Aprendizagem de Máquinas e Mineração de Dados – 2018.1 - DCA 0133

Terceira Lista de Exercícios

1-) Utilize uma rede NARX para fazer a predição de um passo, até predição de três passos da série temporal $x(n) = \ln(1 + \cos(n + sen^2(n)))$. Avalie o desempenho mostrando para cada caso os erros de predição. Solucione o problema considerando a NARX uma rede uma Perpcetron de múltiplas camadas com realimentação.

2-) Desenvolva um sistema para reconhecer vogais escritas a mão fazendo uso de uma:
a-) Rede neural competitiva
b-) Rede neural SOM

Compare o desempenho da duas redes.

Obs. Utilize um banco de dados existente ou gere seu banco de dados usando para cada vogal a escrita de 10 pessoas diferentes.

3-) Pesquise e apresente um trabalho sobre a reconstrução de imagens bidimensional e tridimensional usando a rede SOM e a rede Neuro-GAS.

4-)Um problema para testar a capacidade de uma rede neural atuar como classificador de padrões é o problema das duas espirais intercaladas. Gere os exemplos de treinamento usando as seguintes equações:

para espiral 1 $x = \frac{\theta}{4}\cos\theta \quad y = \frac{\theta}{4}\,\text{sen}\,\theta \quad \theta \geq 0$
para espiral 2 $x = (\frac{\theta}{4} + 0.8)\cos\theta \quad y = (\frac{\theta}{4} + 0.8)\,\text{sen}\,\theta \quad \theta \geq 0$
fazendo θ assumir 51 igualmente espaçados valores entre 0 e 20 radianos. Utilize uma rede competitiva e em seguida uma rede SOM para atuar como classificador auto-supervisionado, isto é, a espiral 1 sendo uma classe e espiral 2 sendo outra classe. Para comparar as regiões de decisões formadas pela rede , gere uma grade uniforme com 100 x 100 exemplos de teste em um quadrado [-5,5]. Esboce os pontos classificados pela rede.

5-) A propriedade de ordenação topológica do algoritmo SOM pode ser usada para formar uma representação bidimensional abstrata de um espaço de entrada de alta dimensionalidade. Para investigar esta forma de representação, considere uma grade bidimensional consistindo de 10x10 neurônios que é treinada tendo como entrada os dados oriundos de quatro distribuições gaussianas, C_1, C_2, C_3, e C_4, em um espaço de entrada de dimensionalidade igual a oito, isto é $\mathbf{x} = (x_1, x_2, \ldots x_8)^t$. Todas as nuvens têm variâncias unitária, mas centros ou vetores média diferentes dados por $\mathbf{m_1} = (0,0,0,0,0,0,0,0)^t, \mathbf{m}_2 = (4,0,0,0,0,0,0,0)^t\, \mathbf{m}_3 = (0,0,0,4,0,0,0,0)^t, \mathbf{m}_4 = (0,0,0,0,0,0,0,4)^t$.
Calcule o mapa produzido pelo algoritmo SOM, com cada neurônio do mapa sendo rotulado com a classe particular mais representada pelos pontos de entrada em sua volta. O objetivo é visualizar os dados de dimensão 8 em um espaço de dimensão 2, constiuido pela grade de neurônios.

6-) Implemente o algoritmo K-means e considere o dados apresentados na tabela abaixo para serem usando no processo de clustering.

Amostra	x_1	x_2	x_3
1	-7.82	-4.58	-3.97
2	-6.68	3.16	2.71
3	4.36	-2.19	2.09
4	6.72	0.88	2.80
5	-8.64	3.06	3.50
6	-6.87	0.57	-5.45
7	4.47	-2.62	5.76
8	6.73	-2.01	4.18
9	-7.71	2.34	-6.33
10	-6.91	-0.49	-5.68
11	6.18	2.81	5.82
12	6.72	-0.93	-4.04
13	-6.25	-0.26	0.56
14	-6.94	-1.22	1.13
15	8.09	0.20	2.25
16	6.81	0.17	-4.15
17	-5.19	4.24	4.04
18	-6.38	-1.74	1.43
19	4.08	1.30	5.33
20	6.27	0.93	-2.78

a-) Considere que existam três clusters e a inicialização dos centros seja dada por $\mathbf{m_1}=(0,0,0)^t$, $\mathbf{m_2}=(1,1,1)^t$, $\mathbf{m_3}=(-1,0,2)^t$.
b-) Repita o item a considerando que os centros iniciais sejam $\mathbf{m_1}=(-0.1,0,0.1)^t$, $\mathbf{m_2}=(0,-0.1,0.1)^t$, $\mathbf{m_3}=(-0.1,-0.1,0.1)^t$. Compare obtido com o item (a) e explique a razão da diferenças, incluindo o número de interações para alcançar a convergência.

7-) 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^2 dado 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.

8-) 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.

9-) 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

10-) Pesquise e apresente um estudo sobre BIG DATA.

Calendário das Atividades do Final do Curso:

Dada de apresentação da lista 3: 19/06/2018

Apresentação do Trabalho Final: 26/06/2018

Data da Quarta-Avaliação: 03/07/208