Roteiro Aula 4 Lab

1. Carregue a base iris e a separe em treino e teste usando 40% dos exemplos para teste. Reduza a base de treino para ter apenas os 24 exemplos iniciais e a de teste para ter apenas os 16 exemplos iniciais. Considere que o valor 1.5 representa um elemento ausente. Substitua as ocorrências de valores ausentes pelo valor mais frequente nas bases de treino e de teste reduzidas.

2. Utilize agora os mesmos valores mais frequentes usados na base de treino na base de teste do exercício 1.

3. Padronize a base de treino do exercício 1 e utilize os mesmos fatores de normalização para padronizar a base de teste.

4. Discretize a base Z_Train usando a estratégia de igual frequência nos intervalos e utilize os mesmos intervalos para discretizar a base Z_Test. Repita depois usando a estratégia de igual tamanho dos intervalos. Atenção: crie novas bases para não alterar a definição original de Z_Train e Z_Test.

5. Leia a base de dados breast cancer e execute 3 rodadas de validação cruzada aninhada estratificada, porém use padronização em todas as características da base. Use o classificador KNN com k = {1, 3, 5}. Apresente a acurácia obtida em cada fold de teste, assim como a média da acurácia, desvio padrão e intervalo de confiança.

6. Crie um dataset com 5 exemplos com dados nominais para condições do dia com temperatura (baixa, média e alta), aparência (ensolarado, nublado, chuvoso) e vento (pouco, muito). Transforme os dados de temperatura e vento para o formato ordinal e o de aparência para o formato binário.

7. Compare o desempenho (acurácia média, desvio padrão e intervalo de confiança a 95%) do método PCA com três componentes com o método de seleção univariada de características usando quiquadrado e selecionando 3 características no dataset wine. Use o classificador Vizinho Mais Próximo (k = 1) e o método de validação cruzada com 10 folds.

8. Obtenha o desempenho (acurácia média, desvio padrão e intervalo de confiança a 95%) do método de seleção multivariada de características com seleção sequencial para frente e selecionando 3 características no dataset wine. Use o classificador Vizinho Mais Próximo (k = 1) e o método de validação cruzada com 10 folds.