Machine Learning
Árvore de decisão: ID3

Prof. Hugo de Paula

Visão geral do algoritmo de ID3 (C4.5)

1. Crie um nó N associado à base de dados **B**
   - SE todos os registros de **B** pertencem à mesma classe **C**
     ENTÃO transforme em nó folha rotulado por **C**.
   - SENÃO SE **CAND = {}** ENTÃO transforme **N** numa folha etiquetada
     com o valor **C = max(count(atributo-classe(A))**
   - SENÃO seleciona atributo-teste **A = max(Ganho(CAND))** e rotule **N** com o nome de atributo-teste **A**

Visão geral do algoritmo de ID3 (C4.5)

2. Partição das amostras de **B**

- PARA cada valor s_i do atributo-teste FAÇA:
- Crie um nó-filho N_i, ligado a **N** por um ramo rotulado pelo valor s_i e associe a este nó uma sub-base B_i tal que o **atributo-teste = s_i**
- SE B_i = {} ENTÃO transforme o nó N_i numa folha etiquetada com o valor **C = max(count(atributo-Classe(A))**
- SENÃO calcule **Arvore(B_i , CAND – (atributo-teste))** e associe ao nó N_i

Seleção de atributos: Ganho de informação

- Dados categóricos (número de categorias = v)
- Entropia: $E(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$
- Usando o atributo **A**, a base de dados **B** será particionada em conjuntos S_i. A quantidade de informação final será:

$$I(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

- Ganho de informação: $G(A) = I(p,n) - I(A)$

Seleção de atributos: Taxa de ganho

- Dados categóricos ou contínuos (C4.5)
- Usado para dados categóricos com muitos valores:
  - Ganho da informação tende a produzir overfitting nesses casos.

Seleção de atributos: Índice Gini

- Dados contínuos (*IBM IntelligentMiner*).
- Se uma base **B** contém amostras de **N** classes:

$$gini(B) = 1 - \sum_{j=1}^{n} {p_j}^2$$

onde p_j é a frequência relativa da classe *j* em **B**.

- Se B é particionada nas subclasses $\mathbf{B_1}$ e $\mathbf{B_2}$ com tamanhos $\mathbf{N_1}$ e $\mathbf{N_2}$:

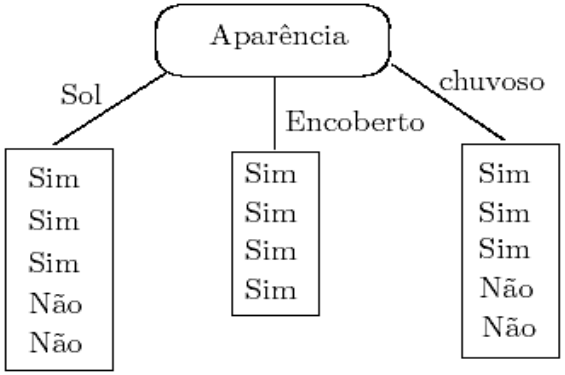
$$gini_{part}(B) = \frac{N_1}{N} gini(B_1) + \frac{N_2}{N} gini(B_2)$$

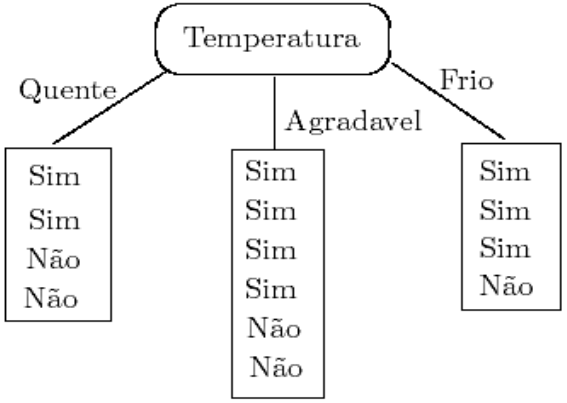
Árvore de decisão ID3: exemplo

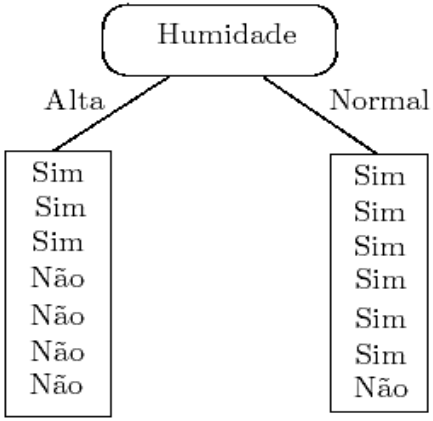
Aparência	Temperatura	Umidade	Vento	Jogar
Ensolarado	Quente	Alta	Fraco	Não
Ensolarado	Quente	Alta	Forte	Não
Nublado	Quente	Alta	Fraco	Sim
Chuvoso	Moderado	Alta	Fraco	Sim
Chuvoso	Frio	Normal	Fraco	Sim
Chuvoso	Frio	Normal	Forte	Não
Nublado	Frio	Normal	Forte	Sim
Ensolarado	Moderado	Alta	Fraco	Não
Ensolarado	Frio	Normal	Fraco	Sim
Chuvoso	Moderado	Normal	Fraco	Sim
Ensolarado	Moderado	Normal	Forte	Sim
Nublado	Moderado	Alta	Forte	Sim
Nublado	Quente	Normal	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Não

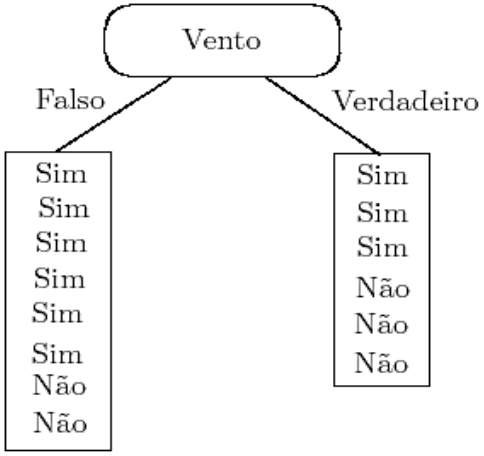
O objetivo é identificar quais as condições ideais para se jogar um determinado jogo.

Árvore de decisão: exemplo









As quatro possibilidades para o atributo do nó raiz.

Critério de escolha intuitivo: atributo que produz os nós mais puros.

Entropia do atributo Aparência

$$I(Aparencia) = \frac{5}{14}E(Folha_1) + \frac{4}{14}E(Folha_2) + \frac{5}{14}E(Folha_3)$$

$$E(Folha_1) = \frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$E(Folha_2) = \frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$E(Folha_3) = \frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

logo: $I(Aparencia) = \frac{5}{14}0.971 + \frac{4}{14}0 + \frac{5}{14}0.971 = 0.693$

Ganho da informação

Entropia do atributo **Temperatura**:

$$I(Temperatura) = \frac{4}{14}E(Folha_1) + \frac{6}{14}E(Folha_2) + \frac{4}{14}E(Folha_3) = 0.911$$

Entropia do atributo **Humidade**:

$$I(Humidade) = \frac{7}{14}E(Folha_1) + \frac{7}{14}E(Folha_2) = 0.788.$$

Ganho da informação

$$I(B) = \frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

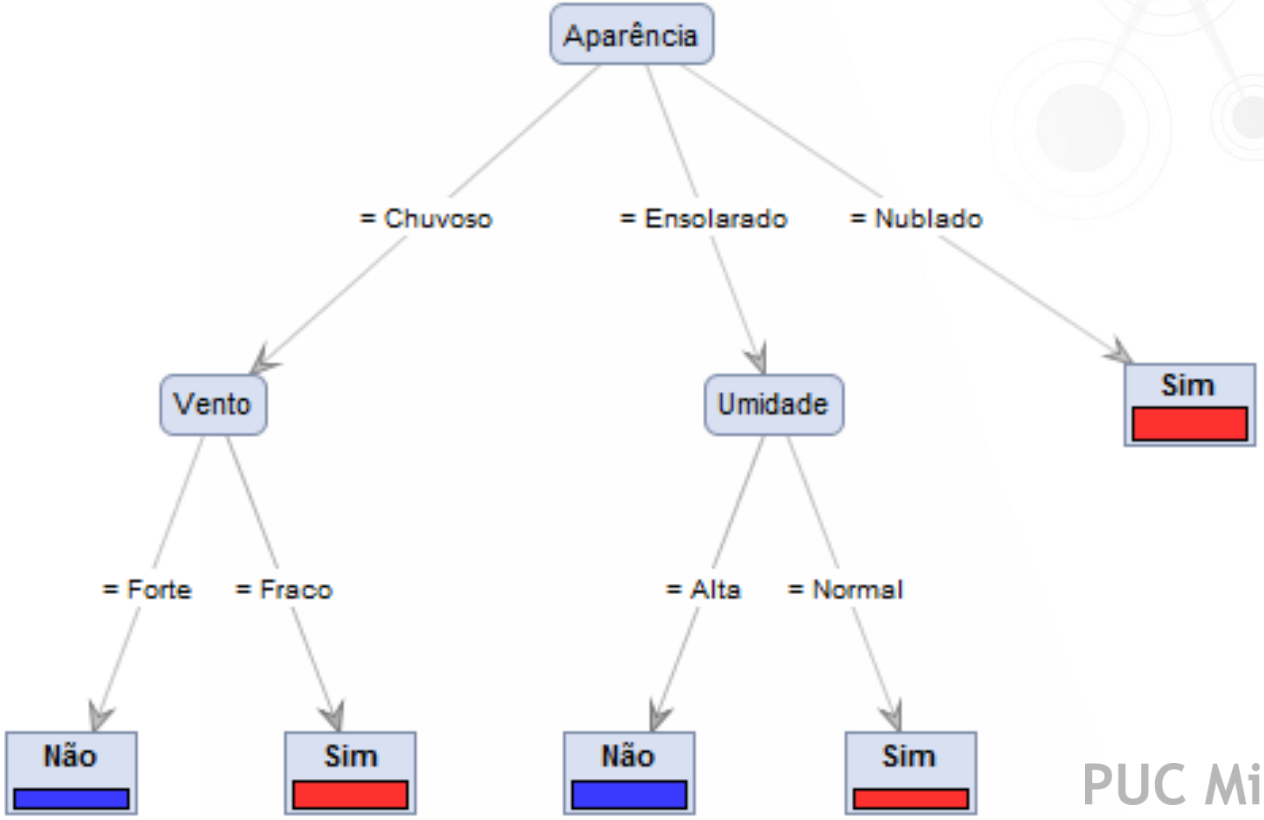
$$\boldsymbol{G(Aparencia) = 0.940 - 0.693 = 0.247}$$

$$G(Tempertura) = 0.940 - 0.911 = 0.029$$

$$G(Humidade) = 0.940 - 0.788 = 0.152$$

$$G(Vento) = 0.940 - 0.892 = 0.020$$

Resultado final da árvore



Aviso legal

O material presente nesta apresentação foi produzido a partir de informações próprias e coletadas de documentos obtidos publicamente a partir da Internet. Este material contém ilustrações adquiridas de bancos de imagens de origem privada ou pública, não possuindo a intenção de violar qualquer direito pertencente à terceiros e sendo voltado para fins acadêmicos ou meramente ilustrativos. Portanto, os textos, fotografias, imagens, logomarcas e sons presentes nesta apresentação se encontram protegidos por direitos autorais ou outros direitos de propriedade intelectual.

Ao usar este material, o usuário deverá respeitar todos os direitos de propriedade intelectual e industrial, os decorrentes da proteção de marcas registradas da mesma, bem como todos os direitos referentes a terceiros que por ventura estejam, ou estiveram, de alguma forma disponíveis nos slides. O simples acesso a este conteúdo não confere ao usuário qualquer direito de uso dos nomes, títulos, palavras, frases, marcas, dentre outras, que nele estejam, ou estiveram, disponíveis.

É vedada sua utilização para finalidades comerciais, publicitárias ou qualquer outra que contrarie a realidade para o qual foi concebido. Sendo que é proibida sua reprodução, distribuição, transmissão, exibição, publicação ou divulgação, total ou parcial, dos textos, figuras, gráficos e demais conteúdos descritos anteriormente, que compõem o presente material, sem prévia e expressa autorização de seu titular, sendo permitida somente a impressão de cópias para uso acadêmico e arquivo pessoal, sem que sejam separadas as partes, permitindo dar o fiel e real entendimento de seu conteúdo e objetivo. Em hipótese alguma o usuário adquirirá quaisquer direitos sobre os mesmos.

O usuário assume toda e qualquer responsabilidade, de caráter civil e/ou criminal, pela utilização indevida das informações, textos, gráficos, marcas, enfim, todo e qualquer direito de propriedade intelectual ou industrial deste material.

PUC Minas

Virtual