PUC Minas

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Lista de Exercícios para a prova 1

Professora: *Marta Noronha*

Disciplina: *Aprendizado de Máquina II*

variância amostral: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

covariância: $cov(x,y) = \frac{\sum_{i=1}^{n}(x-\overline{x})*(y-\overline{y})}{n}$

Coeficiente de variação: $CV = \frac{s}{\overline{x}} * 100$

Família de Minkowski: $d_{min} = (\sum_{i=1}^{n}|x_i - y_i|^m)^{\frac{1}{m}}$ para m > 0

Distância média: $d_{media} = (\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|^2)^{\frac{1}{2}}$

Distância do cosseno: $d_{cos} = \frac{\sum_{i=1}^{n} x_i * y_i}{||x_2|| * ||y_2||}$ com $||x_2|| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$

Distância da corda: $d_{cord} = \left(2 - 2 * \frac{\sum_{i=1}^{n} x_i * y_i}{||x_2|| * ||y_2||}\right)^{\frac{1}{2}}$ com $||x_2|| = (\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$

Distância de Mahalanobis: $d_{mah} = ((x-y) * S^{-1} * (x-y)^T)^{\frac{1}{2}}$ onde S é a matriz de covariância

Correlação de Pearson: $P_{Coef}(X,Y) = \frac{\sum_{i=1}^{p}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{p}(Y_i - \overline{Y})^2}}$

Correlação de Spearman: $r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ com $d_i = R(X_i) - R(Y_i)$, sendo X e Y atributos e os objetos indexados por i.

Soma interna dos quadrados: $WSS = \sum_i \sum_{x \in C_i}(x - m_i)^2$
Entre soma dos quadrados: $BSS = \sum_i |C_i|(m - m_i)^2$

Questão 1: Defina:

a) Hard Clustering

b) Soft Clustering

c) Variável numérica intervalar

d) Variável numérica racional

e) Matriz de similaridade
f) Matriz de dissimilaridade
g) Matriz de correlação
h) Matriz de covariância
i) Simetria e assimetria de medidas (como Jaccard e *Simple Matching*)
j) Maldição da dimensionalidade
k) One hot encoding
j) Label encoding
l) Class encoding
m) Diferença entre classificação e clusterização em relação ao propósito e uso de rótulos de classe
n) Pureza de um cluster
o) Pureza dos clusters
p) Entropia (total e por cluster)
q) Precisão (total e por cluster)
r) Relevância (total e por cluster)
s) F-score (total e por cluster)
t) Medida de validação defeituosa

Questão 02: Quais são as alternativas possíveis, comumente usada por analistas, para realização de agrupamento quando as variáveis são categóricas?

Questão 03: Quais são as propriedades das medidas de distância?

Questão 04: Diferencie as medidas internas, externas e relativas para validação de clusters.

Questão 05: Defina coesão e separação de clusters. Cite dois métodos para descobrir o número ideal de clusters em um conjunto de dados.

Questão 06: Quando a correlação de Pearson deve ser usada?

Questão 07: Em uma empresa foram coletados dados relacionados a variáveis X e Y de uma amostra de funcionários (A a D), conforme mostrado na tabela abaixo.

	X	Y
A	4	1
B	5	4
C	3	5
D	7	3

Calcule:

- A média de cada variável;
- a variância de cada variável;
- o desvio padrão de cada variável;
- o coeficiente de variação de cada variável;
- a matriz de covariância para as variáveis;
- a matriz de correlação de Pearson para as variáveis;
- a matriz de correlação de Spearman para as variáveis;
- a distância de Manhattan entre cada objeto;

- a distância euclidiana entre cada objeto;
- a distância euclidiana média entre cada objeto;
- a distância do cosseno entre cada objeto;
- a distância da corda entre cada objeto;
- a distância de Mahalanobis entre cada objeto.
- Considere a existência de dois grupos no conjunto de dados mostrados na Tabela. Informe quais objetos estão em cada grupo baseado na distância entre eles.

Questão 08: Usando a Tabela abaixo que contém dados de objetos coletados considerando cinco características, V a Z, calcule:

	V	W	X	Y	Z
A	1	1	0	1	1
B	0	1	1	0	1
C	1	1	0	0	1
D	0	1	1	0	0

1. A similaridade de Jaccard entre cada ponto; e,
2. A distância de Jaccard entre cada ponto.

Questão 09: Considere a seguinte tabela de contingência (verdadeiras classes $\downarrow$ × clusters descobertos $\rightarrow$):

Classe \ Cluster	P_A	P_B	P_C
A	6	1	2
B	2	5	1
C	1	2	4
	S_{LA}	S_{LB}	S_{LC}

Total de objetos: $N = 22$.

1. Para cada cluster P_A, P_B e P_C, calcule:
   (a) Pureza;
   (b) Entropia (use $S_{LK} = -\sum_{i=1}^{k} p_i \log_2(p_i)$)
   (c) F-Score (defina precisão e sensibilidade por classe e reporte o F-Score).
2. Calcule as métricas de pureza, entropia (use $E = \sum_1^k \frac{n_k}{N} * S_{LK}$) e F-Score para o conjunto total de clusters.

Questão 10: Considere os seguintes conjuntos de pontos, representando três clusters, em $\mathbb{R}^1$:

$$C_1 = \{5, 7\}, \quad C_2 = \{1, 2, 3\}, \quad C_3 = \{10, 12\}.$$

Calcule:

1. O centro dos dados $m = \frac{1}{N} \sum_i \sum_{x \in C_i} x$, onde N é o total de pontos do conjunto de dados;
2. a coesão WSS;
3. e, a separação BSS.
4. Interprete se os clusters são mais coesos ou mais separados.

Questão 10: Seja o índice de Silhueta definido por:

$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, com $s_i \in [-1, 1]$, em que

$a_i = \frac{1}{|C(i)|-1} \sum_{\substack{j \in C(i) \\ j \neq i}} d(i,j)$ (Coesão intracluster – média das distâncias do ponto i a todos os outros pontos de seu próprio cluster $C(i)$), e

$b_i = \min_{C' \neq C(i)} \left\{ \frac{1}{|C'|} \sum_{j \in C'} d(i,j) \right\}$ (Separação intercluster – menor das médias de distâncias do ponto i a cada um dos demais clusters C'.).

- $s_i \approx 1$: ponto bem encaixado em seu cluster.
- $s_i \approx 0$: ponto na fronteira entre dois clusters.
- $s_i < 0$: ponto possivelmente classificado no cluster errado.

Considere os pontos em $\mathbb{R}^2$:

- cluster 1:
  - P1 = (1,3)
  - P2 = (2,3)
- cluster 2:
  - P3 = (5,6)
  - P4 = (6,5)
  - P5 = (5.5,5.5)
- Calcule o índice de silhueta por ponto.
- Calcule o índice de silhueta global.
- Calcule a distorção. ($Distorcao = \sum_1^k \left(\frac{1}{|c_k|} \sum_1^{|c_k|} dist(ponto, centroide)^2 \right)$)
- Calcule a inérica. ($Inercia = \sum_1^k \sum_1^{|c_k|} dist(ponto, centroide)^2$)

OBSERVAÇÃO: Além destes exercícios, revejam os exercícios que estão nos slides, os quais foram feitos durante as aulas.