Tarefa: Aprendizado simbólico com ID3 (frutas)

Objetivos de aprendizagem

- compreender como utilizar um algoritmo de aprendizado simbólico
- compreender como avaliar o desempenho dos algoritmos de aprendizado na prática

Equipe

Até 2 pessoas

Cenário de contextualização

Um personagem de um jogo anda em um labirinto onde há frutas que lhe dão energia em valores variados. No início do jogo, as frutas com suas características são sorteadas e posicionadas em cada uma das posições no labirinto e, uma vez comidas, a posição fica desocupada. O agente pode comer a fruta desde que estejam na mesma posição ou ignorá-la. O agente ganha o jogo se atingir a fonte de água e perde se ficar sem energia antes de atingi-la ou se comer uma fruta venenosa.

FRUTAS, SUAS CARACTERÍSTICAS E ENERGIA FORNECIDA

As frutas podem fornecer quantias distintas de energia dependendo das suas características as quais o agente desconhece e terá que aprender utilizando ID3 a partir de um *dataset* que contém exemplares de frutas.

Uma fruta é caracterizada por um padrão de 5 cores {c1, c2, c3, c4, c5} e pela energia fornecida E, tal que E ∈ {0, 2, 4}, sendo que 0 corresponde a uma **fruta venenosa**.

a) 

b) 

Figura 1: a) esquema de padrão de fruta de 5 cores onde K=black, W=White, R=red, R=red, B=blue, E=energia fornecida{; b) um exemplar possível de fruta que fornece energia de para andar 4 casas.

OBJETIVO DA TAREFA

Executar o algoritmo ID3 para o dataset fornecido (treinamento e testes), analisar os resultados produzidos e responder o questionário abaixo.

QUESTIONÁRIO

ID3

(NÃO UTILIZE O J48 PARA RESPONDER ESTAS QUESTÕES)

1) Qual conhecimento o ID3 proporcionou que era desconhecido antes de sua execução?
2) Construa um arquivo .arff a partir do dataset fornecido pelo professor. Copie aqui o cabeçalho do arquivo *.arff* utilizado para treinamento no WEKA (definição dos atributos e da classe de saída).
3) Qual o tamanho do arquivo de treinamento (quantas instâncias)?

4) Qual o número de instâncias por classe?
5) Qual o valor de entropia para o dataset datasetFrutasEnergia-training.arff em relação aos valores possíveis para a classe de saída E={0,2,4}? Qual a interpretação que você dá ao valor obtido?
6) Qual foi a árvore de decisão gerada pelo algoritmo? *Copie e cole aqui.*
7) Todos os atributos do datasetFrutasEnergia-training.arff foram utilizados pelo ID3 na geração da árvore de decisão? Caso não, quais ficaram de fora?
8) Para o ramo C1=B e C2=R, explique, por meio do cálculo de entropia, o porquê de o ID3 ter escolhido o atributo C3 como sendo o próximo em vez do C4.
9) Para o ramo C1=B, C2=R e C3=R, explique porque o ID3 não necessitou incluir mais atributos no ramo.
10) Defina o arquivo datasetFrutasEnergia-test.arff (opção *supplied test set*) como sendo de teste para o modelo aprendido anteriormente. Analise o desempenho do modelo para os exemplos contidos em datasetFrutasEnergia-test.arff com base nas medidas abaixo explicando o significado e contextualizando-as para a tarefa em questão:
    a. matriz de confusão: *<intrepretação em linguagem natural>*
    b. para cada classe de saída
        i. TP rate: *<intrepretação em linguagem natural>*
        ii. FP Rate: *<intrepretação em linguagem natural>*
        iii. precision: *<intrepretação em linguagem natural>*
        iv. recall: *<intrepretação em linguagem natural>*
        v. f-measure: *<intrepretação em linguagem natural>*
11) Baseando-se nos resultados acima, qual(is) medida(s) indica(m) a probabilidade de o personagem morrer por engano (comer uma fruta venenosa por engano) ao utilizar o modelo aprendido? Explique.

---- RESPONDER AS QUESTÕES ABAIXO COM O ALGORITMO **J48** ---

12) Abra o arquivo marciano.arff. Treine o algoritmo J48 e visualize a árvore produzida (modelo). Escreva o modelo aprendido em forma de conjunções lógicas. **Importante: utilize os parâmetros confidence fator = 0.7 e minNumObj = 1 (consulte** http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html)

13) Modifique o arquivo marciano.arff substituindo o atributo pernas de simbólico para integer. Escreva o modelo aprendido em forma de conjunções lógica. **Importante: utilize os parâmetros confidence fator = 0.7 e minNumObj = 1.**

14) Qual a diferença entre o modelo da questão 12 e da 13?
15) Qual a entropia para o ramo *pernas <=1* e *pernas > 1*?