Universidade de Brasília
Departamento de Estatística
EST0077 - 2023/1

SOLUÇÃO DOS EXERCÍCIOS PARA ENTREGA 1

Igor de Oliveira Barros Faluhelyi
Prof. George

A solução dos exercícios para entrega 1 - prova 1 - é requisitada como parte do processo avaliativo da disciplina de Tópicos 2 no primeiro semestre de 2023.

Brasília
2023

Sumário

1 Exercício 3 da Lista 1

Fazer os seguintes exercícios do Capítulo 2 de James et al.[1] (2021):

1.1 Exercício 1

Enunciado:

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = Var(\varepsilon)$, is extremely high.

Resposta:

(a) Melhor: Um tamanho de amostra grande significa que um modelo flexível será capaz de se ajustar melhor aos dados

(b) Pior: Um modelo flexível provavelmente teria um *overfit*. Modelos flexíveis geralmente obtêm melhores resultados quando há grandes conjuntos de dados disponíveis.

(c) Melhor: Modelos flexíveis têm um desempenho melhor em conjuntos de dados não-lineares, pois possuem mais graus de liberdade para aproximar uma relação não-linear.

(d) Pior: Um modelo flexível provavelmente terá um *overfit*, devido a ajustar-se de forma mais precisa ao ruído nos termos de erro do que um modelo inflexível. Em outras palavras, seja f a função ideal para descrever os dados, as observações - pontos do $\mathbb{R}^n$ - estarão distantes de f se a variância dos termos de erro for muito alta. Isso sugere que f é linear e, portanto, um modelo mais simples seria capaz de estimar melhor f.

[1] James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning (With applications in R)*, Springer, 2021.

1.2 Exercício 7

Enunciado:

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable:

TABLE

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Resposta:

(a) A distância euclidiana é a distância em linha reta entre dois pontos. Ela pode ser calculada usando o teorema de Pitágoras.

Para o espaço em 3D, temos:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}$$

Utilizando a fórmula acima, obtemos as seguintes distâncias:

$$d(1, test) = 3$$
$$d(2, test) = 2$$
$$d(3, test) = 3.16$$
$$d(4, test) = 2.24$$
$$d(5, test) = 1.41$$
$$d(6, test) = 1.73$$

o cálculo segue no arquivo .R.

(b) Verde, pois a observação mais próxima é verde.

(c) Vermelho. As três observações mais próximas são verde, vermelho e vermelho. A probabilidade do ponto de teste pertencer a vermelho é de 2/3 e verde é de 1/3.

Portanto, a predição é vermelho. O cálculo pela função knn do pacote class do R está no arquivo .R.

(d) Para fronteiras altamente não-lineares, esperamos que o melhor valor de K seja pequeno. Valores menores de K resultam em um modelo KNN mais flexível, o que produzirá uma fronteira de decisão não-linear. Um valor maior de K significa que mais pontos de dados são considerados pelo modelo KNN, o que resulta em uma fronteira de decisão mais próxima de uma forma linear.

2 Exercício 24 da Lista 2

Considere o seguinte conjunto de dados de Pacientes em Tratamento de Hemodiálise:

Idade	**Proteína**	**Energia**	**Albumina**	**IMC**
32	1.59	2738.86	4.2	24.1
61	0.49	824.26	3.9	29.8
51	1.14	1307.03	4.1	20.0
53	0.74	925.47	4.2	25.0
24	1.99	2787.46	3.8	21.5
65	1.00	1222.51	4.2	25.0
35	2.32	2038.28	4.1	18.7
45	0.93	1061.53	4.2	22.0
57	0.81	1657.73	4.2	31.2
32	1.23	1652.76	3.9	24.3
66	0.99	1636.25	4.1	27.7
27	1.40	1845.07	4.0	21.8
54	1.08	1542.30	3.9	29.0
55	1.22	1214.53	4.0	21.1
50	0.57	1451.17	4.0	27.1
48	0.83	1786.95	4.1	24.7
28	1.55	1975.26	3.5	18.8
66	1.10	1248.64	4.0	18.9
66	0.44	987.86	4.0	27.6
48	0.58	1067.10	4.3	26.4
60	0.43	968.62	4.0	35.9
59	0.66	836.94	3.9	25.3
50	1.81	1197.99	3.9	19.5
29	1.21	1818.31	4.2	21.8
40	0.98	1238.91	3.5	21.9
47	1.48	2153.47	3.5	17.3
52	0.98	1720.60	3.6	29.7
54	1.02	1906.30	4.5	31.9
53	0.82	981.85	3.9	26.2
47	0.46	1020.95	4.4	31.2
42	1.34	1028.10	3.6	18.1
79	1.48	1465.91	3.9	18.3
61	1.39	1456.12	3.9	24.9

2.1 Represente graficamente e através de medidas descritivas.

Descrição para variável Idade:

Estatística	**Valor**
Média	49.58
Desvio Padrão	13.23
Mínimo	24
1º Quartil	42
Mediana	51
3º Quartil	59
Máximo	79

Descrição para variável Proteína:

Estatística	**Valor**
Média	1.093
Desvio Padrão	0.46
Mínimo	0.43
1º Quartil	0.81
Mediana	1.02
3º Quartil	1.39
Máximo	2.32

Descrição para variável Energia:

Estatística	**Valor**
Média	1477.7
Desvio Padrão	499.96
Mínimo	824.3
1º Quartil	1061.5
Mediana	1451.2
3º Quartil	1787.0
Máximo	2787.5

Descrição para variável Albumina:

Estatística	**Valor**
Média	3.985
Desvio Padrão	0.25
Mínimo	3.500
1º Quartil	3.900
Mediana	4.000
3º Quartil	4.200
Máximo	4.500

Descrição para variável IMC:

Estatística	**Valor**
Média	24.45
Desvio Padrão	148,63
Mínimo	17.30
1º Quartil	21.10
Mediana	24.70
3º Quartil	27.60
Máximo	35.90

Figura 1: Boxplot das variáveis

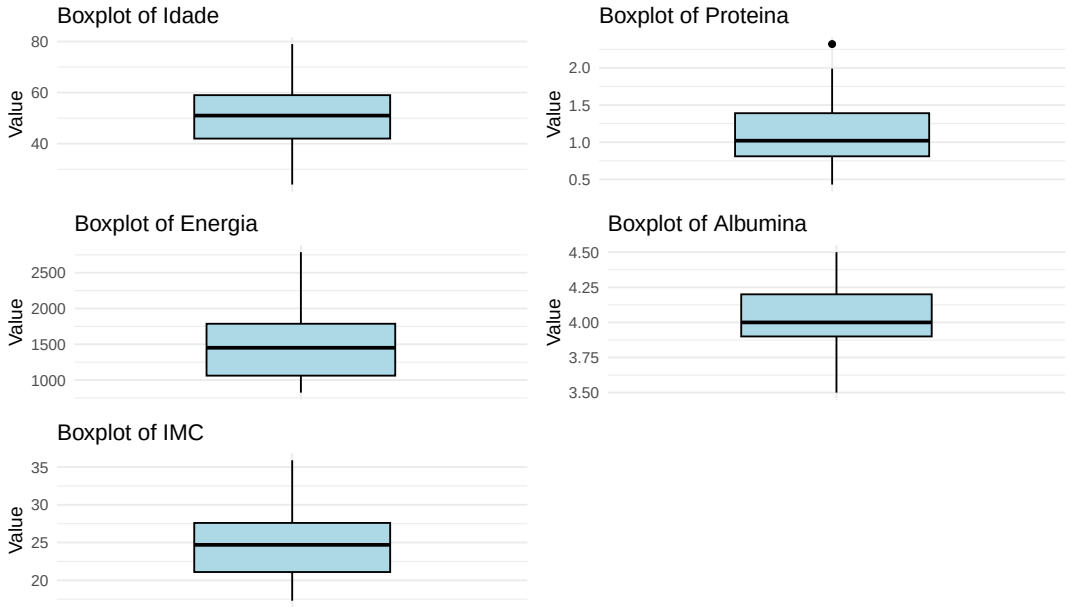
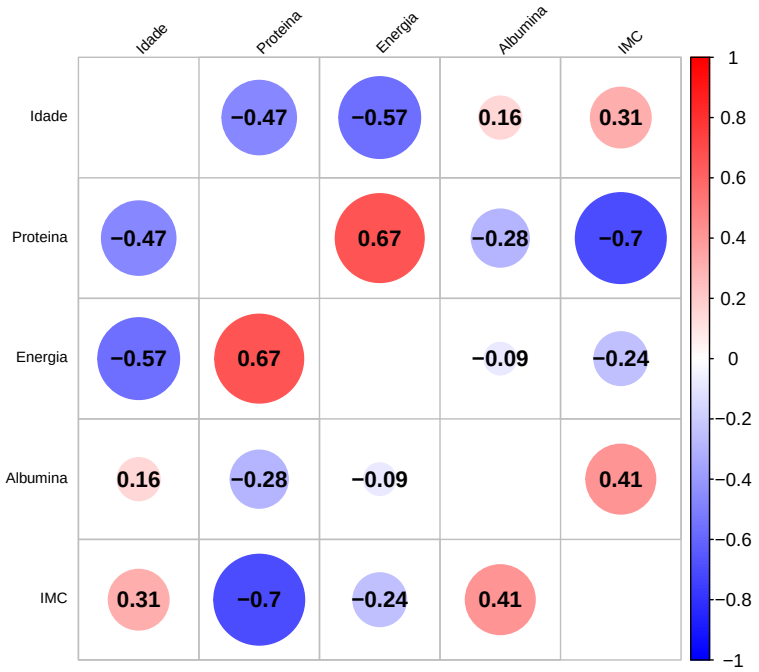


Figura 2: Gráfico de dispersão das variáveis



2.2 Obtenha a decomposição espectral e verifique se existe indicação de uma possível redução da dimensão do estudo em questão. Justifique.