Primeira lista de Exercícios de Aprendizado de Máquina

Data de Entrega: 14/10/2019

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. Faça um relatório explicando como foi resolvido o exercício e envie junto o código fonte. Envie a lista para patrick.ciarelli@ufes.br

Parte I – Pré-Processamento de Dados

1) A base de dados Nebulosa (disponibilizada em anexo) está contaminada com ruídos, redundâncias, dados incompletos (substituídos pelo valor -100), 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. Remova as amostras com 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? Por quê?

2) Dada a base de dados Breast Cancer Wisconsin (Diagnostic) (baixar em https://archive.ics.uci.edu/ml/datasets/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 e selecione o número de componentes até eles corresponderem a 90% da informação de variância dos dados (conforme mostrado nos slides). 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) Outra técnica para redução de dimensionalidade, mas de forma supervisionada, é o Discriminante Linear de Fisher (para duas classes) e a sua versão multiclasse. Quando aplicado este método o tamanho do vetor de características é reduzido para C-1, onde C é o número de classes do problema. Seguindo os slides de http://www.csd.uwo.ca/~olga/Courses/CS434a_541a/Lecture8.pdf (há um exemplo no meio), 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

3) 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 2020;
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. Informe os resultados;
d) Calcule o coeficiente de correlação entre os dados e realize o teste de hipótese de Pearson para 5% e 1% de nível de significância (teste bilateral). Informe os resultados.

4) 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). 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 $F_{1,142} = 3,908$.

5) Para a base de dados Communities and Crime (disponível em https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime) faça:
a) Faça as análises e alterações necessárias na base de dados para predizer a variável ViolentCrimesPerPop usando regressão linear. Observe que essa base de dados possui valores faltantes. Explique as considerações e mudanças propostas.
b) Divida aleatoriamente a base de dados em duas partes: treino, com 70% das amostras, e teste, com 30%. Use a parte de treino para estimar um **modelo linear** que melhor se ajuste aos dados. Obtenha os valores de RMSE e MAPE sobre o conjunto de treino e teste.
c) Aplique PCA sobre os dados de treino para reduzir os dados para 5 atributos. Realize análise gráfica sobre as variáveis e proponha alterações para melhorar o modelo de regressão linear (que poderá ser um modelo polinomial). Com esses atributos, obtenha os valores de RMSE e MAPE sobre o conjunto de treino e teste. Compare com os resultados da letra b).

Parte III – Métodos de Classificação Baseados em Distância

6) 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 6000 amostras de treino selecionadas aleatoriamente e o restante para teste (normalmente o conjunto de treinamento é maior do que de teste, mas para reduzir o custo computacional ele foi reduzido aqui). 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) Use o Edit-kNN para classificar com métrica de distância Euclidiana. Pode ser o mesmo valor de k da letra b).
d) 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?

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

a) Divida a base de dados em três partes de forma estratificada. Selecione 6 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 6 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 de a), 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 a obtida na letra a)? Esse era o resultado esperado? Esse procedimento parece correto? Por quê?

Questões Teóricas

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

2) Comente sobre a veracidade das afirmações:

a) “Quanto mais variáveis de entrada forem usadas em um modelo de aprendizado de máquina, melhor será a qualidade do modelo”.
b) “Independente da qualidade, quanto mais amostras forem obtidas para uma base de dados, maior a tendência de se obter modelos mais adequados”.

c) “Às vezes com simples manipulações na base de dados (limpeza, conversão de valores, etc.) pode-se conseguir melhoras significativas nos resultados, sem fazer nenhuma alteração na técnica de aprendizado de máquina usada”.

3) Em certas tarefas de aprendizado supervisionado as amostras de diferentes classes aparecem com sobreposição, de tal forma que não é possível obter uma superfície que separe de forma adequada as amostras das diferentes classes. O que se poderia fazer nestas situações para tentar melhorar a qualidade de classificação?

4) Quais devem ser as características que uma base de dados deve ter para:
a) Uma regressão linear se ajustar bem aos dados?
b) O classificador Rocchio conseguir um bom resultado de classificação?
c) O classificador Vizinho mais Próximo conseguir um bom resultado de classificação?

5) 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?