# Arboles de decisiones

Introducción a los Árboles de Decisión

Los árboles de decisión son modelos de aprendizaje supervisado utilizados tanto para clasificación como para regresión. Su estructura recuerda a un árbol invertido, donde cada nodo interno representa una "prueba" en un atributo, cada rama representa el resultado de la prueba, y cada hoja representa una clase o un valor de salida.

Conceptos Fundamentales:

1. Nodo raíz: Es el Nodo superior de árbol, que representa la variable inicial sobre la que se hará la primera partición de nuestro conjunto de datos.
2. Nodos Internos: Representan las “Pruebas” sobre los atributos (características) de los datos.
3. Hojas: Representan las predicciones o categorías finales.
4. Rama: son las conexiones entre nodos y representan el resultado de las pruebas de los nodos internos.

Diagrama

Descripción generada automáticamente

Construcción del Árbol

El proceso de construcción de un árbol de decisión implica seleccionar el mejor atributo en cada paso para dividir el conjunto de datos, con el objetivo de crear los subconjuntos más homogéneos posibles. Este proceso se basa en medidas de "pureza", como la ganancia de información o la disminución de la impureza de Gini.

En el campo de la ciencia de datos el concepto de “homogéneos” se refiere a conjuntos de datos que sean consistentes en sus características. Es decir los datos homogéneos tiene una estructura y propiedades similares. Lo que facilita los análisis y procesamientos. Imagina que tiene un conjunto de datos de pacientes de un hospital, un conjunto homogéneo podría ser una lista de pacientes que tiene la misma enfermedad, la misma edad o en un mismo rango de esta. Esto permitiría estudiar el efecto de un tratamiento especifico en ese grupo de manera mas precisa.

Para determinar cortes en nuestros datos que permitan capturar separar y capturar la información tenemos distintas formulas que nos ayudan a determinar esto:

Entropía:

La entropía mide la cantidad de desorden o impureza en un conjunto de datos. Se calcula como:

Donde:

es la proporción perteneciente a la clase 𝑖.

es el número de clases.

Ejemplo:

Imagina que tienes un conjunto de datos con 10 ejemplos, donde:

6 ejemplos son de la clase A

4 ejemplos son de la clase B

La entropía se calcularía como:

1. Calculo de proporciones:

2. Calculo de entropía:

Interpretación y Escala:

La entropía de 0.971 indica que el conjunto de datos es relativamente desordenado, es decir, las clases no están claramente separadas. La entropía varía de 0 (completa pureza, todos los ejemplos pertenecen a una sola clase) a 1 (máxima impureza, los ejemplos están uniformemente distribuidos entre las clases).

Ganancia de Información:

La ganancia de información mide la reducción en la entropía al dividir un conjunto de datos 𝑆 sobre un atributo 𝐴

Donde:

es el subconjunto de 𝑆 para el cual el atributo 𝐴 tiene valor 𝑣.

Ejemplo:

Imagina que tiene un atributa A con dos valores posibles Y . El conjunto S se divide en:

: 4 ejemplos (3 de clase A,1 de clase B)

: 6 Ejemplos (3 de clase A, 3 de clase B)

Primero calculamos la entropía para estos subconjuntos:

Una vez con la entropía podemos calcular la ganancia de información:

Interpretación y Escala:

La ganancia de información de 0.0466 es relativamente baja, lo que sugiere que la división del conjunto de datos usando el atributo 𝐴 no reduce mucho el desorden. Valores más altos indican una mejor división, con una ganancia de información máxima cuando se alcanza una pureza completa en los subconjuntos.

Índice de Gini

El índice de Gini mide la impureza de un conjunto de datos 𝑆 y se define como:

Donde:

es la proporción perteneciente a la clase 𝑖.

Ejemplo:

Usando el mismo conjunto de datos 𝑆 con 10 ejemplos:

6 clase A.

4 clase B.

Primero calculamos las proporciones de las clases.

Ahora podemos calcular el índice de Gini:

Interpretación y Escala:

Un índice de Gini de 0.48 indica que el conjunto de datos tiene una impureza moderada. El índice de Gini varía de 0 (completa pureza) a 0.5 (máxima impureza en un conjunto de dos clases igualmente distribuidas).

Disminución de la Impureza de Gini

La disminución de la impureza de Gini se calcula como:

Con los mismos subconjuntos

Índice de Gini de los subconjuntos:

Disminución de la impureza de Gini:

Interpretación y Escala:

Una disminución de la impureza de Gini de 0.03 es baja, lo que sugiere que el atributo 𝐴 no proporciona una buena división para reducir la impureza del conjunto de datos. Valores más altos indican una mejor división.

Entonces pasemos a la parte conceptual general para entender como es que funcionan los arboles de decisiones:

1. Calcular la métrica de pureza inicial: Antes de dividir los datos, se calcula la métrica de pureza (como la entropía o el índice de Gini) del conjunto de datos completo, considerando la distribución de la variable objetivo en todo el conjunto de datos.

2. Calcular la métrica de pureza después de la división: Cuando se evalúa un atributo para dividir los datos, se calcula la métrica de pureza en los subconjuntos resultantes después de aplicar la división. Aquí, la variable objetivo se utiliza para determinar la distribución de clases en cada subconjunto, lo que a su vez se utiliza para calcular la ganancia de información o la disminución de la impureza de Gini.

3. Seleccionar el mejor corte: El atributo que proporcione la mayor ganancia de información o la mayor disminución de la impureza de Gini se selecciona como el atributo de división. En este punto, la variable objetivo se utiliza para calcular la ganancia de información o la disminución de la impureza de Gini y determinar la división óptima.

4. Dividir los datos y construir el árbol: Una vez que se selecciona un atributo de división, se dividen los datos en subconjuntos basados en el valor de ese atributo. Este proceso de división y selección de atributos se repite recursivamente para construir el árbol de decisiones.

5. Predicción: Una vez que el árbol está construido, se puede utilizar para predecir la variable objetivo de nuevas instancias. Para hacer una predicción, se sigue el camino desde la raíz del árbol hasta una hoja, siguiendo las reglas de división basadas en los valores de las características de la instancia, y se predice la clase mayoritaria en la hoja como la predicción final.

Cuando llega un nuevo registro para predecir, el modelo sigue las reglas establecidas en el árbol para determinar la clase de la nueva instancia. Comienza en la raíz del árbol y se mueve a través de las ramas siguiendo las condiciones establecidas por los cortes en cada nodo (determinados por las fórmulas) hasta llegar a una hoja, donde se hace la predicción final basada en la mayoría de clases de los ejemplos de entrenamiento en esa hoja. Este proceso refleja la forma en que las fórmulas de las ganancias de información o disminución de la impureza de Gini influyen en la estructura y las reglas del árbol para realizar predicciones precisas.

Fundamentos de la Toma de Decisiones

Los árboles de decisión son modelos predictivos simples y fáciles de interpretar que descomponen un conjunto de datos en subgrupos más pequeños basados en decisiones tomadas en cada nodo. Cada nodo del árbol representa una característica del conjunto de datos, y las ramas representan las posibles salidas o valores de esa característica. Entender cómo se construyen y funcionan estos árboles es fundamental para el análisis y la interpretación de datos, así como para la construcción de modelos más complejos.

Transición a Modelos Más Avanzados:

1. Árboles Aleatorios (Random Forests)

Ensamblaje de Árboles: Los árboles aleatorios son un conjunto (ensamble) de múltiples árboles de decisión. Cada árbol en el bosque es construido a partir de una muestra aleatoria del conjunto de datos original, y sus predicciones son combinadas para mejorar la precisión y reducir el sobreajuste.

2. XGBoost (Extreme Gradient Boosting)

Boosting: XGBoost es una técnica de boosting, que construye árboles de decisión secuencialmente, cada uno intentando corregir los errores del anterior. Este enfoque iterativo ajusta los pesos de los ejemplos mal clasificados, mejorando progresivamente el modelo.

Optimización y Eficiencia: XGBoost implementa diversas optimizaciones, como la regularización para evitar el sobreajuste, la ponderación de los ejemplos y técnicas de paralelización para mejorar la velocidad y eficiencia. Comprender la mecánica de cómo se construyen y ajustan los árboles de decisión básicos es crucial para entender cómo estos principios son ampliados y mejorados en XGBoost.

Entender los árboles de decisión no solo te proporciona las herramientas necesarias para construir y evaluar estos modelos básicos, sino que también sienta las bases para abordar y dominar modelos más avanzados como los árboles aleatorios y XGBoost. Esta comprensión integral permite a los científicos de datos mejorar la precisión de sus predicciones, optimizar el rendimiento de sus modelos y abordar problemas de sobreajuste y eficiencia con mayor eficacia.

A continuación, te dejo el enlace de mi GitHub con su implementación en Python en el Notebook realizamos un modelo de árboles de decisiones teniendo en cuenta las fórmulas y conceptos aquí explicados. Y lo comparamos con un árbol de decisiones de la librería sklearn la cual ya está programada. Espero que este contenido te sea de utilidad y no dudes en dejarme su sugerencias e inquietudes. Gracias por tu tiempo.