Exercícios de Aprendizado de Máquina

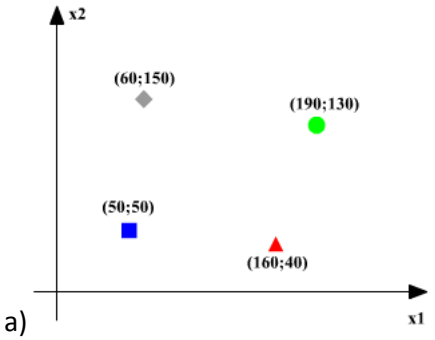
Parte III – Métodos de Classificação Baseados em Distância

1) Para a figura abaixo, obtenha o diagrama de Voronoi das amostras quadrado, triângulo e losango para as métricas de (pode tirar foto da resposta e enviar):

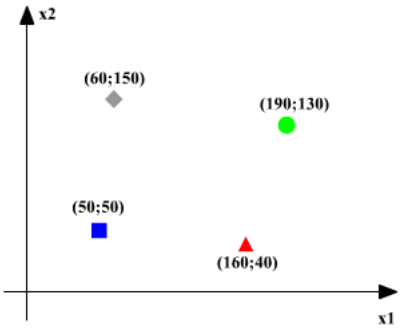
a) Distância Euclidiana;

b) Similaridade Cosseno;

c) Obtenha a classe (quadrado, triângulo ou losango) da amostra círculo para um classificador NN, se for usada a métrica de Distância Euclidiana e a Similaridade Cosseno.



a)



b)

2) Realize a classificação da base de dados Car Evaluation (disponível em http://archive.ics.uci.edu/ml/) usando o kNN. Realize 3-fold cross validation e, para cada rodada, use dois folds para a parte de calibração e um fold para teste. Na parte de calibração o treinamento deve ser realizado usando um fold e a validação do valor de k deve ser realizado usando o outro fold. A calibração deve ser realizada de forma a maximizar a acurácia. Expresse os resultados em forma de acurácia média, macroprecision médio, macrorecall médio, tabela de contingência média (dado em porcentagem).

3) Usando as técnicas de seleção de características SFS e SBE sobre a base de dados Wine (disponível em http://archive.ics.uci.edu/ml/), faça:

a) Divida a base de dados em três partes de forma estratificada. Selecione 5 atributos usando uma parte da base de dados para treinamento e valide os atributos sobre uma outra parte usando a métrica acurácia. Após determinar os 5 atributos, obtenha a acurácia sobre a terceira parte, usando as duas partes como treinamento. Use o classificador Vizinho mais Próximo nesta tarefa. Quais foram os atributos selecionados e a acurácia sobre a terceira parte?

b) Realize o mesmo procedimento, mas agora selecionando 10 atributos;

c) Realize o mesmo procedimento de a) e b), mas agora selecionando os atributos usando duas partes para treinamento e validando o resultado sobre as mesmas duas partes. Após determinar os atributos, obtenha a acurácia sobre a terceira parte. A acurácia sobre a terceira parte foi melhor, igual ou pior do que as obtidas nas letras a) e b). Por quê?

4) A base de dados Nebulosa (disponibilizada em anexo) está contaminada com ruídos, redundâncias, dados incompletos, inconsistências e *outliers*. Para esta base:

a) Obtenha os resultados da classificação (métrica acurácia) usando a técnica do vizinho mais próximo (NN) e Rocchio. Utilize a distância Euclidiana e a base de dados crua, sem pré-processamento. Use o conjunto das 143 amostras para treino e o de 28 amostras para teste. Os dados incompletos podem ser substituídos por uma constante igual a zero.

b) Realize um pré-processamento sobre os dados de forma a reduzir os ruídos, as redundâncias, inconsistências, *outliers* e a interferência dos dados incompletos. Obtenha os resultados da classificação usando a técnica do vizinho mais próximo (NN) e Rocchio usando a distância Euclidiana e a mesma distribuição dos dados.