ACH2016 - Inteligência Artificial
Aula 03 - Aprendizado de Máquina

Valdinei Freire da Silva

valdinei.freire@usp.br - Bloco A1 100-O

Russell e Norvig, Capítulo 18

Tarefa de Aprendizado Supervisionado

Dado um conjunto de treinamento com N exemplos de pares entrada-saída

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N),$$

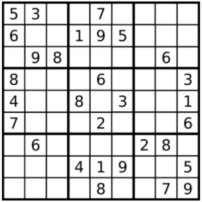
onde cada y_i foi gerado por uma função f desconhecida, isto é, $y_i = f(x_i)$.

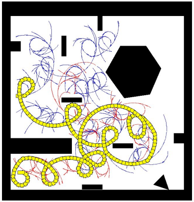
Descubra uma função h que aproxima a verdadeira função f.

x é a entrada e y é a saída.

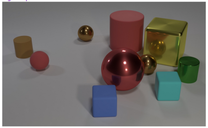
x e y podem ser qualquer valor, números ou categorias, x usualmente é um vetor de valores (atributos).

Qual é a função em um sistema inteligente?

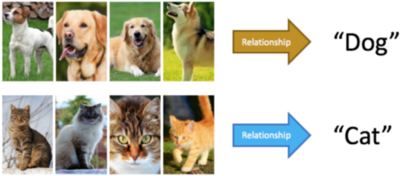


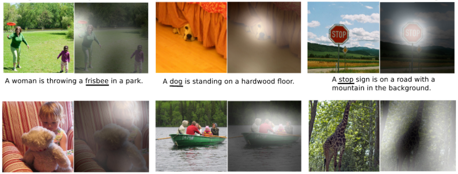


Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or red things?





A woman is throwing a frisbee in a park.

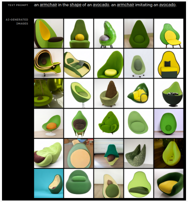
A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.



Tarefa de Aprendizado Supervisionado

A função h é uma hipótese; aprendizado é uma busca no espaço de hipóteses possíveis.

Uma hipótese adequada deve representar adequadamente os exemplos de treinamento, assim como novos exemplos (generalização)

Usualmente considera-se um conjunto de teste, disjunto do conjunto de treinamento, para avaliar a qualidade da hipótese em novos exemplos.

A função f pode ser estocástica, nesse caso deve-se aprender então $\Pr(Y|x)$.

Tarefa de Aprendizado Supervisionado

Se os valores de y são **categóricos**, o problema de aprendizado é chamado de **classificação**.

Se os valores de y são **cardinais**, o problema de aprendizado é chamado de **regressão**.

Overfitting: espaço de hipótese $\mathcal{H}$ pode representar bem os exemplos de treinamento, mas não exemplos novos.

Underfitting: espaço de hipótese $\mathcal{H}$ não pode representar bem os exemplos de treinamento.

Árvore de Decisão

Uma árvore de decisão representa uma função que recebe como entrada um vetor de entrada (atributos) e retorna um único valor, a decisão.

Cada nó interno da árvore representa um teste com relação a um dos valores dos atributos.

As arestas que saem do nó são anotadas com os valores possíveis do atributos.

Cada folha da árvore especifica um valor de retorno.

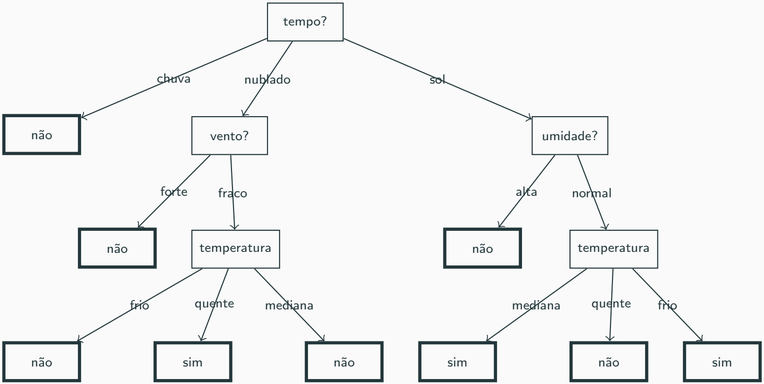
Árvore de Decisão - Exemplo

Predizer se uma pessoa irá jogar tênis com base nas condições meteorológicas

Tempo: sol, nublado ou chuva
Temperatura: quente, mediana ou frio
Umidade: alta ou normal
Vento: forte ou fraco

Se os atributos de entrada e a saída são binários, então uma árvore de decisão pode expressar qualquer expressão de lógica proposicional.

Árvore de Decisão



Exemplo de Amostras

Dia	Tempo	Temperatura	Umidade	Vento	Jogou?
1	sol	quente	alta	fraco	não
2	sol	quente	alta	forte	não
3	nublado	quente	alta	fraco	sim
4	chuva	mediana	alta	fraco	sim
5	chuva	frio	normal	fraco	sim
6	chuva	frio	normal	forte	não
7	nublado	frio	normal	forte	sim
8	sol	mediana	alta	fraco	não
9	sol	frio	normal	fraco	sim
10	chuva	mediana	normal	fraco	sim
11	sol	mediana	normal	forte	sim
12	nublado	mediana	alta	forte	sim
13	nublado	quente	normal	fraco	sim
14	chuva	mediana	alta	forte	não

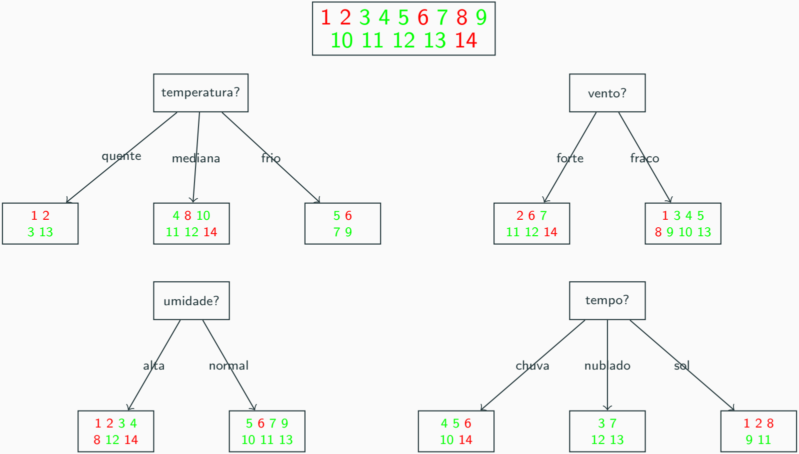
Árvore de Decisão

Como encontrar a melhor árvore?

No caso de n atributos binários, temos 2^{2^n} expressões possíveis.

Quantas funções podem ser geradas no caso geral?

Induzindo a Árvore de Decisão



Entropia

A entropia de uma variável aleatória V com valores v_k onde cada um assume probabilidade $\Pr(v_k)$ é definida por:

$$H(V) = -\sum_{\text{todo } k} \Pr(v_k) \log_2 \Pr(v_k)$$

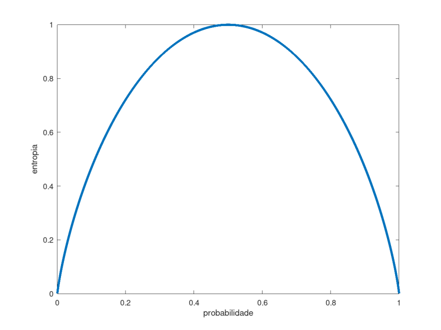
Quando a variável aleatória é booleana, pode-se definir a seguinte função:

$$B(q) = -(q \log_2 q + (1 - q) \log_2 (1 - q))$$

O maior valor de entropia ocorre quando a distribuição de V é uniforme.

O menor valor de entropia ocorre quando a distribuição V assume um único valor.

Entropia



Entropia - Exemplo

Em nosso caso, os valores são `sim` ou `não`.

Entropia do conjunto de dados:

$$\Pr(V = \texttt{sim}) = \frac{9}{14}$$

$$\Pr(V = \texttt{não}) = \frac{5}{14}$$

$$H(V) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) = 0.9403$$

Entropia - Exemplo

Exemplo quando temperatura é quente:

$$\Pr(V = \texttt{sim} | \text{temperatura} = \texttt{quente}) = \frac{2}{4}$$

$$\Pr(V = \texttt{não} | \text{temperatura} = \texttt{quente}) = \frac{2}{4}$$

$$H(V | \text{temperatura} = \texttt{quente}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$

Entropia - Exemplo

Exemplo quando temperatura é mediana:

$$\Pr(V = \texttt{sim}|\text{temperatura} = \texttt{mediana}) = \frac{4}{6}$$

$$\Pr(V = \texttt{não}|\text{temperatura} = \texttt{mediana}) = \frac{2}{6}$$

$$H(V|\text{temperatura} = \texttt{quente}) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.9183$$

Ganho de informação

Um atributo A com d valores distintos particiona o conjunto de treinamento E em $E_1, \ldots, E_d$.

Cada subconjunto E_k possui p_k exemplos positivos e n_k exemplos negativos. Pode-se calcular o ganho de informação para um atributo A:

$$Gain(A) = B\left(\frac{p}{p+n}\right) - \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right)$$

Escolhe-se o atributo com maior ganho para criar um novo nó.

Ganho de informação - Exemplo

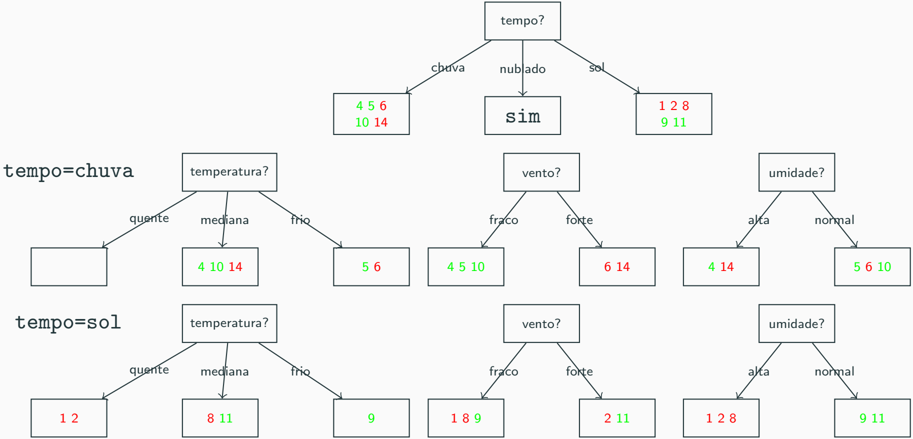
$$gain(\texttt{temperatura}) = B\left(\frac{9}{14}\right) - \left[\frac{4}{14}B\left(\frac{2}{4}\right) + \frac{6}{14}B\left(\frac{4}{6}\right) + \frac{4}{14}B\left(\frac{3}{4}\right)\right] = 0.0292$$

$$gain(\texttt{tempo}) = B\left(\frac{9}{14}\right) - \left[\frac{5}{14}B\left(\frac{3}{5}\right) + \frac{4}{14}B\left(\frac{4}{4}\right) + \frac{5}{14}B\left(\frac{2}{5}\right)\right] = 0.2467$$

$$gain(\texttt{umidade}) = B\left(\frac{9}{14}\right) - \left[\frac{7}{14}B\left(\frac{3}{7}\right) + \frac{7}{14}B\left(\frac{6}{7}\right)\right] = 0.1518$$

$$gain(\texttt{vento}) = B\left(\frac{9}{14}\right) - \left[\frac{6}{14}B\left(\frac{3}{6}\right) + \frac{8}{14}B\left(\frac{6}{8}\right)\right] = 0.0481$$

Ganho de informação - Exemplo



Ganho de informação - Exemplo

`tempo=chuva`

$$gain(\texttt{temperatura}) = B\left(\frac{3}{5}\right) - \left[\frac{0}{5}B\left(\frac{0}{0}\right) + \frac{3}{5}B\left(\frac{2}{3}\right) + \frac{2}{5}B\left(\frac{1}{2}\right)\right] = 0.0200$$

$$gain(\texttt{umidade}) = B\left(\frac{3}{5}\right) - \left[\frac{2}{5}B\left(\frac{1}{2}\right) + \frac{3}{5}B\left(\frac{2}{3}\right)\right] = 0.0200$$

$$gain(\texttt{vento}) = B\left(\frac{3}{5}\right) - \left[\frac{3}{5}B\left(\frac{3}{3}\right) + \frac{2}{5}B\left(\frac{0}{2}\right)\right] = 0.9710$$

Ganho de informação - Exemplo

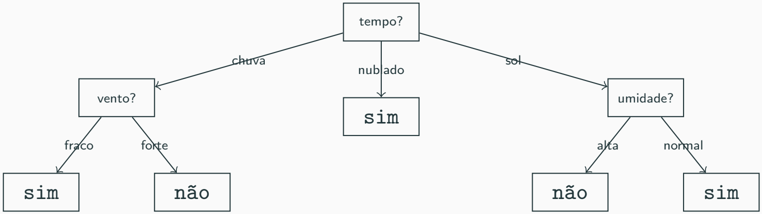
`tempo=sol`

$$gain(\texttt{temperatura}) = B\left(\frac{2}{5}\right) - \left[\frac{2}{5}B\left(\frac{0}{2}\right) + \frac{2}{5}B\left(\frac{1}{2}\right) + \frac{1}{5}B\left(\frac{1}{1}\right)\right] = 0.5710$$

$$gain(\texttt{umidade}) = B\left(\frac{2}{5}\right) - \left[\frac{3}{5}B\left(\frac{0}{3}\right) + \frac{2}{5}B\left(\frac{2}{2}\right)\right] = 0.9710$$

$$gain(\texttt{vento}) = B\left(\frac{2}{5}\right) - \left[\frac{3}{5}B\left(\frac{1}{3}\right) + \frac{2}{5}B\left(\frac{1}{2}\right)\right] = 0.0200$$

Ganho de informação - Exemplo



Overfitting em Árvore de Decisão

Em problemas com respostas probabilística, ou com erros nos dados, pode ocorrer Overfitting.

Ganho de Informação: pode-se limitar a expansão da árvore enquanto o ganho de informação for maior que um limiar ϵ. No entanto, pode ser necessário considerar mais de um atributo para ter um ganho relevante.

Poda: pode-se expandir a árvore inteira e utilizar algum teste de significância para decidir entre podar ou não um ramo da árvore. O teste de significância leva em conta a quantidade de dados (graus de liberdade) e o ganho de informação.

Variações de Árvore de Decisão

valores faltando: considerar a média dos valores em amostras que alcançam um determinado nó.

atributos com muitos valores: utilizar testes booleanos, considerando apenas um valor por vez.

valores de entrada contínuos: utilizar teste booleanos do tipo $x > c$.

valores de saída contínuos: cada folha é representada por uma regressão linear em um subconjunto de atributos.

Exercício

Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

Figure 18.3 Examples for the restaurant domain.