Aprendizagem de Máquina
Prof. Tiago Buarque A. de Carvalho

Exercícios sobre Comparação de Classificadores

Atenção: a palavra significativa é utilizada para indicar uma diferença grande o suficiente que é atestada através de um teste de hipótese.

1. (25 pontos) Realize 100-*fold cross validation* estratificado na base Skin Segmentation utilizando o classificador 1-NN com distância Euclidiana então realize os procedimentos abaixo.

   (a) Mostre a média, o máximo e o mínimo da medida-F.

   (b) Mostre o histograma da medida-F.

   (c) Calcule o intervalo de confiança da medida-F.

   (d) Qual a medida-F mínima que você espera ao aplicar este classificador, sob as mesmas condições de treinamento, para dados nunca vistos?

   (e) Qual a medida-F esperada para o classificador quando aplicada a dados nunca antes vistos.

   A base Skin Segmentation (`archive.ics.uci.edu/ml/datasets/Skin+Segmentation`) tem três 4 colunas, as três primeiras são atributos e a última é a classe.

2. (25 pontos) Realize um experimento pareado com 100 repetições de Holdout 50/50 utilizando o classificador 1-NN com distância Euclidiana. Utilize duas versões da base Wine `archive.ics.uci.edu/ml/datasets/Wine` para este experimento, a primeira versão é a base original, a segunda versão é a base sem a última coluna. Após calcular 100 taxas de acero para cada uma das versões da base, realize os procedimentos abaixo.

   (a) Calcule a diferença das 100 taxas de acerto.

   (b) Calcule o intervalo de confiança destas diferenças.

   (c) Realize o teste de hipótese sobre estas diferenças para verificar se a diferença da taxa de acerto é significativa entre as duas versões. Mostre sua conclusão para o teste.

   (d) Calcule o intervalo de confiança da taxa de acerto para cada versão da base.

   (e) Realize o teste de hipótese de sobreposição dos intervalos de confiança. Mostre sua conclusão para o teste.

3. (25 pontos) Qual o número máximo de características que podem ser removidas da base Iris `archive.ics.uci.edu/ml/datasets/iris` sem reduzir significativamente a taxa de acerto? Defina a metodologia utilizada para justificar sua resposta.

4. (25 pontos) Utilizando o classificador k-NN na base Wine `archive.ics.uci.edu/ml/datasets/Wine`, teste os valores $k = 1, \ldots, 15$. Para qual valor de k o classificador apresenta uma taxa de acerto significativamente maior? Defina a metodologia utilizada para justificar sua resposta.