Arvore de decisão

April 29, 2025

0.1 Árvore de Decisão

Árvores de decisão são uma ferramenta popular e poderosa usada em vários campos, como aprendizado de máquina, mineração de dados e estatística. Elas fornecem uma maneira clara e intuitiva de tomar decisões com base em dados, modelando as relações entre diferentes variáveis. Este artigo aborda o que são árvores de decisão, como elas funcionam, suas vantagens e desvantagens e suas aplicações.

0.1.1 O que é uma Árvore de Decisão?

Uma árvore de decisão é uma estrutura semelhante a um fluxograma usada para tomar decisões ou fazer previsões. Ela consiste em nós que representam decisões ou testes em atributos, ramificações que representam o resultado dessas decisões e nós folha que representam resultados finais ou previsões. Cada nó interno corresponde a um teste em um atributo, cada ramificação corresponde ao resultado do teste e cada nó folha corresponde a um rótulo de classe ou um valor contínuo.

0.1.2 Terminologias de Árvore de Decisão

Existem termos especializados associados às árvores de decisão que denotam vários componentes e facetas da estrutura da árvore e do procedimento de tomada de decisão:

- **Nó Raiz**: O nó raiz de uma árvore de decisão representa a escolha ou característica original da qual a árvore se ramifica. É o nó mais alto.
- **Nós Internos (Nós de Decisão)**: Nós na árvore cujas escolhas são determinadas pelos valores de atributos específicos. Esses nós têm ramificações que levam a outros nós.
- **Nós Folha (Nós Terminais)**: Os terminais dos ramos, onde as escolhas ou previsões são decididas. Não há mais ramificações nesses nós.
- **Ramificações (Arestas)**: Links entre nós que mostram como as decisões são tomadas em resposta a circunstâncias específicas.
- **Divisão**: O processo de dividir um nó em dois ou mais subnós com base em um critério de decisão. Envolve selecionar um atributo e um limite para criar subconjuntos de dados.
- **Nó Pai**: Um nó que é dividido em nós filhos. É o nó original do qual uma divisão se origina.
- **Nó Filho**: Nós criados como resultado de uma divisão de um nó pai.
- **Critério de Decisão**: A regra ou condição usada para determinar como os dados devem ser divididos em um nó de decisão. Envolve comparar valores de atributos com um limite.

- **Poda**: O processo de remover ramos ou nós de uma árvore de decisão para melhorar sua generalização e evitar overfitting.

0.1.3 Funcionamento das Árvores de Decisão

O processo de criação de uma árvore de decisão envolve:

1. **Selecionar o Melhor Atributo**: Usando uma métrica como o indice de Gini, entropia ou ganho de informação, o melhor atributo para dividir os dados é selecionado.
2. **Dividir o Conjunto de Dados**: O conjunto de dados é dividido em subconjuntos com base no atributo selecionado.
3. **Repetir o Processo**: O processo é repetido recursivamente para cada subconjunto, criando um novo nó interno ou nó folha até que um critério de parada seja atendido (por exemplo, todas as instâncias em um nó pertencem à mesma classe ou uma profundidade predefinida é atingida).

0.1.4 Métricas para Divisão

- **Indice de Gini**: Mede a probabilidade de uma classificação incorreta de uma nova instância se ela for classificada aleatoriamente de acordo com a distribuição de classes no conjunto de dados.

$$\text{Índice de Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

onde p_i é a probabilidade de uma instância ser classificada em uma classe específica.

- **Entropia**: Mede a quantidade de incerteza ou impureza no conjunto de dados.

$$\text{Entropia} = 1 - \sum_{i=1}^{n} p_i \log_2 (p_i)$$

onde p_i é a probabilidade de uma instância ser classificada em uma classe específica.

- **Ganho de Informação**: Mede a redução na entropia ou indice de Gini depois que um conjunto de dados é dividido em um atributo.

$$\text{Ganho de Informação} = \text{Entropia}_{\text{pai}} - \sum_{i=1}^{n} \left(\frac{|D_i|}{|D|} * \text{Entropia}(D_i) \right)$$

onde D_i é o subconjunto de D após a divisão por um atributo.

0.1.5 Vantagens das Árvores de Decisão

- **Simplicidade e Interpretabilidade**: Árvores de decisão são fáceis de entender e interpretar. A representação visual espelha de perto os processos de tomada de decisão humana.
- **Versatilidade**: Pode ser usada para tarefas de classificação e regressão.

- **Não há Necessidade de Dimensionamento de Recursos**: As árvores de decisão não exigem normalização ou dimensionamento dos dados.
- **Trabalha com Relacionamentos Não Lineares**: Capaz de capturar relacionamentos não lineares entre atributos e variáveis dpendentes (alvo).

0.1.6 Desvantagens das Árvores de Decisão

- **Sobretreinamento**: Árvores de decisão podem facilmente sofrer overfitting dos dados de treinamento, especialmente se forem profundas com muitos nós.
- **Instabilidade**: Pequenas variações nos dados podem resultar na geração de uma árvore completamente diferente.
- **Distorção em Relação a Recursos com Mais Níveis**: Recursos com mais níveis podem dominar a estrutura da árvore.

0.1.7 Poda

Para superar o sobretreinamento, técnicas de poda são usadas. A poda reduz o tamanho da árvore removendo nós que fornecem pouco poder na classificação de instâncias.

Existem dois tipos principais de poda:

- **Pré-poda (Parada Precoce)**: Impede o crescimento da árvore quando ela atende a certos critérios (por exemplo, profundidade máxima, número mínimo de amostras por folha).
- **Pós-poda**: Remove galhos de uma árvore totalmente crescida que não fornecem energia significativa.

0.1.8 Aplicações de Árvores de Decisão

- **Tomada de Decisão Empresarial**: Usada no planejamento estratégico e alocação de recursos.
- **Assistência Médica**: Auxilia no diagnóstico de doenças e sugere planos de tratamento.
- **Finanças**: Ajuda na pontuação de crédito e na avaliação de risco.
- **Marketing**: Usada para segmentar clientes e prever o comportamento deles.

0.2 Como uma Árvore de Decisão é Formada

O processo de formação de uma árvore de decisão envolve particionar recursivamente os dados com base nos valores de diferentes atributos. O algoritmo seleciona o melhor atributo para dividir os dados em cada nó interno, com base em certos critérios, como ganho de informação ou impureza de Gini. Esse processo de divisão continua até que um critério de parada seja atendido, como atingir uma profundidade máxima ou ter um número mínimo de instâncias em um nó folha.

0.2.1 Por que Usar Árvores de Decisão

Árvores de decisão são amplamente usadas em aprendizado de máquina por várias razões:

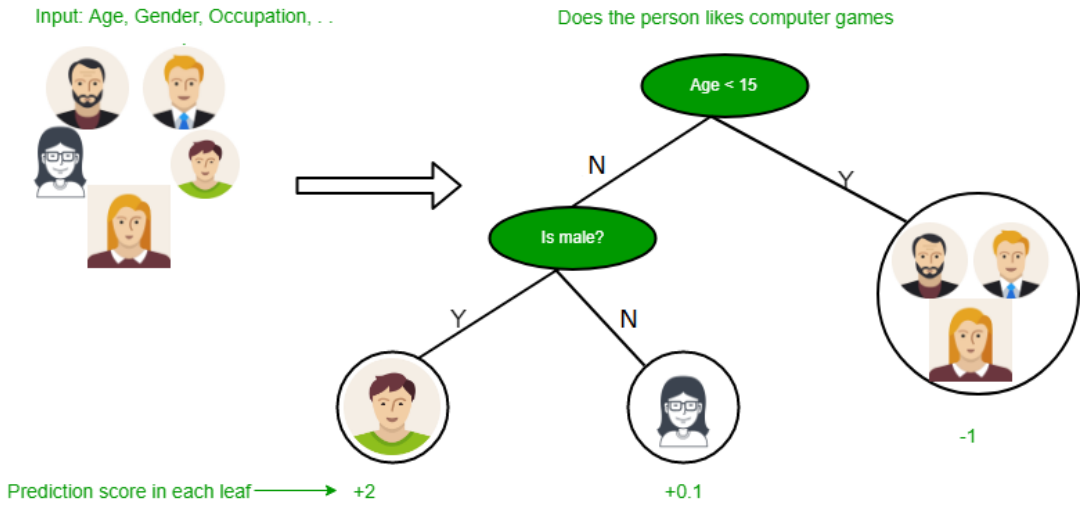
- **Versatilidade**: Árvores de decisão são extremamente versáteis na simulação de processos de tomada de decisão complexos, devido à sua interpretabilidade e flexibilidade.
- **Estrutura Hierárquica**: Sua estrutura hierárquica permite a representação de cenários de escolha complexos que levam em conta uma variedade de causas e resultados.
- **Compreensão do Lógica de Decisão**: Elas fornecem insights compreensíveis sobre a lógica de decisão, sendo especialmente úteis para tarefas de classificação e regressão.
- **Dados Numéricos e Categóricos**: São proficientes com dados numéricos e categóricos, e podem se adaptar facilmente a uma variedade de conjuntos de dados graças à sua capacidade autônoma de seleção de características.
- **Visualização Simples**: Árvores de decisão também oferecem uma visualização simples, o que ajuda a compreender e explicar os processos de decisão subjacentes em um modelo.

0.3 Abordagem de Árvore de Decisão

A árvore de decisão usa a representação em árvore para resolver problemas, onde cada nó folha corresponde a um rótulo de classe e os atributos são representados nos nós internos da árvore. Podemos representar qualquer função booleana em atributos discretos usando a árvore de decisão.

[1]:
```
from IPython.display import Image
Image(filename='decision.png')
```

[1]:



0.3.1 Suposições ao usar a Árvore de Decisão

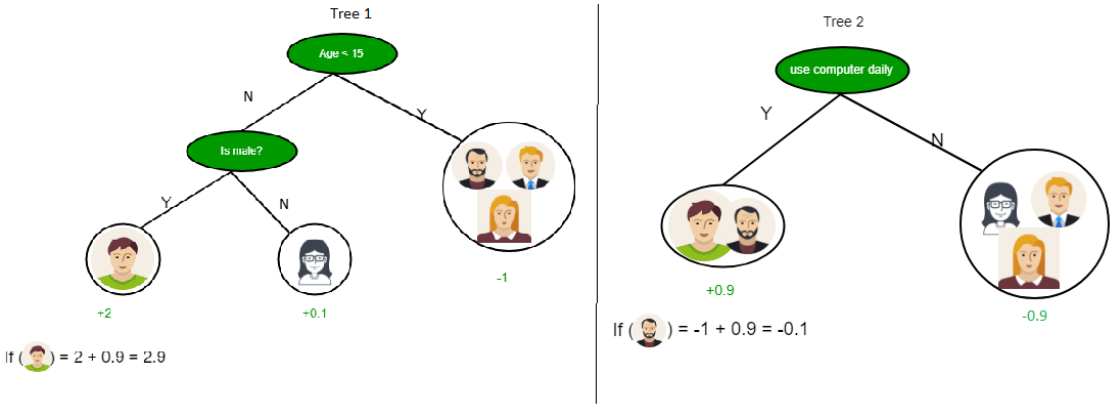
No início, consideramos todo o conjunto de treinamento como a raiz.

- Os valores das características são preferencialmente categóricos. Se os valores forem contínuos, eles são discretizados antes de construir o modelo.
- Com base nos valores dos atributos, os registros são distribuídos recursivamente.

- Usamos métodos estatísticos para ordenar os atributos como raiz ou nó interno.

[2]:
```
Image(filename='decisiontree.png')
```

[2]:



0.4 Funcionamento da Árvore de Decisão

A árvore de decisão funciona na forma de Soma de Produtos, também conhecida como Forma Normal Disjuntiva. No exemplo, estamos prevendo o uso do computador na vida diária das pessoas. O maior desafio na árvore de decisão é a identificação do atributo para o nó raiz em cada nível. Esse processo é conhecido como seleção de atributos. Temos duas medidas populares de seleção de atributos:

1. **Ganho de Informação**: Quando usamos um nó em uma árvore de decisão para particionar as instâncias de treinamento em subconjuntos menores, a entropia muda. O ganho de informação é uma medida dessa mudança na entropia.

- Suponha que S seja um conjunto de instâncias,
- A seja um atributo,
- S_v seja o subconjunto de S
- v representa um valor individual que o atributo A pode assumir e
- Valores (A) é o conjunto de todos os valores possíveis de A, então

$$\text{Ganho}\,(S, A) = \text{Entropia}\,(S) - \sum_{v}^{A} \frac{|S_v|}{|S|} \cdot \text{Entropia}\,(S_v)$$

2. Entropia A entropia é a medida de incerteza de uma variável aleatória. Ela caracteriza a impureza de uma coleção arbitrária de exemplos. Quanto maior a entropia, maior o conteúdo de informação.

Suponha que S seja um conjunto de instâncias, A seja um atributo, S_v seja o subconjunto de S com $A = v$, e $Valores(A)$ seja o conjunto de todos os valores possíveis de A, então:

$$\text{Ganho}\,(S, A) = \text{Entropia}\,(S) - \sum_{v \,\in\, \text{Valores}(A)} \frac{|S_v|}{|S|} \cdot \text{Entropia}\,(S_v)$$

0.4.1 Exemplo

Para o conjunto $X = a, a, a, b, b, b, b, b$

Total de instâncias: 8

Instâncias do tipo b: 5

Instâncias do tipo a: 3

A entropia $H(x)$ é calculada da seguinte forma:

$$H(x) = -\left[\left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right) + \left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right)\right]$$

$$H(x) = -\left[0,375 \cdot (-1,415) + 0,625 \cdot (-0,678)\right]$$

$$H(x) = -\left(-0,53 - 0,424\right) = 0,954$$

0.5 Construindo uma Árvore de Decisão Usando Ganho de Informação

0.5.1 Essenciais:

- Comece com todas as instâncias de treinamento associadas ao nó raiz.
- Use o ganho de informação para escolher qual atributo rotular em cada nó.
- Nota: Nenhum caminho da raiz à folha deve conter o mesmo atributo discreto duas vezes.
- Construa recursivamente cada subárvore no subconjunto de instâncias de treinamento que seriam classificadas por aquele caminho na árvore.
- Se todas as instâncias de treinamento restantes forem positivas ou negativas, rotule o nó como "sim" ou "não", respectivamente.
- Se não houver mais atributos, rotule com uma votação majoritária das instâncias de treinamento restantes naquele nó.
- Se não houver instâncias restantes, rotule com uma votação majoritária das instâncias de treinamento do nó pai.

0.5.2 Exemplo:

Vamos desenhar uma Árvore de Decisão para os seguintes dados usando ganho de informação.

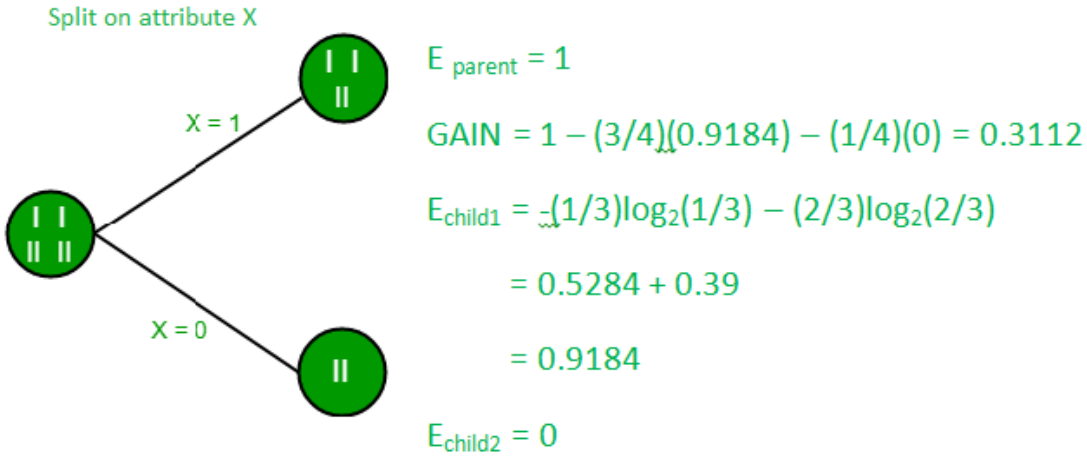
Conjunto de treinamento: 3 características e 2 classes.

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Aqui, temos 3 características e 2 classes de saída. Para construir uma árvore de decisão usando ganho de informação, tomaremos cada uma das características e calcularemos a informação para cada uma.

[4]:
```
Image(filename='tr4.png')
```

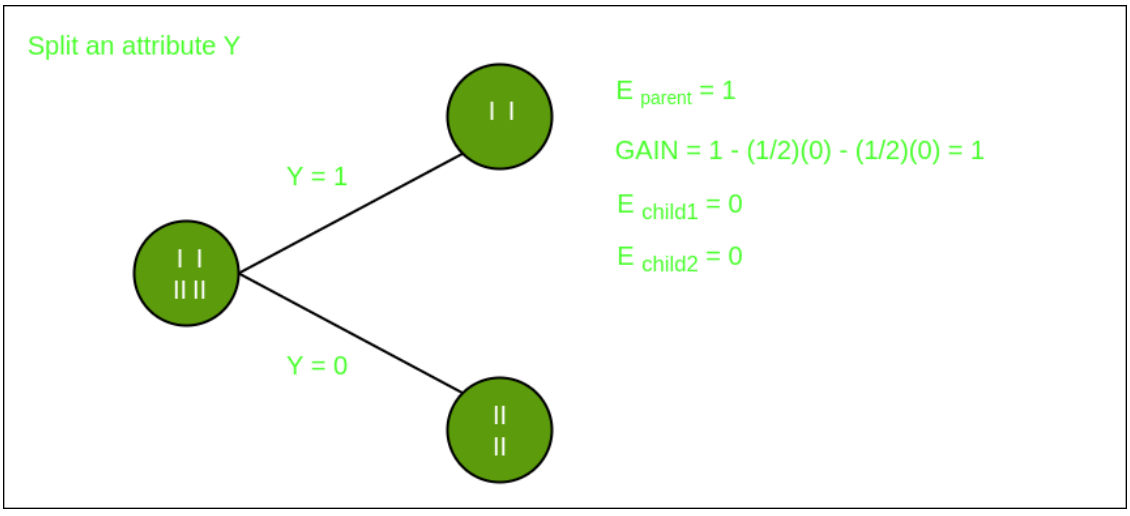
[4]:



Divisão na característica X

[5]:
```
Image(filename='y-attribute.png')
```

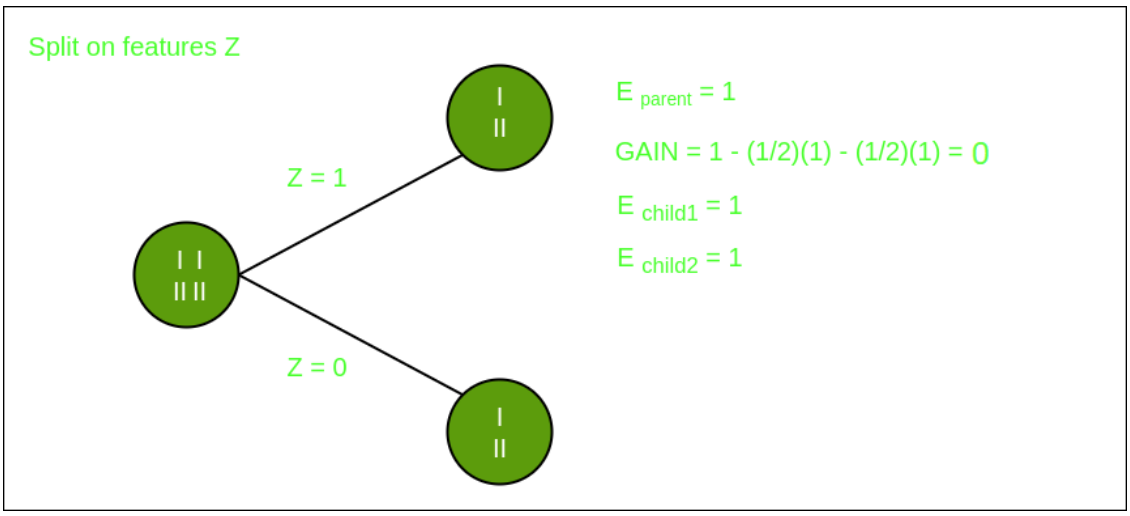
[5]:



Divisão na característica Y

[6]:
```
Image(filename='z-attribute.png')
```

[6]:

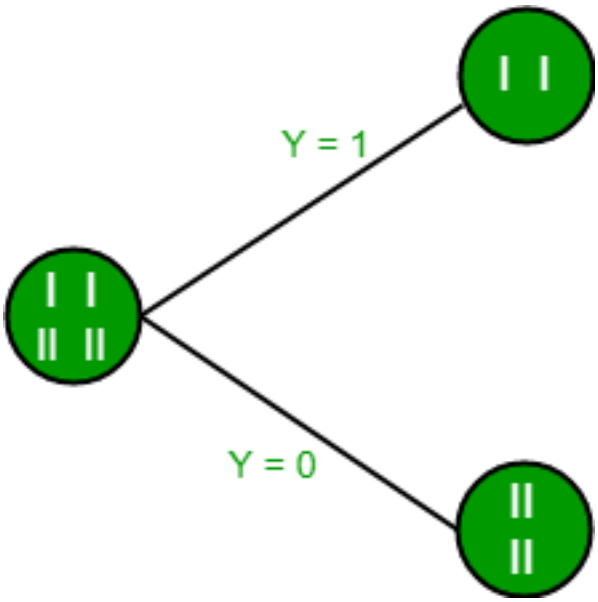


Divisão na característica Z

A partir das imagens acima, podemos ver que o ganho de informação é máximo quando fazemos uma divisão na característica Y. Portanto, para o nó raiz, a característica mais adequada é a Y. Agora podemos ver que, ao dividir o conjunto de dados pela característica Y, o filho contém um subconjunto puro da variável alvo. Então, não precisamos dividir mais o conjunto de dados. A árvore final para o conjunto de dados acima ficaria assim:

[7]:
```
Image(filename='tr6.png')
```

[7]:



0.6 Índice de Gini

O Índice de Gini é uma métrica para medir com que frequência um elemento escolhido aleatoriamente seria identificado incorretamente. Isso significa que um atributo com um índice de Gini mais baixo deve ser preferido. A fórmula para o cálculo do Índice de Gini é dada por:

$$\textit{Índice de Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

O Índice de Gini é uma medida da desigualdade ou impureza de uma distribuição, comumente usada em árvores de decisão e outros algoritmos de aprendizado de máquina. Ele varia de 0 a 0,5, onde 0 indica um conjunto puro (todas as instâncias pertencem à mesma classe) e 0,5 indica um conjunto maximamente impuro (instâncias distribuídas uniformemente entre as classes).

Características Adicionais do Índice de Gini:

- Ele é calculado somando as probabilidades quadradas de cada resultado em uma distribuição e subtraindo o resultado de 1.
- Um Índice de Gini mais baixo indica uma distribuição mais homogênea ou pura, enquanto um Índice de Gini mais alto indica uma distribuição mais heterogênea ou impura.
- Em árvores de decisão, o Índice de Gini é usado para avaliar a qualidade de uma divisão medindo a diferença entre a impureza do nó pai e a impureza ponderada dos nós filhos.
- Comparado a outras medidas de impureza como a entropia, o Índice de Gini é mais rápido de calcular e mais sensível a mudanças nas probabilidades das classes.
- Uma desvantagem do Índice de Gini é que ele tende a favorecer divisões que criam nós filhos de tamanho igual, mesmo que não sejam ótimos para a precisão da classificação.

Na prática, a escolha entre usar o Índice de Gini ou outras medidas de impureza depende do problema específico e do conjunto de dados, e muitas vezes requer experimentação e ajuste.

0.7 Implementando uma Arvore de Decisão em Python