Mineração Estatística de Dados:

Projeto 1: Pré-processamento de dados e classificação

Entrega: O projeto deve ser entregue na forma de um Notebook do Jupyter no formato html.

Objetivo: Predizer o meio de transporte de usuários a partir de dados de celular.

1 – Considere a base de dados (arquivo dataset_5secondWindow%5B1%5D.csv):

https://www.kaggle.com/fschwartzer/tmd-dataset-5-seconds-sliding-window

2 – Realize o pré-processamento dos dados: Verifque se há NaN ou outros erros no dados. Selecione apenas os atributos relevantes e numéricos.

3 - No caso do classificador Knn, verifique o efeito do parâmetro k na classificação.

4 – Compare os classificadores:

- knn (melhor k observado no item anterior)
- Naive Bayes
- Decisão Bayesiana

5 – Verifique o efeito da normalização (atributos em [0,1]) e padronização (atributos com média 0 e variância 1) dos dados. Compare os casos sem processamento, com padronização e com normalização para os classificadores:

- knn (melhor k observado no item anterior)
- Naive Bayes
- Decisão Bayesiana

6 – Mostre a matriz de correlação entre os atributos. Considere os atributos com menor correlação (por exemplo, menor do que 0.5). Realize a classificação novamente apenas com esses atributos. A acurácia melhora?

7 – Verifique qual dos classificadores é mais robusto com relação à presença de ruídos. Para isso:

- Aplique a normalização dos dados para que os atributos apresentem média igual a zero e variância igual a 1.
- Inclua em X% dos atributos, um valor normalmente distribuído com média zero e variância 1. Considere toda a matriz dos dados, sorteando uma posição da matriz de forma aleatória.

- Varie o nível de ruído, de 0 a 50% (em passos de 5%) e avalie como muda a classificação. Construa um gráfico de X% de ruído versus porcentagem de classificação correta. Coloque a média e o desvio padrão calculados a partir de ao menos 10 simulações. Considere 70% dos dados no conjunto de treinamento.
- Discuta os resultados.

8 – No caso do classificador Naive Bayes, é possível considerar diferentes funções para estimar as probabilidades. Compare os casos: (i) Gaussian Naive Bayes, (ii) multinomial Naive Bayes e (iii) Bernoulli Naive Bayes. Considere os casos com e sem padronização.

9 –No caso do Knn, compare as classificações usando diferentes métricas. Varie k e mostre as curvas (em um mesmo plot) para as distâncias euclidiana, Manhattan, Chebyshev e Minkowsky (p=0.5, p=1.5, p = 3).

10 – Faça um gráfico da fração de elementos no conjunto de treinamento (10% até 90% em passos de 10%) versus acurácia para os classificadores:

- knn (melhor k observado anteriormente)
- Naive Bayes
- Decisão Bayesiana

Considere os casos com e sem padronização.