Long Test 2

Data Analytics (CS 61061)

20 November 2021

Instructions:

- There are FOUR questions in this test. Attempt ALL questions.
- You are advised to write down all the intermediate calculations towards the calculation for your final answer. This will help you to get partial credits.
- Write your answer up to four decimal points.
- Maximum time allowed is 60 minutes. You can plan on the average maximum 15 minutes to each question. Full marks is 50.

Question 1

Consider the following set of records, where each record is defined by two ordinal attributes *size* ={S, M, L} and *quality* = {EX, A, B, C} such that $S < M < L$ and $EX > A > B > C$.

Object	Size	Quality
A	S	A
B	M	B
C	L	C
D	L	EX

(a) Compute the rank values to all attribute values.

(b) Write down the similarity matrix.

(*Important: Please write your answers in the form of matrices*).

[(2+2)+4 = 8]

Answer:

(a) Rank values to all attributes are

Object	Size	Quality
A	S(0.0)	A(0.66)
B	L(1.0)	EX(1.0)
C	L(1.0)	C(0.0)
D	M(0.5)	B(0.32)

(b) The similarity matrix

	A	B	C	D
A	0.0	1.056	1.0	1.599
B		0.0	1.0	0.599
C			0.0	0.599
D				0.0

Question 2

The following table shows the confusion matrix (CM) of a classification problem with six classes labelled as C_1, C_2, C_3, C_4, C_5 and C_6.

Class	C_1	C_2	C_3	C_4	C_5	C_6
C_1	50	15	7	1	2	1
C_2	10	52	6	2	1	2
C_3	5	6	16	3	4	2
C_4	1	2	0	21	3	1
C_5	2	1	2	0	47	4
C_6	1	3	2	1	2	29

(a) **Transform the CM of multiclass classification into a CM of size 2×2 considering the class C_2 as the positive (+) class and classes C_1, C_3, C_4, C_5 and C_6 combined together as negative (-) class. (Important: Please write your answers in the form of 2×2 matrix).**

(b) **Calculate the predictive accuracy to classify a record belongs to class C_2.**

(c) **Calculate the mean error rate of the classification to classify a record belongs to class C_2.**

(d) **Calculate the standard error rate of the classification to classify a record belongs to class C_2.**

(e) **Calculate the range of true accuracy. Assume τ_α with confidence level α = 95% is 1.96.**

[4+3+2+3+3=15]

Answer:

(a) The transformed CM of size 2×2 is:

	+	-
+	52	21
-	27	207

(b) The predictive accuracy is

$$\varepsilon = \frac{52 + 207}{52 + 21 + 27 + 207} = \frac{259}{307} = 0.8436$$

(c) The mean error rate is:

Error is = 0.1546
= error x number of test data
= 0.1546 x 307
= 48%

(d) Standard error rate (σ) = $\sqrt{\in (1-\in)/N}$ = $\sqrt{\frac{0.8436\times0.1546}{307}}$ = 0.0207

(e) True accuracy, $\tilde{\in} = \in \pm \tau_\alpha \times \sqrt{\in (1-\in)/N}$ = 0.8436±0.0207×1.96 = 0.8031 to 0.8842 with τ_α =1.96 and α = 0.95.

Question 3

Consider a training data set as shown in the table given below.

Person	Gender	Height	Class
1	F	1.6	S
2	M	2.0	M
3	F	1.9	M
4	F	1.88	M
5	F	1.7	S
6	M	1.85	M
7	F	1.6	S
8	M	1.7	S
9	M	2.2	T
10	M	2.1	T
11	F	1.8	M
12	M	1.95	M
13	F	1.9	M
14	F	1.8	M
15	F	1.75	S

(a) Calculate the entropy of the data set.

(b) Suppose, you select "Gender" as the splitting attribute. Calculate the following.

- **i. Information gain**
- **ii. Gini index**
- iii. **Gain ratio**

[3+3×3 = 12]

Answer:

(a) Entropy:

$$E = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$p_1 = \frac{5}{15} = 0.3333 \; p_2 = \frac{8}{15} = 0.5333 \qquad p_3 = \frac{2}{15} = 0.1333$$

Entropy =- $\sum_1^3 p_i log_2 p_i$ = 0.3333x0.4771+0.5333x0.2730+0.1333x0.8751 = 1.3996

(b) Information gain = $\alpha(Gender, D)$ = *E(D)* - $E_{Gender}(D)$

Here, E(D) = 1.3996 and $E_{Gender}(D) = 9/15 * \{-4/9\log(4/9) - 5/9\log(5/9)\} + 6/15\{-1/6\log(1/6) - 3/6\log(3/6) - 2/6\log(2/6)\} = 1.17829$

Information gain = $\alpha(Gender, D)$ = 1.3996 - 1.17829 = 0.2213

(c) Gini index = $\gamma(A, D) = G(D) - G_A(D)$

G(D) = 1 - (5/15)^2 - (8/15)^2 -(2/15)^2= 0.5867

and $G_{Gender}(D)$ = 9/15* (1-(4/9)^2-(5/9)^2) + 6/15*(1-(1/6)^2-(3/6)^2-(2/6)^2)

= 0.5407

Gini index = 0.5867-0.5407 = 0.046

(d) Gain ratio = $\beta(Gender, D) = \frac{\alpha(Gender,D)}{E^*_{Gender}(D)}$,

$$E^*_{Gender}(D) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} . \log \frac{|D_j|}{|D|}$$

E^*(gender) = -9/15log(9/15) - 6/15log(6/15)= 0.97

Gain Ratio = 0.2213/0.97 = 0.2281

Question 4

A data set with three attributes A1, A2 and A3 is given below.

	A_1	A_2	A_3
O1	1	3	4
O2	12	8	3
O3	2	4	1
O4	10	5	7
O5	6	6	5
O6	19	20	8
O7	2	4	6
O8	4	5	5
O9	5	5	6
O10	10	10	10
O11	2	1	2
O12	7	8	5
O13	3	1	4
O14	12	10	6
O15	6	12	10
O16	8	6	7

At the beginning of the k-Means algorithm with k = 3, the three cluster centroids O_1, O_2, and O_{16} are selected as shown int the table (in shaded row entries). Assume L_2 norm for the distance measurement.
An initial cluster is created.
A cluster can be represented as, for example, [6,1,5,12], when the cluster with centroid O6 and objects O1, O5, and O12 are in it. Note that the first object should be the cluster centroid and other objects in the cluster are in the ascending order of their numbers. In comma separated value (CSV) format, and without any blank space between them. Use the start and closing square brackets [and].

Answer the following:

(a) List the objects which are under the cluster whose cluster centroid is O6.
(b) List the objects which are under the cluster whose cluster centroid is O11.
(c) List the objects which are under the cluster whose cluster centroid is O16.
Hint: You are advised to obtain the contingency table storing d1, d2, and d3 the three distances from three cluster centroids and then decides the assignment.
(d) Calculate the SSE (intra-cluster similarity) of the cluster you have obtained.

[4 + 4 + 4 + 3 = 15]

Answer

The contingency table calculating the Euclidean distances of each object from the three cluster centroids and the assignment of objects are shown below:

Object	F_1	F_2	F_3	d1	d2	d3	Assignment
O1	1	3	4	25.0798	**3.0000**	8.1853	C2
O2	12	8	3	14.7648	12.2474	**6.0000**	C3
O3	2	4	1	24.3721	**3.1622**	8.7177	C2
O4	10	5	7	17.5214	10.2469	**2.2360**	C3
O5	6	6	5	19.3390	7.0710	**2.8284**	C3
O7	2	4	6	23.4307	**5.0000**	**6.4031**	C2
O8	4	5	5	21.4242	5.3851	**4.5825**	C3
O9	5	5	6	20.6155	6.4031	**3.3166**	C3
O10	10	10	10	13.6014	14.4568	**5.3851**	C3
O12	7	8	5	17.2336	9.1104	**3.000**	C3
O13	3	1	4	25.1594	**2.2360**	7.6811	C2
O14	12	10	6	12.3693	14.0356	**5.7445**	C3
O15	6	12	10	15.3948	14.1774	**7.0000**	C3

(a) The objects which are under the cluster whose cluster centroid C1 are: [6,]

(b) The objects which are under the cluster whose cluster centroid O_{11} are: [11,1,3,7,13]

(c) The objects which are under the cluster whose cluster centroid O_{16} are: [16,2,4,5,8,9,10,12,14,15]

(d) Calculation of SSE of the cluster

SSE of the cluster is = $\sum_{i=1}^{k} \sum_{x \in C_i} dist^2\,(m_i, x)$

m_i Corresponds to the centre (mean) of the cluster C_i and x is a data point in clusterC_i.

Mean of the centroids in three clusters are:

C1: [19.0000,20.0000,8.0000]

C2=[2.7143,3.2857,4.0000]

C3=[8.8750,8.1250,6.6250]

The SSE is calculated as :

SSE = 0 + 15.2998 + 28.1487

= 43.4485

The table below shows the calculations of intra-similarity measures:

Object	F_1	F_2	F_3	Intra-similarity measure			Assignment
O1	1	3	4		1.737944		C2
O2	12	8	3			4.78768	C3
O3	2	4	1		3.165509		C2
O4	10	5	7			3.342435	C3
O5	6	6	5			3.92707	C3
O7	2	4	6		2.240636		C2
O8	4	5	5		2.364709		C2
O9	5	5	6		3.487585		C2
O10	10	10	10			4.021427	C3
O12	7	8	5			2.484326	C3
O13	3	1	4		2.303486		C2
O14	12	10	6			3.69755	C3
O15	6	12	10			5.888283	C3