Fundamentos Matemáticos dos Algoritmos de Machine Learning: Árvores de Decisão

Seu Nome

21 de junho de 2025

Resumo

Este documento investiga os fundamentos das Árvores de Decisão, um algoritmo de aprendizado supervisionado amplamente utilizado para tarefas de classificação e regressão. A análise se concentra nos critérios matemáticos de divisão de nós, como o Índice de Gini e a Entropia, que são usados para construir a árvore de forma recursiva, buscando maximizar a pureza dos nós.

1 Introdução às Árvores de Decisão

As Árvores de Decisão são modelos de aprendizado de máquina que se assemelham a um fluxograma, onde cada "nó" interno representa um teste em um atributo, cada "ramo"representa o resultado do teste, e cada "nó folha" (terminal) representa uma decisão final (uma classe ou um valor).

Sua principal vantagem é a alta interpretabilidade. A lógica da árvore pode ser visualizada e entendida facilmente, tornando-a uma ferramenta poderosa para análise exploratória de dados e para explicar as decisões do modelo a stakeholders não-técnicos. Elas podem lidar tanto com dados categóricos quanto numéricos.

2 A Construção da Árvore: Divisão Recursiva

Uma árvore de decisão é construída a partir do topo (nó raiz) para baixo, através de um processo de divisão recursiva. O algoritmo busca, a cada passo, encontrar o melhor atributo e o melhor valor de corte para dividir o conjunto de dados em subconjuntos que sejam os mais "puros" possíveis.

Mas como medir a "pureza" de um nó? A matemática por trás das árvores de decisão reside nas funções que quantificam a impureza (ou a mistura de classes) em um conjunto de dados. O objetivo do algoritmo é encontrar a divisão que resulta na maior redução da impureza.

As duas métricas mais comuns para medir a impureza em problemas de classificação são o **Índice de Gini** e a **Entropia**.

3 Critérios de Divisão

Seja S um conjunto de dados (um nó) com amostras de C classes diferentes. Seja p_i a proporção de amostras da classe i no conjunto S.

$$p_i = \frac{\text{Número de amostras da classe } i \text{ em } S}{\text{Número total de amostras em } S}$$

A soma de todas as proporções é, obviamente, 1: $\sum_{i=1}^{C} p_i = 1$.

3.1 Índice de Gini (Gini Impurity)

O Índice de Gini mede a probabilidade de um elemento, escolhido aleatoriamente do conjunto, ser classificado incorretamente se sua classe fosse atribuída aleatoriamente de acordo com a distribuição de classes no conjunto.

A fórmula para o Índice de Gini é:

$$Gini(S) = 1 - \sum_{i=1}^{C} (p_i)^2$$

- **Pureza Máxima:** Se o nó é puro (todas as amostras pertencem a uma única classe), existe um $p_i = 1$ e todos os outros são 0. Nesse caso, $Gini(S) = 1 - 1^2 = 0$.
- **Impureza Máxima:** Para um problema com C classes, a impureza é máxima quando as amostras estão uniformemente distribuídas, ou seja, $p_i = 1/C$ para todas as classes. Nesse caso, o valor do Gini se aproxima de $1 - 1/C$. Para 2 classes, o máximo é 0.5.

3.2 Entropia (Entropy)

A Entropia é um conceito da Teoria da Informação que mede o grau de incerteza ou desordem em um conjunto. Em nosso contexto, ela mede a impureza de um nó.

A fórmula para a Entropia é:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Por convenção, $0 \log_2(0)$ é definido como 0.

- **Pureza Máxima:** Se o nó é puro ($p_i = 1$), a entropia é $H(S) = -1 \log_2(1) = 0$.
- **Impureza Máxima:** A entropia é máxima quando as amostras estão uniformemente distribuídas ($p_i = 1/C$). Para 2 classes, o máximo é $-2 \cdot (0.5 \log_2 0.5) = 1$.

3.3 Ganho de Informação (Information Gain)

Quando usamos Entropia, o critério de decisão para a divisão é o **Ganho de Informação**. Ele mede a redução na entropia que obtemos ao dividir um nó S em subconjuntos S_v com base em um atributo A.

O Ganho de Informação é calculado como:

$$IG(S, A) = H(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} H(S_v)$$

Onde:

- $H(S)$ é a entropia do nó pai (antes da divisão).
- $\sum \frac{|S_v|}{|S|} H(S_v)$ é a **média ponderada da entropia** dos nós filhos.
- $|S_v|$ é o número de amostras no subconjunto v e $|S|$ é o número total de amostras.

O algoritmo calcula o Ganho de Informação para todos os atributos possíveis e escolhe aquele que **maximiza** essa medida, pois uma maior redução na entropia significa uma divisão mais "informativa". Uma lógica similar de cálculo de "ganho"pode ser aplicada ao Índice de Gini.

4 O Processo de Construção e Poda

O algoritmo de construção (como o ID3 ou o CART) segue estes passos:

1. Começa com todo o conjunto de dados no nó raiz.
2. Calcula a impureza (Gini ou Entropia) para todas as divisões possíveis.
3. Escolhe a divisão (atributo e valor) que resulta na maior redução de impureza (maior Ganho de Informação ou menor Gini ponderado).
4. Divide o nó em dois ou mais nós filhos.
5. Repete os passos 2 a 4 para cada nó filho de forma recursiva.

A recursão para quando uma condição de parada é atingida, como:

- O nó se torna puro (impureza = 0).
- A profundidade máxima da árvore é alcançada.
- O número de amostras em um nó é menor que um limiar mínimo.

Se a árvore crescer demais, ela pode se ajustar perfeitamente aos dados de treino, incluindo ruídos, um fenômeno chamado **overfitting**. Para evitar isso, técnicas de **poda (pruning)** são aplicadas, removendo ramos que fornecem pouco poder preditivo.

5 Conclusão

As Árvores de Decisão são modelos intuitivos e poderosos, cuja base matemática reside na otimização de métricas de pureza, como o Índice de Gini e a Entropia. Ao buscar recursivamente as divisões que mais reduzem a impureza dos dados, o algoritmo constrói uma estrutura hierárquica de regras de decisão. Compreender como essas métricas funcionam é fundamental para ajustar os hiperparâmetros do modelo, controlar sua complexidade e extrair insights valiosos e interpretáveis dos dados.