Trabalho 1 - INF01017 – 2018/01

Exemplo de aplicação do algoritmo de indução de árvores de decisão

Juntamente com o enunciado do trabalho, será disponibilizado no Moodle um pequeno conjunto de dados a ser utilizado como *benchmark,* para validação e verificação de corretude da implementação do algoritmo de indução de uma árvore de decisão (utilizando o Ganho de Informação como critério de seleção de atributos).

No relatório final, os alunos devem demonstrar que a saída da função de treinamento de uma árvore de decisão a partir destes dados gera a árvore de decisão esperada e os Ganhos de Informação corretos em cada divisão de nós (ver abaixo). Neste processo, não deve ser aplicada a validação cruzada, e os dados devem ser utilizados para treinar uma árvore de decisão simples, e não uma Floresta Aleatória.

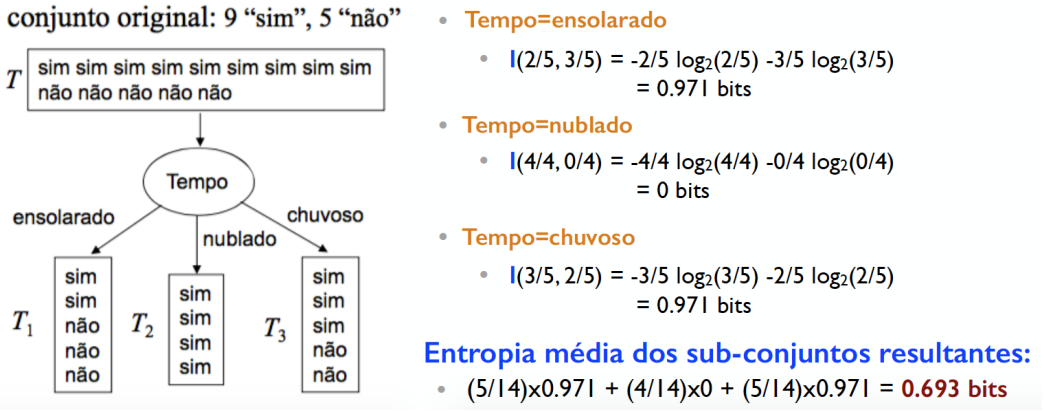
Dados de treinamento (Joga é a classe a ser predita)

Tempo	Temperatura	Umidade	Ventoso	Joga
ensolarado	quente	alta	falso	não
ensolarado	quente	alta	verdadeiro	não
nublado	quente	alta	falso	sim
chuvoso	amena	alta	falso	sim
chuvoso	fria	normal	falso	sim
chuvoso	fria	normal	verdadeiro	não
nublado	fria	normal	verdadeiro	sim
ensolarado	amena	alta	falso	não
ensolarado	fria	normal	falso	sim
chuvoso	amena	normal	falso	sim
ensolarado	amena	normal	verdadeiro	sim
nublado	amena	alta	verdadeiro	sim
nublado	quente	normal	falso	sim
chuvoso	amena	alta	verdadeiro	não

Entropia do conjunto original (9 instâncias “sim”, 5 instâncias “não”): **0.940 bits**

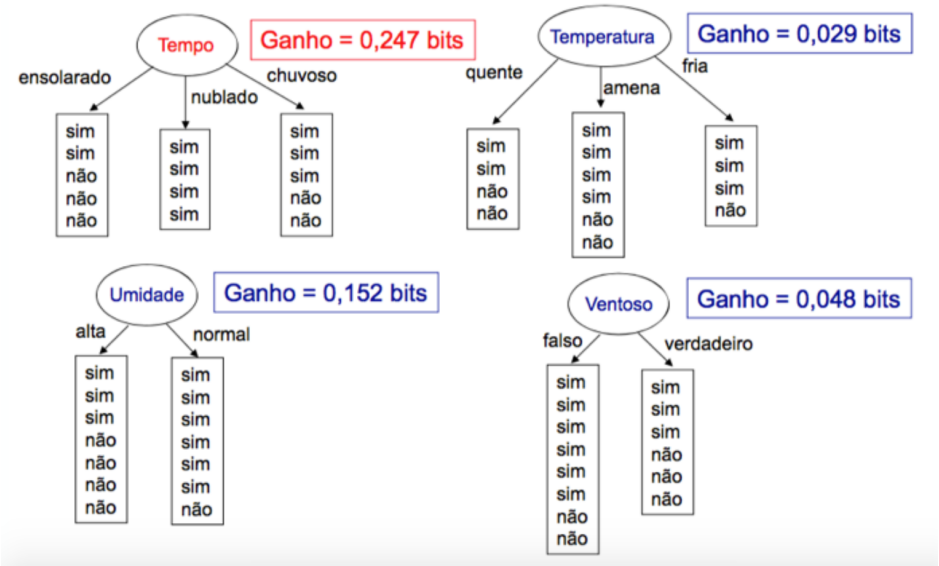
Seleção de atributo para o nó raiz:

Assumindo avaliação do atributo “Tempo”, teríamos



Ganho (Tempo) = 0.940 – 0.693 = 0.247 bits

Repetindo esta análise para os 4 atributos, verifica-se que o maior ganho é proporcionado pelo atributo Tempo, sendo este selecionado para o nó raiz:



Seleção de atributo para a partição gerada por Tempo=Ensolarado:

Entropia da partição após seleção do atributo Teempo (2 instâncias “sim”, 3 instâncias “não”): **0.971 bits**

A árvore de decisão esperada, obtida repetindo-se iterativamente o processo acima até o critério de parada ser satisfeito, é dada abaixo: