EA072 –Exercícios de Fixação de Conceitos 3

Curvas *Precision-Recall* e ROC + Árvores de Decisão + Redes Bayesianas + *TensorFlow*

Peso da Lista: 2 || Data de entrega: 30/11/2015

Nota: O relatório pode ser apresentado individualmente, por grupo de 2 alunos OU por grupo de 3 alunos.

1 Curvas *Precision-Recall* e ROC (0,5 pontos)

Com base na leitura do paper de Davis & Goadrich (2006), intitulado "The Relationship Between Precision-Recall and ROC Curves" e fornecido pelo professor, explique os seguintes conceitos:

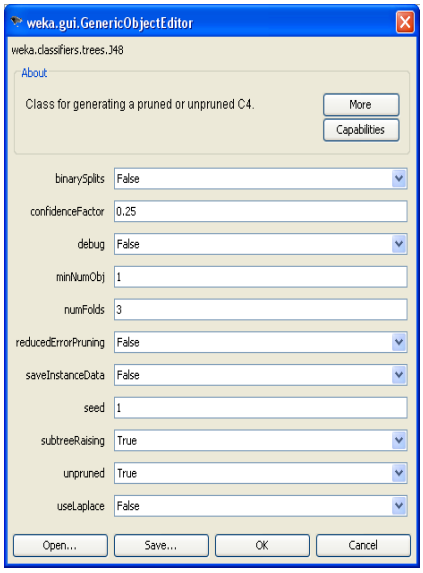
(a) Matriz de confusão;
(b) *Recall*;
(c) *Precision*;
(d) Taxa de verdadeiros positivos;
(e) Taxa de falsos positivos;
(f) Curvas *Precision-Recall* e ROC;
(g) Critério de desempenho AUC-ROC.

Por fim, explique o motivo pelo qual, em um problema de classificação binária, pode não ser suficiente monitorar apenas a taxa de acerto do classificador.

2 Árvores de Decisão (0,5 pontos)

Utilizando o pacote de software WEKA (http://www.cs.waikato.ac.nz/ml/weka/), construa uma árvore de decisão capaz de classificar os dados fornecidos no arquivo ⟨dados_AD_EFC3.arff⟩. Os passos são os seguintes:

(1) Entre com a opção Weka Explorer.
(2) Na aba Preprocess, abra o arquivo fornecido (Open file). O arquivo ⟨dados_AD_EFC3.arff⟩ contém dados reais vinculados à presença ou não de um espécime em certos locais de observação, sob certas condições ambientais.
(3) Reporte quais são os atributos e estatísticas acerca de seus intervalos de excursão. Existem atributos numéricos? Existem atributos categóricos?
(4) Vá para a aba Classify.
(5) Em Choose, escolha trees/J48. Esta é a versão em Java do C4.5. Clique com o botão direito do mouse logo à frente de J48 e você vai abrir uma janela de parâmetros. Procure definir a configuração de parâmetros apresentada no quadro ao lado.
(6) Em Test options, escolha Cross validation Folds 10.
(7) Pressione o botão start e analise os resultados. Indique a porcentagem de classificação correta e apresente a árvore de decisão. Apresente também a matriz de confusão, indicando o que ela informa.
(8) Apresente também a árvore de decisão na forma de regras SE ⟨?⟩ ENTÃO ⟨classe é ?⟩. Isso não envolve copy-and-paste.
(9) Altere ao menos os parâmetros **minNunObj** e **unpruned** (um de cada vez) e analise os resultados em comparação com o obtido na configuração anterior. Qual é o significado de **minNunObj**? Qual é o significado de **unpruned**?



(10) **Tome algum outro conjunto de dados de classificação, de preferência envolvendo atributos categóricos**, e ajuste os parâmetros do J48 visando bom desempenho da ferramenta. Repare que você vai precisar elaborar um cabeçalho nos moldes do arquivo <*.arff> fornecido. Analise os resultados, indicando a porcentagem de classificação correta e apresentando a árvore de decisão obtida.

3 Redes Bayesianas (0,5 pontos)

Utilizando o pacote de software **Bayes Net Toolbox for Matlab**, em sua versão mais simplificada disponibilizada junto ao roteiro deste EFC3, siga o Manual do Usuário e construa uma rede bayesiana para o arquivo de dados fornecido no arquivo zip associado ao EFC3 (dados_BN_EFC3.mat) e, em seguida, faça:

(1) Apresente as características dos dados fornecidos: número de amostras, número de atributos, valores assumidos pelos atributos (no caso de atributos binários, 1 é falso e 2 é verdadeiro);

(2) Apresente a rede bayesiana resultante, incluindo as tabelas de probabilidades associadas a cada nó da rede. O acesso às tabelas de probabilidade deveria ser feito com o uso do mouse, ao clicar sobre o número da variável. Mas como passou a haver alguma incompatibilidade de recursos gráficos entre versões do Matlab (o toolbox recomendado corresponde a uma versão antiga e mais simples), a melhor opção é salvar a rede gerada (Network → Save), carregar o arquivo *.mat salvo (`load .\network\filename.mat`) e digitar [tables.cpt] na linha de comando do Matlab. As tabelas vão aparecer na ordem do número de cada variável da rede.

(3) Procure justificar por que os atributos 1 e 7 não possuem conexões a outros nós da rede.

(4) Qual é a probabilidade de 8 ser verdade?

(5) Qual é a probabilidade de 6 ser verdade dado que 5 é falso e 8 é verdade?

(6) Qual é a probabilidade de 5 ser verdade, dado que 3 é verdade? Observação: Aqui são necessários cálculos a partir das tabelas de probabilidades, enquanto que os itens (d) e (e) saem direto de campos dessas tabelas.

4 *TensorFlow* (0,5 pontos)

O Google é um conglomerado sustentado por soluções de inteligência artificial. No contexto específico de aprendizado de máquina, que pode ser considerada uma das áreas de Inteligência Artificial, o Google está baseando suas soluções e disponibilizando com código aberto para os usuários a sua infraestrutura de aprendizado de máquina, denominada *TensorFlow* Faça uma busca relacionada a esta que é considerada a melhor infraestrutura de aprendizado de máquina atualmente disponível e descreva suas principais características, incluindo (mas não restrito a): (1) Que soluções do Google hoje já usam o *TensorFlow*? (2) Como usuários comuns podem desenvolver soluções de aprendizado de máquina empregando o *TensorFlow* e trabalhando em rede? (3) Qual é o paradigma de fluxo de informação empregado no *TensorFlow*?

Nota: Consultar o enunciado da lista 3 de 2014 para contato com um **software de plotagem e análise de redes complexas**. Na verdade, dada a elevada demanda por visualização e análise de redes, existe uma quantidade elevada de softwares e toolboxes voltados para modelagem, visualização e análise de redes, incluindo soluções específicas para redes complexas.