Primeira lista de Exercícios de Aprendizado de Máquina

Lista Individual

Data de Entrega: 22/10/2023

Não utilize funções prontas de algoritmos aprendidos em sala de aula (a não ser quando informado o contrário). Implemente as suas e apresente-as na lista. Procure se informar sobre as bases de dados que são usadas na lista. Faça um relatório explicando como foi resolvido o exercício e envie junto com o código fonte. **Não serão toleradas cópias de trabalhos ou de questões (plágio).** Envie pelo classroom da disciplina.

Parte I – Pré-Processamento de Dados

1) Dada a base de dados Haberman's Survival (disponibilizada em http://archive.ics.uci.edu/dataset/43/haberman+s+survival), 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) A matriz de coeficientes de correlação;

d) O histograma com 8 bins de cada um dos atributos para cada uma das classes (gere gráficos dos histogramas com cores diferentes para cada classe);

e) Gere um gráfico 3D das amostras, identificando cada classe. Analise o gráfico e informe se a tarefa de classificação é fácil ou difícil e justifique a sua resposta.

2) Para a base de dados Car Evaluation (disponibilizada em http://archive.ics.uci.edu/dataset/19/car+evaluation), calcule a informação mútua entre os atributos de entrada (as 6 primeiras colunas) e o atributo de saída (a última coluna). Informe os resultados e comente a sua solução.

3) 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). Pode usar a função *eig* do Matlab ou do Python.

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 covariância dos dados seja uma matriz identidade). O que tem de diferente entre os gráficos de a) e b)?

c) gere um gráfico 2D usando o t-SNE (pode usar o código disponível em https://lvdmaaten.github.io/tsne/ com os parâmetros *default*), identificando cada classe (a base possui 5 classes). Lembre-se de não usar a coluna de rótulos para obter a redução de dimensão.

d) Utilize as primeiras 480 amostras para treinar o classificador vizinho mais próximo (NN) (utilize a distância Euclidiana) e as demais 120 para teste. Calcule a métrica acurácia e informe o valor obtido. Pelo resultado obtido, qual dos gráficos (t-SNE ou PCA) você acha que melhor representou a "realidade" da distribuição dos dados? Por quê?

4) Dada a base de dados Breast Cancer Wisconsin (Diagnostic) (baixar em http://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic):

a) Obtenha a acurácia de classificação quando usando o classificador vizinho mais próximo (NN) (utilize a distância Euclidiana). Use os dados do arquivo wdbc.data, sendo as primeiras 300 amostras para treino e as demais para teste. Antes, repare os atributos da base de dados e a posição dos rótulos. Quais atributos você pode eliminar da base de dados antes do experimento? Por quê?

b) Aplique o PCA sobre os dados de treino (use a matriz de coeficientes de correlação) e selecione o número de componentes até eles corresponderem a 90% da informação de variância dos dados (conforme mostrado nos slides) (lembre-se que no PCA não entra o rótulo das amostras). Quantos componentes foram selecionados? Calcule a nova acurácia do NN usando as componentes selecionadas. O resultado alterou de forma significativa em relação ao obtido em a)? Qual foi a vantagem observada usando PCA?

c) Aplique o Discriminante Linear de Fisher sobre os dados de treino (lembre-se que ele é supervisionado). Obtenha os novos dados após a aplicação de Fisher sobre os dados de treino e obtenha a acurácia do NN sobre o conjunto de teste. Quais as vantagens desta abordagem sobre o PCA?

Parte II – Regressão Linear

5) Para a base regressao1.csv, sabendo que x é a entrada e t é a saída, obtenha um **modelo linear de regressão** usando as primeiras 100 amostras para treino e o resto para teste. Obtenha novamente um modelo de regressão linear, mas usando agora RANSAC para eliminar os outliers dos dados de treino. Calcule os valores de RMSE e MAPE de ambos os modelos sobre os dados de teste e compare os resultados.

6) Para a base de dados Auto MPG (disponibilizada em https://archive.ics.uci.edu/ml/datasets/Auto+MPG) faça:

a) Baixe o arquivo auto-mpg.data, remova as linhas que tem interrogação (?) e remova a última coluna (por quê?). Com as 150 primeiras linhas obtenha um modelo de regressão linear multivariada para predizer o valor da primeira variável (mpg). Avalie o resultado sobre o restante da base de dados, usando a métrica RMSE.

b) Verifique quais são os atributos que estão relacionados com a saída: A partir dos coeficientes obtidos, aplique o teste F de Snedecor sobre cada variável individualmente (conforme nos slides). Faça este teste usando os próprios dados de treinamento usados para obter o modelo. Indique quais foram os atributos que podem ser desconsiderados. Obtenha sobre o restante da base de dados a métrica RMSE com o modelo sem considerar esses atributos (não precisa estimar um novo modelo, só

considere os valores dos coeficientes deles iguais a zero). Compare os resultados obtidos em a) e em b). Considere que os resíduos do modelo possui distribuição aproximadamente normal e que $F_{1,142} = 3{,}908$.

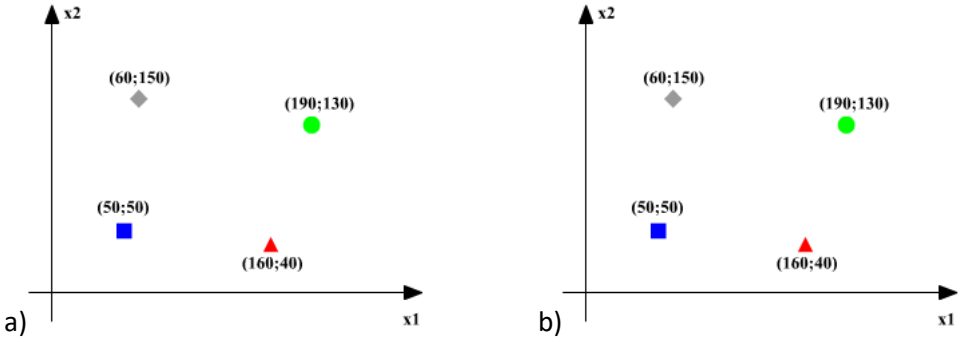
Parte III – Métodos de Classificação Baseados em Distância

7) 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.



8) Realize a classificação da base de dados HTRU2 (disponível em https://archive.ics.uci.edu/ml/datasets/HTRU2) usando o esquema de validação hold-out. Para cada execução, use 70% das amostras como treinamento (selecionadas aleatoriamente) e o restante para teste. Execute 5 vezes o treinamento e teste e retorne a acurácia, recall e precisão média para cada algoritmo. Faça a classificação usando:

   a) Rocchio com métrica de distância Mahalanobis;

   b) kNN com métrica de distância Euclidiana. Para selecionar o melhor valor de k divida a base de treinamento em duas partes iguais: uma para treinar e a outra para validar e encontrar o melhor valor de k;

   c) Compare os resultados, tempos de execução e número de protótipos usados por cada algoritmo. Considerando a distribuição das classes, você considera o valor da acurácia média relevante? Por quê?

9) Usando as técnicas de seleção de características SFS e SBS sobre a base de dados Wine (disponível em https://archive.ics.uci.edu/ml/datasets/Wine) faça:

a) Divida a base de dados em três partes de forma estratificada. Selecione 4 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 4 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 8 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)? Esse era o resultado esperado? Por quê?

Questões Teóricas

1) 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*:

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

2) Explique o dilema entre bias e variância e o seu relacionamento com *underfitting* e *overfitting*.

3) Em uma empresa é adotado um método de Aprendizado de Máquina para detectar defeito de fabricação de peças mecânicas, sendo que raramente acontece este tipo de problema na fábrica. Um funcionário anuncia empolgado que o sistema alcançou uma acurácia de 99%, porém seu gerente não achou o resultado tão relevante. Responda:
a) Por que o gerente não ficou empolgado com o resultado achado?
b) O que o funcionário poderia fazer para confirmar se o método empregado é adequado para o problema?