Tópicos Especiais em Engenharia de Computação – 2017.2 - DCA0301

Terceira Lista de Exercícios

1-) Considere o processo de identificação de aglomerados ("clusters") com base em uma técnica hierárquica aglomerativa. Neste problema considere o método de Ward resumido abaixo. Considere também dois critérios para parada do processo aglomerativo no dendograma e identificação do número de aglomerados. O critério R^2 e o critérioo Pseudo T^2.
Para o problema considere a tabela de índices de desenvolvimento de paíse (Fonte ONU-2002, Livro – Análise de dados através de métodos de estatística multivariada – Sueli A. Mingoti) abaixo.

Método de Ward:

a-) Inicialmente, cada elemento é considerado como um único conglomerado

b-) Em cada passo do algoritmo de agrupamento (formação do dendograma) calcule a similaridade fazendo uso da distância Euclidiana ao quadrado entre os conglomerados formados, isto é

$d(C_l, C_i) = \frac{n_l n_i}{n_l + n_i} \left\| \mathbf{m}_l - \mathbf{m_i} \right\|^2$ onde,

n_i é o número de elementos no conglomerado C_i

$\mathbf{m}_i$ é o centroide do conglomerado C_i dado por $\mathbf{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x_{ij}}$

Junte os aglomerados com menor distância.

Critério de parada pelo coeficiente R^2

Calcule o coeficiente R^2 em função do número de passos e pare o processo quando for observado um salto elevado no valor do coeficiente. Este ponto determina o número de aglomerados.

$$R^2(g_k) = \frac{SSB}{SST_c}$$

$$SST_c = \sum_{i=1}^{g_k} \sum_{j=1}^{n_i} \left\| \mathbf{x_{ij}} - \mathbf{m_i} \right\|^2$$

$$SSB = \sum_{i=1}^{g_k} n_i \left\| \mathbf{m_i} - \mathbf{m} \right\|^2$$

m: vetor média global
$g_{k:}$ número de conglomerados

Critério do Pseudo T^2

Busca-se determinar o número de agrupamento que resulte no maior valor do coeficiente Pseudo T^2dado por

$$Pst^2 = \frac{B_{il}}{\left[\sum_{j \in C_i} \left\|\mathbf{x_{ij}} - \mathbf{m_i}\right\|^2 + \sum_{j \in C_l} \left\|\mathbf{x_{lj}} - \mathbf{m_l}\right\|^2\right](n_i + n_l - 2)^{-1}}$$

$$B_{il} = \frac{n_i n_l}{n_i + n_l} \left\|\mathbf{m_i} - \mathbf{m_l}\right\|^2$$

Países	Expectativa de Vida	Educação	PIB	Estabilidade Política
Reino Unido	0.88	0.99	0.91	1.10
Austrália	0.90	0.99	0.93	1.26
Canadá	0.90	0.98	0.94	1.24
Estados Unidos	0.87	0.98	0.97	1.18
Japão	0.93	0.93	0.93	1.20
França	0.89	0.97	0.92	1.04
Cingapura	0.88	0.87	0.91	1.41
Argentina	0.81	0.92	0.80	0.55
Uruguai	0.82	0.92	0.75	1.05
Cuba	0.85	0.90	0.64	0.07
Colômbia	0.77	0.85	0.69	-1.36
Brasil	0.71	0.83	0.72	0.47
Paraguai	0.75	0.83	0.63	-0.87
Egito	0.70	0.62	0.60	0.21
Nigéria	0.44	0.58	0.37	-1.36
Senegal	0.47	0.37	0.45	-0.68
Serra Leoa	0.23	0.33	0.27	-1.26
Angola	0.34	0.36	0.51	-1.98
Etiópia	0.31	0.35	0.32	-0.55
Moçambique	0.24	0.37	0.36	0.20
China	0.76	0.80	0.61	0.39
Média	0.69	0.75	0.68	0.16
Desvio Padrão	0.24	0.249	0.229	1.056

Construa dendondograma e indique o ponto de corte ou de parada determinado com isto os clusters ou aglomerados.

2-) Repita o problema acima considerado agora o método do K-means ou k-médias que é uma técnica de clusterização para determinação de clusters por particionamento.

Compare os resultados com os obtidos pelo método da questão 1.

3-) Repita o problema da questão 1 considerado agora a solução do problema pela rede de Kohonen ou como é conhecido SOM que é uma técnica de clusterização e também uma técnica de visualização de dados de alta dimensão em baixa dimensão. Compare com os resultados obtidos nas questões 1 e 2.

4-) Considere o problema de análise de componentes principais (PCA), isto é, determinar em uma distribuição de dados as componentes que tenham associadas a elas a maior variância e representar as mesmas no espaço de dados formado pelos autovetores da matriz de correlação. Neste sentido considere o seguinte problema.
A tabela abaixo apresenta os dados relativos a amostras de solo. Para cada amostra, tem-se as medidas das porcentagens de areia (X1), sedimentos (X2), argila (X3) e a quantidade de material orgânico (X4). Da referida tabela obtenha as estatísticas descritivas de cada variável, isto é, a média, a mediana, o desvio padrão, os valores máximo e mínimo. Sob estas condições :

a-) Obtenha desta tabela a matriz de covariância.

b-)Desta matriz determine os autovalores ordenados do máximo ao mínimo e os autovetores correspondentes.

c-) Apresente as equações da componentes principais, isto é, cada componente é dada por

$Y_i = \mathbf{e}^t{}_i X_{\mathbf{i}} = e_{1i}X_1 + e_{2i}X_2 + e_{3i}X_3 + e_{4i}X_4 \qquad i = 1,2,3,4$, onde e_{ji} é a componente i do autovetor j.

d-) Calcule os percentuais de variância para cada componente e ordene a classificação das variáveis segundo este critério.

Tabela: Dados das amostras de solo (Livro – Análise de dados através de métodos de estatística multivariada – Sueli A. Mingoti)

Amostra	Areia (%):X_1	Sedimentos(%):X_2	Argila(%):X_3	Mat. Orgân(%):X4
1	79,9	13,9	6,2	3,3
2	78,5	16,3	7,2	2,5
3	68,9	22,6	8,5	3,6
4	62,2	20,2	17,6	2,8
5	69,2	23,7	7,1	0,9
6	67,8	19,8	12,4	3,8
7	61,3	24,9	13,8	2,2
9	71,6	19,2	9,2	3,6
10	83,7	10,5	5,8	4,4
11	67,1	26,5	6,4	1,4
12	59,8	27,9	12,3	3,5
13	66,7	23,2	10,1	2,9

Amostra	Areia (%):X_1	Sedimentos(%):X_2	Argila(%):X_3	Mat. Orgân(%):X4
14	72,8	14,5	12,7	1,9
15	60,9	28,9	10,2	1,5
16	61,4	29,2	9,4	2,5
17	75,0	16,8	8,2	3,1
18	80,5	11,9	7,6	3,8
19	71,3	18,5	10,2	2,6
20	56,6	28,9	14,5	2,8
21	55,9	32,8	11,3	3,1
22	61,5	28,1	10,4	2,7
23	59,2	28,4	12,4	2,8
24	76,9	16,3	6,8	2,9
25	58,0	27,6	14,4	3,4

5-) Considere os problemas abaixo relativos a aproximação por mínimos quadrado (regressão linear, regressão não linear) de dados. Para cada problema após a obtenção dos coeficientes do modelo de regressão calcule as estimativas estatísticas dos dados (média, variância, desvio padrão), assim como o coeficiente de correlação dado por $r = \sqrt{\frac{S_t - S_r}{S_t}}$ onde $S_t = \sum_{i=1}^{N} (y_i - \bar{y})^2$ e $S_r = \sum_{i=1}^{N} (y_i - \hat{y}_i(x))^2$ sendo $\hat{y}(x)$ o modelo de aproximação.

a-) Para a tabela de dados

x	4	6	8	10	14	16	20	22	24	28	34	36	38
y	30	18	22	28	14	22	16	8	20	14	14	0	8

Aproxime os dados por um modelo regressivo linear, isto é, $\hat{y}(x) = a_0 + a_1 x$.

b-) Para a tabela de dados

x	2.5	3.5	5	6	7.5	10	12.5	15	17.5	20
y	5	3.4	2	1.6	1.2	0.8	0.6	0.4	0.3	0.3

Aproxime os dados por um modelo regressivo não linear dados por $\hat{y}(x) = a_0 x^{a_1}$

Sugestão: Linearize inicialmente o modelo.

c-) Para a tabela de dados

x_1	1	1	2	2	3	3	4	4
x2	1	2	1	2	1	2	1	2
y	18	12.8	25.7	20.6	35.0	29.8	45.5	40.3

Aproxime os dados por um modelo regressivo linear múltiplo dado por $\hat{y}(x) = a_0 + a_1 x_1 + a_2 x_2$.

Considere as duas questões abaixo associadas ao Controle Estatístico da Qualidade.

6-) Uma matriz de ejeção é usada para fabricação de hastes de alumínio. O diâmetro das hastes é uma característica crítica da qualidade. Abaixo, são mostrados valores de $\overline{X}$ e R para 20 amostras de cinco hastes cada. As especificações sobre as hastes são $0{,}5035 \pm 0{,}0010$ polegadas. Os valores apresentados na tabela são os três últimos dígitos das medidas; isto é, 34,2 é lido como 0,50342.

a-) Estabeleça gráficos de controle $\overline{X}$ e R, revisando os limites de controle experimentais, admitindo que se possa encontrar causas atribuíveis.
b-) Calcule a RCP (Razão da Capacidade do Processo)
c-) Qual a porcentagem de hastes defeituosas produzidas por este processo.

Amostra	$\overline{X}$	R	Amostra	$\overline{X}$	R
1	34,2	3	11	35,4	8
2	31,6	4	12	34,0	6
3	31,8	4	13	36,0	4
4	33,4	5	14	37,2	7
5	35,0	4	15	35,2	3
6	32,1	2	16	33,4	10
7	32,6	7	17	35,0	4
8	33,8	9	18	34,4	7
9	34,8	10	19	33,9	8
10	38,6	4	20	34,0	4

7-) Os dados seguintes representam o número de defeitos de solda observados em 24 amostras de cinco placas de circuitos impresso: 7, 6, 8, 10, 24, 6, 5, 4, 8, 11, 15, 8, 4, 16, 11, 12, 8, 6, 5, 9, 7, 14, 8, 21. Podemos concluir que o processo esteja sob controle usando o gráfico c? Se não, suponha que se possa encontrar causas atribuíveis e revise os limites de controle.

Considere as duas questões abaixo associadas a Engenharia de Confiabilidade.

8-) Cinco unidades idênticas são arranjadas em uma redundância ativa para formar um subsistema. As falhas das unidades são independentes, e pelo menos duas das unidades devem sobreviver 1000 horas para que o subsistema realize sua missão.

a-) Se as unidades tem distribuição exponencial do tempo de falha, com taxa de falhas 0,002, qual é a confiabilidade do subsistema?

b-) Qual é a confiabilidade se for exigida a sobrevivência de apenas uma das unidades?

9-) Para um teste sem reposição que termina após 200 horas de operação, nota-se que as falhas ocorrem nos seguintes tempos: 9, 21, 40,55 e 85 horas. Supõe-se que as unidades tenham uma distribuição exponencial do tempo de falha, com 100 unidades em teste incialmente.
a-) Estime o tempo médio de falha
b-) Construa um limite de confiança inferior a 95% de confiança para o tempo médio de falha.

Sugestão de temas para o trabalho escrito sob forma de um artigo científico:

1-) Redes Bayesiana Dinâmicas e Aplicações

2-) Árvore de Decisões

3-) Métodos de Aprendizagem de Máquinas para Classificação de Padrões

4-) Big Data

5-) Mineração de Dados

6-) Deep Learning

7-) Aprendizagem por Reforço

Avaliação 3: 0,5*Lista 3+0,5*Trabalho

23/11/2018 – Entrega da Lista
28/11/2018 - Apresentação dos Trabalhos