

Aula 1 de Laboratório

Exercícios

OBS: Para cada exercício, implemente o código, chame o professor e explique a implementação.

- 1) Dada a base de dados Iris (iris.data), e sabendo que as 4 primeiras colunas são os atributos de entrada e a quinta coluna é a coluna de classe, obtenha:
 - a) A média e variância de cada um dos atributos;
 - b) A média e variância do primeiro e terceiro atributos para cada uma das classes;
 - c) A visualização dos dados usando o comando plotmatrix. Você acha que é possível separar pelo menos amostras de uma classe das demais? Por quê?
- 2) Dada a base de dados Haberman (haberman.data), e sabendo que as 3 primeiras colunas são os atributos de entrada e a quarta coluna é a coluna de classe, faça:
 - a) Leia a descrição da base de dados em <https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival>;
 - b) Visualize os dados usando o comando plotmatrix. Observe a distribuição dos dados;
 - c) Calcule os momentos de cada distribuição de atributos usando o comando moment. É possível diferenciar as diferentes distribuições só pelo valor dos momentos?
 - d) Visualize os dados usando o comando boxplot. Observe os resultados;
 - e) Compare os resultados de média e mediana para o terceiro atributo. Compare também os valores de desvio padrão e intervalo max-min para o terceiro atributo;
 - f) Faça um algoritmo para tratar a presença de “outliers”.
- 3) O arquivo AirQualityUCI.txt contém 4 atributos da base de dados original Air Quality (<https://archive.ics.uci.edu/ml/datasets/Air+quality>). Essa base de dados possui dados incompletos (*missing values*) que foram completados com o valor -200.
 - a) Visualize o gráfico de cada atributo usando o comando plot. Visualize a presença dos dados faltantes e do comportamento das variáveis;
 - b) Proponha e implemente um algoritmo para tratar os dados faltantes.
- 4) Dada a base de dados Perfume (perfume.data), e sabendo que a primeira coluna é o atributo de entrada e a segunda coluna é a coluna de classe, faça:
 - a) Visualize os dados usando o comando plotmatrix. Observe a disposição dos dados;
 - b) Proponha um método para identificar e remover dados redundantes.