Aula 06

Implementação da **validação cruzada** para avaliação de modelos de classificação

O objetivo deste exercício é implementar o processo de ***k-fold cross validation*** (**validação cruzada**), técnica predominantemente utilizada em Aprendizado de Máquina para avaliar o poder de generalização de modelos.

Esta técnica será aplicada para avaliação de modelos de k-nearest neighbors (KNN) para classificação com o dataset Pima Indian Diabetes (disponível para download no Moodle), o qual tem como objetivo tentar predizer se um paciente tem diabetes baseado em um pequeno conjunto de atributos demográficos e clínicos. Este conjunto de dados é composto por 768 instâncias, sendo 500 da classe negativa (0, não possui diabetes) e 268 da classe positiva (1, possui diabetes). Além da classe (coluna “Outcome”), existem 8 atributos preditivos numéricos a partir do qual a predição deverá ser realizada. Observe que os atributos possuem escalas diferentes, sendo necessário normalizar os valores.

O código deve primeiramente realizar a divisão dos dados originais em *k* folds de forma estratificada (mantendo a proporção de exemplos por classe em cada fold), e então iterativamente treinar um modelo utilizando k-1 folds e testá-lo no fold restante, variando o fold de teste a cada repetição deste processo. A cada teste realizado, deverão ser calculadas e armazenadas as medidas de acurácia e F1-measure (F-measure com β=1). Ao final, o código deverá reportar a média e desvio padrão para estas medidas de desempenho dentre todos os folds testados.

Os alunos deverão se organizar em grupos de até três pessoas - estes grupos, preferencialmente, devem ser os mesmos para implementação dos dois trabalhos práticos que compõem a avaliação da disciplina.

Requisitos básicos da implementação:

- Normalização dos atributos preditivos numéricos
- Estratégia de k-fold cross-validation estratificada para divisão de dados em treinamento e teste, sugere-se utilizar k=5 ou k=10.
- Treinamento e teste de modelos com o algoritmo KNN, utilizando a distância Euclidiana (sugere-se implementar o algoritmo utilizando os 5-vizinhos mais próximos)

- Avaliação do desempenho do modelo para cada fold de teste utilizando as métricas de acurácia e F1-measure, reportando ao final a média e desvio padrão obtidos no processo de k-fold cross-validation

Itens extras, recomendados:

- Realizar o repeated cross-validation, isto é, repetir todo o processo de k-fold cross validation “r” vezes, aumentando o número de modelos avaliados para estimativa mais robusta de desempenho do algoritmo. Utilizar, por exemplo, r=10 e k=10.
- Executar o treinamento e teste do KNN com o repeated cross-validation para diferentes valores de k-vizinhos mais próximos, como 3, 5 e 7, comparando os desempenhos.

Após a conclusão do exercício, cada grupo deverá entregar via Moodle:

- O código da implementação do cross-validation e KNN
- Um breve relatório (em **.pdf** ou **.html**) com o desempenho médio do algoritmo KNN, estimado com a estratégia de k-fold cross-validation estratificada. Deverão ser apresentadas as medidas de média e desvio padrão, com gráficos que demonstram a distribuição destes valores. Se itens extras forem implementados, os resultados destes também devem ser reportados.

O prazo final de entrega deste exercício é dia **29 de março às 23:55h**.