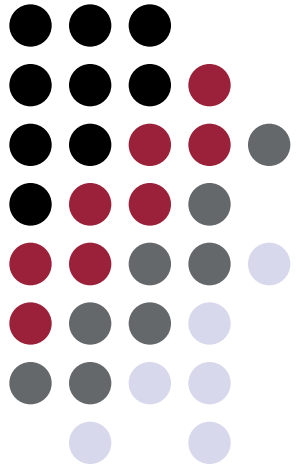
Mineração de Dados: Classificação II

CEA462 – Sistemas de Apoio à Decisão

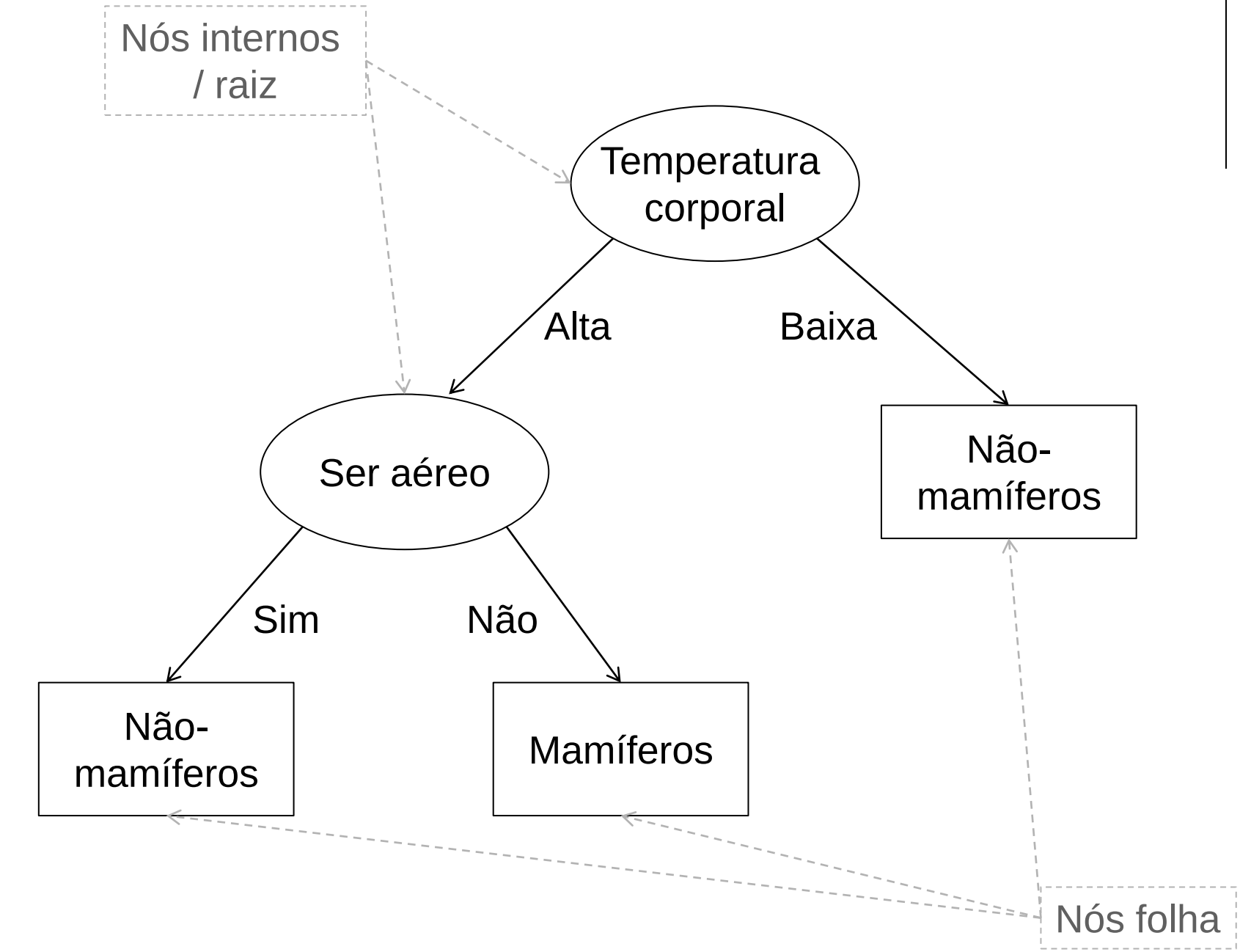
Prof. MSc. George H. G. Fonseca
Universidade Federal de Ouro Preto



Árvore de Decisão

- Dois tipos de nós
  - Nós internos / raiz
    - Representam um atributo, que divide o conjunto de dados
  - Nós folha
    - Representam um rótulo

Árvore de Decisão



Árvore de Decisão

- Mas, qual a melhor ordem dos atributos na árvore?
  - Há várias métricas, mas considerar-se-á **Entropia** (E(*n*))
    - “Impureza” de valor do atributo
    - *C*: Conjunto de classes
    - *v(a)*: Valor de atributo
    - *p*(*i* | *v*(*a*)): Probabilidade de elemento *i* pertencer à classe *c*
  - $E(v(a)) = \sum_{c \in C}\left(-\, p(i \mid v(a)\right) * \; log_2\, p(i \mid v(a)) \,)$

Árvore de Decisão

- E a entropia de cada atributo?
  - $E(A) = \sum_{v(a) \in V(A)} (\frac{N_{v(a)}}{N} * E(v(a)))$
  - De forma menos formal:
    - “Média ponderada das entropias de cada valor possível”

Árvore de Decisão

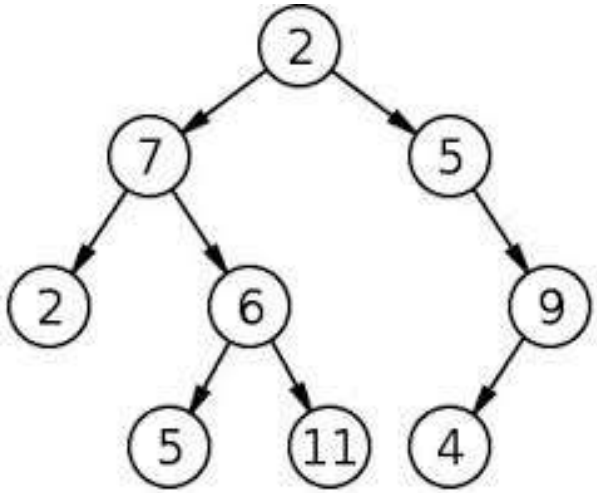
- Mas, qual a melhor ordem dos atributos na árvore?
  - **Ganho de informação**
    - Compara-se a impureza do nó pai com a impureza do atributo (nó) candidato
    - *v(a)*: Valor possível para um atributo *A*
    - *N*: Número de elementos no nó pai
    - *N*(*v*(*a*)): Número de elementos no nó pai que possui valor *v* p/ o atributo *a*
  - $\Delta = E(pai) - \sum_{v(a) \in V(a)} \frac{N(v(a))}{N} E(v(a))$

Árvore de Decisão

- Encontrar a árvore de decisão ótima (e.g. que melhor classifica os dados) é um problema *NP*-completo
  - Há de se testar todas as possíveis configurações de árvore para garantir que uma é a melhor
  - A abordagem dada é heurística e gulosa (caminha sempre para o melhor nó) e consegue resultados satisfatórios

Árvore de Decisão

- Construída a árvore, a classificação é feita em O(*h*), onde *h* é a altura da árvore
  - Muito eficiente



Árvore de Decisão
Exemplo

- Liberar empréstimo (2) ou não (1)
  - Considerar-se-á
    - Renda mensal [0, 1,5k] = baixa
    - Renda mensal]1,5k, 3k] = média
    - Renda mensal]3, ∞] = alta

ID	Renda mensal	Casa própria	Classe
1	1,0k	Não	1
2	6,0k	Sim	2
3	1,3k	Não	1
4	2,0k	Não	1
5	1,1k	Não	1
6	1,8k	Sim	1
7	4,3k	Sim	2
8	0,9k	Não	1
8	2,7k	Sim	2

Árvore de Decisão
Exemplo

- Calcular as entropias
  - $E(rm_b) = -p_1 * log_2(p_1) - p_2 * log_{2(p_2)}$

    $$E(rm_b) = -\frac{4}{4} * log_2\left(\frac{4}{4}\right) - \frac{0}{4} * log_2\left(\frac{0}{4}\right)$$

    $$E(rm_b) = 0$$

    Renda mensal = baixa

  - $E(rm_m) = -p_1 * log_2(p_1) - p_2 * log_{2(p_2)}$

    $$E(rm_m) = -\frac{2}{3} * log_2\left(\frac{2}{3}\right) - \frac{1}{3} * log_2\left(\frac{1}{3}\right)$$

    $$E(rm_m) = 0{,}92$$

    Renda mensal = média

  - $E(rm_a) = -p_1 * log_2(p_1) - p_2 * log_{2(p_2)}$

    $$E(rm_a) = -\frac{0}{2} * log_2\left(\frac{0}{2}\right) - \frac{2}{2} * log_2\left(\frac{2}{2}\right)$$

    $$E(rm_a) = 0$$

    Renda mensal = alta

Árvore de Decisão
Exemplo

- Calcular as entropias
  - $E(cp_s) = -p_1 * log_2(p_1) - p_2 * log_{2(p_2)}$

$$E(cp_s) = -\frac{1}{4} * log_2\left(\frac{1}{4}\right) - \frac{3}{4} * log_2\left(\frac{3}{4}\right)$$

$$E(cp_s) = 0{,}82$$

Casa própria = sim

  - $E(cp_n) = -p_1 * log_2(p_1) - p_2 * log_{2(p_2)}$

$$E(cp_n) = -\frac{5}{5} * log_2\left(\frac{5}{5}\right) - \frac{0}{5} * log_2\left(\frac{0}{5}\right)$$

$$E(cp_n) = 0$$

Casa própria = não

ID	Renda mensal	Casa própria	Classe
1	1,0k	Não	1
2	6,0k	Sim	2
3	1,3k	Não	1
4	2,0k	Não	1
5	1,1k	Não	1
6	1,8k	Sim	1
7	4,3k	Sim	2
8	0,9k	Não	1
9	2,7k	Sim	2

Árvore de Decisão
Exemplo

- Calcular as entropias
  - $E(rm) = \frac{N_b}{N} * E(rm_b) + \frac{N_m}{N} * E(rm_m) + \frac{N_a}{N} * E(rm_a)$

$$E(rm) = \frac{4}{9} * 0 + \frac{3}{9} * 0{,}92 + \frac{2}{9} * 0$$

$$E(rm) = 0{,}31$$

Renda mensal

  - $E(cp) = \frac{N_s}{N} * E(cp_s) + \frac{N_n}{N} * E(cp_n)$

$$E(cp) = \frac{4}{9} * 0{,}82 + \frac{5}{9} * 0$$

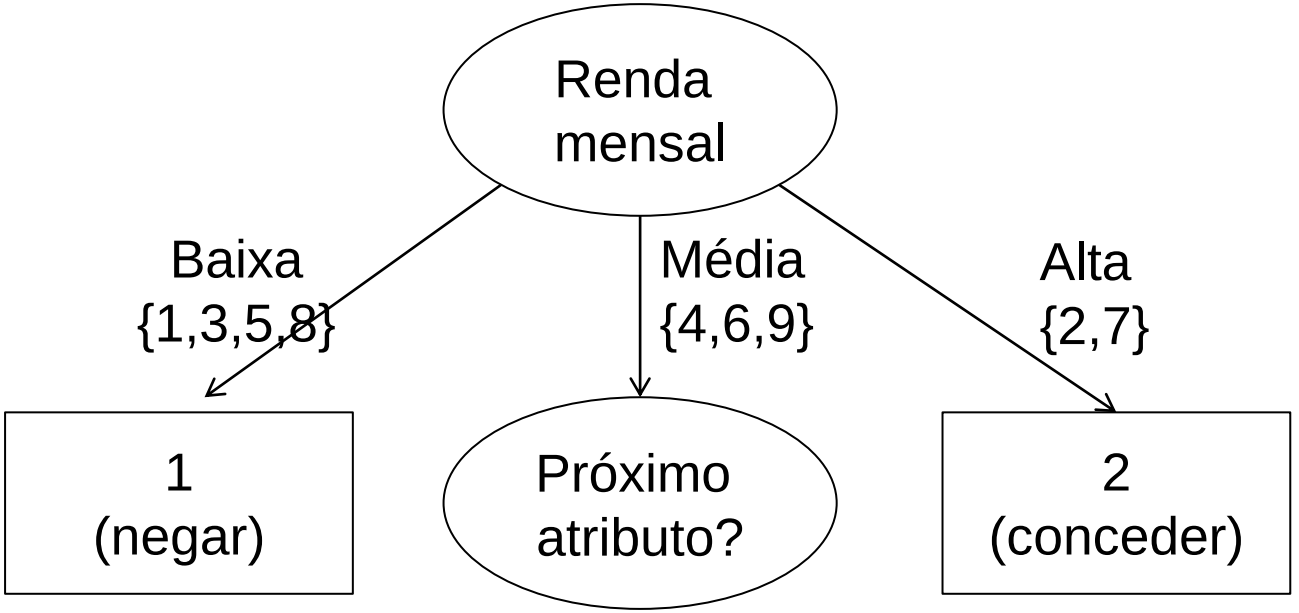
$$E(cp) = 0{,}36$$

Casa própria

ID	Renda mensal	Casa própria	Classe
1	1,0k	Não	1
2	6,0k	Sim	2
3	1,3k	Não	1
4	2,0k	Não	1
5	1,1k	Não	1
6	1,8k	Sim	1
7	4,3k	Sim	2
8	0,9k	Não	1
9	2,7k	Sim	2

Árvore de Decisão
Exemplo

- Nó raiz
  - Renda mensal comporá o nó raiz uma vez que possui a menor entropia
  - Caso todos os registros figurem na mesma classificação, o nó filho é folha



Árvore de Decisão
Exemplo

- Próximo atributo...
  - $\Delta(cp) = E(pai) - \sum_{v(a) \in V(A)} \frac{N(v(a))}{N} E(v(a))$

$$\Delta(cp) = 0{,}92 - \left(\frac{N(v_s)}{N} * E(v_s) + \frac{N(v_n)}{N} * E(v_n) \right)$$

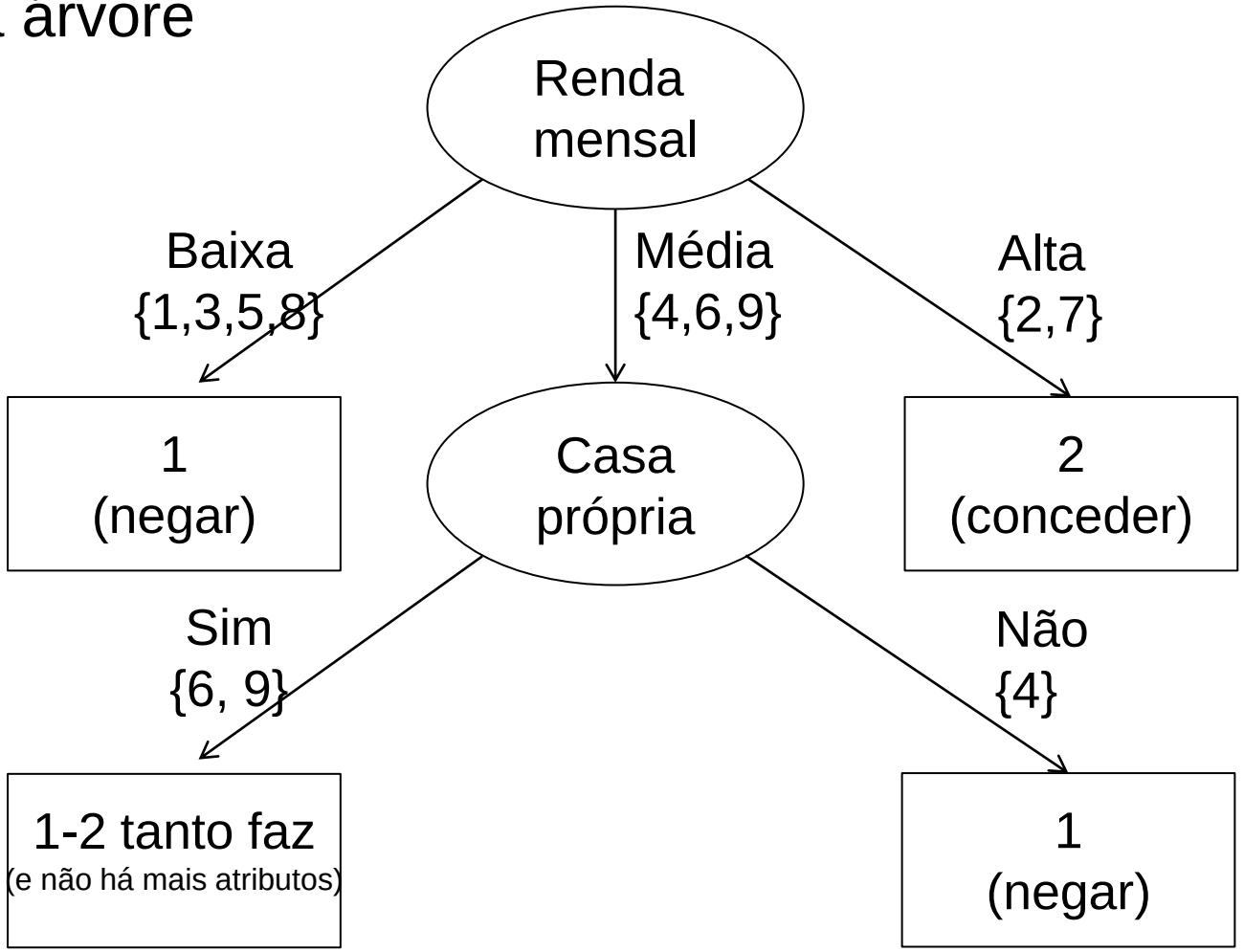
$$\Delta(cp) = 0{,}92 - \left(\frac{2}{3} * 0{,}82 + \frac{1}{3} * 0 \right)$$

$$\Delta(cp) = 0{,}37$$

ID	Renda mensal	Casa própria	Classe
1	1,0k	Não	1
2	6,0k	Sim	2
3	1,3k	Não	1
4	2,0k	Não	1
5	1,1k	Não	1
6	1,8k	Sim	1
7	4,3k	Sim	2
8	0,9k	Não	1
9	2,7k	Sim	2

Árvore de Decisão
Exemplo

- Nova árvore



Conceitos adicionais sobre Classificação

- Capacidade de Generalização
  - Capacidade de um modelo, além de adaptar bem aos dados de treinamento, deve classificar com precisão exemplos novos

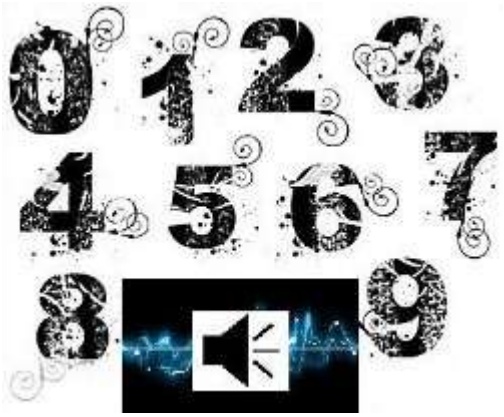
- *Overfitting*
  - Classificador não aprende um modelo, e sim “memoriza” exemplos
  - Exemplo: árvore de decisão onde praticamente cada valor numérico leva a um único nó folha

- *Underfitting*
  - Classificador não aprende o modelo

Sobre o Trabalho

- Trabalhos de classificação
  - Usar a base de dados Spoken Arabic Digits para validar o trabalho
  - http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit

  - Dados valores de frequência de áudio, identificar números arábicos



Exercício

- Gere a árvore de decisão para a seguinte base de dados

ID	Temp. Corporal	Voa	Classe
Homem	Quente	Não	1
Pardal	Quente	Sim	2
Jacaré	Fria	Não	2
Cobra	Fria	Não	2
Macaco	Quente	Não	1
Boi	Quente	Não	1
Sapo	Fria	Não	2
Gato	Quente	Não	1
Gavião	Quente	Sim	2

Valor	$\log_2$
1/9	-3,17
2/9	-2,17
3/9	-1,58
4/9	-1,17
5/9	-0,85
6/9	-0,58
7/9	-0,36
8/9	-0,17
9/9	0,00

Bibliografia

- *Introdução ao Data Mining*. Steinbach, Michael; Kumar, Vipin; Tan, Pang-ning, Rio de Janeiro: Ed. Ciência Moderna, 2009. Capítulo 4.