

IA006 – Exercícios de Fixação de Conceitos

EFC 2 – 2s2019

Parte 1 – Classificação binária

Problema: identificação do gênero do locutor a partir de trechos de voz

Base de dados: dados_voz_genero.csv

https://www.mldata.io/dataset-details/gender_voice/

Você dispõe de um conjunto de dados contendo 3168 amostras rotuladas. Cada amostra é descrita por 19 atributos acústicos extraídos de trechos gravados de voz, considerando a faixa de frequências de 0 a 280 Hz. A última coluna corresponde ao rótulo associado a cada padrão, sendo igual a '1' para o gênero masculino, e '0' para o gênero feminino.

- Faça uma análise das características dos atributos de entrada considerando os respectivos histogramas e as medidas de correlação entre eles.
- Construa, então, o modelo de regressão logística para realizar a classificação dos padrões. Para isso, reserve uma parte dos dados (e.g., 20%) para teste, usando todas as demais amostras para o treinamento do modelo. Pensem na pertinência e na possibilidade de realizar algum pré-processamento nos dados (e.g., normalização).

Apresente e discuta os seguintes resultados com relação ao conjunto de teste:

- ✓ A curva ROC;
 - ✓ A curva de evolução da F_1 -medida em função do valor do *threshold* de decisão.
- Indique qual seria o valor mais adequado para o *threshold* de decisão e por quê. Empregando, então, esse *threshold*, obtenha a matriz de confusão e a acurácia do classificador para o conjunto de teste. Comente os resultados obtidos.

Parte 2 – Classificação multi-classe

Problema: identificação de atividade humana usando dados de *smartphones*

Base de dados: har_smartphone.zip

<https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones>

Nesta atividade, o conjunto de dados contém atributos nos domínios do tempo e da frequência extraídos de sinais de acelerômetro e giroscópio de um *smartphone*. Os rótulos correspondentes aos dados indicam qual a atividade realizada por um

voluntário humano durante a aquisição dos sinais: 0 – caminhada; 1 – subindo escadas; 2 – descendo escadas; 3 – sentado; 4 – em pé; 5 – deitado.

O conjunto de dados já está separado em uma parte para treinamento e outra para teste. Ao todo, temos 7352 amostras de treinamento e 2947 amostras de teste; cada amostra é descrita por 561 atributos temporais ou espectrais.

Dois métodos de classificação serão explorados nesta aplicação: regressão logística e *k-nearest neighbors*.

- a) Construa uma solução para este problema baseada no modelo de regressão logística. Descreva a abordagem escolhida para resolvê-lo (softmax, classificadores binários combinados em um esquema um-contra-um ou um-contra-todos). Obtenha, então, a matriz de confusão para o classificador considerando os dados do conjunto de teste. Além disso, adote uma métrica global para a avaliação do desempenho (médio) deste classificador. (Consulte a referência: M. SOKOLOVA & G. LAPALME, A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009). Discuta os resultados obtidos.
- b) Considere, agora, a técnica *k-nearest neighbors* (*k*NN). Varie o parâmetro *k* e analise as matrizes de confusão obtidas junto aos dados de teste e o desempenho médio (computado com a mesma métrica adotada no item (a)). Comente os resultados obtidos, inclusive estabelecendo uma comparação com o desempenho da regressão logística.

Considerações gerais:

- A implementação dos classificadores explorados neste EFC – regressão logística e *k*NN – deve ser autoral e original, i.e., não deve se basear em pacotes prontos (*scikit-learn*, *toolboxes* do Matlab, etc.).
- Por outro lado, as demais funcionalidades (cálculo de métricas de desempenho, separação de conjunto de dados, pré-processamento, etc.) podem explorar tais recursos.
- No relatório, não é necessário descrever a teoria sobre os modelos de regressão logística e *k-nearest neighbors*. Não obstante, todas as escolhas feitas com respeito às características dos modelos, dos dados e dos experimentos devem ser apresentadas e justificadas, de modo a possibilitar a reprodução da metodologia.

Instruções para a entrega do trabalho

Cada aluno deve entregar por e-mail um **único documento**, em formato PDF, contendo a descrição do procedimento experimental e toda a análise dos resultados obtidos para os exercícios propostos.

- ✓ **Título do e-mail:** Entrega_EFC2_IA006
- ✓ **Título do documento PDF:** meu_RA.pdf (e.g., 034021.pdf)
- ✓ Enviar para lboccato@dca.fee.unicamp.br, attux@dca.fee.unicamp.br