Primeira lista de Exercícios de Aprendizado de Máquina

Data de Entrega: 20/06/2017

Não utilize funções prontas de algoritmos aprendidos em sala de aula. Implemente as suas e apresente-as na lista. Faça um relatório explicando como foi resolvido o exercício e envie junto com o código fonte.

Parte I – Pré-Processamento de Dados

1) Dada a base de dados Iris (disponibilizada em http://archive.ics.uci.edu/ml/), obtenha:
a) A média e variância de cada um dos atributos;
b) A média e variância de cada um dos atributos para cada uma das classes;
c) O histograma com 16 bins de cada um dos atributos para cada uma das classes (gere um único gráfico de histograma de 16 bins, ou seja, dividido em 16 segmentos, para cada atributo diferenciando as classes, para mostrar como estão distribuídos os valores para diferentes classes);
d) Gere um gráfico 2D com os dois componentes principais (uso de PCA) das amostras, identificando cada classe. Pode usar a função *eig* do Matlab.

2) Dada a base de dados CNAE-9_reduzido (em anexo):
a) gere um gráfico 2D com os dois componentes principais (uso de PCA) das amostras, identificando cada classe (a base possui 5 classes. O rótulo das amostras está na primeira coluna. Essa coluna não deve ser usada no PCA).
b) gere um gráfico 2D com os dois componentes principais (uso de PCA) das amostras, identificando cada classe (a base possui 5 classes). Para este gráfico realize o branqueamento dos dados (isto é, após a aplicação do PCA garantir que a matriz de correlação dos dados seja uma matriz identidade). O que tem de diferente entre os gráficos de a) e b)?
c) Após a visualização dos gráficos, é possível identificar ao menos uma classe que possa ser facilmente separada das outras classes? Seria possível separar essa classe (sem grandes perdas) se fosse usada somente uma componente principal? Gere o gráfico de uma componente para explicar.

3) 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 de 143 amostras para treino e o de 28 amostras para teste. Proponha um tratamento para os dados incompletos.
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 divisão dos dados.

c) Compare os resultados obtidos em a) e b). Qual deles retornou o melhor resultado?

Parte II – Regressão Linear

4) Prove que, para uma quantidade de *N* amostras com entradas *x* e saídas *t*, as equações

$$w_0 = \bar{t} - w_1 \bar{x} \qquad\qquad w_1 = \frac{\overline{xt} - \bar{x}\bar{t}}{\overline{x^2} - \left(\bar{x}\right)^2}$$

sendo:

$$\bar{a} = \frac{1}{N}\sum_{k=1}^{N} a_k$$

minimizam a função de perda quadrática média L:

$$L = \frac{1}{N}\sum_{k=1}^{N}\left(t_k - \left(w_0 + w_1 x_k\right)\right)^2$$

5) Para a base de dados Runner (disponibilizada em anexo) obtenha:

a) A equação linear que se ajusta aos dados e a RMSE;

b) Predizer o resultado para o ano de 2016;

c) Utilize o teste de hipótese de Kendall para verificar se existe dependência entre os atributos. Realize o teste para 5% e 1% de nível de significância;

d) Calcule a correlação entre os dados. Se o módulo da correlação for acima de 0,85 realize o teste de hipótese de Pearson para 5% e 1% de nível de significância (teste bilateral).

6) Para a base de dados Polinômio (disponibilizada em anexo), faça:

a) Divida aleatoriamente a base de dados em duas partes: treino, como 70% das amostras, e teste, com 30%. Use a parte de treino para estimar um **modelo linear** que melhor se ajusta aos dados, sendo a entrada do modelo a primeira coluna e a saída a segunda coluna. Informe os parâmetros do modelo encontrado e obtenha os valores de RMSE e MAPE sobre o conjunto de treino e teste. Mostre um gráfico do modelo estimado e dos pontos da base de dados. O modelo conseguiu se ajustar bem aos dados? Por quê?

b) Estimar o **modelo polinomial** que melhor se ajusta aos dados usando os dados de treinamento. Informe os parâmetros do modelo encontrado. Use os fatores de determinação de complexidade do modelo para auxiliar a encontrar o modelo. Obtenha os valores RMSE e MAPE do modelo obtido sobre os dados de treino e teste. Mostre um gráfico com o novo modelo. O modelo conseguiu se ajustar melhor aos dados? Por quê?

c) Utilize o método de Ransac sobre os dados de treinamento para remover os *outliers* e obter o modelo polinomial. Informe os parâmetros do modelo encontrado. Obtenha o RMSE e MAPE do modelo obtido sobre os dados de treino e teste. Mostre um gráfico com o novo modelo. O Ransac conseguiu ajustar melhor o modelo aos dados?

7) Explique o dilema entre bia e variância e o seu relacionamento com *underfitting* e *overfitting*.

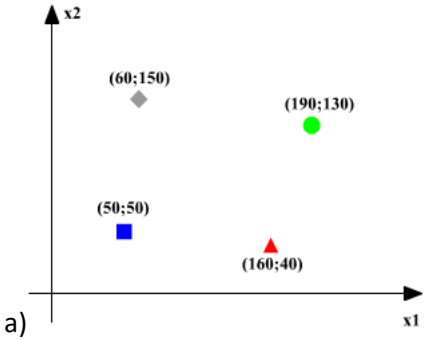
Parte III – Métodos de Classificação Baseados em Distância

8) Para a figura abaixo, obtenha o diagrama de Voronoi das amostras quadrado, triângulo e losango para as métricas de:

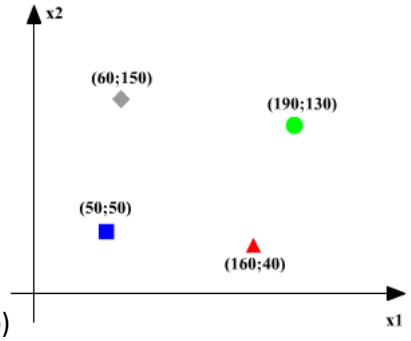
a) Distância Euclidiana;

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)

9) 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 e tabela de contingência médio (dado em porcentagem).

10) Usando as técnicas de seleção de características SFS e SBS 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 como 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 anteriores como treinamento. Use o classificador Vizinho mais Próximo nesta tarefa. Quais foram os atributos selecionados?

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 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?