



CLASSIFICAÇÃO

Prof. Julio Cesar dos Reis

<u>ireis@ic.unicamp.br</u>

www.ic.unicamp.br/~jreis

<u>Vídeo</u>

Objetivos da aula

 Aprender terminologia e conceitos sobre a tarefa de classificação

□ Estudar uma técnica básica para a classificação

Classificação

Objetiva a partir de um banco de dados contendo objetos pré-classificados (objetos cuja classe é conhecida), construir um modelo que seja capaz de classificar automaticamente novos objetos cuja a classe é desconhecida) em função de suas características

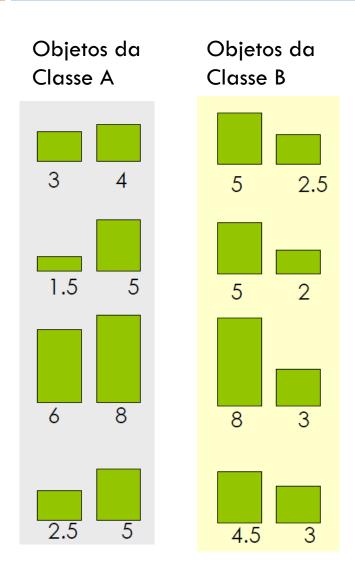
 O modelo criado (via treinamento) é chamado de modelo classificador

Tarefa de classificação

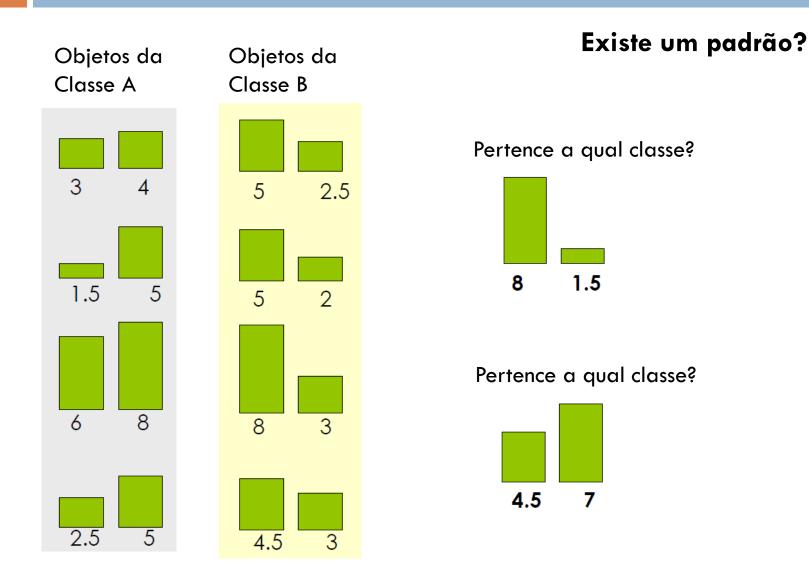
- Cada exemplo pertence a uma classe pré-definida
- Cada exemplo consiste de
 - Um conjunto de atributos preditores
 - Um atributo classe

- Objetivo
 - Predizer a classe do exemplo dado seus valores de atributos preditores

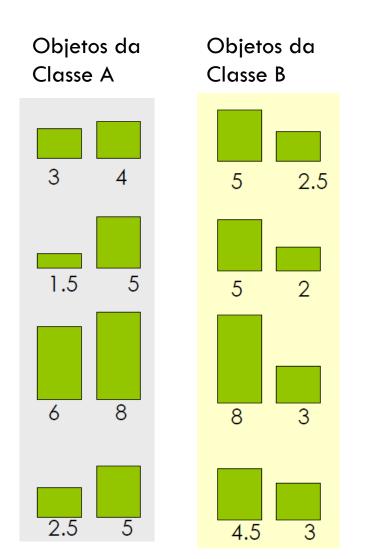
Exemplo 1 de classificação



Exemplo 1 de classificação



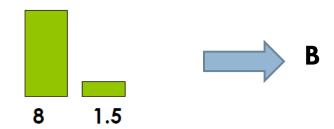
Exemplo 1 de classificação



Existe um padrão?

barra esquerda maior => B barra esquerda menor => A

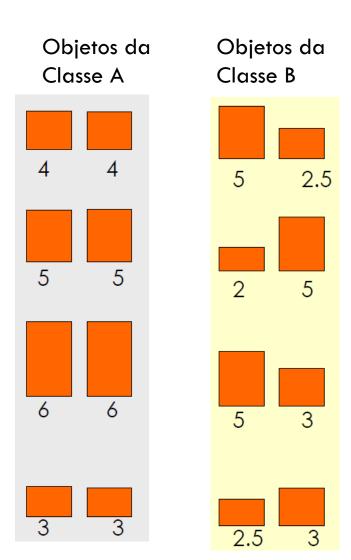
Pertence a qual classe?



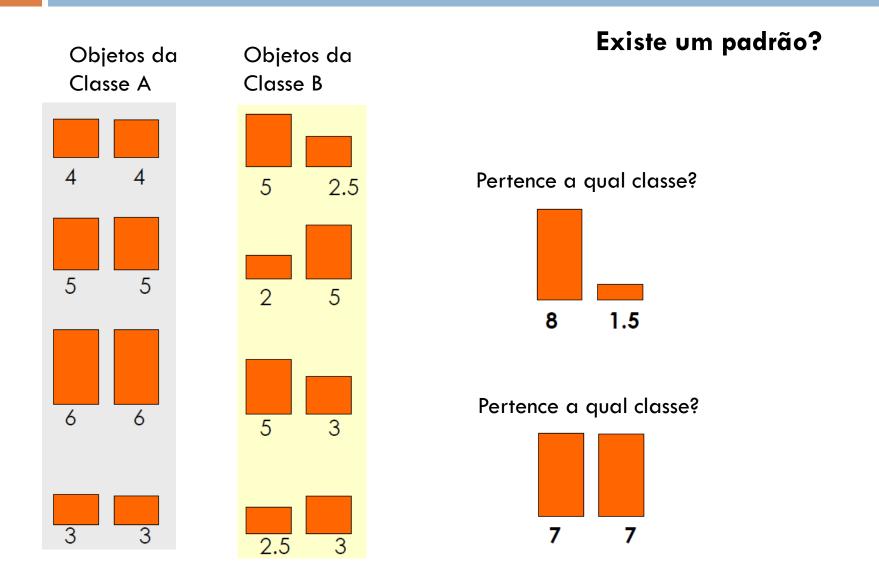
Pertence a qual classe?



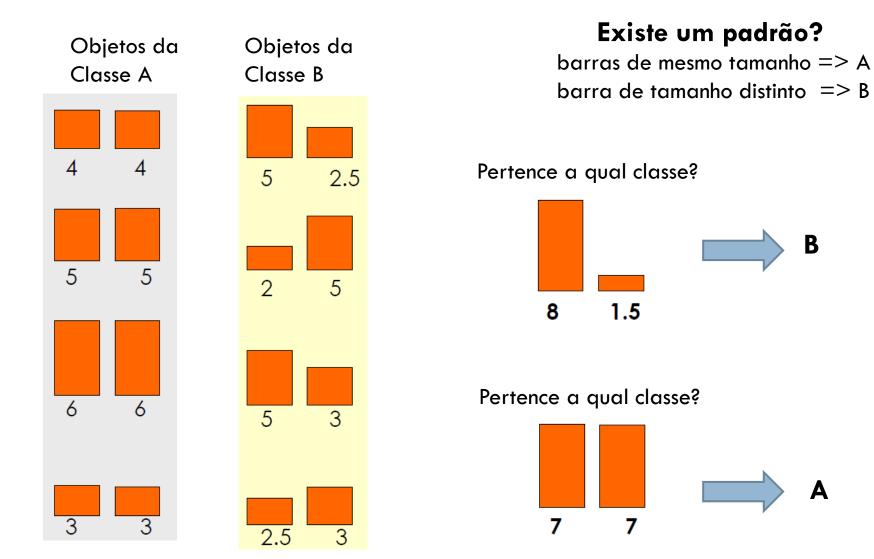
Exemplo 2 de classificação



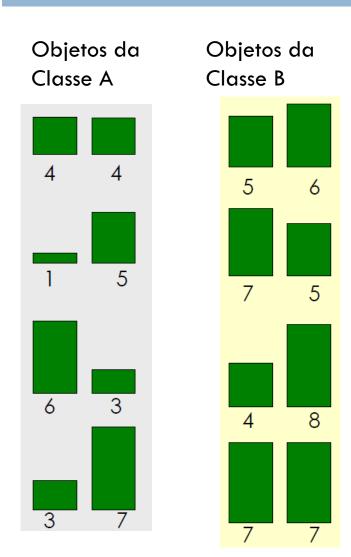
Exemplo 2 de classificação



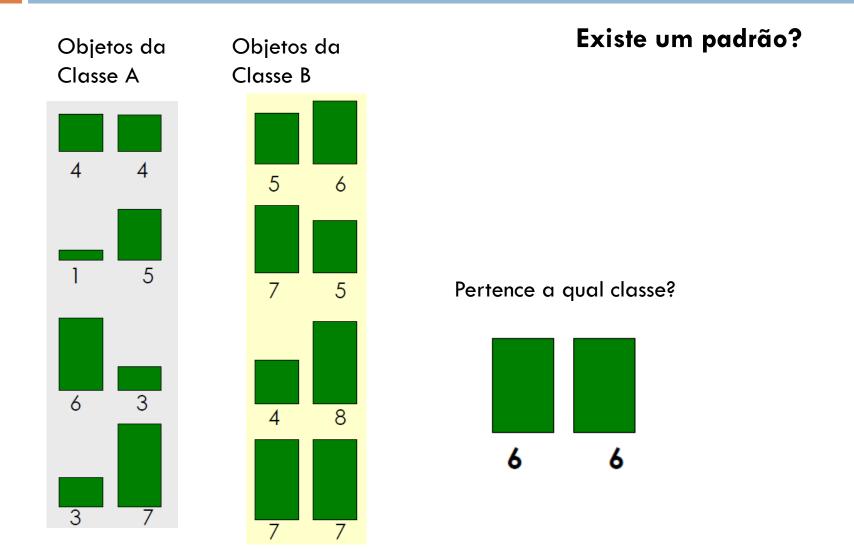
Exemplo 2 de classificação



Exemplo 3 de classificação

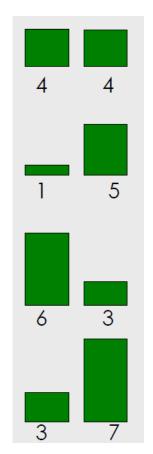


Exemplo 3 de classificação

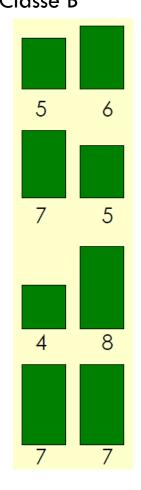


Exemplo 3 de classificação





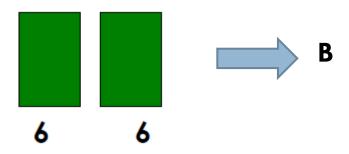
Objetos da Classe B



Existe um padrão?

se
o quadrado da soma das
duas barras for menor ou igual a 100
(a + b)² <=100
=> A
Senão => B

Pertence a qual classe?



Exemplo de classificação*

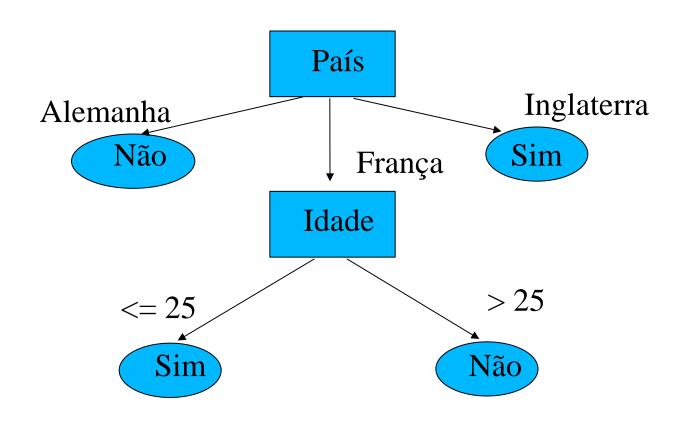
- Uma editora internacional publica o livro "Guia de Restaurantes Franceses na Inglaterra" em 3 países: Inglaterra, França e Alemanha
- A editora tem um banco de dados sobre clientes nesses
 3 países
- Deseja-se saber quais clientes são mais prováveis compradores do livro (para fins de mala direta direcionada)
 - Atributo meta (classe): compra (sim/não)

Exemplo de classificação

Sexo	País	Idade	Compra
М	França	25	Sim
М	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
М	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
М	França	55	Não

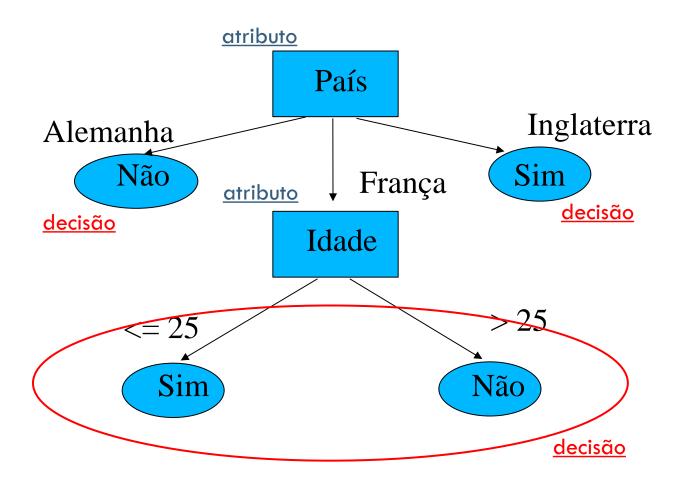
Modelo de classificação

□ Árvore de decisão



Modelo de classificação

□ Árvore de decisão



Terminologia

- Exemplo, um objeto, um caso, um registro, uma tupla
- Atributo, variável, feature, característica
- Conjunto de treinamento (aprendizado)
- Conjunto de teste (avaliação)

Fases na classificação

□ Predição de uma categoria ou classe discreta

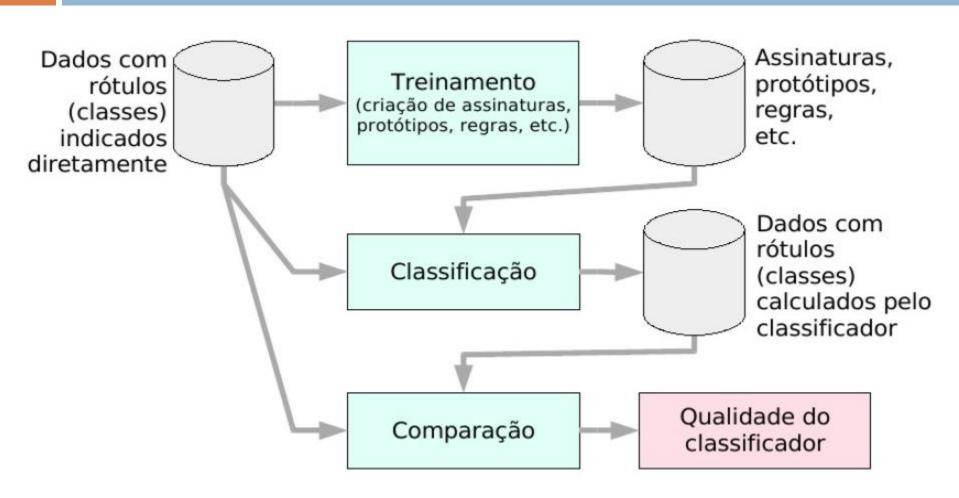
Primeira fase

- Entrada: Instancias para as quais as classes são conhecidas
- Cria-se um classificador ou modelo (fase de treinamento)

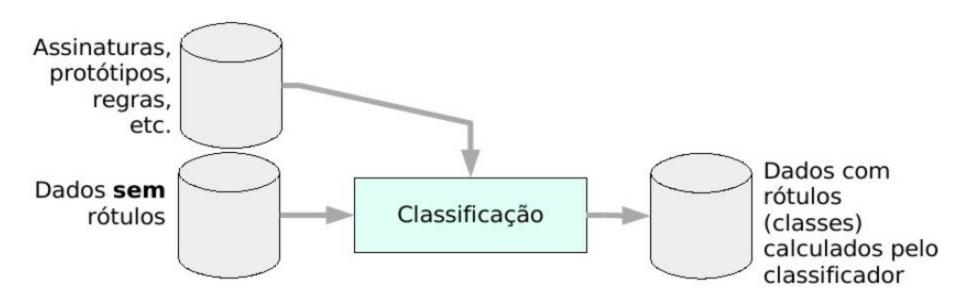
Segunda fase

- Entrada: vários dados para os quais as classes não são conhecidas
- Usa-se o classificador para indicar classes para esses dados
- Podemos avaliar o modelo classificando instancias com classes conhecidas

Fases na classificação



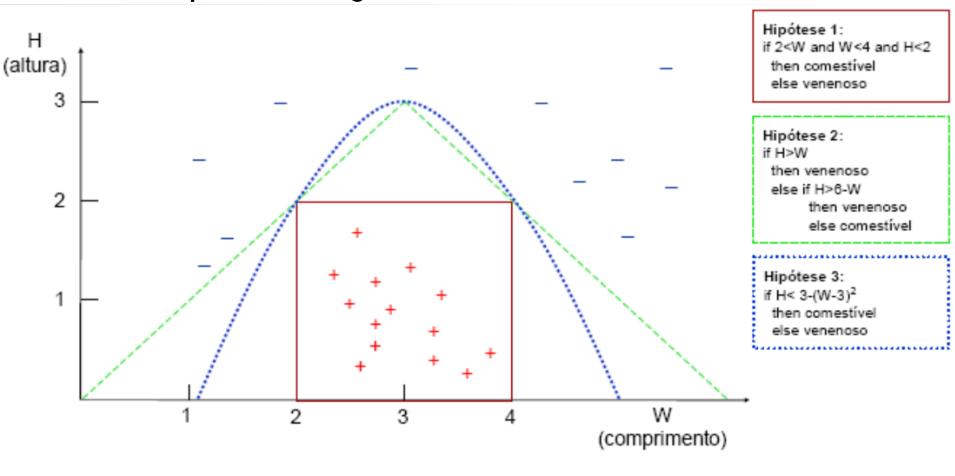
Uso do classificador



Hipóteses de classificação

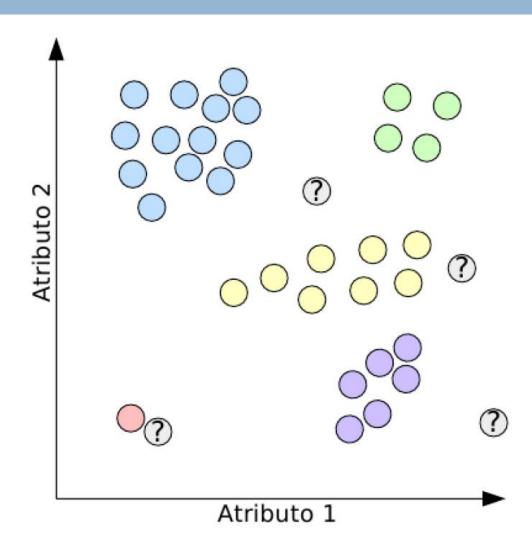
Comestível

Exemplo dos cogumelos comestíveis e venenosos

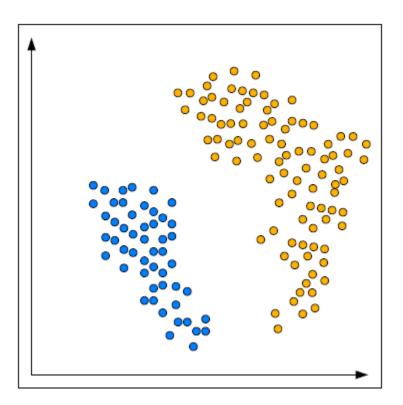


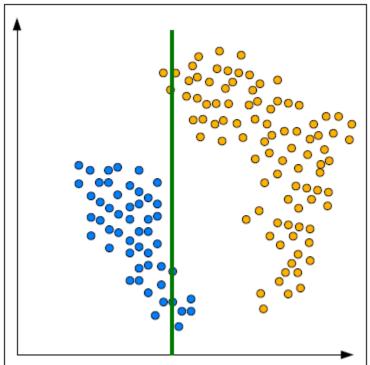
Venenoso

Problema da classificação

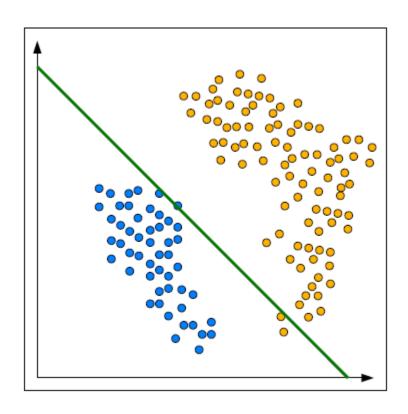


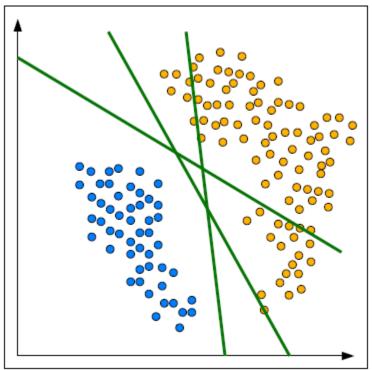
Classificação e espaço de atributos



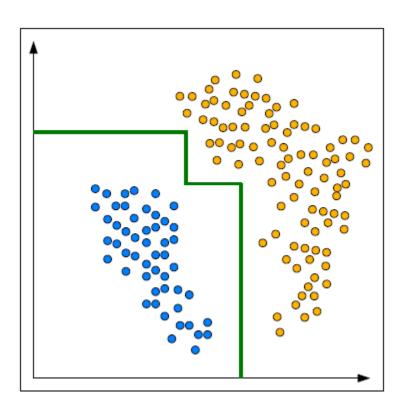


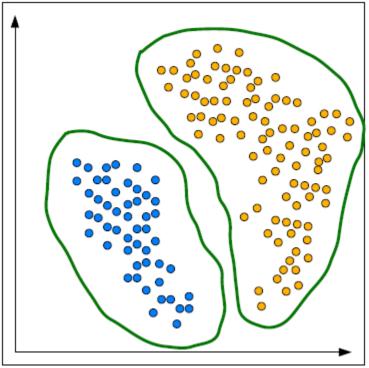
Classificação e espaço de atributos





Classificação e espaço de atributos





Hipótese do aprendizado indutivo

- □ A tarefa de classificação é não determinística
- Qualquer hipótese que aproxime bem o conceito alvo num conjunto de treinamento, suficientemente grande, aproximará o conceito alvo para exemplos não observados

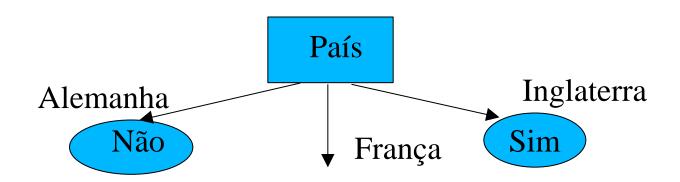
Métodos de classificação supervisionada

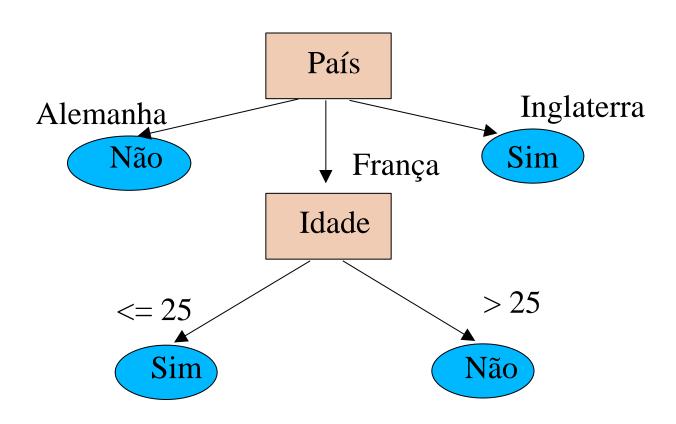
- Baseado em separabilidade (entropia)
 - Árvores de decisão
- □ Baseados em distância e diferenças
 - Mínima distância euclidiana e variantes
- Baseado nos dados vizinhos
 - Algoritmo dos vizinhos mais próximos e similares (KNN)
- Baseado em particionamento
 - Redes neurais
 - SVM (support vector machines)

Exemplo de classificação

Sexo	País	Idade	Compra
М	França	25	Sim
М	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
М	Alemanha	21	Não
М	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
М	França	55	Não

Árvores de decisão (AD)





- Um dos métodos práticos mais usados
- Induz funções discretas
 - Robustas a ruído
- Capaz de aprender expressões <u>disjuntivas</u>

Se país = Inglaterra

OU

Se país = França e idade <= 25

Então

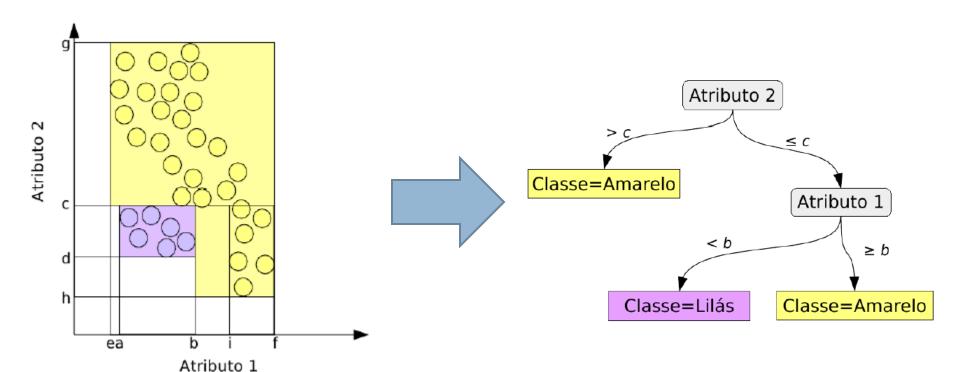
Comprar = sim

□ Classificação tem como base um conjunto de atributos

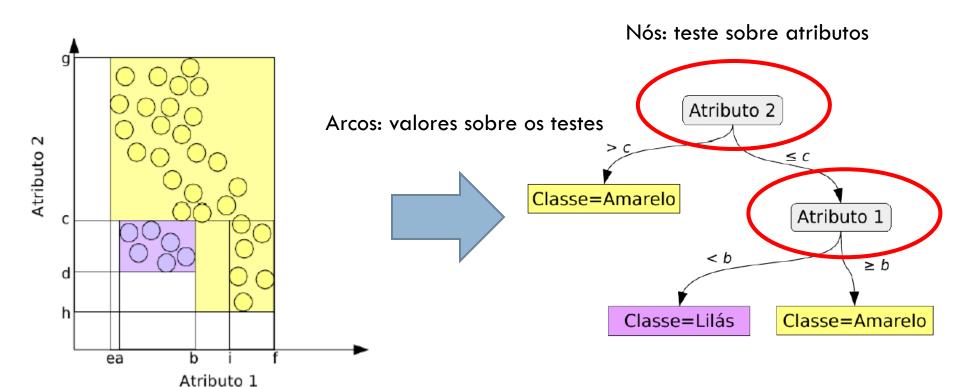
 Cada nó interno corresponde a um teste sobre os valores dos atributos

 Os arcos são rotulados com os valores possíveis do teste

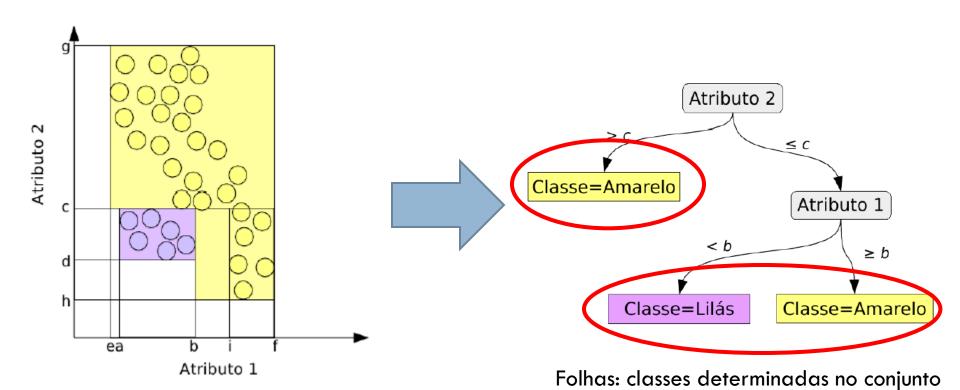
Cada folha na árvore especifica a classificação



Árvores de decisão

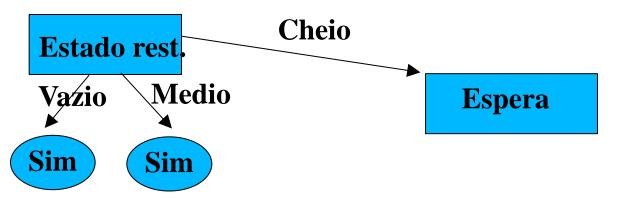


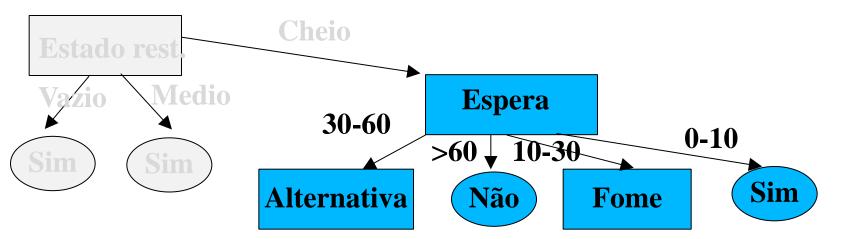
Árvores de decisão

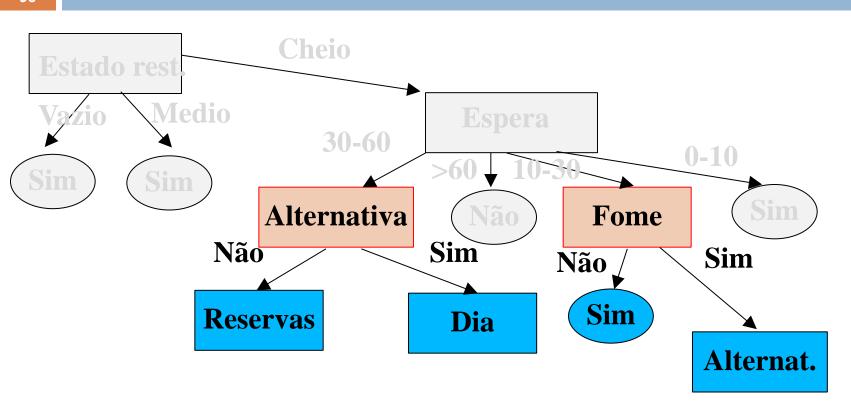


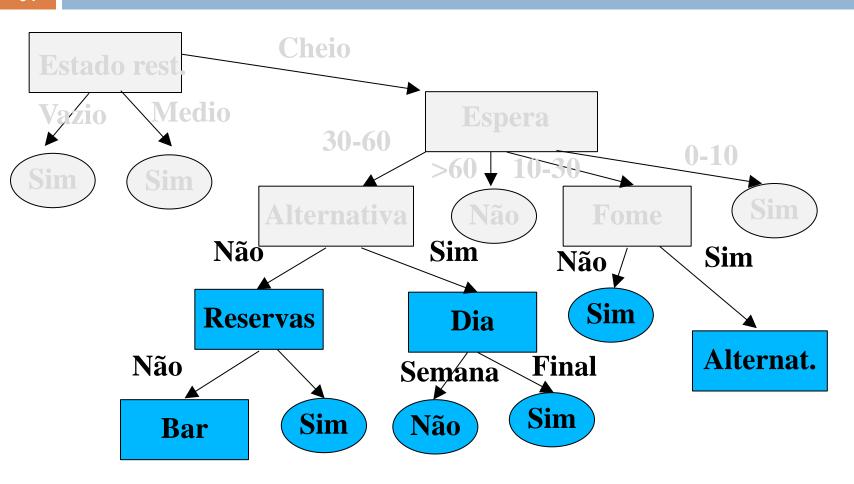
Exemplo do restaurante

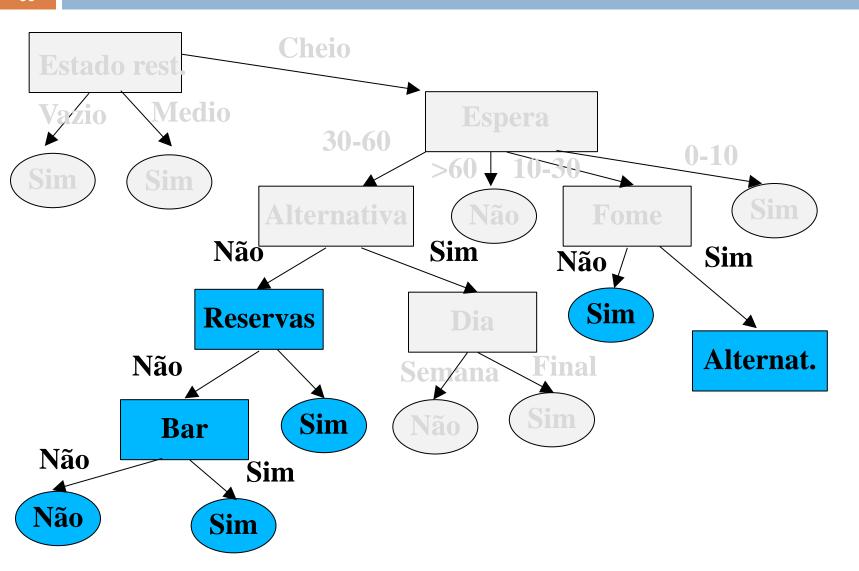
- □ Espera por uma mesa num restaurante?
 - Decidir que propriedades (ou atributos) estão disponíveis para descrever os exemplos do domínio
- Alternativas
 - Existe um bar no local?
 - Qual o "tamanho" da fome?
 - Restaurante está cheio?
 - Qual o tempo de espera?
 - Qual o nível de sua paciência hoje?

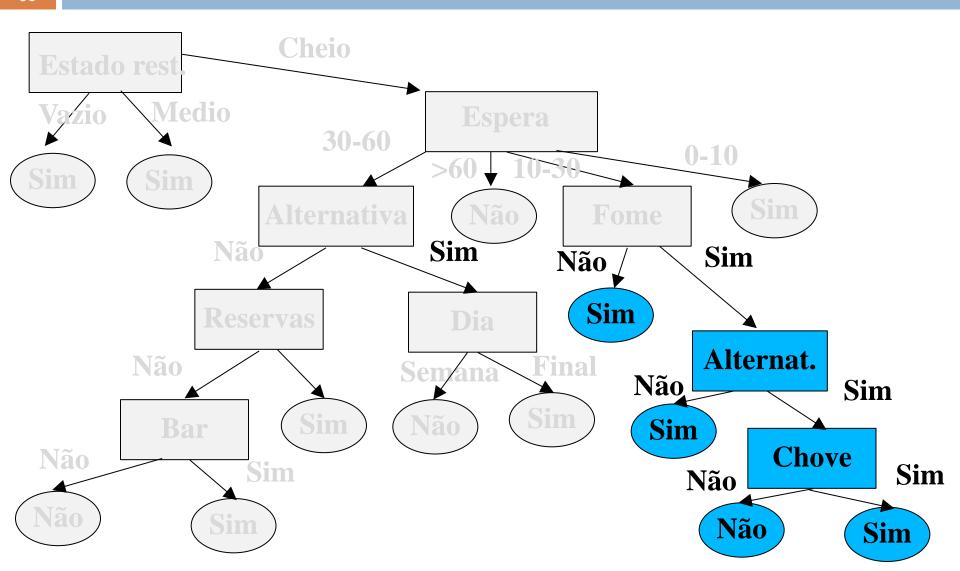


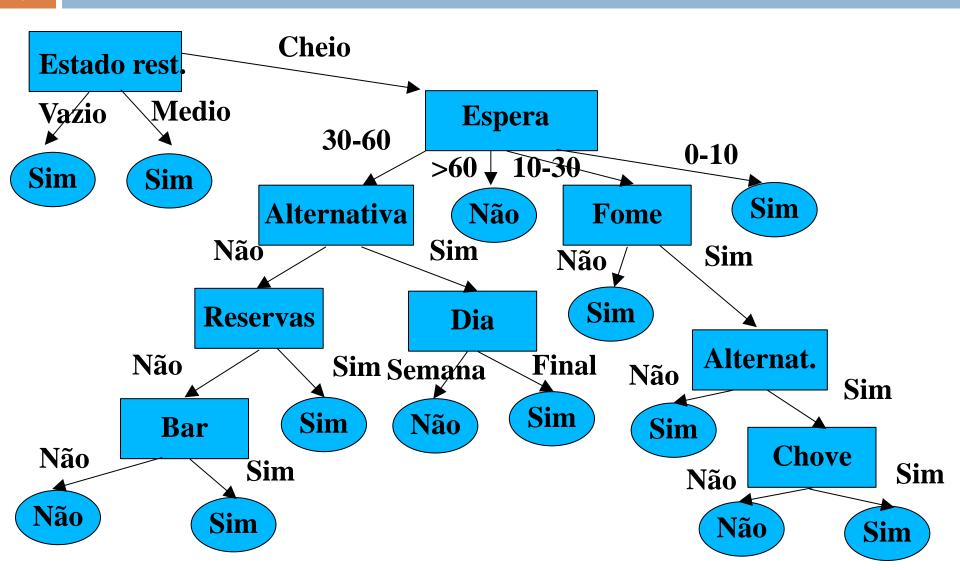






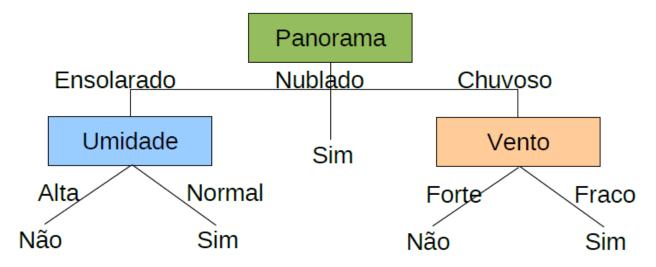






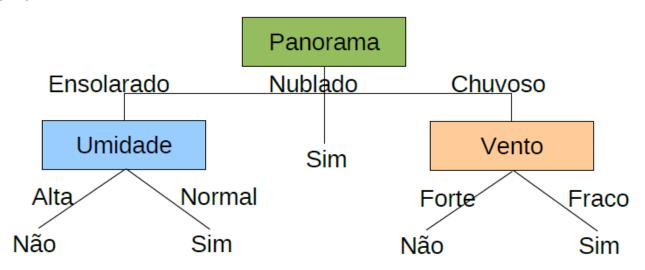
Tomada de decisões

- Tomada de decisão para o problema "jogar tênis"
 - Classificar se condições de um dia são adequadas ou não para se jogar tênis



Tomada de decisões

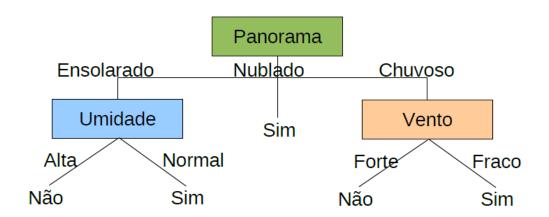
- □ Tomada de decisão para o problema "jogar tênis"
 - Classificar se condições de um dia são adequadas ou não para se jogar tênis



- Instância: <Panorama=Ensolarado, Temperatura=Quente, Umidade=Alta>
- Saída: Não

Tomada de decisões

 Pode-se montar uma expressão para verificar quando é possível jogar tênis:



<Panorama=Ensolarado AND Umidade=Normal>
OR <Panorama=Nublado>
OR <Panorama=Chuvoso AND Vento=Fraco>

Problemas apropriados para AD

- Instâncias são representadas por pares atributo-valor
 - Há um conjunto fixo de atributos (ex: Umidade) e seus valores (ex: Alta, Normal)
 - Situação ideal é quando cada atributo pode assumir poucos valores, no entanto, as árvores de decisão também podem trabalhar com valores reais

Problemas apropriados para AD

- Instâncias são representadas por pares atributo-valor
 - Há um conjunto fixo de atributos (ex: Umidade) e seus valores (ex: Alta, Normal)
 - Situação ideal é quando cada atributo pode assumir poucos valores, no entanto, as árvores de decisão também podem trabalhar com valores reais
- A função a ser aproximada tem valores discretos
 - No exemplo a função deve produzir "Sim" ou "Não"
 - Pode-se facilmente estendê-las para produzir mais de dois valores de saída
 - Tornam-se mais complexas e menos utilizadas abordagens que buscam produzir valores reais como saída

Criação da árvore

 Um exemplo é descrito pelo valor dos atributos e o valor do predicado objetivo (classificação)

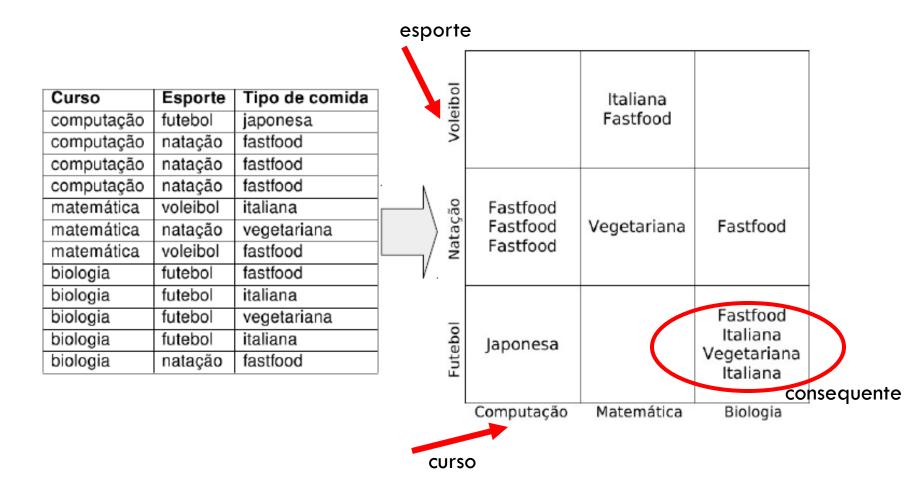
- Solução trivial
 - Uma folha para cada exemplo

Memorização das observações sem extrair padrão

 Extrair padrões significa descrever um grande número de casos de uma maneira concisa

Criação com "força bruta"

□ Tabela de decisão



Criação com "força bruta"

- □ Desvantagem principal é o número de combinações
 - □ 3 cursos, 3 esportes: 9 células
 - 8 cursos, 6 esportes, 4 preferências musicais, 3 referências por filmes: 576 células em quatro dimensões

- Mistura de consequentes
 - Como tratar?

Indução de árvores

- Encontrar a árvore de decisão menor é um problema intratável
- □ Solução
 - Heurísticas simples criam boas árvores
 - Classificação correta com o menor número de teste

Indução de árvores

- Encontrar a árvore de decisão menor é um problema intratável
- □ Solução
 - Heurísticas simples criam boas árvores
 - Classificação correta com o menor número de teste

Ideia básica

- Testar o atributo mais importante primeiro
- Separar o maior número de casos, a cada vez

Algoritmo para criação da árvore

 Uma árvore de decisão é construída de forma "top-down" usando o princípio de dividir-para-conquistar

 Inicialmente, todas as tuplas (registros) são alocadas na raiz da árvore

- Seleciona-se um atributo e divide o conjunto de tuplas
 - Objetivo: separar as classes

Esse processo é repetido recursivamente

Algoritmo ID3

- Considere um conjunto de dados para treinamento
- Constrói a árvore em uma abordagem top-down considerando a questão:
 - "Qual atributo é o mais importante e, portanto, deve ser colocado na raíz da árvore?"

- Cada atributo é testado e sua capacidade para se tornar nó raíz é avaliada
 - Seleção via métrica de ganho de informação do atributo

Algoritmo ID3

 Cria-se tantos nós filhos da raíz quantos valores possíveis esse atributo puder assumir (caso discreto)

 Repete-se o processo <u>recursivamente</u> para cada nó filho da raíz

 Algoritmo para quando todas os atributos foram testados

Ganho de informação no ID3

- Mede a efetividade de um atributo em classificar um conjunto de treinamento
 - Quão bom um atributo é para classificar um conjunto de treinamento
- Ganho de Informação de um <u>atributo A</u>:
 - É a redução na *Entropia*, causada pelo particionamento de exemplos de acordo com este atributo

$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

O problema "jogar tênis"

	CLASSE				
Dia	Perspectiva	Temperatura	Umidade	Vento	JogarTênis
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Moderada	Alta	Fraco	Sim
D5	Chuvoso	Fresca	Normal	Fraco	Sim
D6	Chuvoso	Fresca	Normal	Forte	Não
D7	Nublado	Fresca	Normal	Forte	Sim
D8	Ensolarado	Moderada	Alta	Fraco	Não
D9	Ensolarado	Fresca	Normal	Fraco	Sim
D10	Chuvoso	Moderada	Normal	Fraco	Sim
D11	Ensolarado	Moderada	Normal	Forte	Sim
D12	Nublado	Moderada	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Moderada	Alta	Forte	Não

Calculando ganho de informação

- Considere S um conjunto de treinamento contendo o atributo Vento (Fraco ou Forte)
 - □ S contém 14 exemplos [9+, 5-] com relação a classe
- Considere:
 - Vento=Fraco (8 no total)
 - 6 exemplos positivos e 2 exemplos dos negativos
 - Vento=Forte (6 no total)
 - 3 exemplos positivos e 3 negativos

$$S = [9+, 5-]$$

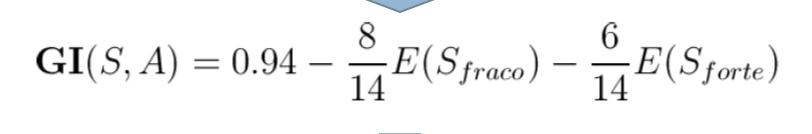
 $S_{fraco} \leftarrow [6+, 2-]$
 $S_{forte} \leftarrow [3+, 3-]$

Calculando ganho de informação

$$S = [9+, 5-]$$

 $S_{fraco} \leftarrow [6+, 2-]$
 $S_{forte} \leftarrow [3+, 3-]$

$$GI(S, A) = E(S) - \sum_{v \in Valores(A)} \frac{S_v}{S} E(S_v)$$



$$\mathbf{GI}(S, A) = 0.94 - \frac{8}{14}0.811 - \frac{6}{14}1.00 = 0.048$$

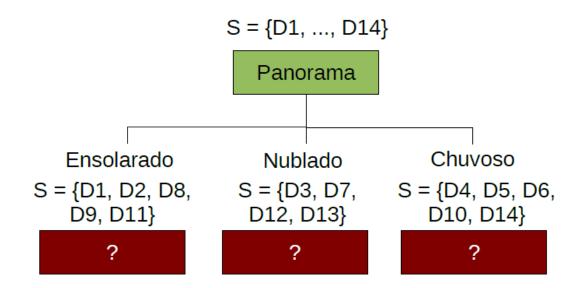
$$E(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Escolher a raiz (calculo de ganho de informação)

GI(S, Panorama) = 0.246 GI(S, Umidade) = 0.151 GI(S, Vento) = 0.048GI(S, Temperatura) = 0.029

- Atributo com maior ganho é selecionado para ser raíz da árvore de decisão
 - □ É o que mais reduz o nível de incerteza!
 - Panorama é escolhido
 - Criamos nós filhos a partir da raíz de acordo com os possíveis valores assumidos pelo atributo Panorama

- Proceder da mesma maneira para os demais ramos que surgem a partir da raíz
- Em cada ramo consideramos somente os exemplos nele contidos
 - Desde que haja divergência entre as classes de saída

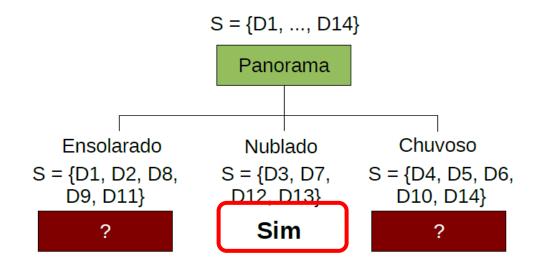


O problema "jogar tênis"

Análise do Nublado

		INSTÂNCIAS			CLASSE	
Dia	Perspectiva	Temperatura	Umidade	Vento	JogarTênis	
D1	Ensolarado	Quente	Alta	Fraco	Não	
D2	Ensolarado	Quente	Alta	Forte	Não	
D3	Nublado	Quente	Alta	Fraco	Sim	
D4	Chuvoso	Moderada	Alta	Fraco	Sim	
D5	Chuvoso	Fresca	Normal	Fraco	Sim	
D6	Chuvoso	Fresca	Normal	Forte	Não	
D7	Nublado	Fresca	Normal	Forte	Sim 🖊	
D8	Ensolarado	Moderada	Alta	Fraco	Não	
D9	Ensolarado	Fresca	Normal	Fraco	Sim	
D10	Chuvoso	Moderada	Normal	Fraco	Sim	
D11	Ensolarado	Moderada	Normal	Forte	Sim	
D12	Nublado	Moderada	Alta	Forte	Sim	
D13	Nublado	Quente	Normal	Fraco	Sim	
D14	Chuvoso	Moderada	Alta	Forte	Não	

 Um dos ramos não tem divergência entre as classes de saída, ou seja, Entropia é igual a zero

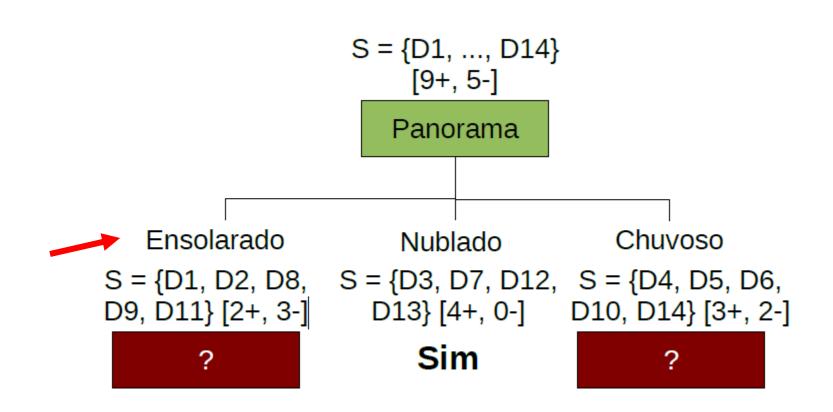


- Atributos existentes incorporados acima de determinado nó não entram na avaliação de Ganho de Informação desse nó
 - Neste caso dois novos nós serão criados
 - O atributo panorama não será mais avaliado

O problema "jogar tênis"

	CLASSE				
Dia	Perspectiva	Temperatura	Umidade	Vento	JogarTênis
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Moderada	Alta	Fraco	Sim
D5	Chuvoso	Fresca	Normal	Fraco	Sim
D6	Chuvoso	Fresca	Normal	Forte	Não
D7	Nublado	Fresca	Normal	Forte	Sim
D8	Ensolarado	Moderada	Alta	Fraco	Não
D9	Ensolarado	Fresca	Normal	Fraco	Sim
D10	Chuvoso	Moderada	Normal	Fraco	Sim
D11	Ensolarado	Moderada	Normal	Forte	Sim 🚄
D12	Nublado	Moderada	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Moderada	Alta	Forte	Não

Análise do Ensolarado

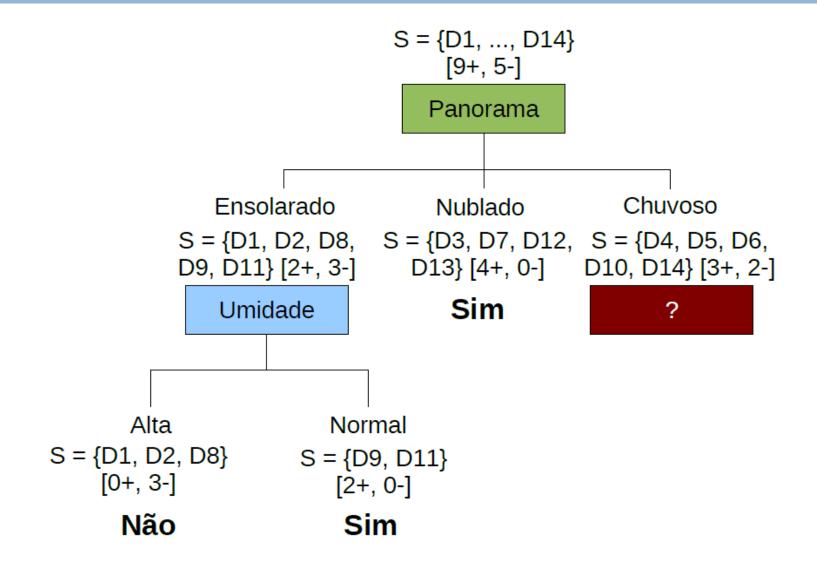


- Ganho de Informação para o ramo Ensolarado
 - Calcula-se a Entropia para E(S = Ensolarado)

$$E(S = Ensolarado) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

Escolher a raiz (calculo de ganho de informação)

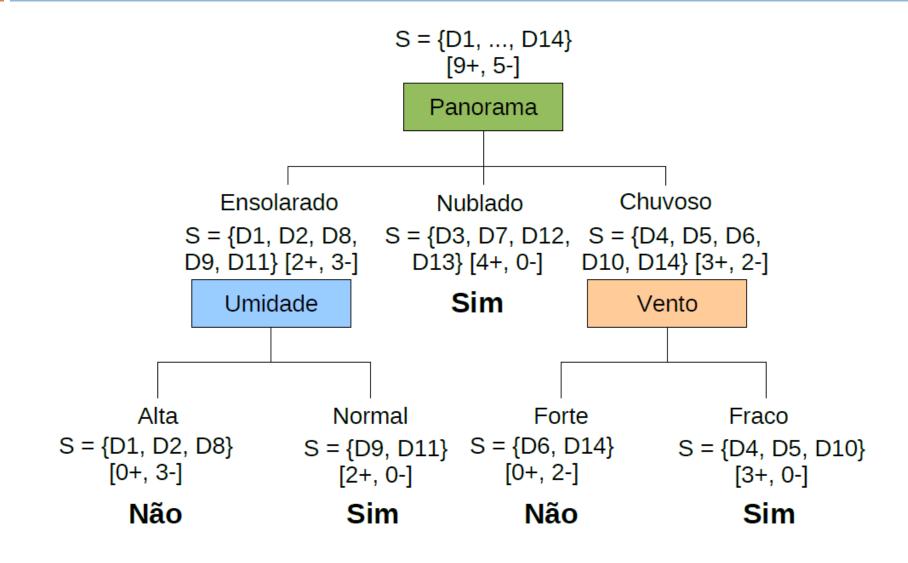
$$GI(S, Umidade) = 0.97 - \frac{3}{5}0.0 - \frac{2}{5}0.0 = 0.97$$
 $GI(S, Temperatura) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1.0 = 0.57$
 $GI(S, Vento) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.019$



Umidade alta

INSTÂNCIAS CLASSE Dia Perspectiva Temperatura Umidade Vento JogarTênis D1Ensolarado Ouente Alta Fraco Não D2Ensolarado Ouente Alta Forte Não D3Sim Nublado Quente Alta Fraco Fraco Sim D4 Chuvoso Moderada Alta D5 Normal Fraco Sim Chuvoso Fresca D6 Chuvoso Fresca Normal Forte Não D7Sim Nublado Fresca Normal Forte D8 Moderada Fraco Não Ensolarado Alta D9 Ensolarado Normal Fraco Sim Fresca Sim D10 Moderada Chuvoso Normal Fraco D11 Ensolarado Moderada Normal Forte Sim D12 Nublado Moderada Sim Alta Forte D13 Nublado Normal Fraco Sim Quente D14 Moderada Forte Não Chuvoso Alta

Análise do Ensolarado



Umidade baixa

Análise do Ensolarado

	INSTÂNCIAS						
Dia	Perspectiva	Temperatura	Umidade	Vento	JogarTênis		
D1	Ensolarado	Quente	Alta	Fraco	Não		
D2	Ensolarado	Quente	Alta	Forte	Não		
D3	Nublado	Quente	Alta	Fraco	Sim		
D4	Chuvoso	Moderada	Alta	Fraco	Sim		
D5	Chuvoso	Fresca	Normal	Fraco	Sim		
D6	Chuvoso	Fresca	Normal	Forte	Não		
D7	Nublado	Fresca	Normal	Forte	Sim		
D8	Ensolarado	Moderada	Alta	Fraco	Não		
D9	Ensolarado	Fresca	Normal	Fraco	Sim		
D10	Chuvoso	Moderada	Normal	Fraco	Sim		
D11	Ensolarado	Moderada	Normal	Forte	Sim		
D12	Nublado	Moderada	Alta	Forte	Sim		
D13	Nublado	Quente	Normal	Fraco	Sim		
D14	Chuvoso	Moderada	Alta	Forte	Não		

Outros critérios

 Há vários outros critérios que podem ser usados para selecionar atributos quando construindo uma árvore de decisão

Nenhum critério é superior em todas as aplicações

 A eficácia de cada critério depende dos dados sendo minerados

Vantagens e desvantagens da AD

- Pontos positivos
 - Modelo de simples interpretação
 - □ Possível variar precisão vs. concisão
 - Semelhança com sistemas especialistas

- Pontos negativos
 - Estrutura (ordem) da árvore depende dos dados
 - Em muitos casos a árvore pode ser extensa

Síntese da aula

 Tarefa de classificação é relevante para várias aplicações e problemas de mineração de dados

- □ Há diversos métodos para a classificação
- Árvores de decisão é um método muito utilizado principalmente quando classes tem valores discretos

Algoritmo recursivo de indução de árvores de decisão

Fontes*

- Alguns slides e imagens foram retirados do conteúdo em:
 - http://www.lac.inpe.br/~rafael.santos/Docs/ELAC/20 12/Elac01 DM Dia2.pdf
 - Slides Prof. Rodrigo Fernandes de Mello (IME/USP)