ACH2016 - Inteligência Artificial - 1o Sem /2018

Profa. Patrícia R. Oliveira

Especificação do Trabalho Prático

Entrega no Tidia: até a meia-noite do dia 10/06/2018. Devem ser postados o código fonte; o código executável (junto com eventuais instruções de execução do mesmo); o conjunto de dados após discretização e remoção de missing values (ver item 1ª); o relatório de resultados (extensão .pdf ou .doc) e a apresentação (extensão .ppt). Basta uma postagem por grupo.

Observação 1: Este trabalho deve ser realizado por grupos de no mínimo 3 (três) e no máximo 4 (quatro) alunos.

Observação 2: Pode ser utilizada qualquer linguagem de programação (Java, C, C++, etc). Fica vedado o uso de pacotes utilitários que já contenham algoritmos de aprendizagem de máquina implementados (por exemplo, WEKA).

Observação 3: Plágios, mesmo que parciais, serão punidos com nota zero para os copiadores e copiados envolvidos. O mesmo acontecerá caso seja detectado plágio com arquivos disponíveis na internet e outras fontes.

Tarefa 1)

Implemente o algoritmo ID3 para o aprendizado de uma árvore de decisão. Teste o seu programa com o conjunto de treinamento "PlayTennis", descrito no Capítulo 3 do livro do Mitchell.

Nessa etapa, o grupo deve implementar métodos (procedimentos) para:

- a) Construir uma árvore de decisão utilizando o algoritmo ID3.
- b) Apresentar a árvore construída.

Passo 2)

Adapte o programa desenvolvido no Passo1 para a base de dados *Adult census* (http://archive.ics.uci.edu/ml/datasets/Adult), com o objetivo de predizer se a renda anual de um indivíduo ultrapassa o montante de U$50.000,00 baseando-se em dados de censo.

Nessa etapa, o grupo deve implementar métodos (procedimentos) para as seguintes atividades:

• a) Pre-processar os dados.

Cada exemplo no conjunto *Adult* apresenta 14 atributos (6 contínuos e 8 discretos), além do rótulo de classe. É preciso, portanto, discretizar todos os atributos contínuos antes de executar o ID3. Ainda é necessário remover do conjunto todos os exemplos para os quais alguns valores de atributos não tenham sido observados (conhecidos como *missing values*).

O grupo deve incluir no seu relatório as seguintes informações sobre essa atividade:

- i) A descrição de cada um dos atributos do conjunto: significado no domínio, se for contínuo, deve-se relatar o resultado da sua discretização, se for discreto, deve-se apontar os seus possíveis valores.
- ii) O levantamento de quantos exemplos havia no conjunto original e quantos tiveram que ser excluídos por apresentarem *missing values*.

• b) Avaliar a técnica para a classificação do conjunto de dados.

Utilizar o método de amostragem 10-fold-crossvalidation para estimar o erro do modelo, considerando um intervalo de confiança de 95%.

Passo 3) Podar a árvore construída no Passo 2

Nesse estágio, o grupo deve executar o processo de pós-poda da árvore e plotar um gráfico de evolução do desempenho da técnica com relação ao número de nós que árvore possui.

• a) Dividir novamente o conjunto de dados *Adult census*

Criar, aleatoriamente, uma nova partição do conjunto, com três subconjuntos disjuntos, de aproximadamente mesmo tamanho:

- Conjunto de treinamento: para construir o modelo (Árvore de Decisão)
- Conjunto de teste: para medir o desempenho do modelo
- Conjunto de validação: para efetuar a poda da árvore

• b) Executar o algoritmo de pós-poda e criar gráfico de comparação

Executar o algoritmo descrito na Seção 3.7.1.1 (pág. 69) do livro do Mitchell e reproduzir, para o conjunto de dados analisado neste trabalho, o gráfico apresentado na Figura 3.7. Incluam também, neste gráfico, o desempenho do modelo para o conjunto de validação.

• c) Apresentar o ranqueamento das regras resultantes da poda

Apresente as regras mantidas pela árvore de decisão (no formato IF-THEN) após o processo de poda. Estas devem estar ordenadas de forma decrescente com relação à acurácia, considerando a classificação do conjunto de teste.