Lista de Exercícios 12

(prazo final para entrega: 16/06/2019 - DOMINGO

1) Avalie o desempenho dos algoritmos **DecisionTree, RandomForest** e **Gradient Boosting** para o dataset "Wisconsin Diagnostic Breast Cancer (WDBC)". Separe os dados em treino (60%), validação (20%) e teste (20%).
    a) Use um valor constante para o parâmetro **random_state** e teste os resultados com as seguintes combinações de hiper-parâmetros para RandomForest e Gradient Boosting usando Grid-Search:
        i) **learning_rate**: 0.1, 0.05, 0.01 (somente para o Gradient Boosting)
        ii) **n_estimators**: 50, 100, 200
        iii) **max_depth**: 3, 5, 7

    b) Mostre a **importância das features** de acordo com o **melhor modelo de classificação** e o **melhor modelo de regressão** encontrados dentre os 3 usados nesta lista de exercícios (DecisionTree, RandomForest e Gradient Boosting).

2) Faça a clusterização do dataset deste link usando o algoritmo K-Means++.
    a) O dataset possui os seguintes dados de motoristas: a distância média dirigida por dia e a média percentual do tempo que um motorista estava 5mph acima do limite de velocidade.
    b) Portanto, agrupe os motoristas pela similaridade das features acima.
    c) Use o método do cotovelo (Elbow Method) para identificar o melhor valor de k para o K-Means++.
    d) Mostre o resultado graficamente.

3) Use Clusterização Hierárquica no mesmo dataset da questão 2 usando "complete" como critério de ligação (linkage). Mostre o dendograma.
    a) Use o parâmetro n_clusters da função scipy.cluster.hierarchy.cut_tree para obter o número de clusters igual ao melhor resultado obtido com o K-Means (Questão 2). Exemplo:

    …
    distance_matrix = scipy.spatial.distance.pdist(X, metric='euclidean')
    cluster_model = scipy.cluster.hierarchy.complete(distance_matrix)
    dendogram = scipy.cluster.hierarchy.dendrogram(cluster_model)
    sensor_cluster_list = scipy.cluster.hierarchy.cut_tree(cluster_model, n_clusters=8)

    b) Mostre o resultado graficamente.

4) Use o DBScan para clusterizar o mesmo dataset da questão 2 e mostre o resultado graficamente.