MAE0399 - Análise de Dados e Simulação - primeiro semestre de 2020
Professora: Márcia D'Elia Branco

LISTA 1

1) Para cada um dos cenários a seguir indique se o problema pode ser considerado como Classificação ou Regressão. Também, indique se o principal interesse é Inferência ou Predição. Finalmente, tente identificar n (tamanho da amostra) e p (numero de preditores).

a) Nós coletamos informações das 500 maiores empresas da brasileiras. Para cada uma delas anota-se o lucro anual, número de empregados, tipo de empresa e salário médio anual. Estamos interessado em entender que fatores afetam o salário.

b) Nós estamos interessados em lançar um novo produto e desejamos avaliar se ele será bem sucedido ou não. Selecionamos uma amostra de 20 produtos lançados anteriormente com caracteristicas similares. Para cada um deles anota-se se obteve ou não sucesso, o preço de venda, o valor gasto em propaganda e mais 10 outras variáveis.

c) Estamos interessados em prever a taxa de juros estabelecida pelo Banco Central no próximo mês. Para isso analisamos a série histórias das taxas de juros dos 12 últimos meses e mais outros 5 fatores economicos que podem afetar essa taxa.

2) Descreva três situações práticas que poderiam ser classificadas como um problema de Classificação, Regressão ou Agrupamento (*Clustering*) .

3) Explique a diferença entre uma abordagem paramétrica e uma não-paramétrica de aprendizado estatístico. Quais as vantagens e desvantagens de cada uma delas?

4) Explique a diferença entre aprendizado supervisionado e não-supervisionado.

5) Os dados representam velocidades do vento (km/h) num determinado aeroporto para os primeiros 15 dias de dezembro de 2008.

(a) Calcule a média, a mediana, o desvio padrão e os quartis da velocidade e desenhe o gráfico *boxplot*. Faça manualmente.

Dia	1	2	3	4	5	6	7	8
Velocidade	22,2	61,1	13,7	27,8	22,7	7,4	8,7	6,3
Dia	9	10	11	12	13	14	15	
Velocidade	20,4	25,6	23,2	11,1	13,0	7,2	14,8	

(b) Existe algum valor atípico? Em caso afirmativo, qual? Remova esse valor e refaça o item anterior. Comente as diferenças encontradas nas medidas calculadas. Quais medidas são mais afetadas pela presença de valores discrepantes?

6) Considere os dados da planilha "dadosdomiciliosCEA15P02.xlsx". Use o programa **R** para responder os itens a seguir.

(a) Construa os gráficos *boxplot* para a variável "consumo de gás anual domiciliar per capita (em Kg)" , segundo o tipo de domicílio (Casa, Apartamento, Comodo).

(b) Obtenha as medidas resumos (média, mediana, min, max, Q1, Q3 e dp) da variável "consumo de gás anual domiciliar per capita (em Kg)", segundo o tipo de domicílio (Casa, Apartamento, Comodo). Apresente esses resultados em uma tabela.

(c) Analise os resultados (comente) obtidos com base nos resultados de (a) e (b). Compare os grupos em relação a tendência central, dispersão e existência de valores discrepantes.

(d) Construa os gráficos *boxplot* para a variável "consumo de gás anual domiciliar per capita (em Kg)" , segundo a presença de rede geral de energia elétrica. Compare os grupos em relação a assimetria dos dados.

(e) Determine os gráficos histograma e *boxplot* para a variável " renda domiciliar per capita mensal (em reais)". A suposição de normalidade para essa variável é rasoável? Por que?

(f) Categorize a variável renda. Considere as seguintes classes: até 1 salário mínimo(s.m.), de 1 a 2 s.m., de 2 a 4 s.m. e mais que 4 s.m. . Para essa nova variável, obtenha um gráfico de setores (pizza).

(g) Construa os gráficos *boxplot* para a variável "consumo de gás anual domiciliar per capita (em Kg)" , segundo a variável renda categorizada. Obtenha as medidas resumos por categoria e faça uma análise dos resultados (comente).

7) A tabela a seguir apresenta um conjunto de treinamento com 6 observações, 3 preditores e uma variável resposta qualitativa.

Obs	X_1	X_2	X_3	Y
1	0	3	0	Vermelho
2	2	0	0	Vermelho
3	0	1	3	Vermelho
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Vermelho

Deseja-se usar esses dados para prever Y usando o método dos K vizinhos mais próximos, quando $X_1 = X_2 = X_3 = 0$. Use como medida de vizinhança a distância euclidiana. Obtenha a predição de Y considerando $K = 1$ e $K = 3$.

8) Considere o modelo de regressão linear sem o intercepto

$$Y = \beta X + \epsilon$$

e uma amostra de treinamento de tamanho n. Obtenha o estimador de mínimos quadrados para β.

9) Uma indústria farmacêutica vende um remédio para combater resfriado. Após dois anos de operação, ela coletou as seguintes informações trimestrais:

Trimestre	Y	X	Z
1	96	20	90
2	92	20	70
3	99	25	90
4	104	25	70
5	117	30	80
6	106	30	80
7	112	35	100
8	105	35	90

Sendo Y: Vendas (10.000 unidades) ; X: Temperatura média do trimestre (graus Celsius) e Z: Despesas com propaganda (1.000 reais).

a) Qual seria a variável resposta e quais as preditoras?

b) Construa os diagramas de dispersão das variáveis *Vendas* versus *Despesas com propaganda* e *Vendas* versus *Temperatura média do trimestre*. Obtenha os coeficientes de correlação linear associados aos gráficos construídos e interprete os valores obtidos. Qual das duas variáveis explica melhor as *Vendas* ?

c) Obtenha a reta de regressão das vendas em função da variável escolhida em (b). Qual é o significado prático do coeficiente **b** encontrado?

d) Ajuste um modelo de regressão linear multipla considerando as duas variáveis preditoras. Compare o valor de R^2 desse ajuste com o ajuste simples obtido no item (c). Comente. Qual a interpretação da medida R^2 ?