Avaliação sobre Aprendizado de Máquina

1 message

Google Forms <forms-receipts-noreply@google.com> Fri, Sep 15, 2017 at 4:59 PM
To: renvmorales@gmail.com

Thanks for filling out Avaliação sobre Aprendizado de Máquina

Here's what we got from you:

Avaliação sobre Aprendizado de Máquina

Observações:
- A interpretação das questões é parte integrante da avaliação.
- Sempre que julgar apropriado, você pode usar softwares (e.g., R, Python, Weka, etc.) para resolver as questões.
- O tempo de realização da prova será usado em sua avaliação. Inicie a prova imediatamente após o recebimento e envie suas respostas o mais rapidamente possível. O tempo máximo de prova é de 3 horas.

Obs: Utilizar a vírgula como separador de decimais. Ex: 0,15 ao invés de 0.15

Email address *

renvmorales@gmail.com

Nome Completo *

Renzo Angelo Viloche Morales

Email *

renvmorales@gmail.com

Módulo 1 – Agrupamento de Dados

Levando-se em conta a matriz de distâncias entre cinco objetos abaixo, esboçar o dendrograma obtido pelo método hierárquico aglomerativo conhecido como vinculação simples (single linkage), no qual a distância entre dois clusters é dada pela menor distância entre dois objetos (um de cada cluster).

$$M_D = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

Questão 1

Considerando o dendrograma obtido, assinale a(s) alternativa(s) correta(s):

- [ ] Inicialmente, cada objeto não forma um singleton (cluster unitário).
- [ ] Os objetos 1 e 2 formam o primeiro cluster não unitário.
- [ ] Os objetos 1 e 5 nunca pertencerão ao mesmo cluster.
- [ ] Os objetos 4 e 5 não podem pertencer ao mesmo cluster.
- [ ] O algoritmo produz uma sequência de partições aninhadas.

Questão 2

Considere que os seguintes vetores representam tuplas de um banco de dados: [1,1];[1,2];[2,1];[2,2];[5,1];[6,1];[5,2]. Simular a execução de 5 iterações do algoritmo k-means, para k=2, inicializando o algoritmo nos pontos [3,0] e [5,0]. Quais são os centróides obtidos?

Questão 3

Considere uma partição de referência formada por duas categorias P={P1,P2}, sendo P1={x1,x3,x6} e P2={x2,x4,x5,x7}, e um conjunto de grupos (clusters) obtidos por meio de um algoritmo de agrupamento C={C1,C2}, sendo C1={x1,x3,x4,x5} e C2={x2,x6,x7}. Calcular a aderência entre as categorias e os grupos usando o Rand Index.

Questão 4

Considerando que p=[p1,...,pn] e q=[q1,...,qn] são versores:

- ( ) A distância euclidiana entre p e q, dist_euclidiana(p,q), é igual ao cosseno do ângulo entre os versores.
- ( ) É possível demonstrar que dist_euclidiana(p,q)² = 2(1 - cos θ), onde θ é o ângulo entre p e q.
- ( ) Não se pode usar p e q como entradas para um algoritmo de agrupamento de documentos textuais.
- ( ) Nenhuma das alternativas está correta.

Módulo 2 – Classificação

Considere a seguinte base de dados formada por quatro atributos previsores e pelo atributo meta para a classificação (classe).

Tabela em csv: http://goo.gl/52BJpm

Aparência	Temperatura	Umidade	Vento	Classe
Ensolarado	Quente	Alta	Falsa	Não
Ensolarado	Quente	Alta	Verdadeiro	Não
Chuvoso	Frio	Normal	Verdadeiro	Não
Ensolarado	Morna	Alta	Falsa	Não
Chuvoso	Morna	Alta	Verdadeiro	Não
Nublado	Quente	Alta	Falsa	Sim
Chuvoso	Morna	Alta	Falsa	Sim
Chuvoso	Frio	Normal	Falsa	Sim
Nublado	Frio	Normal	Verdadeiro	Sim
Ensolarado	Frio	Normal	Falsa	Sim
Chuvoso	Morna	Normal	Falsa	Sim
Ensolarado	Morna	Normal	Verdadeiro	Sim
Nublado	Morna	Alta	Verdadeiro	Sim
Nublado	Quente	Normal	Falsa	Sim

Questão 5

Classificar a tupla [ensolarado, quente, normal, verdadeiro, ?] pelo classificador bayesiano simples (Naive Bayes).

- (●) Sim
- ( ) Não
- ( ) As duas classes são igualmente prováveis

Considere a seguinte base de dados, na qual o atributo meta (classe) é “Espera”

Tabela em csv: http://goo.gl/3nkfBh

Exemplo	Sexta / Sábado	Faminto	Clientes	Tipo	Espera
1	Não	Sim	Cheio	Tailandês	Não
2	Sim	Não	Cheio	Francês	Não
3	Não	Não	Nenhum	Hambúrguer	Não
4	Sim	Não	Cheio	Hambúrguer	Não
5	Sim	Sim	Cheio	Italiano	Não
6	Não	Não	Nenhum	Tailandês	Não
7	Não	Sim	Alguns	Francês	Sim
8	Não	Não	Alguns	Hambúrguer	Sim
9	Sim	Sim	Cheio	Tailandês	Sim
10	Não	Sim	Alguns	Italiano	Sim
11	Não	Sim	Alguns	Tailandês	Sim
12	Sim	Sim	Cheio	Hambúrguer	Sim

Questão 6A

Obter a árvore de decisão, sem nenhuma poda e selecionando-se os atributos pelo critério do ganho da informação, para classificar a tupla [Sim,Sim,Cheio,Francês,?].

Questão 6B

Deseja-se classificar um novo exemplo xt=[Sim,Sim,Alguns,Italiano,?]. Considere que todos os seus atributos sejam nominais. Pede-se classificar xt de acordo com o método dos vizinhos mais próximos (k-NN, com k=3). Utilize a medida de dissimilaridade baseada no coeficiente de casamento simples para um espaço p-dimensional

$$d_{SM}(i,j) = \sum_{k=1}^{k=p} s_k \qquad s_k = \begin{cases} (x_{ik} = x_{jk}) \Rightarrow s_k = 0; \\ (x_{ik} \neq x_{jk}) \Rightarrow s_k = 1; \end{cases}$$

Módulo 3 – Regressão

Resolva as questões 7 - 9 considerando a base de dados abaixo (cujas variáveis são todas contínuas):
Tabela em csv: http://goo.gl/LTN1ca

X1	X2	X3	X4	X5	Y
180	8	3070	1300	3504	120
150	8	3500	1650	3693	115
180	8	3180	1500	3436	110
160	8	3040	1500	3433	120
170	8	3020	1400	3449	105
150	8	4290	1980	4341	100
140	8	4540	2200	4354	90
140	8	4400	2150	4312	85
140	8	4550	2250	4425	100
150	8	3900	1900	3850	85
150	8	3830	1700	3563	100
140	8	3400	1600	3609	80
150	8	4000	1500	3761	95
140	8	4550	2250	3086	100
240	4	1130	9500	2372	150
220	6	1980	9500	2833	155
180	6	1990	9700	2774	155
210	6	2000	8500	2587	160
270	4	9700	8800	2130	145
260	4	9700	4600	1835	205

Questão 7

Estime a capacidade de generalização de um regressor linear, Y=f(X1, X2, ..., X5), sem usar regularização, via validação cruzada Leave-One-Out.

RMSE = 40,10

Questão 8

Construa uma árvore de regressão para estimar o valor de Y na tupla [245,4,9700,4600,1835,?]. Use todos os dados disponíveis e faça com que a altura máxima da árvore seja igual a 3.

128

Questão 9

Utilizando um regressor (não paramétrico) k-NN, com k=5 vizinhos e baseado em distâncias euclidianas, obter Y para a tupla [245,4,9700,4600,1835,?].

128

Módulo 4 – Regras de Associação

Tabela em csv: http://goo.gl/GBVjhV

T	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	N	S	N	S	S	N	N
2	S	N	S	S	S	N	N
3	N	S	N	S	S	N	N
4	S	S	N	S	S	N	N
5	N	N	S	N	N	N	N
6	N	N	N	N	S	N	N
7	N	N	N	S	N	N	N
8	N	N	N	N	N	N	S
9	N	N	N	N	N	S	S
10	N	N	N	N	N	S	N

Questão 10

Considerando a tabela acima, contendo 10 transações, para as quais S (sim) e N (não) significam respectivamente a ocorrência ou não de um determinado item numa transação, obter o suporte e a confiança da regra “Se {manteiga, pão} então {café}”.

Suporte = 0,3 e Confiança = 0,75