# ÁRVORES DE DECISÃO

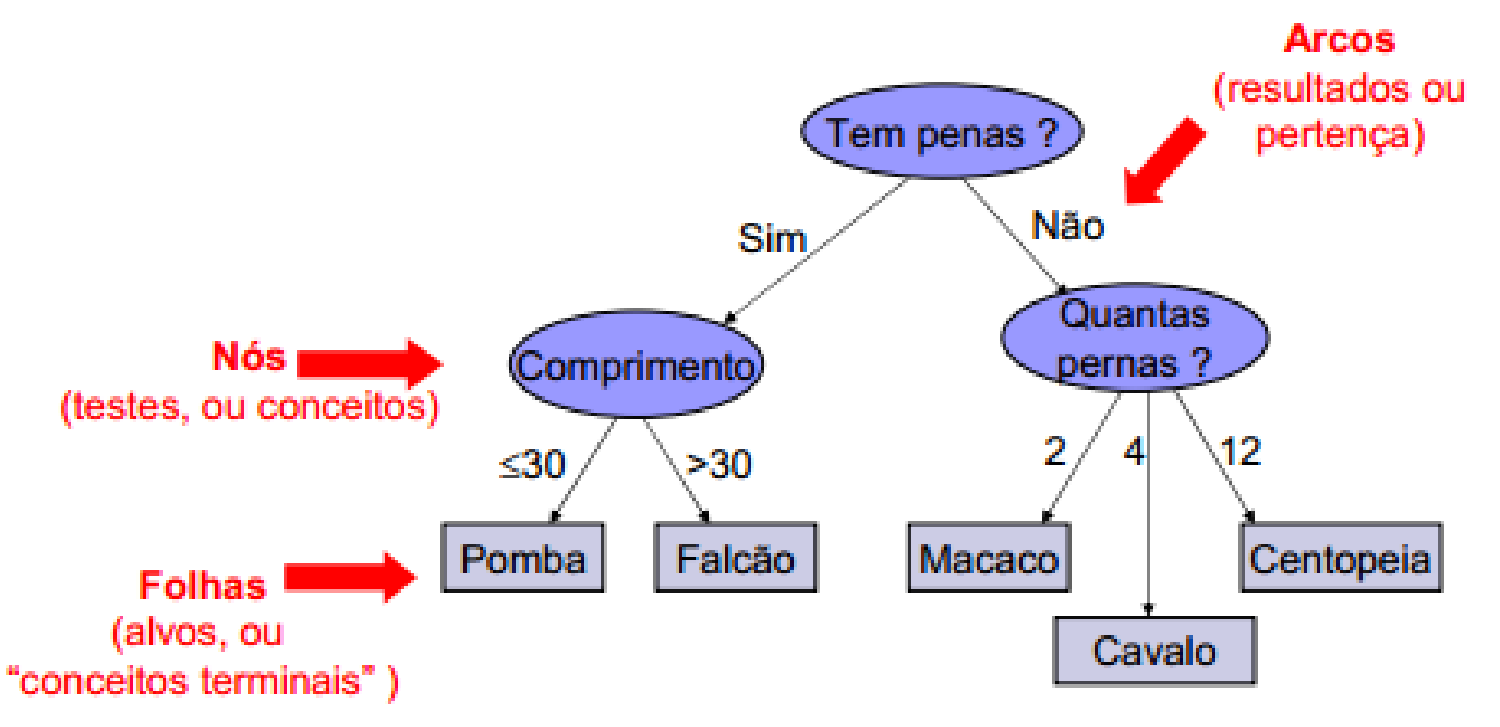
Benjamin Grando Moreira

# O QUE É

- Modo de representar conhecimento
- Uma das formas de algoritmo de aprendizado mais simples e de maior sucesso
- Uma árvore de decisão tem como entrada um objeto ou situação descritos por um conjunto de atributos e como saída uma “decisão” (previsão do valor de saída dada a entrada)
- Uma árvore de decisão toma as suas decisões através de uma sequência de testes
- Forma mais simples:
  - Lista de perguntas: respostas “sim” ou “não”
- Hierarquicamente arranjadas

# O QUE É

- Estrutura da árvore determinada por meio de aprendizado




- Árvores de decisão também podem ser representadas como um conjunto de regras SE-ENTÃO

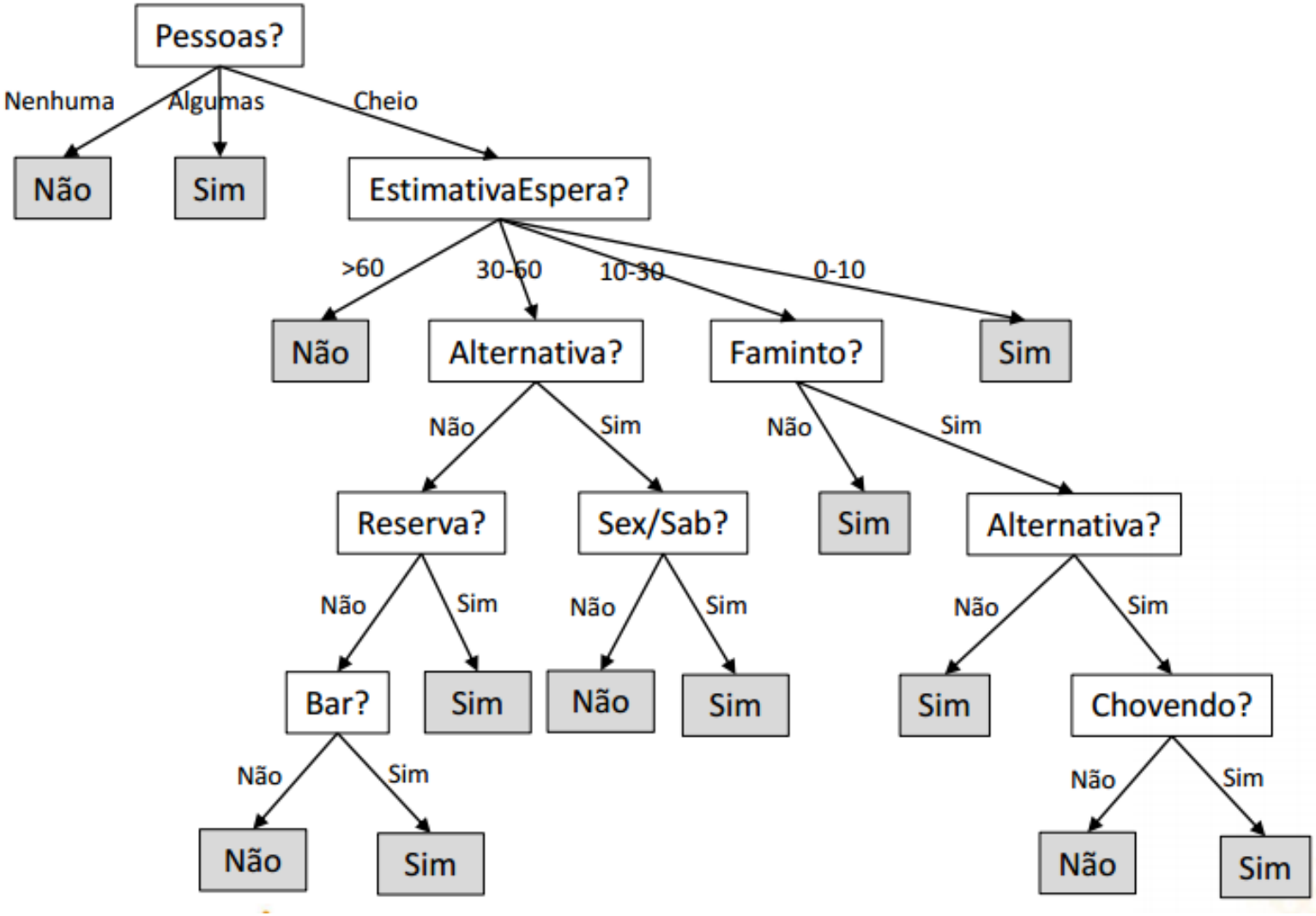
# EXEMPLO - RESTAURANTE

- Problema: Esperar por uma mesa em um restaurante
- O objetivo é aprender uma definição para o predicado “vai esperar”
- Primeiramente é necessário definir quais atributos estão disponíveis para descrever alguns exemplos nesse domínio

# EXEMPLO - ATRIBUTOS

- **Alternativa**: Verdadeiro se existe um restaurante alternativo adequado nas proximidades.
- **Bar**: Verdadeiro se o restaurante tem uma área de bar confortável para ficar esperando.
- **Sex/Sab**: Verdadeiro se o dia da semana for sexta ou sábado.
- **Faminto**: Verdadeiro se estamos com fome.
- **Pessoas**: Quantas pessoas estão no restaurante (os valores são Nenhuma, Algumas e Cheio).
- **Preço**: Preço do restaurante de ($, $ $, $$$).
- **Chuva**: Verdadeiro se está chovendo lá fora.
- **Reserva**: Verdadeiro se nós fizemos uma reserva.
- **Tipo**: Tipo de restaurante (Francês, Italiano, Tailandês, Hambúrguer).
- **EstimativaEspera**: Tempo de espera estimado (00-10, 10-30, 30-60, > 60 minutos).

# EXEMPLO - RESTAURANTE



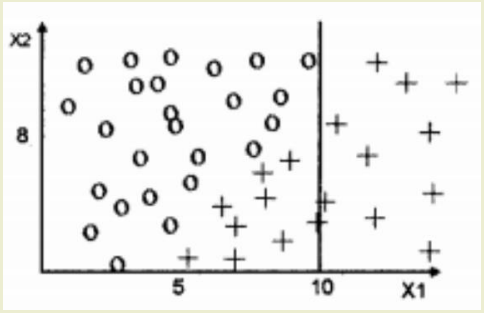
# GERAÇÃO DA ÁRVORE DE DECISÃO

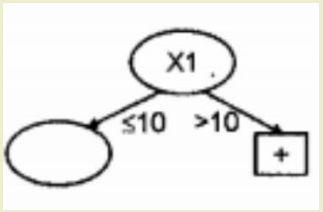
- É possível gerar uma árvore de decisão a partir de um conjunto de exemplos.
- Exemplos positivos são aqueles que levam a uma resposta positiva.
  - Exemplo: “vai esperar” = Sim.
- Exemplos negativos são aqueles que levam a uma resposta negativa.
  - Exemplo: “vai esperar” = Não.

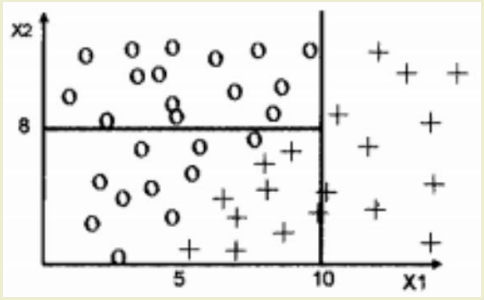
# CONJUNTO DE TREINAMENTO

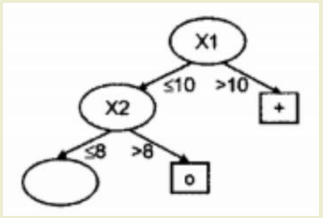
| | Atributos | | | | | | | | | | Obj. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Exemplo** | **Alt.** | **Bar** | **S/S** | **Fam.** | **Pes.** | **Pre.** | **Chov.** | **Res.** | **Tipo** | **Est.** | **Esp.** |
| $X_1$ | Sim | Não | Não | Sim | Algumas | $$$ | Não | Sim | Fran. | 0-10 | Sim |
| $X_2$ | Sim | Não | Não | Sim | Cheio | $ | Não | Não | Tai. | 30-60 | Não |
| $X_3$ | Não | Sim | Não | Não | Algumas | $ | Não | Não | Ham. | 0-10 | Sim |
| $X_4$ | Sim | Não | Sim | Sim | Cheio | $ | Sim | Não | Tai. | 10-30 | Sim |
| $X_5$ | Sim | Não | Sim | Não | Cheio | $$$ | Não | Sim | Fran. | >60 | Não |
| $X_6$ | Não | Sim | Não | Sim | Algumas | $$ | Sim | Sim | Ital. | 0-10 | Sim |
| $X_7$ | Não | Sim | Não | Não | Nenhuma | $ | Sim | Não | Ham. | 0-10 | Não |
| $X_8$ | Não | Não | Não | Sim | Algumas | $$ | Sim | Sim | Tai. | 0-10 | Sim |
| $X_9$ | Não | Sim | Sim | Não | Cheio | $ | Sim | Não | Ham. | >60 | Não |
| $X_{10}$ | Sim | Sim | Sim | Sim | Cheio | $$$ | Não | Sim | Ital. | 10-30 | Não |
| $X_{11}$ | Não | Não | Não | Não | Nenhuma | $ | Não | Não | Tai. | 0-10 | Não |
| $X_{12}$ | Sim | Sim | Sim | Sim | Cheio | $ | Não | Não | Ham. | 30-60 | Sim |

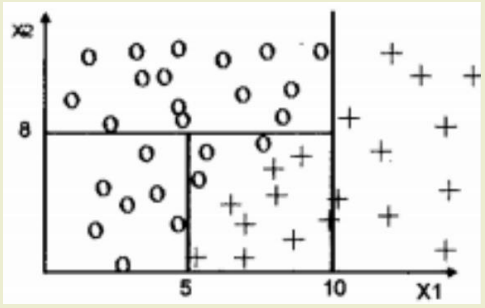
# RAZÕES GEOMÉTRICAS

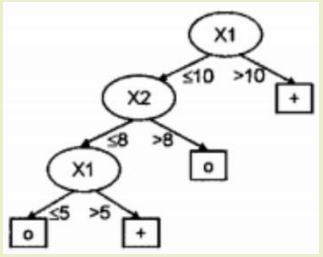


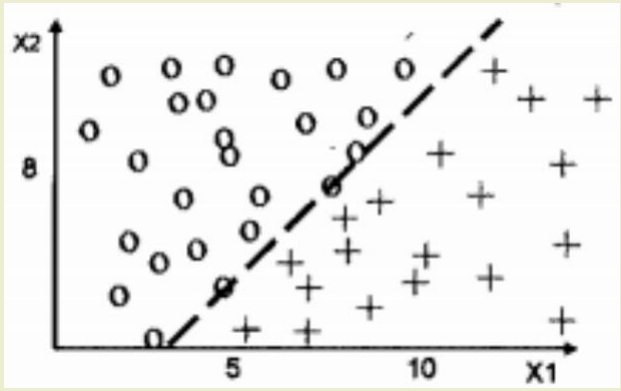



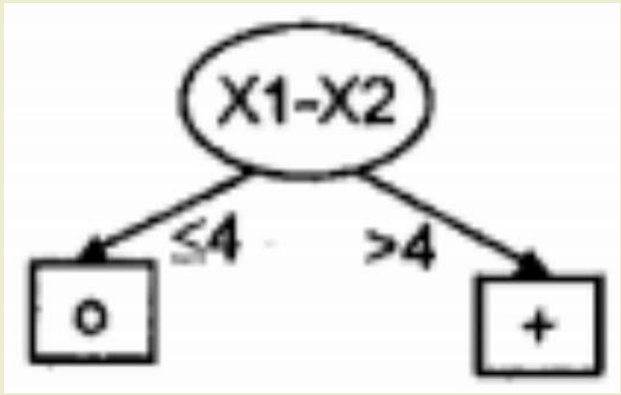



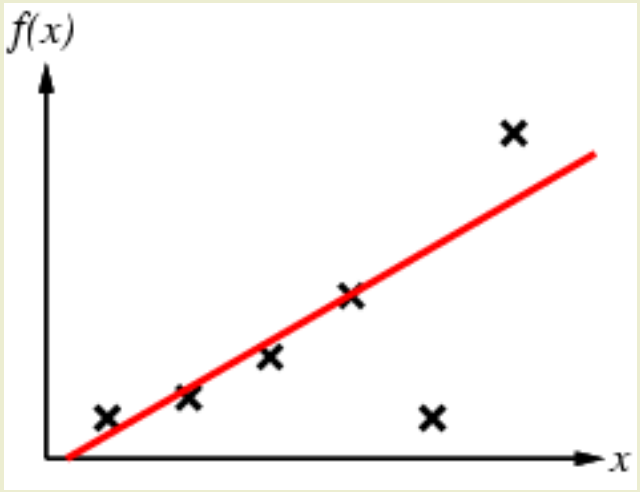



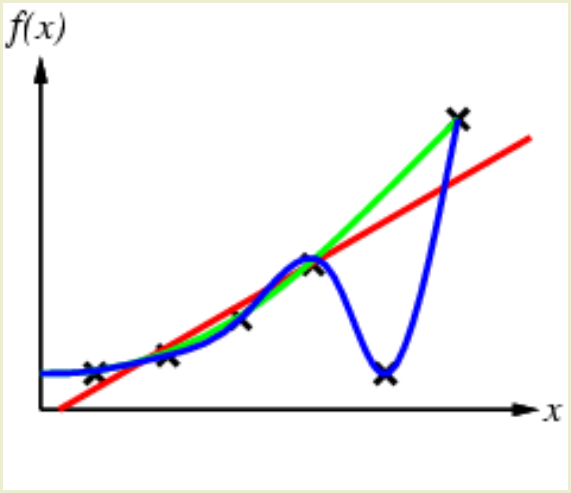



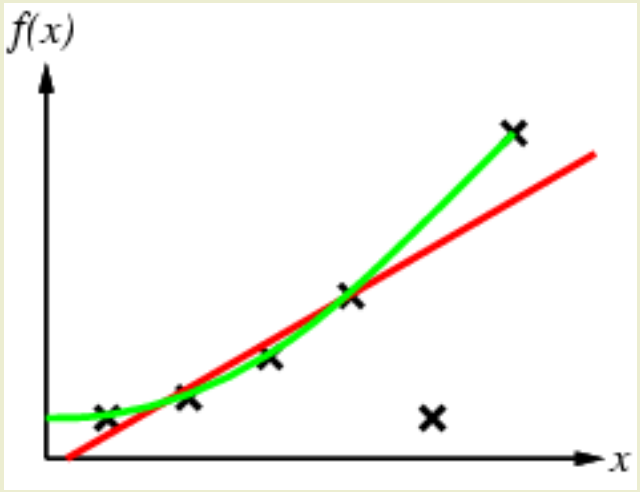
# EXERCÍCIO

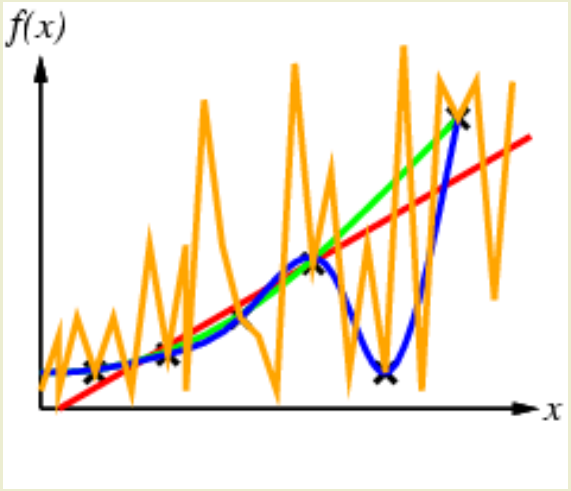
| Tempo | Temperatura | Umidade | Vento | Joga |
|---|---|---|---|---|
| **Sol** | Alta | Media | Não | Não |
| **Sol** | Alta | Alta | Sim | Não |
| **Nublado** | Alta | Alta | Não | Sim |
| **Chuva** | Baixa | Alta | Não | Sim |
| **Chuva** | Baixa | Media | Não | Sim |
| **Chuva** | Baixa | Baixa | Sim | Não |
| **Nublado** | Baixa | Baixa | Sim | Sim |
| **Sol** | Media | Alta | Não | Não |
| **Sol** | Baixa | Baixa | Não | Sim |
| **Chuva** | Media | Media | Não | Sim |
| **Sol** | Media | Baixa | Sim | Sim |
| **Nublado** | Media | Alta | Sim | Sim |
| **Nublado** | Alta | Baixa | Não | Sim |
| **Chuva** | Baixa | Alta | Sim | Não |

# GERAÇÃO DA ÁRVORE DE DECISÃO



# GERAÇÃO DA ÁRVORE DE DECISÃO

- Seguindo o principio de **Ockham**, devemos encontrar a menor árvore de decisão que seja consistente com os exemplos de treinamento:
  - "Qualquer fenómeno deve assumir apenas as premissas estritamente necessárias à explicação do fenómeno e eliminar todas as que não causariam qualquer diferença aparente nas predições da hipótese ou teoria."
- A ideia básica do algoritmo é testar os atributos mais importantes primeiro
  - O atributo mais importante é aquele que faz mais diferença para a classificação de um exemplo
- Dessa forma, esperamos conseguir a classificação correta com um pequeno número de testes.

# VANTAGENS

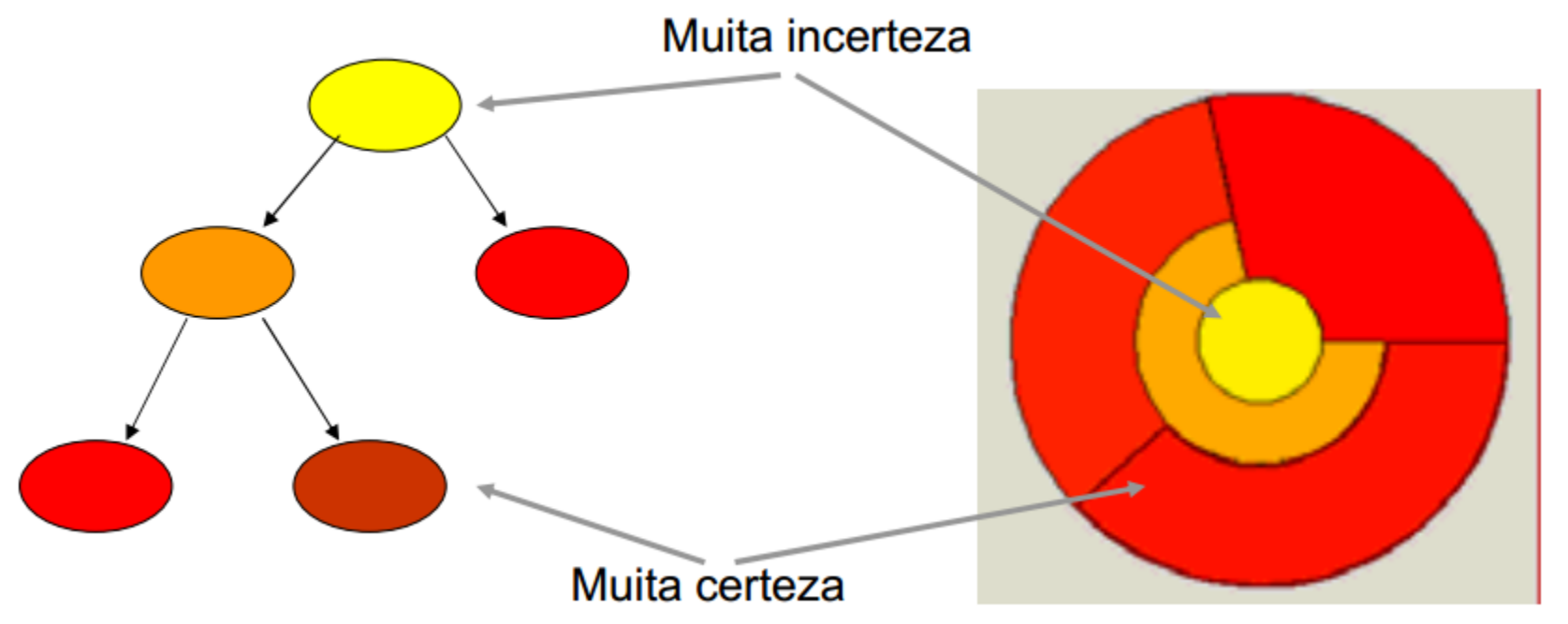
- Interpretação: percebe-se a razão da decisão
- Facilidade em lidar com diversos tipos de informação
  - Real, nominal, ordinal, etc
  - Não é necessário definir “importância relativa”
- Insensível a fatores de escala
- Escolha automática dos atributos mais relevantes em cada caso
  - Atributos mais relevantes aparecem mais acima na árvore
- Adaptável também a problemas de regressão
  - Modelos locais lineares como folhas
- O conjunto de dados do treinamento pode conter erros ou valores de atributos faltando

# DESVANTAGENS

- Fronteiras lineares e perpendiculares aos eixos
- Sensibilidade a pequenas perturbações no conjunto de treino (geram redes muito diferentes)

# OUTRA MANEIRA DE VER ÁRVORES

- Vista de cima: permite ver também o número de dados abrangidos, e o poder discriminante da pergunta

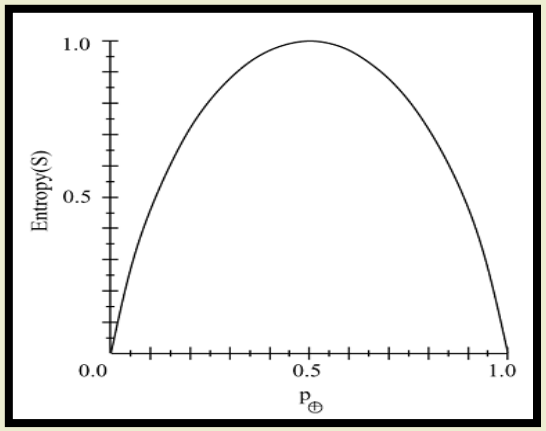


# ENTROPIA

- A Entropia é uma medida que caracteriza a aleatoriedade (impureza) de uma coleção arbitrária de exemplos.
- Dado uma coleção **S** de exemplos **positivos** e **negativos** de um conceito alvo, a entropia de S a esta classificação booleana é:

$$\text{Entropia}(S) = - p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

- $p\oplus$ é a proporção de exemplos positivos em S
- $p\ominus$ é a proporção de exemplos negativos em S



# EXEMPLO NO JOGO DE TÊNIS

- Uma coleção S com 14 exemplos sendo 9 positivos (sim) e 5 negativos (não) [9+, 5-] o valor da entropia é:

$$Entropia\ ([9+,5-]) = -(\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14})$$
$$= 0.940$$

# GANHO DE INFORMAÇÃO

- A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia, ou seja, a aleatoriedade - dificuldade de previsão - da variável que define as classes.

- Ganho de Informação é a redução esperada na entropia causada pela partição dos exemplos de acordo com o teste no atributo A.

$$Gain\,(S,A) \equiv Entropia\ (S) - \sum_{v \in Valores\,(A)} \frac{|S_v|}{|S|} Entropia\ (S_v)$$

# EXEMPLO NO JOGO DE TÊNIS

- Ganho de Informação para o atributo **Tempo = Sol**

| Tempo | Temperatura | Umidade | Vento | Joga |
|---|---|---|---|---|
| Sol | Alta | Media | Não | Não |
| Sol | Alta | Alta | Sim | Não |
| Sol | Media | Alta | Não | Não |
| Sol | Baixa | Baixa | Não | Sim |
| Sol | Media | Baixa | Sim | Sim |

- p (sim | tempo = sol) = 2/5
- p (não | tempo = sol) = 3/5
- Entropia ( joga | tempo = sol) =
  - (2/5) * log2 (2/5) – (3/5) * log2 (3/5) = 0.971

# EXEMPLO NO JOGO DE TÊNIS

- Ganho de Informação para o atributo **Tempo = Chuva**

| Tempo | Temperatura | Umidade | Vento | Joga |
|---|---|---|---|---|
| Chuva | Baixa | Alta | Não | Sim |
| Chuva | Baixa | Media | Não | Sim |
| Chuva | Baixa | Baixa | Sim | Não |
| Chuva | Media | Media | Não | Sim |
| Chuva | Baixa | Alta | Sim | Não |

- p (sim | tempo = Chuva) = 3/5
- p (não | tempo = Chuva) = 2/5
- Entropia ( joga | tempo = chuva) =
  - (3/5) * log2 (3/5) – (2/5) * log2 (2/5) = 0,971

# EXEMPLO NO JOGO DE TÊNIS

- Ganho de Informação para o atributo **Tempo = Nublado**

| Tempo | Temperatura | Umidade | Vento | Joga |
|---|---|---|---|---|
| Nublado | Alta | Alta | Não | Sim |
| Nublado | Baixa | Baixa | Sim | Sim |
| Nublado | Media | Alta | Sim | Sim |
| Nublado | Alta | Baixa | Não | Sim |

- p (sim | tempo = Nublado) = 1
- p (não | tempo = Nublado) = 0
- Entropia ( joga | tempo = nublado) =
  - (1/4) * log2 (1/4) – 0 * log2 (0) = 0

# EXEMPLO NO JOGO DE TÊNIS

- Ganho de Informação obtida no atributo Tempo
- Informação (tempo) =
  5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971 = 0.693
- Ganho (S, Tempo) = Entropia (joga) – Informação (tempo)
- Ganho (S, Tempo) = 0.940 – 0.693 = 0.247

| Ganho (S, Umidade) = 0,057 | Ganho (S, Vento) = 0,048 | Ganho (S, Temperatura) = 0,029 | Ganho (S, Tempo) = 0,247 |
|---|---|---|---|

# PROBLEMA COM O GANHO

- Dá preferência a atributos com muitos valores possíveis
- Um exemplo desse problema ocorreria ao utilizar um atributo totalmente irrelevante;
- Nesse caso, seria criado um nó para cada valor possível, e o número de nós seria igual ao número de identificadores;
- Essa divisão geraria um ganho máximo, embora seja totalmente inútil;

# RAZÃO DE GANHO

- A Razão de Ganho é o ganho de informação relativo (ponderado) como critério de avaliação;
- A razão não é definida quando o denominador é igual a zero, ou seja, quando o valor da entropia do nó é zero;
- Além disso, a razão de ganho favorece atributos cujo o valor da entropia é pequeno.

$$Razão\ de\ ganho = \frac{Ganho}{Entropia}$$

# EXEMPLO NO JOGO DE TÊNIS

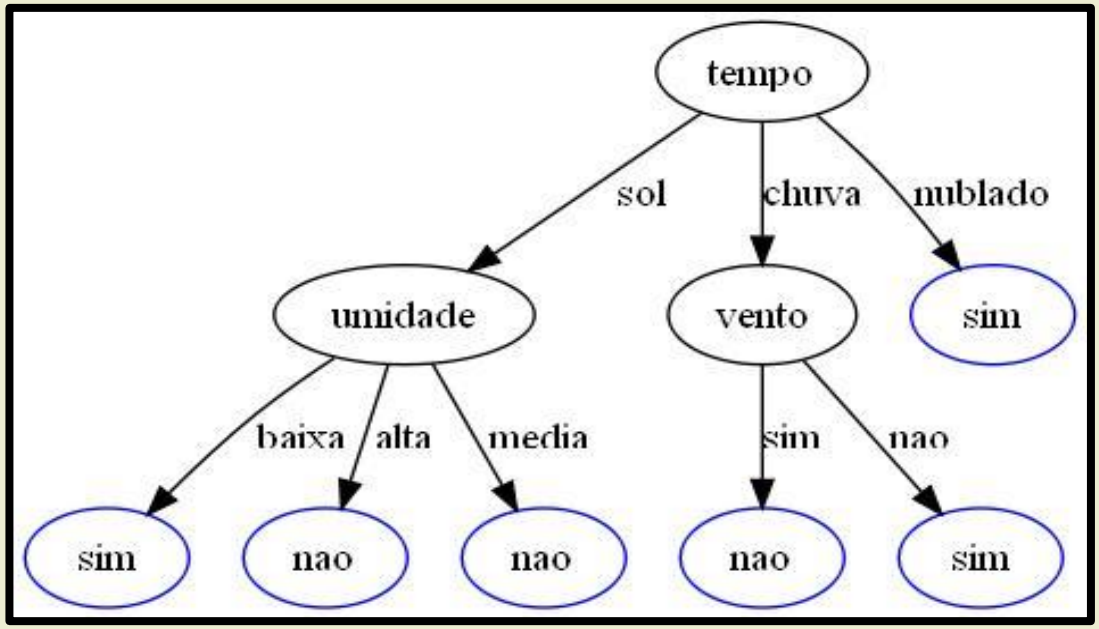
$$Razão\ de\ ganho(tempo) = \frac{Ganho\ (s, tempo)}{Entropia\ (tempo)}$$

$$Razão\ de\ ganho(tempo) = \frac{0{,}247}{0{,}940}$$

$$= \quad 0{,}263$$

# EXEMPLO NO JOGO DE TÊNIS

| Atributo | Ganho de Informação | Razão de Ganho |
|---|---|---|
| Tempo | 0,247 | 0,263 |
| Temperatura | 0,029 | 0,031 |
| Umidade | 0,057 | 0,06 |
| Vento | 0,048 | 0,051 |



# ALGORITMOS POPULARES

- ID3
- C4.5
- CART

# TRABALHOS

- **Avaliação de risco no transporte urbano: uma aplicação ao metrô do Rio de Janeiro:** http://www.scielo.br/scielo.php?pid=S1415-65552005000100006&script=sci_arttext
- ANÁLISE DA SUSTENTABILIDADE DE TRENS TURÍSTICOS NO BRASIL: http://www.pet.coppe.ufrj.br/index.php/producao/teses-de-dsc/doc_download/108-analise-da-sustentabilidade-de-trens-turisticos-no-brasil

# REFERÊNCIAS


- http://www.dcc.fc.up.pt/~ines/aulas/MIM/arvores_de_decisao.pdf
- http://www.isegi.unl.pt/docentes/vlobo/escola_naval/SAD/SAD_EN_8_arvores.pdf
- http://mestrado.deinfo.uepg.br/docs/Aula%20Arvore-decisao.pdf
- ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf
- http://www.slideshare.net/myrmonica/rvores-de-deciso
- http://edirlei.3dgb.com.br/aulas/ia_2012_2/IA_Aula_14_Arvores_de_Decisao_2012.pdf
- http://kdbio.inesc-id.pt/~atf/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id=199