Universidade Federal de Campina Grande
Projeto HP-FRH-Analytics

Lab 1 - Introdução ao Reconhecimento de Padrões

Objetivo: O objetivo deste Lab é verificar, tanto na teoria quanto na prática, alguns dos conceitos básicos de Reconhecimento de Padrões, a saber: Extração de Características, Classificação e Avaliação. Como um objetivo paralelo, deseja-se estimular o aprendizado da ferramenta Weka.

Entrega: Até as 23h59min do dia 19/04/2012.
Formato da entrega:

- Para todos os labs em PRNN, a entrega deverá ser realizada em um arquivo compactado denominado nome_do_aluno-PRNN-LabXX.tar.gz (ou .zip), onde XX é o **número do lab** com dois dígitos. Deverá ser criada uma estrutura de pastas no formato LABXX\NOME_DO_ALUNO*.
- Todas as análises devem ser apresentadas em **um único** documento no formato PDF, sempre acompanhadas de justificativas textuais e, sempre que pertinente, de artefatos visuais.
- **Todos** os scripts R precisam ser executados por intermédio do comando *Rscript <nome_do_script> (parâmetros, caso hajam)*.
- A base de dados **sempre** deve ser indicada como um parâmetro.

Dados: Iris é uma base de dados clássica utilizada na pesquisa e ensino de reconhecimento de padrões. A mesma possui 4 características de 3 classes/espécies das flores Iris. O arquivo "flores-iris" possui 5 colunas com os seguintes dados: "Tamanho_sépala", "Largura_sépala", "Tamanho_pétala", "Largura_pétala", "Classe". Utilizaremos estes dados para cumprirmos os objetivos do laboratório.

1 - A **Análise de Caracteríticas** é um passo importante no processo de RP, pois busca identificar os melhores dados - dentre os coletados na fase anterior - para realizar a classificação. Nesta questão você deverá utilizar ferramentas do R (sugerimos o comando "splom" da biblioteca "lattice") para avaliar as melhores características dos dados para realizar a separação das 3 classes.

- Você deve entregar o script R utilizado.
- No seu documento de análise deve estar o raciocínio para a escolha realizada (utilize-se do gráfico gerado).

2- O processo de **Classificação** serve para determinar, com base nas características e uma dada função, a qual classe um conjunto de vetores de características pertence.

- No seu documento de análise, devem ser comentadas as suas escolhas para as seguintes questões (utilize os gráficos gerados):

  a - Usando as características escolhidas na questão anterior, defina as equações das retas (no formato y=ax+b) que possam ser utilizadas para melhor separar as classes. *Observação*: Podem ser utilizadas quantas retas você achar necessário.
    - Você deve entregar um script R que gera o(s) gráfico(s) com as características escolhidas e as respectivas retas de separação.

  b - Usando a distribuição de probabilidade, justifique porque as características que você escolheu, são de fato as melhores.
    - Você deve entregar um script R que gera o(s) gráfico(s) com as distribuições de probabilidade que você julgou importante(s) para sua resposta.

3- **Avalie** os resultados de classificação concernentes à questão 2.a por intermédio das seguintes medidas de erro:

a. Taxas de Verdadeiro Positivo, Falso Positivo, Verdadeiro Negativo e Falso Negativo,
b. Precisão, Revocação e F-Measure,
c. Matriz de confusão.
    - Você deve entregar um script R que calcule cada uma das medidas acima.
    - No seu documento de análise, devem ser descritos e analisados os resultados obtidos.

4 - Tomando por base os dados utilizados nos labs de **Fundamentos de Analytics,** responda às seguintes perguntas:

a. **Não sendo possível modificar as base em nenhum aspecto**, cite qual (ou quais bases) poderiam ser utilizadas para problemas de (1) classificação, (2) regressão e (3) agrupamento? Justifique!
b. Caso seja possível alterar a **estrutura dos dados**, cite uma base para cada problema (classificação, regressão e agrupamento), justificando sua resposta. Caso já tenha citado uma base na primeira resposta, cite uma base diferente.
    - Esta resposta deve constar no seu documento de análise.