Objectivos

- Adquirir a noção de árvore de decisão
- Descrever o algoritmo ID3 para aprendizagem de árvores de decisão

- Sumário:
  - Árvores de decisão
    - Introdução
    - Algoritmo ID3-C4.5

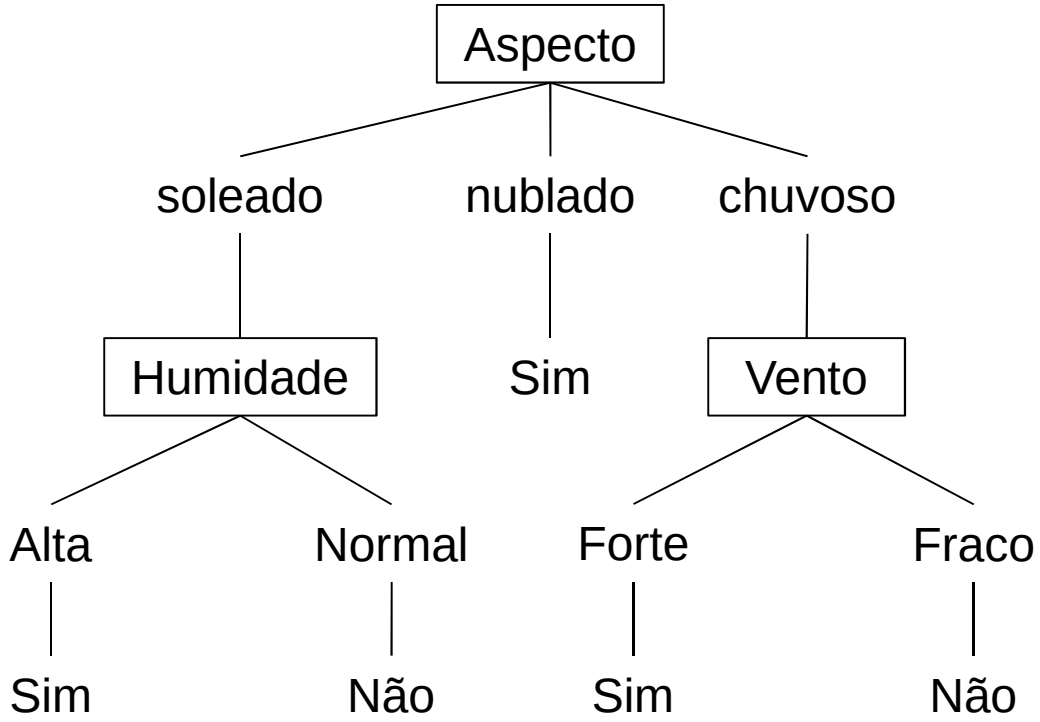
Árvores de decisão (*decision trees*)

- Encontram-se entre os métodos de inferência inductiva mais populares
- Têm sido aplicadas com sucesso a uma ampla gama de problemas de aprendizagem que vai desde o diagnóstico médico até à análise de risco na concessão de créditos

Árvores de decisão (*decision trees*)

- As saídas são apresentadas na forma de ***árvores de decisão***
- Uma ***árvore de decisão*** é uma estrutura que pode ter dois tipos de nós:
  - *Nós de decisão*: consiste numa pergunta (ou teste) relativa ao valor de um atributo. Cada nó possui tantos ramos quanto as possíveis respostas à pergunta
  - *Nós folha*: em cada um só pode haver instâncias pertencentes à mesma classe

Árvores de decisão (*decision trees*)



Árvores de decisão (*decision trees*)

- Em geral uma árvore de decisão corresponde a uma *disjunção de conjunções*
- Cada trajecto *raiz – folha* corresponde a uma *conjunção* de testes sobre os atributos e a própria árvore a uma *disjunção* das referidas conjunções
- Pode ser representada através de um *conjunto de regras*

Árvores de decisão (*decision trees*)

- Se adaptam a problemas com as seguintes características
  - Enquadrados no marco da aprendizagem supervisionada
  - A função objectivo possui valores de saída discretos (classificação)
  - Os valores dos atributos podem ser discretos ou numéricos. Valores numéricos tratados nas extensões do algoritmo básico
  - Requerem descrições disjuntivas
  - Os dados de treino podem conter erros
  - Os dados de treino podem conter valores de atributos omissos

Árvores de decisão (*decision trees*)

- Existem vários algoritmos para a aprendizagem de árvores de decisão
- Um dos mais representativos é o ***ID3***, introduzido por Ross Quinlan (1943 - …), e a sua extensão ***C4.5***

Descrição ID3

- A aprendizagem se realiza através de uma busca descendente (*top-down*) no *espaço* formado pelas *possíveis árvores de decisão*
- Se realiza uma *busca em subida de encosta* que começa com um conjunto vazio e avança de forma recursiva até a elaboração de uma árvore que classifique adequadamente os exemplos analisados
- Na busca se aplica o princípio de "dividir para vencer". A árvore é dividida em forma recursiva em subárvores, buscando-se a maior homogeneidade possível nestas
- O processo se realiza até que cada partição contenha exemplos pertencentes a uma única classe

Descrição ID3

- A heurística seguida é a de escolher em cada nodo de decisão o atributo que tenha maior capacidade de discriminação sobre os exemplos de treino
- Isto favorece
  - A selecção de árvores curtas em detrimento das mais longas
  - A selecção de árvores que colocam atributos com maiores ganhos de informação próximos da raiz

Descrição ID3

- O algoritmo inicia determinando, através dum teste estatístico, que atributo deve ser testado primeiro, sendo o melhor atributo utilizado como nó raiz da árvore
- É criado um descendente do nó raiz para cada possível valor do correspondente atributo e os exemplos de treino são associados ao nodo descendente apropriado
- O processo é repetido utilizando-se os exemplos de treino associados a cada nó descendente para determinar o melhor atributo a ser testado nesse ponto da árvore

Ganho de informação

- A questão central no algoritmo tem a ver com a selecção de qual atributo testar em cada nó da árvore
- Se utiliza uma propriedade estatística, designada *ganho de informação*, que mede quão bem um atributo dado separa os exemplos de treino, de acordo ao objectivo da classificação
  - Mede a efectividade de um atributo ao classificar os dados de treino
- É testado o atributo correspondente ao maior *ganho de informação*

Entropia

- Medida utilizada na Teoria da Informação que caracteriza o grau de (im)pureza de uma colecção arbitrária de exemplos

$$Ent(E) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad p_i = \frac{n_i}{n}$$

  - Por definição se utiliza nos cálculos $0.\log_2 0 = 0$
  - A entropia é igual a 0 se todos os exemplos pertencem à mesma classe
  - A entropia é igual a 1 se no conjunto existe a mesma quantidade de exemplos positivos e negativos

Ganho de informação

- É a redução esperada na entropia devido à partição dos exemplos de acordo a um atributo dado
- O ganho de informação de um atributo A com relação a uma colecção de exemplos E se define como

$$G(E,A) = Ent(E) - \sum_{v \in valores(A)} \frac{n_v}{n} Ent(E_v)$$

$$E_v = \{e \in E \mid A(e) = v\}$$

Algoritmo

- Função ID3
- Entradas
  - E : conjunto de exemplos de treino
  - AO : atributo objectivo
  - A : lista dos atributos a ser testados pela árvore aprendida
- Saída
  - R : árvore de decisão que classifica correctamente os exemplos apresentados
- Criar a raiz da árvore R
- Se todos os exemplos forem positivos, retorna R, etiqueta = +
- Se todos os exemplos forem negativos, retorna R, etiqueta = -
- Se $A \neq \phi$, retorna R, etiqueta = valor mais comum de AO em E

Algoritmo

- Senão inicio
  - $A_i \leftarrow$ melhor-atributo(A, E)
  - Atributo de decisão para R $\leftarrow A_i$
  - Para cada possível valor, v_i, de A_i,
    - Adicionar novo ramo sob R, correspondente ao teste $A_i = v_i$
    - E_{vi} = subconjunto de E, tal que $A_i = v_i$
    - Se E_{vi} = Vazio
      - Sob o novo ramo, adicionar nó folha, etiqueta = valor mais comum de AO em E
    - Senão
      - Sob o novo ramo, adicionar subárvore ID3(E_{vi}, AO, A – {A_i})
- Fim
- Devolver R

Exemplo

- Se dispõe de dados acerca de uma loja na Internet. A informação existente é relativa aos atributos:
  - Sítio de acesso: 0 – internacional, 1 – nacional, 2 – local
  - Primeira quantidade gasta: 0 – menos de 1000 kz, 1 – entre 1000 e 10000 kz, 2 – mais de 10000 kz
  - Zona de vivenda: 0 – internacional, 1 – nacional, 2 – local
  - Última compra: livro, disco
  - Classe: bom, mau

Exemplo

#	Aces so	1º Gasto	Vivenda	Compra	Classe
1	1	0	2	Livro	Bom
2	1	0	1	Disco	Mau
3	1	2	0	Livro	Bom
4	0	2	1	Livro	Bom
5	1	1	1	Livro	Mau
6	2	2	1	Livro	Mau

$$G(E,A) = Ent(E) - \sum_{v \in valores(A)} \frac{n_v}{n} Ent(E_v)$$

$$I(E,A_1) = \sum_{v \in (0,1,2)} \frac{n_v}{n} Ent(E_v)$$

$$E = [3B, 3M] \qquad E_1 = [2B, 2M]$$

$$E_0 = [1B, 0M] \qquad E_2 = [0B, 1M]$$

$$I(E,A_1) = \frac{1}{6} Ent(E_0) + \frac{4}{6} Ent(E_1) + \frac{1}{6} Ent(E_2)$$

$$Ent(E_0) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$$

$$Ent(E_1) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Ent(E_2) = -\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1} = 0$$

$$I(E,A_1) = \frac{1}{6}.0 + \frac{4}{6}.1 + \frac{1}{6}.0 = 0{,}66$$

Exemplo

#	Acesso	1º Gasto	Vivenda	Compra	Classe
1	1	0	2	Livro	Bom
2	1	0	1	Disco	Mau
3	1	2	0	Livro	Bom
4	0	2	1	Livro	Bom
5	1	1	1	Livro	Mau
6	2	2	1	Livro	Mau

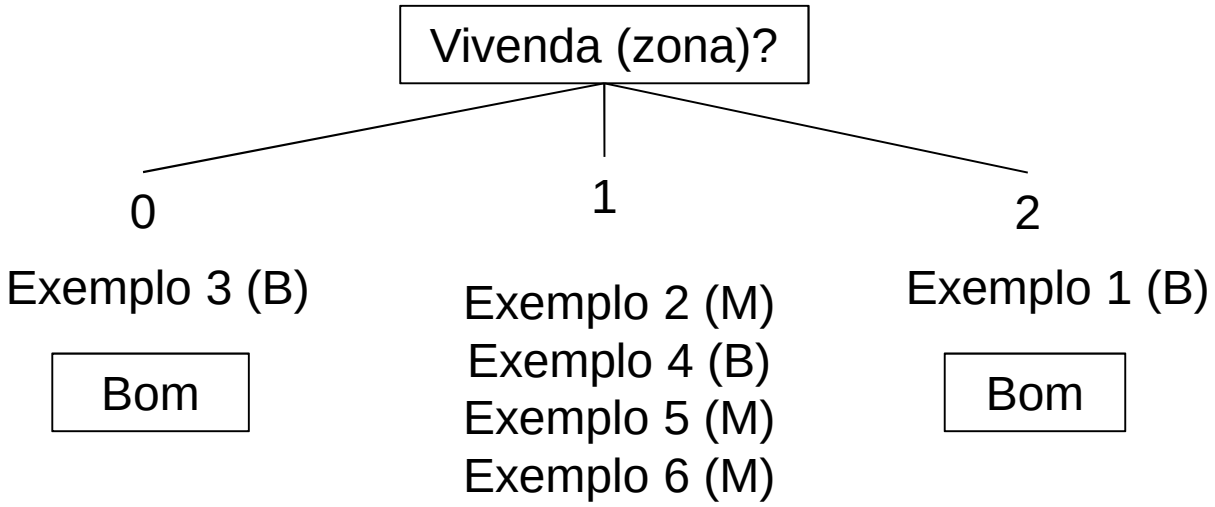
$$I(E, A_1) = 0{,}66$$

$$I(E, A_2) = 0{,}79$$

$$I(E, A_3) = 0{,}54$$

$$I(E, A_4) = 0{,}81$$

Exemplo


Vivenda (zona)?
0
1
2
Exemplo 3 (B)
Bom
Exemplo 2 (M)
Exemplo 4 (B)
Exemplo 5 (M)
Exemplo 6 (M)
Exemplo 1 (B)
Bom

Exemplo

#	Acesso	1º Gasto	Compra	Classe
2	1	0	Disco	Mau
4	0	2	Livro	Bom
5	1	1	Livro	Mau
6	2	2	Livro	Mau

$$I(E, A_1) = ?$$

$$I(E, A_2) = ?$$

$$I(E, A_4) = ?$$

Questões práticas

- Várias questões de ordem prática são colocadas ao aprender árvores de decisão
  - Até que profundidade fazer crescer a árvore de decisão para evitar o sobreajuste (*overfitting*) aos dados de treino
  - Como manejar atributos com valores contínuos
  - Escolha de uma medida adequada para a selecção dos atributos
  - Como manejar atributos com diferentes custos
  - Melhoria da eficiência computacional

Variantes

- Varias melhorias ao método ID3 foram introduzidas por Quinlan e incluídas no método designado C4.5
  - Tratamento de valores contínuos de atributos
  - Manejo de atributos com muitos valores possíveis
  - Controlo do sobreajuste (*overfitting*) através de mecanismos de poda

Variantes

- Outras variantes permitiram tratar questões como:
  - Tratamento de atributos com diferentes custos (Tan e Schlimmer, 1990), (Tan, 1993), (Nuñez, 1988)
  - Aprendizagem em forma incremental (Schlimmer e Granger, 1986), (Utgoff, 1989)

Tarefa

- Seguir exemplo e verificar qual deve ser a árvore final

Bibliografia

- Mitchell, pg. 52 - 77
- Borrajo Millán