Trabalho 4

Universidade Federal do Ceará

Aprendizado de Máquina

Prof. Victor Farias

Entrega: 21/10/2020

Entrega Código + Relatório via Moodle

**Conjunto de dados**

1. **Usar conjunto de dados trab4.data**
2. **Primeiras 4 colunas são as *features* das instâncias**
3. **Última coluna é a variável alvo**

**k-means**

1. **Implemente o k-means usando a distância euclidiana.**
2. **Execute o k-means para k ={2,3,4,5}**
   1. **Plote a distância média de cada ponto para o seu centroide em um gráfico linha em função de k (média sobre 20 rodadas)**
   2. **Discuta qual seria o k ideal a ser usado**

**PCA**

1. **Implemente o PCA**
   1. **Você deve implementar a função de calcular a matriz de covariância**
   2. **A função de achar os autovetores e os autovalores pode ser usado pronto do numpy  
      https://numpy.org/doc/stable/reference/generated/numpy.linalg.eig.html**
2. **Reduza o conjunto de dados original em um conjunto com apenas duas variáveis (2 componentes principais de maior autovalor)**
   1. **Reporte quanto de variância foi preservado**
   2. **Plote cada ponto do conjunto transformado em um gráfico de dispersão 2d atribuindo uma cor para cada uma das classes (3 classes no total).**

**Árvores de decisão**

1. **Implemente a árvore de decisão usando o coeficiente de Gini como mostrado em sala**
2. **Reporte o erro de classificação para o k-fold com k=5**
   1. **Pode usar o k-fold que foi implementado em atividades passadas ou pode usar pronto do scikit-learn**
   2. **Erro de classificação pode usar pronto do scikit-learn também**

