Major Test – II

Data Analytics

(Probabilistic Classifier, Support Vector Machine, Sensitivity Analysis, Decision Tree Induction, Similarity Measures, Clustering Techniques)

1. For a 5-class classification problem, the performance of a classifier is recorded in the form of a confusion matrix.

	C1	C2	C3	C4	C5
C1	78	6	5	8	11
C2	6	54	2	6	5
C3	4	2	44	3	1
C4	8	7	3	105	3
C5	11	5	1	3	19

Table 1

With reference to the data in Table 1, answer the following questions.

(a) What is the observed accuracy of the classifier?
(b) What is the mean error rate?
(c) What is the standard error rate?
(d) What is the upper and lower bounds of the true accuracy with the confidence level $\alpha = 0.95$? Given that the mean value of the lower and upper bounds of a confidence level $\alpha = 095$ is 1.96.

[4 + 2 + 4 + 5 = 15]

Solution:

(a) Observed accuracy of the classifier is calculated as below:

$p = total\ number\ correct\ classification$

$= 78 + 54 + 44 + 105 + 19 = 300$

$N =$ Total number of test data = 400

Observed accuracy $\in = \frac{p}{N} = \frac{300}{400} = 0.75 = 75\%$

(b) Calculation of mean error rate.

Mean error rate = Percentage error = 0.25 = 25%

(c) Calculation of standard error rate

Standard error rate (σ) = $\sqrt{\in (1-\in)/N} = \sqrt{\frac{0.75\times0.25}{400}} = 0.0216$

(d) Calculation of true accuracy

True accuracy = $\tilde{\in}$ $= \in \pm\tau_\alpha \times \sqrt{\in (1-\in)/N} = 0.75 \pm 0.0216 \times 1.96$

$= 0.75 \pm 0.0423$ with $\tau\alpha$=1.96 and α = 0.95.

Upper bound of true accuracy = 0.7923

Lower bound of true accuracy = 0.7077

2. D1 and D2 are two paragraphs given below.

D1: This is a document. Similarity of a document with other document can be measured with cosine similarity. For cosine similarity a document can be represented as a vector. Term frequencies are the components of the vector. For a term, we consider the unique words only.

D2: This is another document. For cosine similarity measurement, create a vector form of the document. For this, identify unique term in the document. The frequencies of each term gives a component of the vector.

In D1 and D2, consider only the root words as the terms. For example, the root word of measured, measurement, etc. is "measure". Ignore articles, prepositions, pronouns, conjunction, adverbs, etc. Ignore the plural form. The terms are underlined and to be considered for the counting of term frequencies in this problem. A document can be represented in the form of a vector, where a term represents a component and the frequency of occurrence of the term is its value. You should arrange the terms in the vector in lexicographic (dictionary) order. For example, a document vector looks like [5,6,1], where there are tree root words, say [aab,bcx,cyz] and their frequencies are 5, 6, and 1, respectively. Note the CSV (comma separated value) representation of the vector within [and].

(a) List all the root words in alphabetical (i.e., dictionary) order. Your answer should be the collection of root words and in CSV, no space(s), and within [and].
(b) Obtain the vector representation of D1. Let it be V1.
(c) Obtain the vector representation of D2. Let it be V2.
(d) Calculate the similarity of V1 and V2 using Euclidean distance (L_2 norm) measure.
(e) Calculate the cosine similarity of V1 and V2.
(f) Which distance calculation is computationally fast?
    - L2 norm calculation is fast.
    - Cosine similarity calculation is fast.

[2+3+3+3+3+1 = 15]

Solution:

The unique words in the two documents (arranged in dictionary order) are listed below:

(a) The list of root words in alphabetical sequence is
[another,cosine,component,consider,create,document,frequency,identify,measure,other,represent,similarity,term,unique,vector,word]

(b) The vector representation of Di is
V1: [0,2,1,1,0,4,1,0,1,1,1,3,2,1,2,1]

(c) The vector representation of D2 is
V2: [1,1,1,0,1,3,1,1,1,0,0,1,2,1,2,0]

(d) Similarity calculation using Euclidean distance (L2 norm):

$$L2(V1, V2) = \sqrt{1+1+0+1+1+1+0+1+0+1+1+4+0+0+0+1}$$

$$= \sqrt{13} = 3.6055$$

(e) Similarity calculation using Cosine similarity:

$$COS(D1, D2) = \frac{D1 \cdot D2}{[D1] \times [D2]}$$

$$D1.D2 = 0 + 2 + 1 + 0 + 0 + 12 + 1 + 0 + 1 + 0 + 0 + 3 + 4 + 1 + 4 + 0 = 29$$

$$[D1] = \sqrt{0+4+1+1+0+16+1+0+1+1+1+9+4+1+4+1}$$

$$= 6.7082$$

$$[D2] = \sqrt{1+1+1+0+1+9+1+1+1+0+0+1+4+1+4+0}$$

$$= 5.0990$$

$$COS(D1, D2) = \frac{29}{6.7082 \times 5.0990} = 0.8478$$

(f) Which distance calculation is computationally fast?

- L2 norm calculation is fast.
- Cosine similarity calculation is fast.

3. A scheme of a training data is stated as below.

Symptom	Duration	Treatment	class
S1, S2, S3	S (Short), M (Medium), L (Large)	A (Allopathy), H (Homeopath), U (Unani)	Cure (Y), Not Cure (N)

A contingency table is prepared with 400 records of patients, which is shown below (Table 2).

		Class		
		Y	N	Totals
Symptom	S1	30	10	40
	S2	14	12	26
	S3	24	14	38
Duratio	S	22	18	40
	M	36	26	62
	L	32	22	54
Treatment	A	30	24	54
	H	28	18	46
	U	24	16	40
	Totals	240	160	400

Table 2

A test data is given below:

S2	M	H	?

You have to classify the test data using the Naïve Bayes' classifier.

(a) What is the probability that the test data is in class Cure?
(b) What is the probability that the test data is in class Not Cure?
(c) In which class the test data will be?
- The test belongs to class Cure.
- The test data belongs to class Not Cure.

(d) What is the entropy of the input data?

[4+4+2+5=15]

Solution:

(a) Calculation of the probability that the test data is in class Cure (Y):

$P(Y|X) = 240/400 * 14/240 * 36/240 * 28/240$

$= 0.6 * 0.0583 * 0.15 * 0.1166 = 0.0006$

(b) Calculation of the probability that the test data is in class Not Cure (N):

$P(N|X) = N = 160/400 * 12/160 * 26/160 * 18/160$

$= 0.4 * 0.075 * 0.1625 * 0.1125 = 0.0005$

(c) The test data belongs to:
- The test belongs to class Cure.
- The test data belongs to class Not Cure.

(d) Calculation of the entropy

Here, $p_1 = \frac{240}{400} = 0.6$ and $p_2 \frac{160}{400} = 0.4$

Entropy = $-p_1 log p_1 - p_2 log p_2$ // $\log_2$ is considered

= -0.6x-0.7369 + -0.4x-1.3219

= 0.9709

4. A data set with three attributes A1, A2 and A3 is given below (Table 3).

	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

Table 3

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 in Table 3 (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]

Solution:

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 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

Summing up all the values, the SSE will be calculated as :

SSE = 0 + 15.2998 + 28.1487

= 43.4485