Como gerar uma árvore de decisão

Assuntos abordados:

- **Parte 2:** Entendimento geral
- **Parte 3:** Como calcular

Entendimento geral:

Tabela para exemplos de como gerar uma árvore de decisão.

Exemplo	Alternativo	Bar	Sex/Sab	fome	Cliente	Preço	Chuva	Res	Tipo	Tempo	Vai esperar?
X1	Sim	Não	Não	Sim	Alguns	RRR	Não	Sim	Francês	0-10	Sim
x2	Sim	Não	Não	Sim	Cheio	R	Não	Não	Tailandês	30-60	Não
x3	Não	Sim	Não	Não	Alguns	R	Não	Não	Hamburger	0-10	Sim
x4	Sim	Não	Sim	Sim	Cheio	R	Sim	Não	Tailandês	10-30	Sim
X5	Sim	Não	Sim	Não	Cheio	RRR	Não	Sim	Francês	>60	Não
X6	Não	Sim	Não	Sim	Alguns	RR	Sim	Sim	Italiano	0-10	Sim
X7	Não	Sim	Não	Não	Nenhum	R	Sim	Não	Hamburger	0-10	Não
X8	Não	Não	Não	Sim	Alguns	RR	Sim	Sim	Tailandês	0-10	Sim
X9	Não	Sim	Sim	Não	Cheio	R	Sim	Não	Hamburger	>60	Não
X10	Sim	Sim	Sim	Sim	Cheio	RRR	Não	Sim	Italiano	10-30	Não
X11	Não	Não	Não	Não	Nenhum	R	Não	Não	Tailandês	0-10	Não
X12	Sim	Sim	Sim	Sim	Cheio	R	Não	Não	Hamburger	30-60	Sim

Como calcular:

Na maioria dos indutores das árvores, as funções de divisão discreta são univariadas, isto é, um nó interno é dividido de acordo com o valor de um único atributo. Com isso, o indutor procura o melhor atributo sobre o qual realizar a divisão.

Entropia: é o cálculo do ganho de informação baseado em uma medida utilizada na teoria da informação. Caracterizada pela impureza dos dados, em um conjunto de dados, é uma medida de falta de homogeneidade dos dados de entrada em relação a sua classificação.

Dado um conjunto de dados de entrada (S) que pode ter c classes distintas, a entropia de S será dada por:

$$Entropia(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Assim, a entropia é uma medida de **aleatoriedade** de uma variável.

Escolha de testes de atributos:

- O esquema de escolha dos atributos é projetado para minimizar a profundidade da árvore final. A ideia é escolher o atributo que vá o mais longe possível na tentativa de fornecer uma classificação exata dos exemplos. Um atributo perfeito divide os exemplos em conjuntos que são todos positivos ou todos negativos.
  - Clientes → bastante bom
  - Tipo → realmente inútil
- A função ESCOLHER-ATRIBUTO deverá ter seu valor máximo quando o atributo for perfeito, e seu valor mínimo quando o atributo for absolutamente inútil.
- Uma medida apropriada é a quantidade esperada de informações fornecidas pelo atributo, que é calculada através de uma expressão matemática.
- Sendo assim, após escolher um atributo para a raiz, a conjunto é dividido em subconjuntos onde o passo a passo é reiniciado, criando-se assim uma árvore de decisão.
- **Ganho de informação:** a partir do teste de atributo é a diferença entre o requisito de informação original e o novo requisito. Para calcular o ganho de determinado atributo utiliza-se a fórmula:

$$ganho(atributo) = Entropia(classe) - Entropia(atributo)$$

Após isso, a heurística usada na função ESCOLHER-ATRIBUTO é simplesmente escolher o atributo com o maior ganho.

Quais são os passos para calcular o ganho de um determinado atributo:

Deve-se levar em consideração as seguintes informações:

- Entropia(0, 1) = 0
- Entropia (1/2, 1/2) = 1

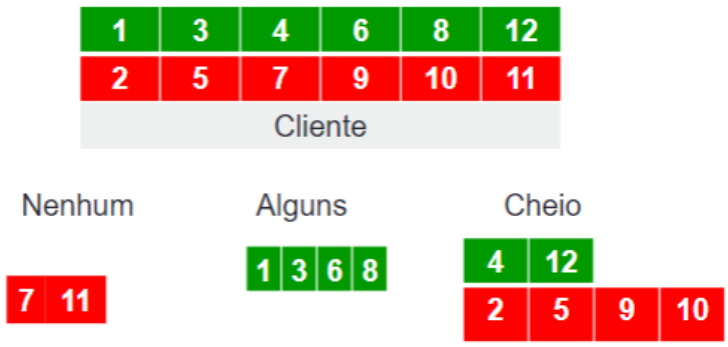
Calcular **Entropia(classe):**

- Pegar todas as instâncias → TOTAL
- Pegar a quantidade de unidades de cada instância → UNIDADE_i
- Calcular a entropia da quantidade de unidades/total →

  Entropia(UNIDADE_i/TOTAL, UNIDADE_i+1/TOTAL)

Calcular **Entropia(atributo):**

- Pegar a quantidade de instâncias de cada ramo. Por exemplo, suponha que tenha uma aresta nenhum, alguns e cheio saindo do atributo cliente e tem a possibilidade de instâncias como Sim ou Não, como mostra a figura a baixo. Deve-se então calcular para cada aresta a quantidade de instâncias relacionadas a ela pelo total.



INCIDÊNCIAS:

Nenhum → 2/12

Alguns → 4/12

Cheio → 6/12

- Após isso, deve-se calcular em uma fórmula para descobrir a entropia(cliente), para cada instância, aresta que saí do atributo, calcular a -***INCIDÊNCIA * Entropia(classificação)***. Para o exemplo vamos utilizar a imagem a cima, suponha que você tem no Nenhum 2 instâncias e as duas são dadas como não então você tem uma entropia a ser a calculada da seguinte forma → ***Entropia(0, 1)***, pois você só tem ***1 atributo a ser avaliado***. Você poderia também ter a possibilidade dos atributos serem igual, sendo assim → ***Entropia(1/2, 1/2)***, para a mesma situação, caso tivesse ***1 Sim e 1 Não***. Porém, caso a quantidade de ***instancias/total_instancias_atributo*** sejam diferentes, como o exemplo do cheio deve-se utilizar uma fórmula para calcular a Entropia:

  X = instancia_1/total_instancias_atributo

  Y = instancia_2/total_instancias_atributo

  *-X * log2(X) - Y * log2(Y)*, quantidade que for necessária basta apenas adicionar a parte em vermelho a mais.

Após calcular todas as Entropias necessárias - classe e atributo - basta substituir os valores na fórmula de ganho que você terá o ganho de determinado atributo.

E para escolher o atributo a ser avaliado no nível dois da árvore?

Para isso, você terá que fazer tudo novamente, porém você vai retirar da tabelas aquelas linhas que já foram avaliadas. Porém caso aconteça de os valores de ganho serem iguais nos atributos você deve levar em consideração a quantidade de instâncias relacionadas.

A ideia base do algoritmo:

- Escolher um atributo
- Estender a árvore adicionando um ramo para cada valor do atributo.

- Passar o exemplo para as folhas (tendo em conta o valor do atributo escolhido)
- Para cada folha:
  - Se todos os exemplos são da mesma classe, associar essa classe à folha 2.
  - Senão, repetir os passos 1 a 4.

Algoritmo → ID3:

```
ID3(Exemplos, Atributo-objetivo, Atributos)
    // ID3 retorna uma árvore de decisão que classifica corretamente os Exemplos determinados
    // Exemplos são os exemplos de treinamento.
    // Atributo-objetivo é o atributo cujo valor deve ser predito pela árvore.
    // Atributos são uma lista de outros atributos que podem ser testados pela árvore de decisão.
    Início
        Crie um nodo Raiz para a árvore
        Se todos os Exemplos são positivos
            Então retorna a Raiz da árvore com o rótulo = sim
        Se todos os Exemplos são negativos
            Então retorna a Raiz da árvore com o rótulo = não
        Se Atributos for vazio
            Então retorna a Raiz da árvore com o rótulo = valor mais comum do Atributo-objetivo em Exemplos
            Senão
                A ← um atributo de Atributos que melhor classifica Exemplos  (atributo de decisão)
                Raiz ← A   (rótulo = atributo de decisão A)
                Para cada possível valor v_i de A  faça
                    Acrescenta um novo arco abaixo da Raiz, correspondendo à resposta A = v_i
                    Seja Exemplos_vi o subconjunto de Exemplos que têm valor v_i para A
                    Se Exemplos_vi for vazio
                        Então acrescenta na extremidade do arco um nodo folha
                                com rótulo = valor mais comum do Atributo-objetivo em Exemplos
                        Senão acrescenta na extremidade do arco a sub árvore
                                ID3(Exemplos_vi, Atributo-objetivo, Atributos - [A])
        Retorna Raiz  (aponta para a árvore)
    Fim
```

Limitações algoritmo ID3:

- Só lida com atributos discretos. Neste caso, os atributos devem ser discretizados.
- Não apresenta alguma forma de tratar valores desconhecidos, ou seja, todos os exemplos do conjunto de treinamento deve ter valores conhecidos para todos os seus atributos.
- O algoritmo não lida com nenhum mecanismo pós-poda, o que poderia amenizar em árvores mais complexas.

Algoritmo C4.5:

É uma extensão do Algoritmo ID3, ele apresenta uma série de melhorias.

- Lidar com atributos contínuos e discretos.
- Lidar com dados de treinamento com atributos incompletos.
- Poda de árvore após a criação. Para aquelas ramificações que não ajudam no processo de decisão e substitui estes ramos por nós folha.
- Implementa o ‘gain ratio’ ao invés do ‘ganho de informação’ tradicional.

O que é Gain Ratio: expressa a proporção de informação gerada pela partição que é útil, ou seja, que aparenta ser útil para a classificação.