Árvores de Decisão

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

Principais métodos preditivos

- ✓ Métodos baseados em distâncias
  - ✓ Algoritmo K-NN
- ✓ Métodos probabilísticos
  - ✓ Naive Bayes
  - ✓ Redes Bayesianas
- ✓ **Métodos baseados em procura**
  - ✓ **Árvores de decisão e regressão**
- ✓ Métodos baseados em otimização
  - ✓ Redes neurais artificiais
  - ✓ Máquinas de vetores de suporte (SVM)

Métodos baseados em procura

- ✓ Ideia central
  - ✓ Procura num espaço de soluções possíveis
  - ✓ Árvores de **classificação**
    - ✓ Atributo alvo é categórico
  - ✓ Árvores de **regressão**
    - ✓ Atributo alvo **não é** categórico
  - ✓ Regras de decisão

Métodos baseados em distâncias

Indução
"memorização" dos dados de treinamento

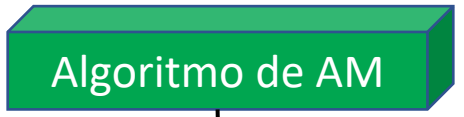
Conjunto de treinamento

Dedução
Comparação da distância

Conjunto de teste

Comparação da nova instância se dá com os dados de treinamento em memória

Outros métodos



Indução
"aprender o modelo"

Modelo

- Uma rede neural
- Árvore decisão
- Rede bayesiana
- ...

Dedução
"aplicar o modelo"

Comparação da nova instância se dá com o modelo

Árvores de decisão

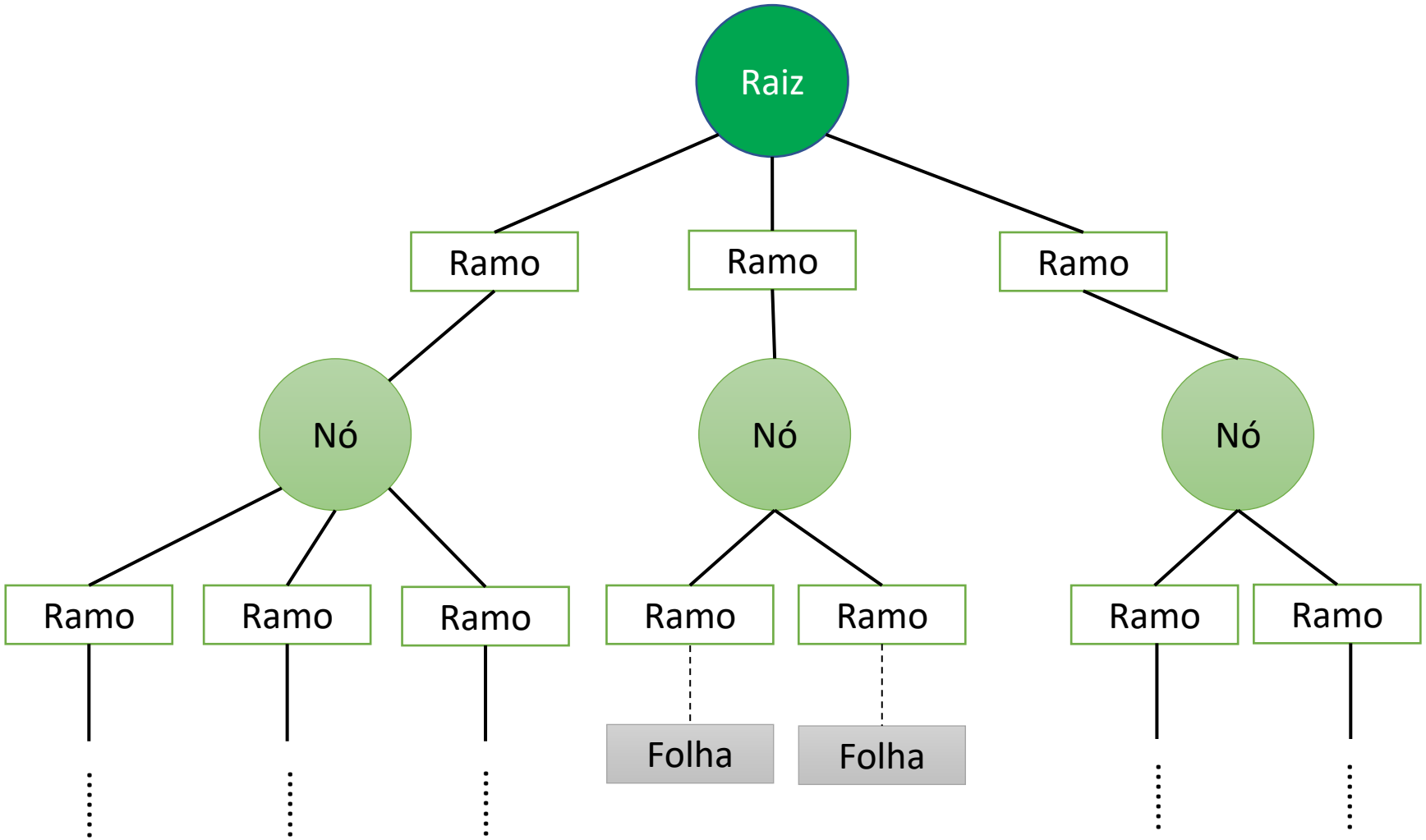
- ✓Tomam como entrada um conjunto de atributos e retornam uma decisão.
- ✓Podem ser representadas como um conjunto de regras SE ENTÃO.
- ✓As instâncias são representadas por pares atributo-valor.
- ✓O conjunto de treinamento **pode conter erros ou valores faltantes**.

Árvores de decisão

- ✓Cada nó de decisão compreende um teste em um atributo.
- ✓Cada ramo descendente corresponde a um possível valor desse atributo.
- ✓Cada nó folha está associado a um valor de uma classe.
- ✓Cada percurso na árvore (do nó raiz até um nó folha) corresponde a uma regra de classificação.
- ✓Podemos ver a árvores como várias sub-árvores que são construídas recursivamente.

Árvores de decisão

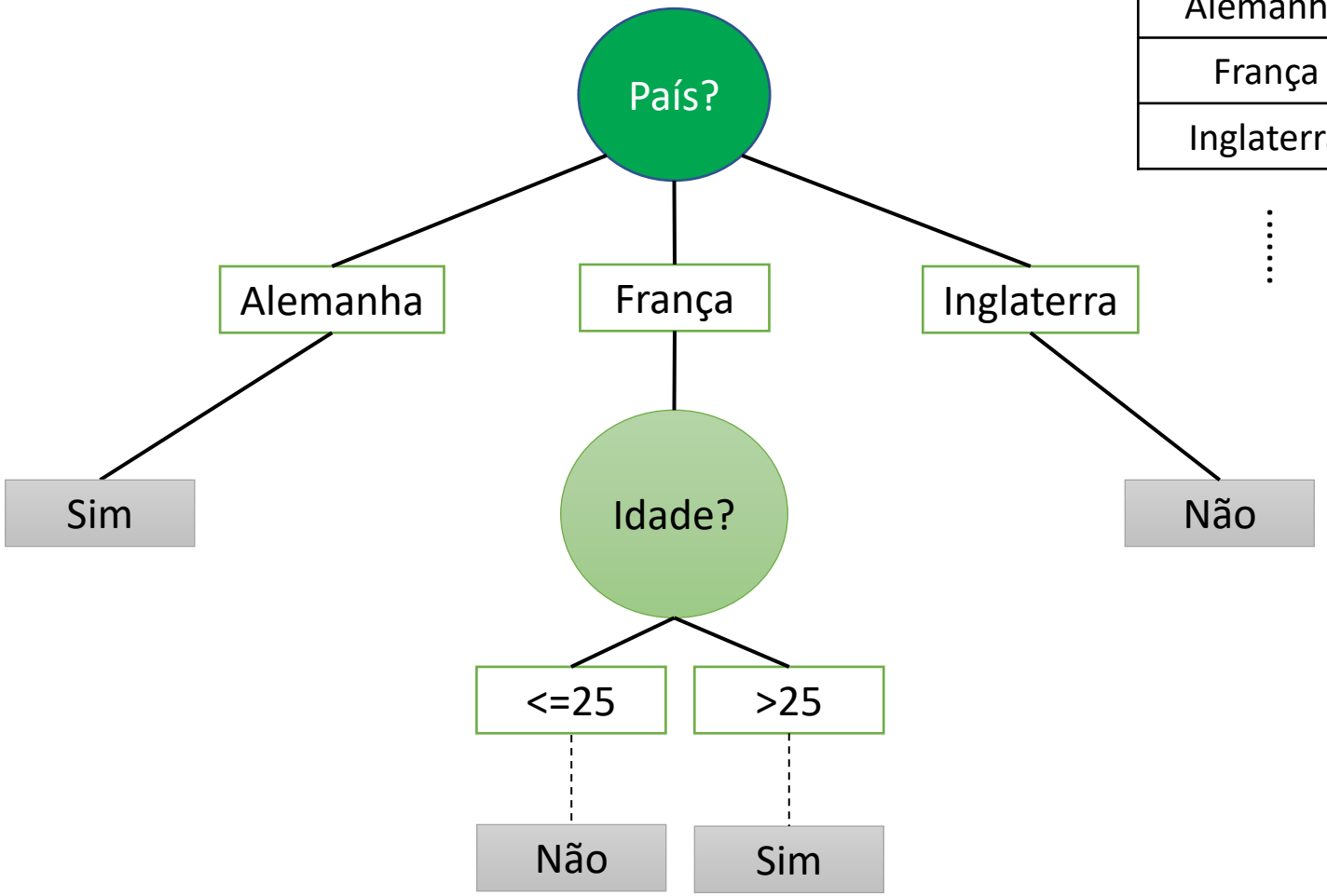
- ✓ Anatomia:



Árvores de decisão

✓Anatomia: exemplo

País	Idade	Bebe cerveja
Alemanha	30	Sim
França	20	Não
Inglaterra	40	Não
⋮	⋮	⋮



Árvores de decisão

- ✓ Daqui em diante, vamos considerar a seguinte base de dados de exemplo
  - ✓ Joga ou não joga tênis a partir dos seguintes critérios (atributos)
  - ✓ A árvore criada ajudará a decidir, para uma observação desconhecida se ela pertence à classe “joga” ou “não joga” de acordo com os valores dos seus atributos

Tempo: Sol, nublado, Chuva

Temperatura: Baixa, média, alta

Umidade: Baixa média e alta

Vento: Sim, não

Joga: Sim, não

Árvores de decisão

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Árvores de decisão

- Como construir? (ideia geral que os algoritmos usam)
  - Árvore é construída de maneira top-down, recursivamente e usando a técnica de programação dividir para conquistar. (divide o espaço de busca e porções menores e, a seguir, combina)
  - 1. Todos os exemplos de treinamento são posicionados na raiz da árvore.
  - 2. Os exemplos são particionados recursivamente, com base nos atributos selecionados, com o objetivo de separar os exemplos por classes.
  - 3. Condições de parada
    - Todos os exemplos de um dado nó (atributo) pertencerem à mesma classe.
    - Não existirem mais atributos para continuar o particionamento
    - Todos os exemplos de treinamento estiverem classificados.

Árvores de decisão

Grande questão que vem à mente:

Como o algoritmo escolhe qual o atributo que será o nó de cada sub-árvore?

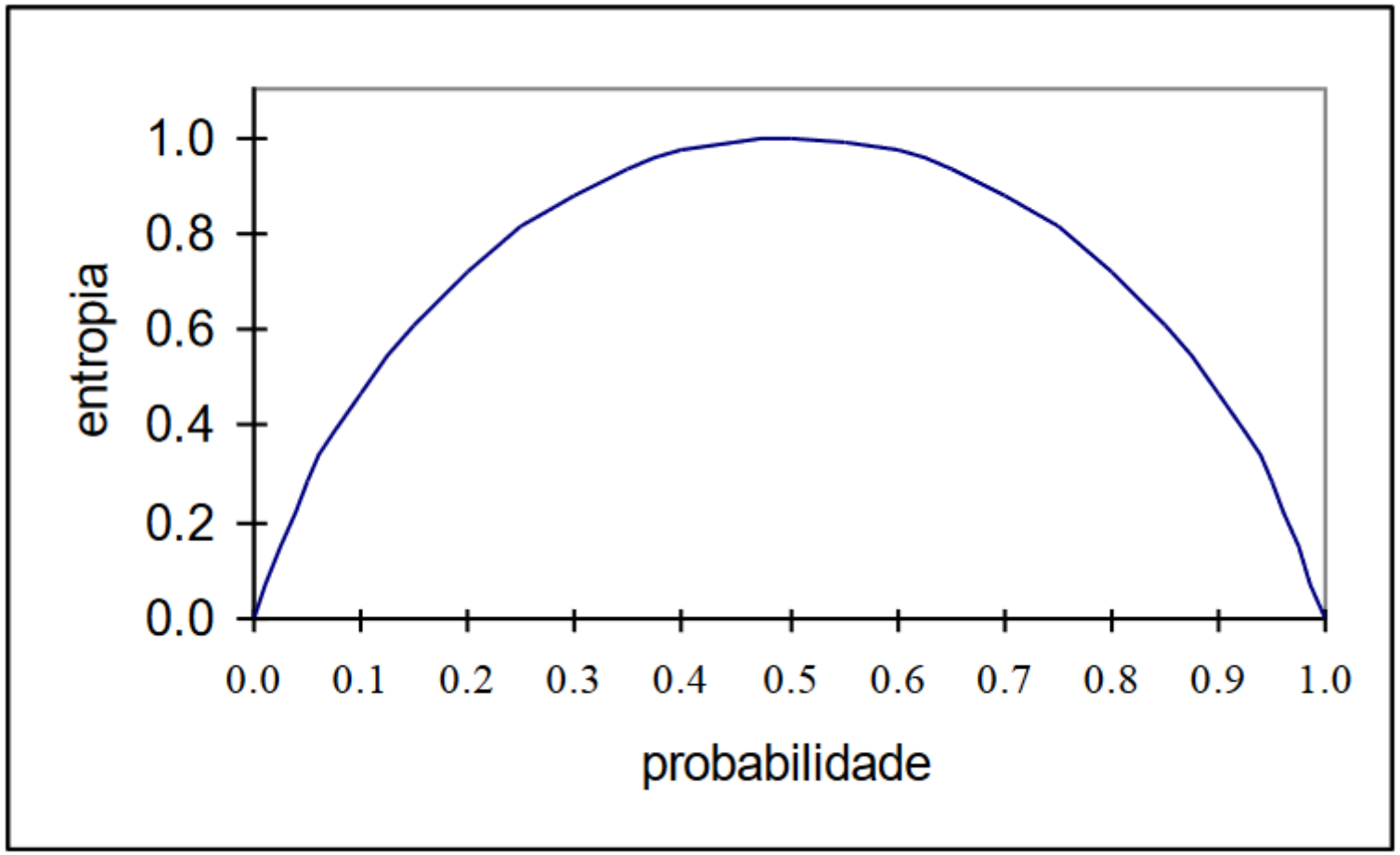
Através da medida do ganho de informação (GI).
Para calcular GI, precisamos entender primeiro o que é a métrica de Entropia (E).

Entropia (E)

- ✓ É uma medida que caracteriza a aleatoriedade (impureza) de um conjunto de exemplos (dados).
- ✓ Pode ser entendida como a dificuldade do atributo em questão prever os valores da variável alvo (classe).
- ✓ Considerando uma variável (atributo) qualquer com N valores possíveis, a probabilidade de acontecimento de cada um desses valores é igual a $\boldsymbol{p_n}$.
- ✓ É medida em bits, por isso a presença do logaritmo na base 2.

$$E(atributo) = -\sum_i p_i \times log_2 p_i$$

Entropia (E)


1.0
0.8
0.6
0.4
0.2
0.0
entropia
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
probabilidade

Entropia (E)

- Exemplo
  - Vamos medir a entropia para o atributo “Joga”.
  - Qual a probabilidade de jogar = “sim”?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Entropia (E)

- ✓ Exemplo
  - ✓ Vamos medir a entropia para o atributo “Joga”.
  - ✓ Qual a probabilidade de jogar = “sim”?

$$P(joga = sim) = \frac{9}{14}$$

Entropia (E)

- Exemplo
  - Vamos medir a entropia para o atributo “Joga”.
  - Qual a probabilidade de jogar = “sim”?

$$P(joga = sim) = \frac{9}{14}$$

  - Qual a probabilidade de jogar = “não” (de não jogar)?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Entropia (E)

- Exemplo
  - Vamos medir a entropia para o atributo “Joga”.
  - Qual a probabilidade de jogar = “sim”?

$$P(joga = sim) = \frac{9}{14}$$

  - Qual a probabilidade de jogar = “não” (de não jogar)?

$$P(joga = não) = \frac{5}{14}$$

Entropia (E)

- Exemplo
  - Vamos medir a entropia para o atributo “Joga”.
  - Qual a probabilidade de jogar = “sim”?

$$P(joga = sim) = \frac{9}{14}$$

  - Qual a probabilidade de jogar = “não” (de não jogar)?

$$P(joga = não) = \frac{5}{14}$$

$$E(atributo) = -\sum_i p_i \times log_2 p_i$$

$$E(Joga) = -\frac{9}{14} \times log_2 \frac{9}{14} + \left(-\frac{5}{14} \times log_2 \frac{5}{14}\right) = 0{,}940\ bit$$

Ganho de Informação (GI)

- ✓ A construção da árvore é guiada pela diminuição da entropia (E), ou seja, da aleatoriedade. Em outras palavras, a aleatoriedade pode ser entendida como a dificuldade de previsão das classes considerando um atributo em questão.

- ✓ E o GI é justamente a redução esperada da entropia que cada atributo causa em relação à variável alvo (classe).

- ✓ Logo, escolhemos o atributo que possuir maior valor de GI para compor o nó da sub árvore sendo gerada.

$$GI(S, A) = E(S) - E(A)$$

S é o atributo alvo (classe)
A é o atributo sendo avaliado

Ganho de Informação (GI)

- Exemplo: GI para o atributo “tempo”
  - Primeiro, precisamos calcular a entropia desse atributo, considerando todas suas combinações.
  - Qual a probabilidade de jogar dado que o tempo está com sol?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com sol?

$$P(joga = sim \mid tempo = sol) = \frac{2}{5}$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com sol?

$$P(joga = sim \mid tempo = sol\ ) = \frac{2}{5}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está com sol?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com sol?

$$P(joga = sim \mid tempo = sol\ ) = \frac{2}{5}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está com sol?

$$P(joga = não \mid tempo = sol\ ) = \frac{3}{5}$$

$$E(Joga \mid tempo = sol) = -\frac{2}{5} \times log_2 \frac{2}{5} + \left(-\frac{3}{5} \times log_2 \frac{3}{5}\right) = 0{,}971\ bit$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está nublado?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está nublado?

$$P(joga = sim \mid tempo = nublado) = \frac{4}{4}$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está nublado?

$$P(joga = sim \mid tempo = nublado\ ) = \frac{4}{4}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está nublado?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está nublado?

$$P(joga = sim \mid tempo = nublado) = \frac{4}{4}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está nublado?

$$P(joga = não \mid tempo = nublado) = \frac{0}{4}$$

$$E(Joga \mid tempo = nublado) = -\frac{4}{4} \times log_2 \frac{4}{4} + (-0 \times log_2 0) = 0$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com chuva?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com chuva?

$$P(joga = sim \mid tempo = chuva) = \frac{3}{5}$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com chuva?

$$P(joga = sim \mid tempo = chuva) = \frac{3}{5}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está com chuva?

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Primeiro, precisamos calcular a entropia desse atributo
  - ✓ Qual a probabilidade de jogar dado que o tempo está com chuva?

$$P(joga = sim \mid tempo = chuva) = \frac{3}{5}$$

  - ✓ Qual a probabilidade de não jogar dado que o tempo está com chuva?

$$P(joga = não \mid tempo = chuva) = \frac{2}{5}$$

$$E(Joga \mid tempo = chuva) = -\frac{3}{5} \times log_2 \frac{3}{5} + \left(-\frac{2}{5} \times log_2 \frac{2}{5}\right) = 0{,}971$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para o atributo “tempo”
  - ✓ Agora que temos a entropia para todas as combinações, vamos calcular a **entropia ponderada**, isto é, considerando as combinações de valores, “sol”, “chuva”, “nublado”.

$$E(Joga \mid tempo = sol) = 0{,}971$$
$$E(Joga \mid tempo = nublado) = 0$$
$$E(Joga \mid tempo = chuva) = 0{,}971$$

$$E(Joga) = 0{,}940$$

Soma das contagens para “sim” e para “não” para cada valor do atributo tempo

$$EP(tempo) = \frac{5}{14} \times 0{,}971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0{,}971 = \mathbf{0,693}$$

$$GI(S, A) = E(S) - E(A)$$

$$GI(Joga, tempo) = E(Joga) - EP(tempo)$$
$$GI(Joga, tempo) = 0{,}940 - 0{,}693$$
$$\boldsymbol{GI(Joga, tempo) = 0,247}$$

Ganho de Informação (GI)

- ✓ Exemplo: GI para os demais atributos
  - ✓ Aplicando a mesma estratégia para todos os demais atributos, obteremos o GI para cada um deles e, a partir disso, poderemos escolher qual o atributo mais adequado para compor o nó raiz da árvore.

Atributo	GI
Tempo	0,247
Temperatura	0,029
Umidade	0,057
Vento	0,048

Qual o atributo que será o nó raiz da árvore?

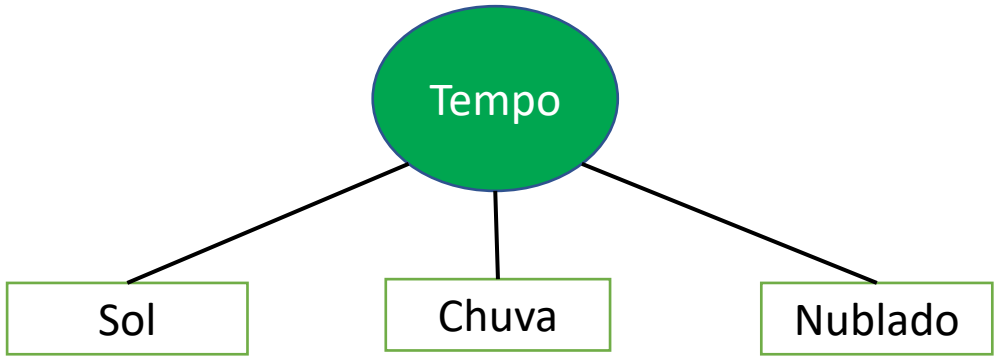
Ganho de Informação (GI)

- ✓ Exemplo: GI para os demais atributos
  - ✓ Aplicando a mesma estratégia para todos os demais atributos, obteremos o GI para cada um deles e, a partir disso, poderemos escolher qual o atributo mais adequado para compor o nó raiz da árvore.

Atributo	GI
Tempo	0,247
Temperatura	0,029
Umidade	0,057
Vento	0,048

Qual o atributo que será o nó raiz da árvore?

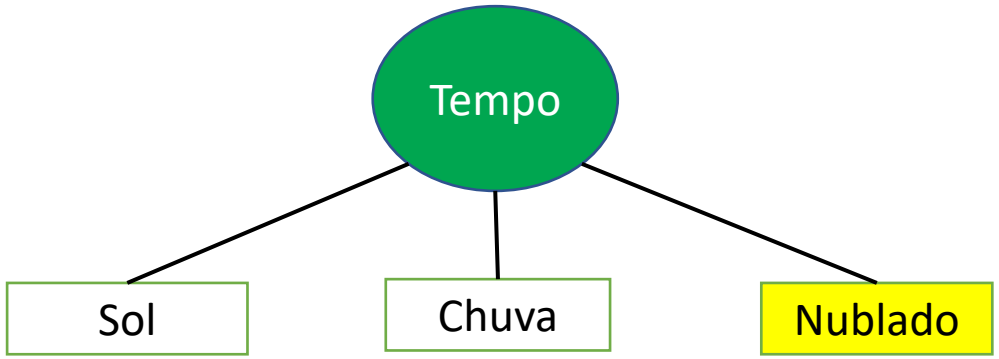
Árvore



O algoritmo vai repetindo o processo recursivamente até chegar em algum dos **critérios de parada**.

Ex: quando não houverem mais atributos no conjunto de treinamento.

Árvore



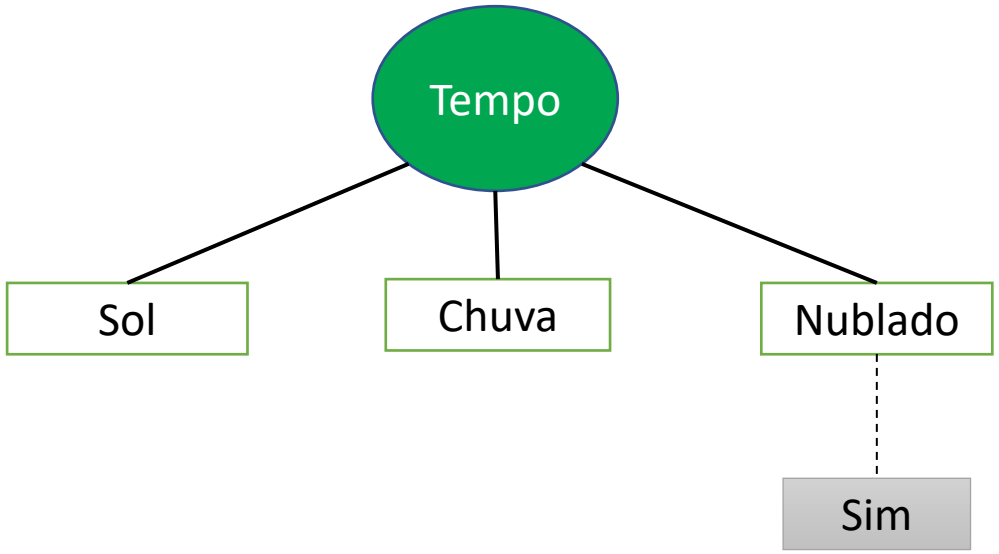
Vamos fazer uma chamada recursiva agora para o ramo “Nublado” para buscar uma nova sub árvore.

Base de dados “joga ou não joga”

Preditores				Classe
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	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Chuva	Média	Media	Não	Sim
Sol	Média	Baixa	Sim	Sim
Nublado	Média	Alta	Sim	Sim
Nublado	Alta	Baixa	Não	Sim
Chuva	Baixa	Alta	Sim	Não

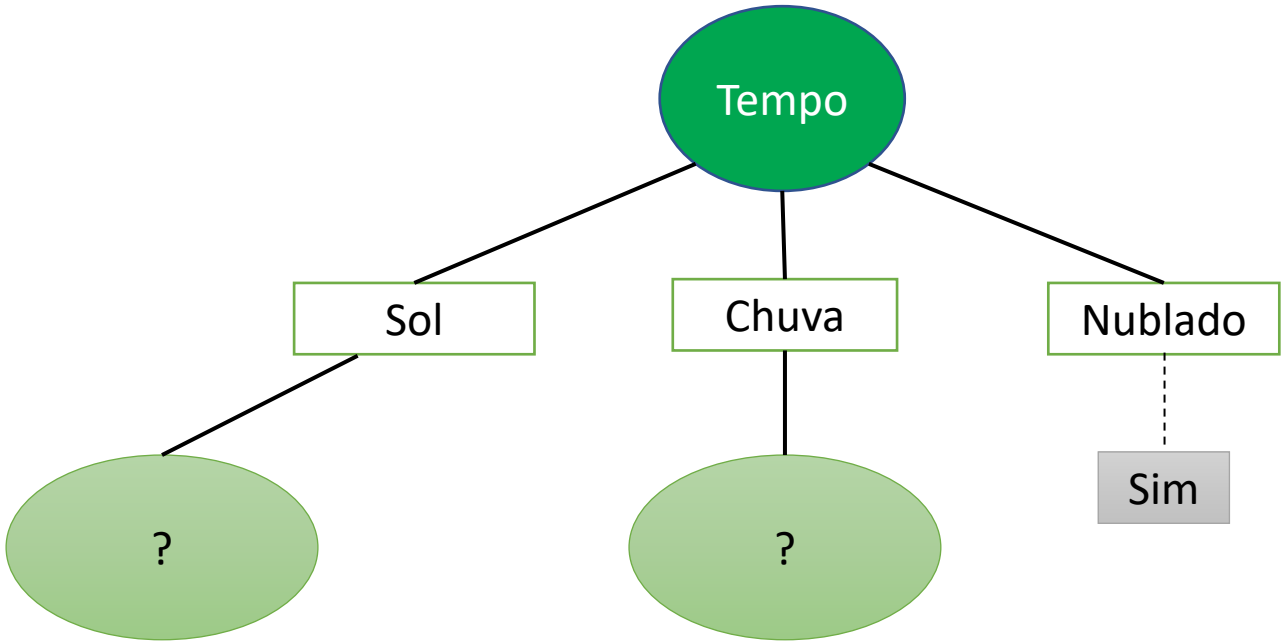
Observe como todos os valores de tempo “nublado” culminam numa mesma classificação “Sim” do atributo joga. Esse é um critério de parada que vai resultar em um nó folha. Veja como ficará a árvore agora.

Árvore



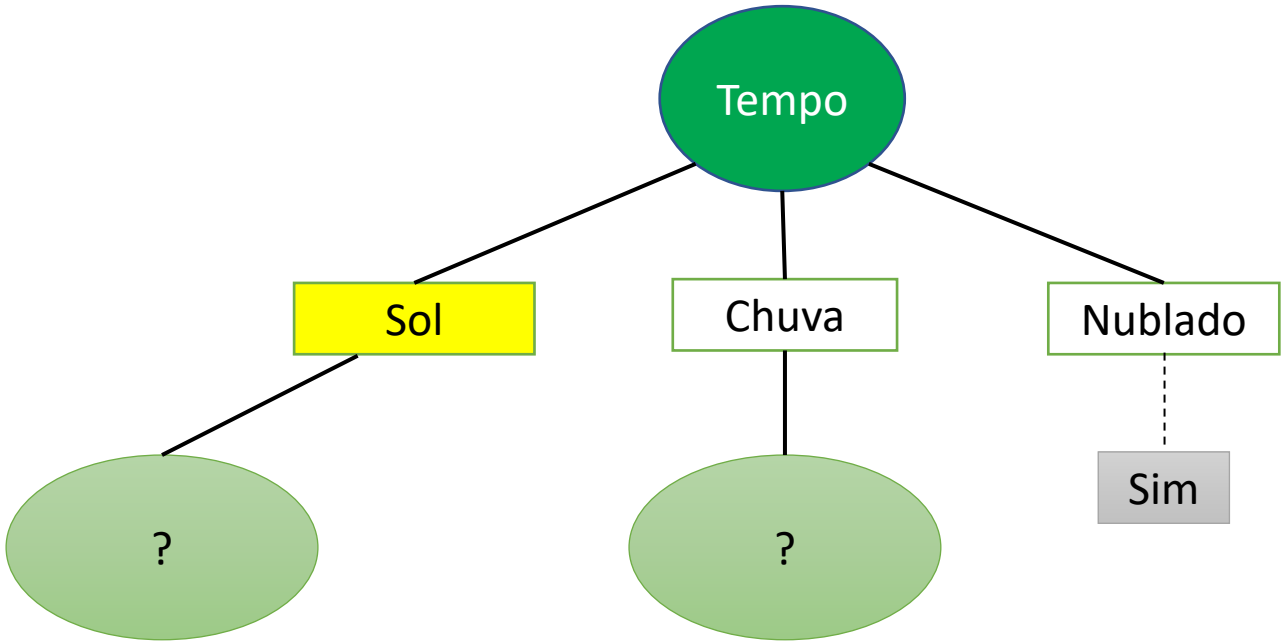
Chegamos em um nó folha, então, paramos a recursão para esse ramo da árvore.

Árvore



Os ramos “Sol” e “Chuva” ainda estão indefinidos. O processo continua, pois temos atributos que podem ser colocados nos nós ainda.

Árvore



Vamos analisar o ramo “Sol”. Para isso, as amostras do conjunto de treinamento são divididas em subconjuntos de acordo com os valores do atributo tempo.

Base de dados “joga ou não joga”

Preditores				Classe
Tempo	**Temperatura**	**Umidade**	**Vento**	**Joga**
Sol	Alta	Media	Não	Não
Sol	Alta	Alta	Sim	Não
Sol	Média	Alta	Não	Não
Sol	Baixa	Baixa	Não	Sim
Sol	Média	Baixa	Sim	Sim

$$GI(S, A) = \boldsymbol{E(S)} - \boldsymbol{E(A)}$$

Observe que E(S) agora não é mais da classe

$$\boldsymbol{E(S)} = E(Joga \mid tempo = sol) = 0{,}971$$

$$\boldsymbol{E(A)} = EP(temperatura) = 0{,}4$$

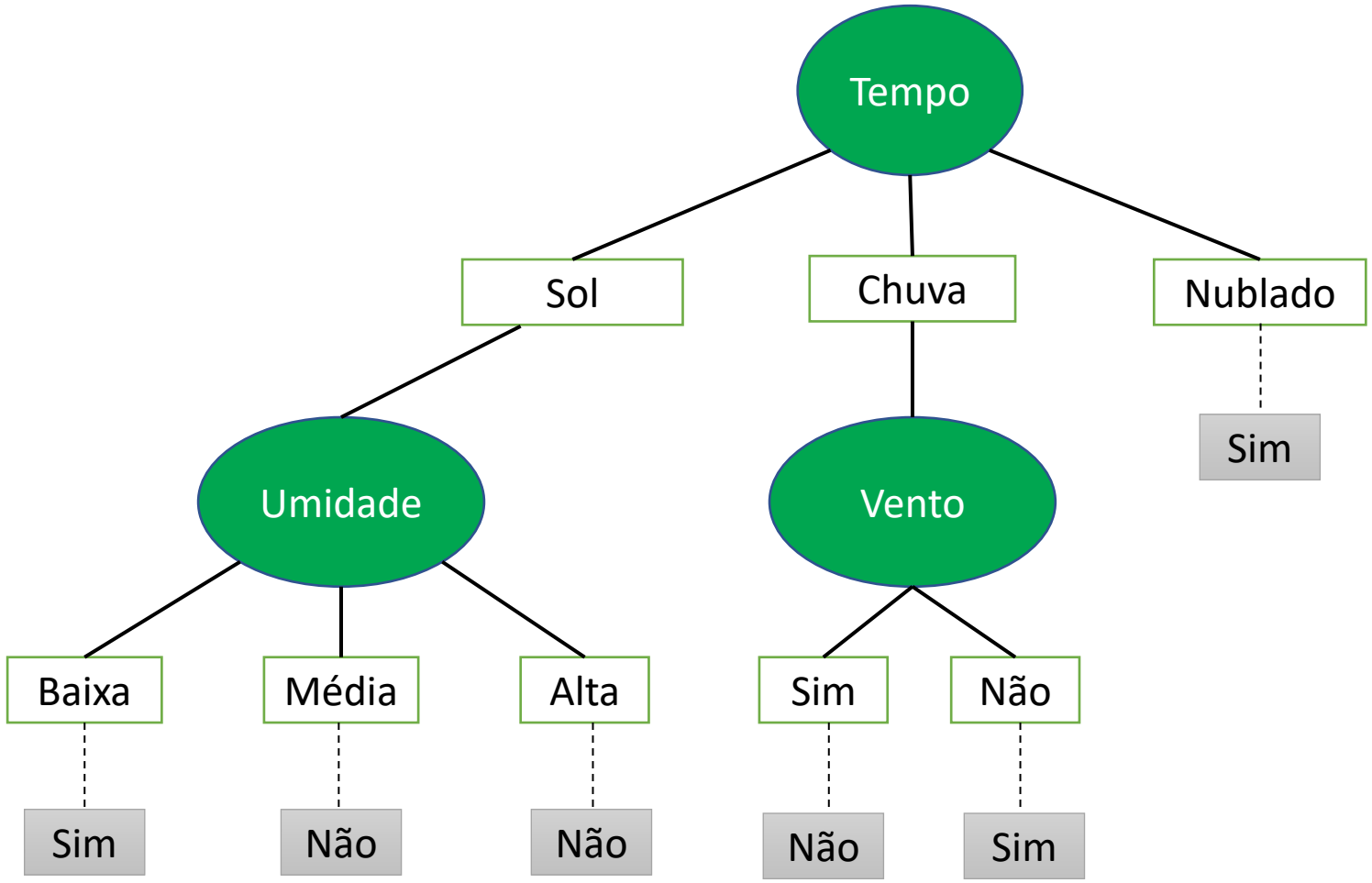
Calculada da mesma forma que fizemos para o nó raiz (tempo)

$$GI(tempo = sol, temperatura) = E(Joga \mid tempo = sol) = -EP(temperatura)$$

$$GI(tempo = sol, temperatura) = 0{,}971 - 0{,}4$$

$$\boldsymbol{GI(tempo = sol, temperatura) = 0{,}571}$$

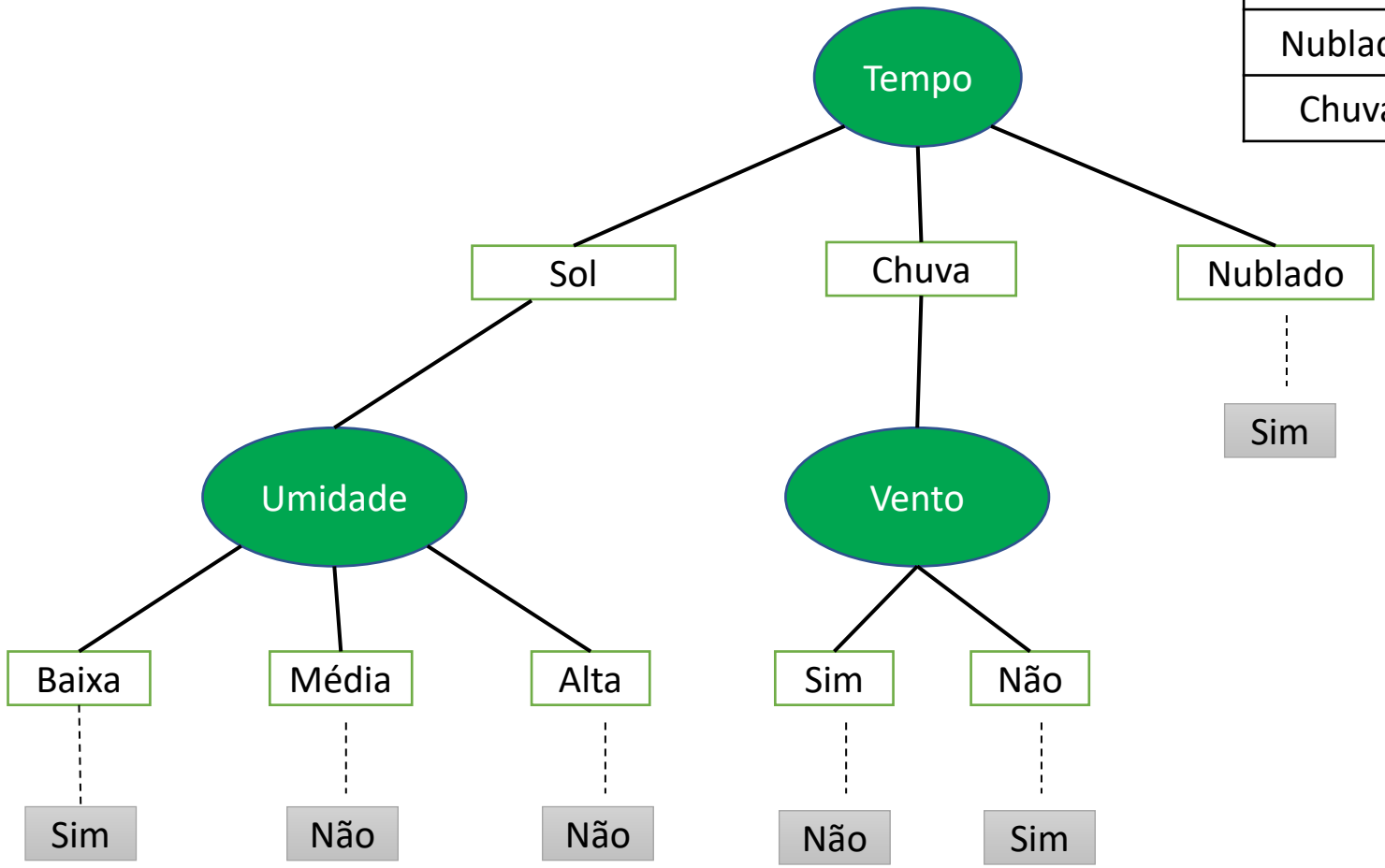
Árvore



Observe que pela temperatura ter apresentado o menor valor de GI, não a consideramos na construção da árvore.

Exercício: qual a classe dos exemplos?

Tempo	Temp.	Umidade	Vento	Joga
Sol	Alta	Media	Não	
Sol	Alta	Alta	Sim	
Nublado	Alta	Alta	Não	
Chuva	Baixa	Alta	Não	



Pergunta

Até agora trabalhamos com um conjunto de dados com atributos apenas categóricos. E se tivermos **atributos numéricos**, como fazemos para determinar o GI?

Prestemos atenção na base do próximo slide.

Base de dados “joga ou não joga”

Preditores				Classe
Tempo	**Temperatura**	**Umidade**	**Vento**	**Joga**
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	78	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	80	Sim	Não

Ganho de Informação (GI)

- E quando tivermos um atributo contínuo? Como fica a E e o GI?
  - Temperatura: 40, 48, 60, 72, 80, 90
  - Joga: Não, Não, Sim, Sim, Sim, Não

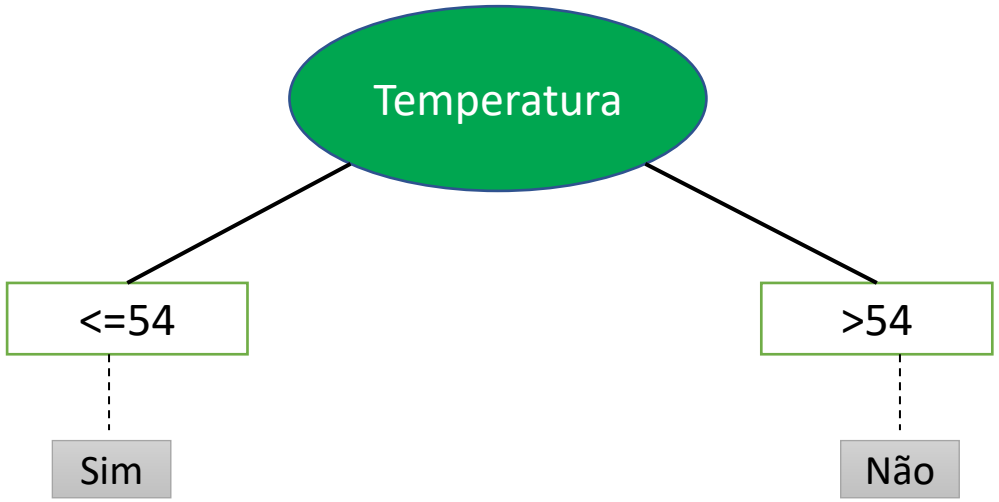
- Escolhemos um limiar C que produza o maior GI.
- Identificar exemplos que diferem na classificação do alvo.
  - C = (48+60)/2 = 54

40, 48,	60, 72, 80, 90

Ganho de Informação (GI)

- ✓Temperatura e umidade agora são valores contínuos e devem ser analisados de maneira ordenada, antes de qualquer análise em termos de GI.
- ✓Seja $v = (v_1, v_2, \ldots, v_n)$ o conjunto de valores possíveis para um atributo. Devemos ordená-los em ordem crescente.

Ganho de Informação (GI)



Ganho de Informação (GI)

- ✓ Para chegar na estrutura de árvore do slide anterior, teríamos que fazer os cálculos da mesma forma que fizemos para os atributos categóricos anteriormente, isto é, da entropia, da entropia ponderada e, por fim, do ganho de informação

$$P(joga = sim \mid Temperatura \leq 50)$$
$$P(joga = não \mid Temperatura \leq 50)$$
$$P(joga = sim \mid Temperatura > 50)$$
$$P(joga = não \mid Temperatura > 50)$$

$$E(Joga \mid temperatura \leq 50)$$

$$E(Joga \mid temperatura > 50)$$

$$EP(temperatura)$$

$$GI(joga, temperatura) = E(joga) - EP(temperatura)$$

$$\boldsymbol{GI(joga, temperatura) = 0,940 - EP(temperatura)}$$

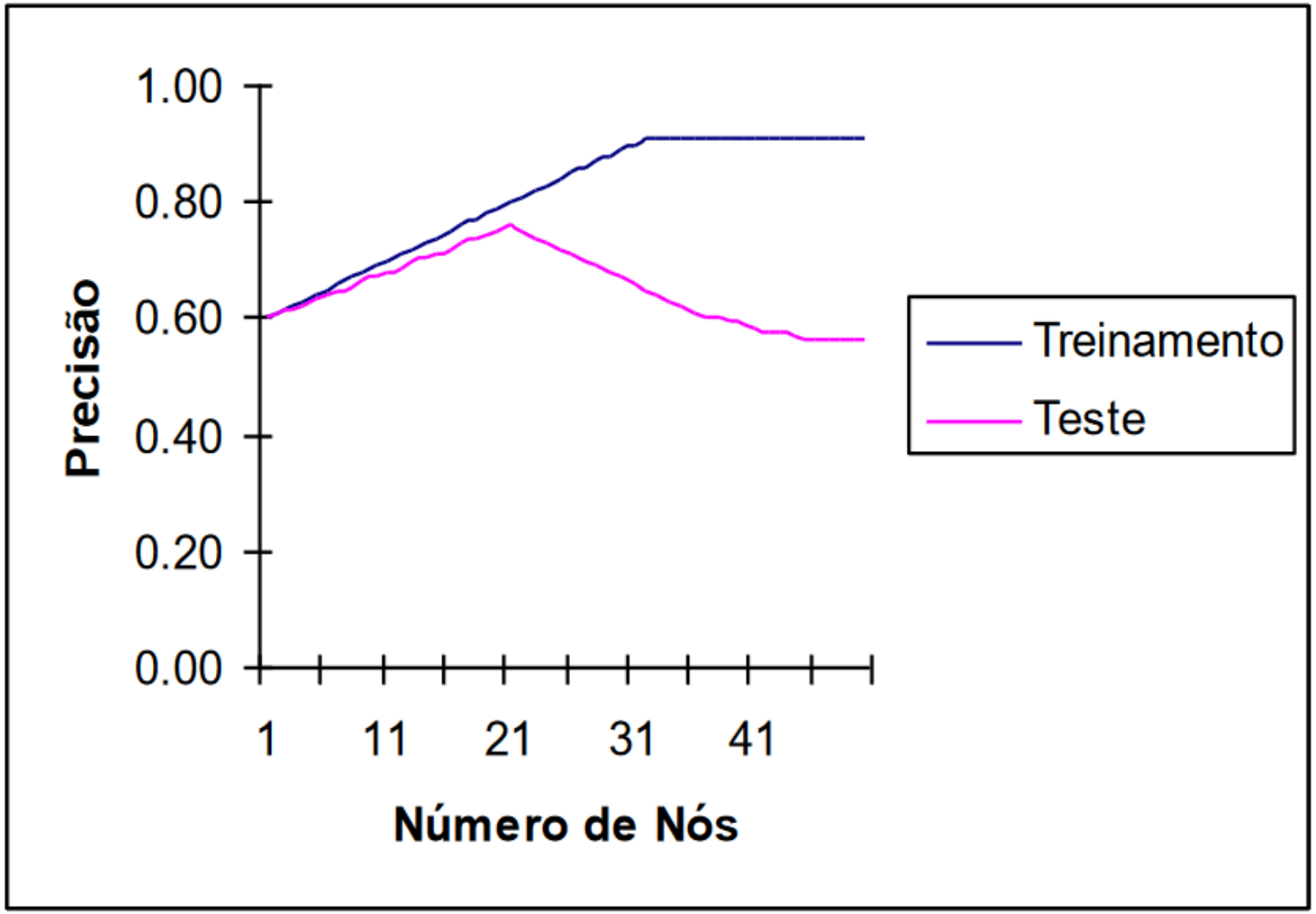
Sobre-ajustamento (Overfitting)

- ✓ Aprendizado muito específico em relação ao conjunto de treinamento, não permitindo ao modelo (árvore) generalizar, isto é, prever com boa acurácia exemplos desconhecidos.

- ✓ Sub árvores podem refletir ruídos ou erros dos dados de treinamento.

- ✓ Para diminuir esse efeito, é preciso detectar e excluir essas sub árvores
  - ✓ Métodos de poda

Sobre-ajustamento (Overfitting)



Métodos de poda

- Melhorar a taxa de acerto do modelo para novos exemplos que não foram utilizados para treinar (gerar) a árvore. Uma árvore podada é mais fácil de ser interpretada.

- **Pré-poda:** ocorre ao mesmo tempo que a árvore está sendo gerada (treinada). Quando a partição não é estatisticamente significativa esse nó é excluído e transformado em nó folha.

- **Pós-poda:** gera uma árvore completa e só depois faz a poda. Tudo aquilo que está abaixo de um nó interno é excluído e esse nó é transformado em folha.
  - Demora mais, mas é mais confiável.

Algoritmos: ID3

- ✓ Iterative Dichotomiser 3.
- ✓ É um algoritmo recursivo, baseado em busca exaustiva.
- ✓ Utiliza o ganho de informação para selecionar o melhor atributo.
- ✓ Sua principal limitação é que só trabalha com atributos categóricos.
  - ✓ Logo, se aplicarmos a dados numéricos, haverá um ramo para cada valor possível distinto. Por exemplo: Um atributo altura que assume valores dentro do conjunto dos número reais terá tantos nós quanto forem os valores distintos para o atributo.
- ✓ Não apresenta nenhuma forma para tratar valores desconhecidos.
- ✓ Não apresenta nenhum método de pós-poda.

Algoritmos: C4.5

- ✓ Usa a técnica “dividir para conquistar”.
- ✓ Aceita tanto atributos categóricos como atributos numéricos.
- ✓ Trata valores desconhecidos.
- ✓ Utiliza a medida de razão de ganho para selecionar o atributo que melhor divide os exemplos.
- ✓ Apresenta um método de pós-poda das árvores geradas.

Algoritmos: CART

- ✓ Induz tanto árvores de classificação (atributo classe é nominal) e de regressão (atributo classe é contínuo).
- ✓ Define limiares a serem utilizados nos nós para dividir os atributos contínuos.
- ✓ As árvores geradas são sempre binárias.
- ✓ Realizada o método de pós-poda.

Dúvidas?

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br