Árvore de decisão

Prof. Hugo de Paula

Estrutura da árvore de decisão

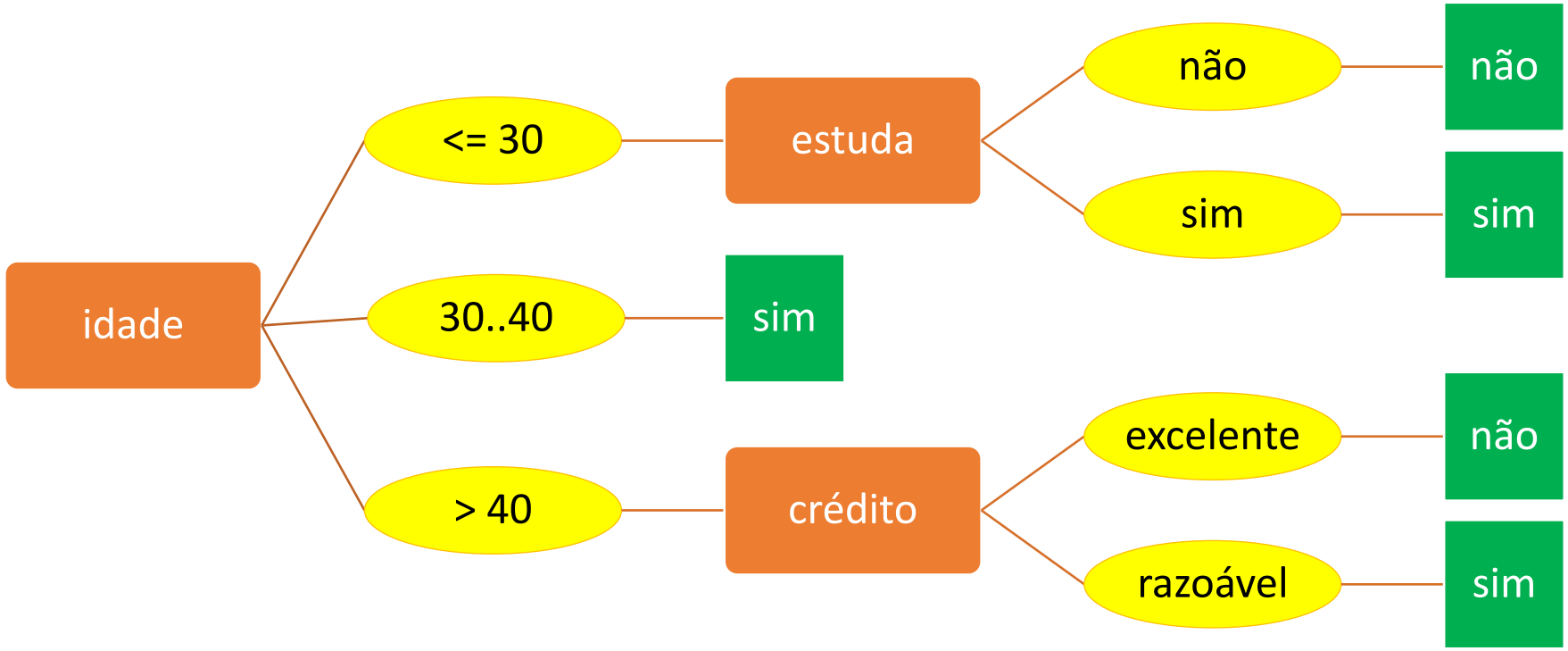
- cada nó é um atributo da base de dados.
- nós folha são do tipo do atributo-classe (ou rótulo, label),
- cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai.
- um atributo que aparece num nó não pode aparecer em seus nós descendentes.

Árvore de decisão: exemplo
Exemplo de Quinlan's ID3

Idade	Renda	Estuda	Crédito	Compra computador
<=30	alta	não	razoável	não
<=30	alta	não	excelente	não
31…40	alta	não	razoável	sim
>40	média	não	razoável	sim
>40	baixa	sim	razoável	sim
>40	baixa	sim	excelente	não
31…40	baixa	sim	excelente	sim
<=30	média	não	razoável	não
<=30	baixa	sim	razoável	sim
>40	média	sim	razoável	sim
<=30	média	sim	excelente	sim
31…40	média	não	excelente	sim
31…40	alta	sim	razoável	sim
>40	média	não	excelente	não

Árvore de decisão: exemplo

- Uma possível árvore de decisão criada pelo algoritmo: o usuário é um potencial comprador ou não.



Principais algoritmos de indução de árvore de decisão

- ID3 (final dos anos 1970)
  Iterative Dichotomiser
- C45 (sucessor do ID3)
- CART (1984)
  Classification and Regression Trees

PUC Minas Virtual

Métodos de seleção de atributos

- Ganho de informação (ID3).
- Taxa de ganho (C4.5, J48).
- Índice GINI - impureza (CART).
- Redução de variância (CART).

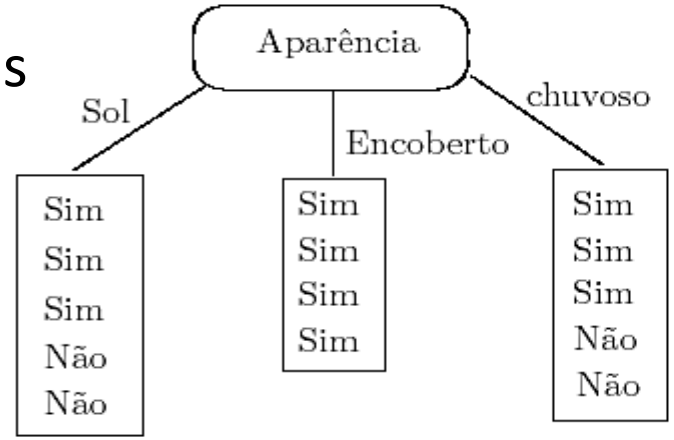
Árvore de decisão: exemplo

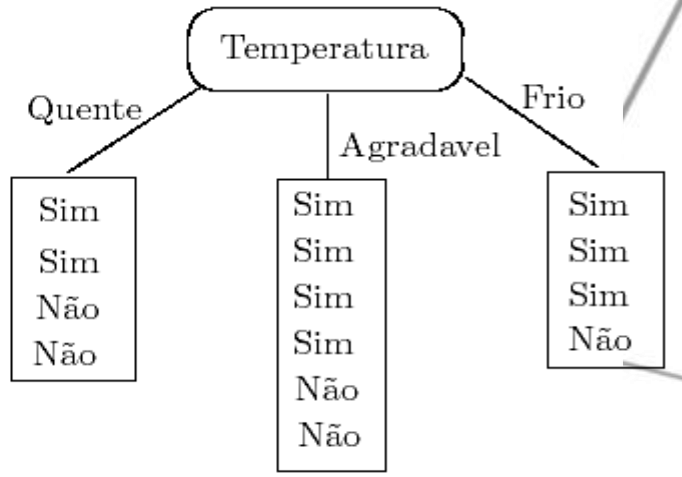
- Considere a base de treinamento.
- O objetivo é identificar quais as condições ideais para se jogar um determinado jogo.

Aparência	Temperatura	Umidade	Vento	Jogar
Sol	Quente	Alta	Fraco	Não
Sol	Quente	Alta	Forte	Não
Encoberto	Quente	Alta	Fraco	Sim
Chuvoso	Moderado	Alta	Fraco	Sim
Chuvoso	Frio	Normal	Fraco	Sim
Chuvoso	Frio	Normal	Forte	Não
Encoberto	Frio	Normal	Forte	Sim
Sol	Moderado	Alta	Fraco	Não
Sol	Frio	Normal	Fraco	Sim
Chuvoso	Moderado	Normal	Fraco	Sim
Sol	Moderado	Normal	Forte	Sim
Encoberto	Moderado	Alta	Forte	Sim
Encoberto	Quente	Normal	Fraco	Sim
Chuvoso	Moderado	Alta	Forte	Não

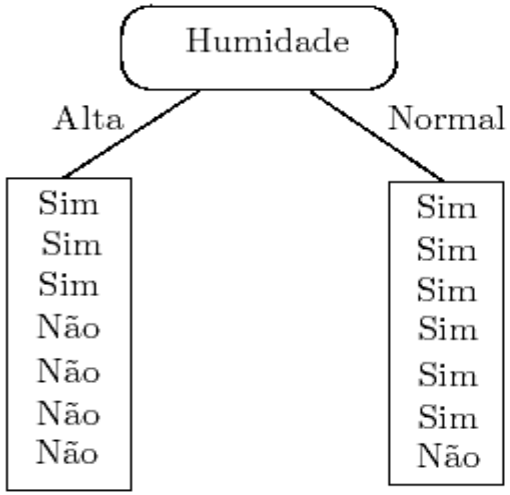
Árvore de decisão: exemplo

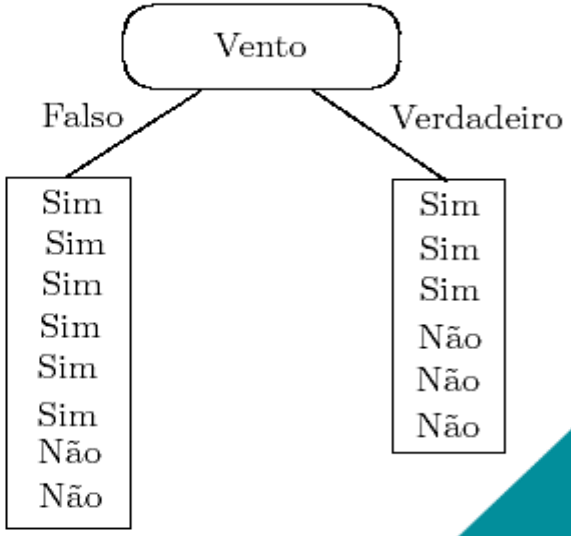
As quatro possibilidades para o atributo do nó raiz.





Critério de escolha intuitivo:
atributo que produz os nós mais puros.





Árvore de decisão: exemplo

Entropia do atributo **Aparência**:

$$I(Aparencia) = \frac{5}{14}E(Folha_1) + \frac{4}{14}E(Folha_2) + \frac{5}{14}E(Folha_3)$$

$$E(Folha_1) = \frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$E(Folha_2) = \frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$E(Folha_3) = \frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5} = 0.971$$

logo

$$I(Aparencia) = \frac{5}{14}0.971 + \frac{4}{14}0 + \frac{5}{14}0.971 = 0.693$$

Entropia do atributo **Temperatura**:

$$I(Temperatura) = \frac{4}{14}E(Folha_1) + \frac{6}{14}E(Folha_2) + \frac{4}{14}E(Folha_3) = 0.911$$

Entropia do atributo **Humidade**:

$$I(Humidade) = \frac{7}{14}E(Folha_1) + \frac{7}{14}E(Folha_2) = 0.788.$$

Ganho da informação:

$$I(B) = \frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

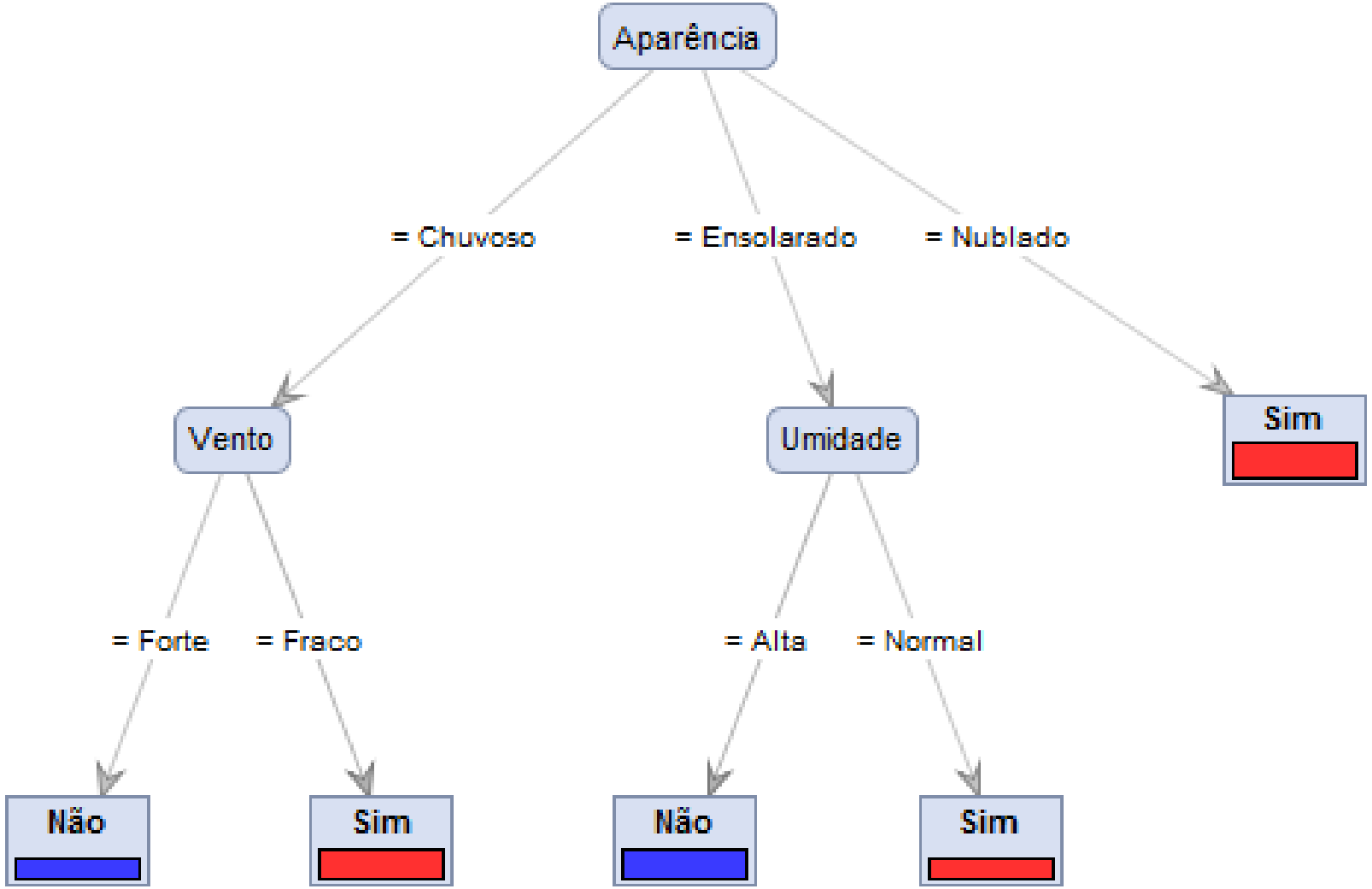
$$\boldsymbol{G(Aparencia) = 0.940 - 0.693 = 0.247}$$

$$G(Tempertura) = 0.940 - 0.911 = 0.029$$

$$G(Humidade) = 0.940 - 0.788 = 0.152$$

$$G(Vento) = 0.940 - 0.892 = 0.020$$

Resultado final da árvore

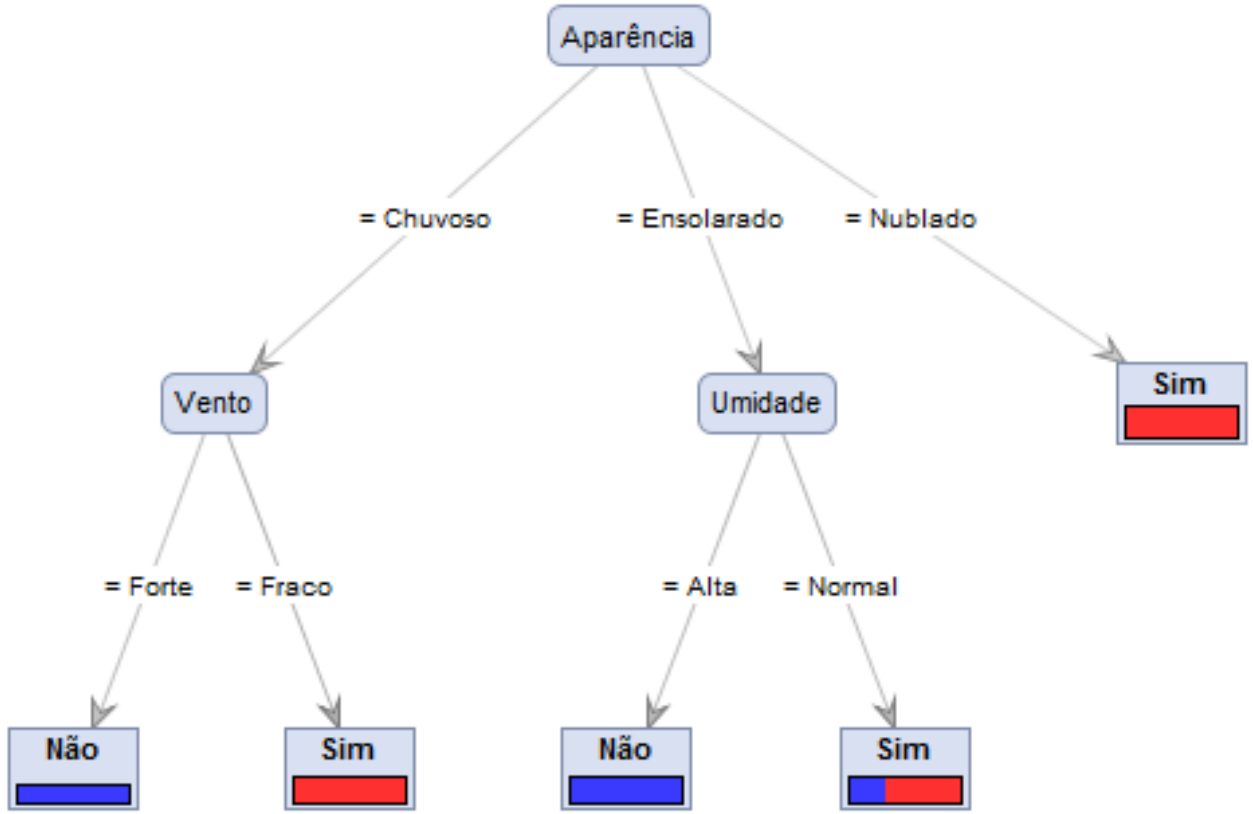


Overfitting (superajustamento) em árvores de decisão

Considere o seguinte ruído na base de treinamento:

<Ensolarado, Quente, Normal, Forte, Não>

Nova árvore:



Overfitting

Considere uma hipótese *h* e:

Taxa de erro sobre o conjunto de treinamento: $err_{train}(h)$

Erro real sobre todo conjunto de dados: $err_{real}(h)$

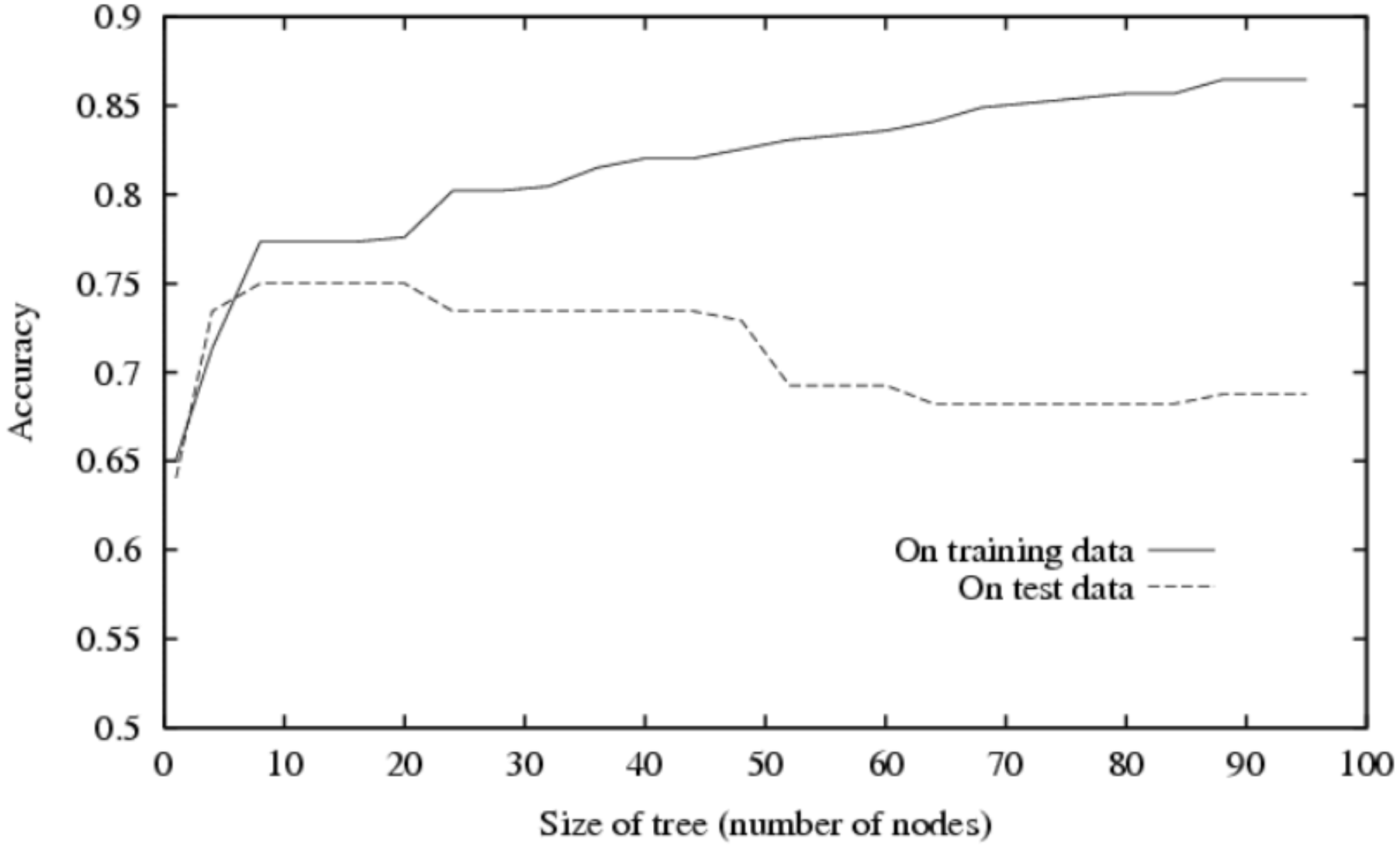
Diz-se que *h* sofre *overfitting* referente aos dados de treinamento se:

$$err_{real}(h) > err_{train}(h)$$

Quantidade de *overfitting*

$$err_{real}(h) - err_{train}(h)$$

Overfitting



Evitando *overfitting*

- Parar de crescer a árvore quando não for estatisticamente relevante.
- Gerar a árvore completa e depois podá-la.

Poda com erro reduzido

Dividir dado em *treinamento* e *validação*.

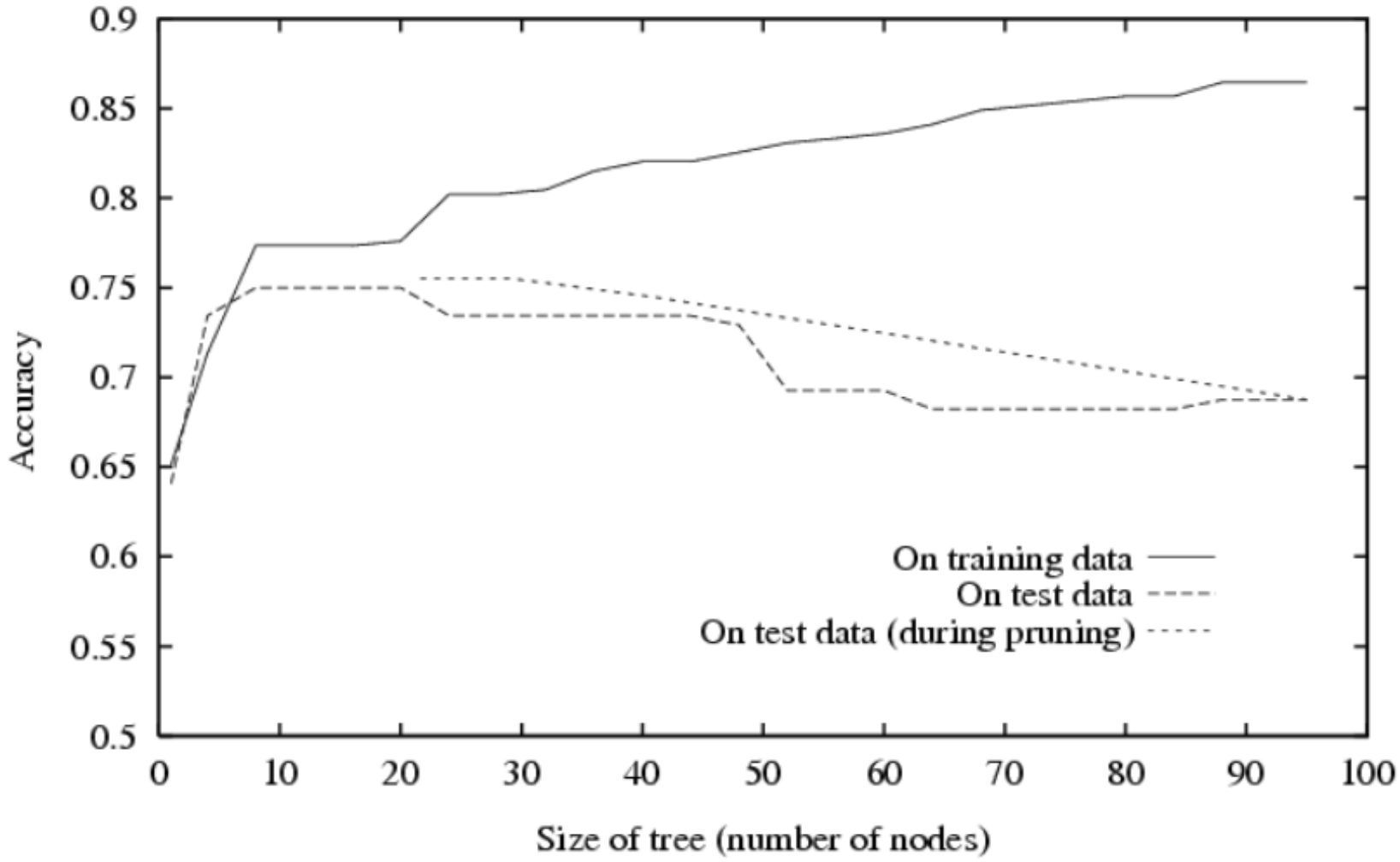
Criar árvore que classifica *treinamento* corretamente

Repetir até que seja prejudicial ao modelo

Avaliar o impacto da poda de cada nó (e seus descendentes) da árvore na *validação*.

Remover nó que mais aumenta a acurácia na *validação* (algoritmo guloso).

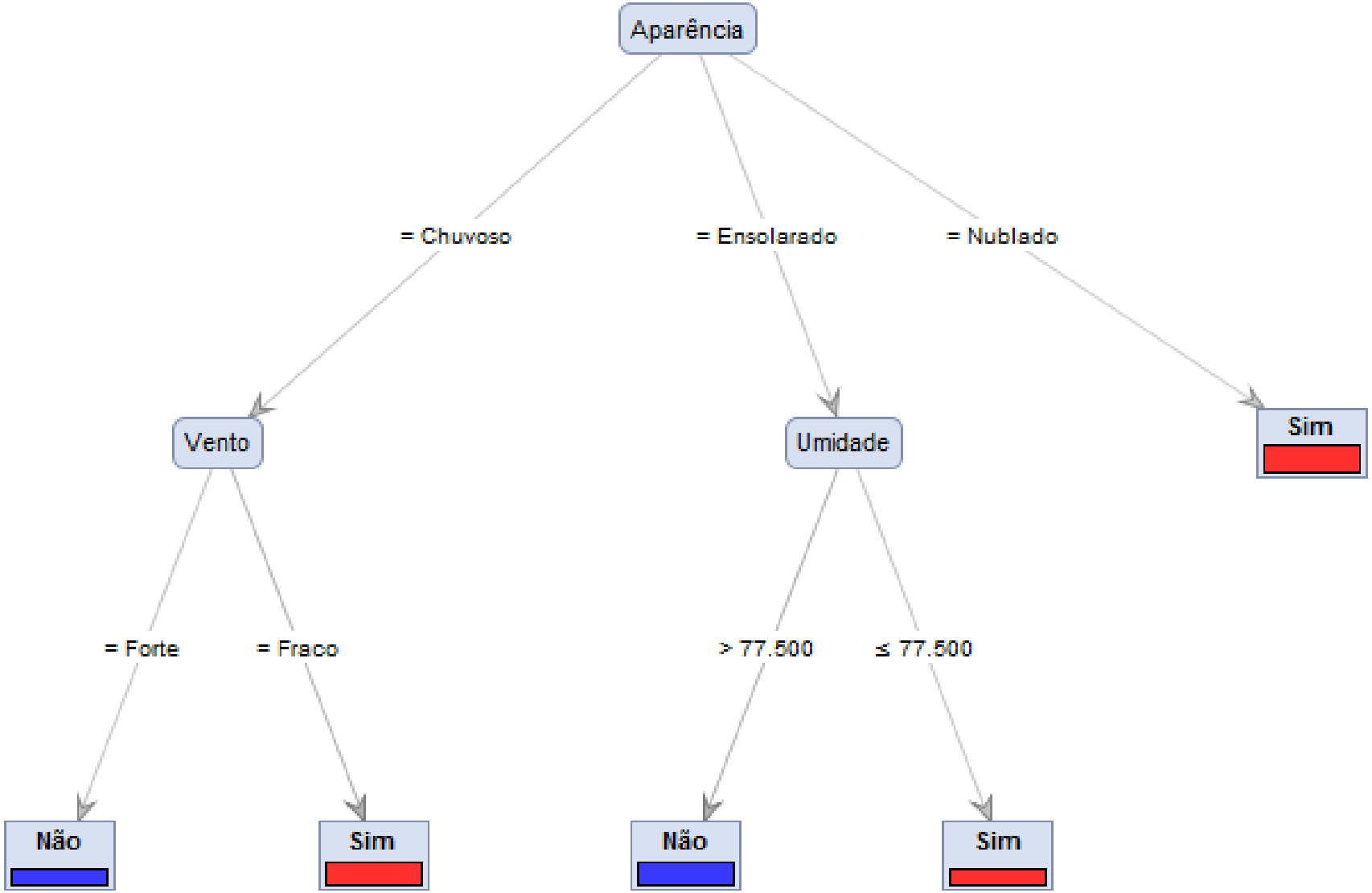
Efeito da poda com erro reduzido no *Overfitting*



Árvores com atributos contínuos

Aparência	Temperatura	Umidade	Vento	Jogar
Sol	Quente	85	85	Não
Sol	Quente	80	90	Não
Encoberto	Quente	83	86	Sim
Chuvoso	Moderado	70	96	Sim
Chuvoso	Frio	68	80	Sim
Chuvoso	Frio	65	70	Não
Encoberto	Frio	64	65	Sim
Sol	Moderado	72	95	Não
Sol	Frio	69	70	Sim
Chuvoso	Moderado	75	80	Sim
Sol	Moderado	75	70	Sim
Encoberto	Moderado	72	90	Sim
Encoberto	Quente	81	75	Sim
Chuvoso	Moderado	71	91	Não

Árvores com atributos contínuos



PUC Minas
Virtual