmulas independientes de estos segmentos. Al dividir los datos en segmentos diferentes, el modelo puede hacer un trabajo mejor de aproximación de datos.

En el siguiente diagrama se representa el diagrama de árbol del modelo en el gráfico de dispersión anterior. Para predecir el resultado, el modelo proporciona dos fórmulas diferentes: una para la bifurcación izquierda, con la fórmula y = .5x x 5 y otra para la bifurcación derecha, con la fórmula y = .25x + 8,75. El punto donde las dos líneas se unen en el gráfico de dispersión es el punto de no linealidad y donde se dividiría un nodo de un modelo de árbol de decisión.



Este es un modelo sencillo con solo dos ecuaciones lineales; por consiguiente, la división en el árbol se encuentra inmediatamente después del nodo **All.** En cambio, una división puede producirse en cualquier nivel del árbol. Eso significa que en un árbol que contenga varios niveles y nodos, donde cada nodo se caracteriza por una colección diferente de atributos, puede que se comparta una fórmula en varios nodos o se aplique solo a un único nodo. Por ejemplo, puede obtener una fórmula de un nodo que se defina como "clientes por encima de una determinada edad e ingresos" y otra en un nodo que represente "clientes que viajan largas distancias.

**Ejemplo:**

El departamento de marketing de la empresa Adventure Works Cycles desea identificar las características de los clientes antiguos que podrían indicar si es probable que realicen alguna compra en el futuro. La base de datos AdventureWorks2012 almacena información demográfica que describe a los clientes antiguos. Mediante el algoritmo de árboles de decisión de Microsoft que analiza esta información, el departamento puede generar un modelo que predice si un determinado cliente va a comprar productos, basándose en el estado de las columnas conocidas sobre ese cliente, como la demografía o los patrones de compra anteriores

**Datos requeridos para los modelos de árboles de decisión**

Cuando prepare los datos para su uso en un modelo de árboles de decisión, conviene que comprenda qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos se utilizan.

Los requisitos para un modelo de árbol de decisión son los siguientes:

**Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una columna de predicción. Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.

Columnas de entrada. Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

Para obtener información más detallada sobre los tipos de contenido y los tipos de datos compatibles con los modelos de árboles de decisión, vea la sección Requisitos de Referencia técnica del algoritmo de árboles

**El caso se estudio**

El caso se estudió hacia ventas y El ejemplo práctico y lo datos requeridos para los arboles de decisión a ventas serán adjuntados al documento con el video.