Test 4

Total marks: 100 Time: 2 hours

Part-I: ***20 X 2 = 40***

1. Which is NOT true about the *k*-Mean clustering algorithm?
   a) The algorithm is applicable to numerical data only.
   b) The algorithm is insensitive to noise/outlier in the training data set.
   c) The algorithm is not scalable.
   d) The algorithm is applicable to non-convex data distribution only.

Correct answer: b
Explanation: The k-Mean algorithm is sensitive to noise in the training data set.

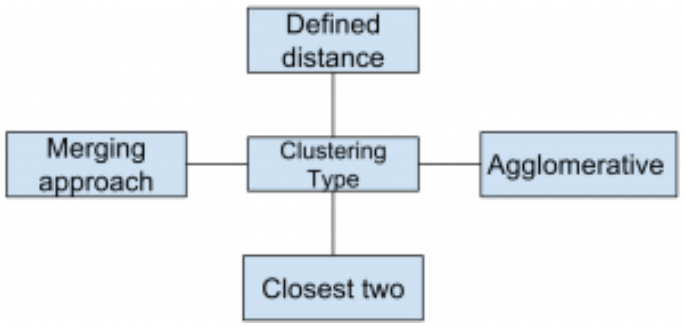
2. We are to measure the similarity between two text documents. Which of the following metric(s) is(are) best suitable for the purpose?
   (a) Jaccard coefficient.
   (b) Euclidean distance.
   (c) Cosine similarity.
   (d) Set difference.

Correct answer: c
Explanation: Cosine similarity gives better result to measure the similarity between two documents. Here, a document should be represented in terms of TDFs (term-document frequencies)

3. Which of the following statement(s) is (are) not correct?
   (a) *k*-Means clustering algorithm does not work on categorical data.
   (b) PAM algorithm requires the value of *k* (the number of clusters) a priori.
   (c) Agglomerative clustering techniques do not require the number of clusters to be specified.
   (d) DIANA is a partition based, whereas AGNES is a hierarchical-based clustering method.

Correct answer: d
Explanation: PAM is a modified version of k-Mean clustering algorithm and hence it requires the value ok k a priori as in k-Means algorithm. Both DIANA and AGNES are hierarchical clustering algorithm.

4. Which of the following clustering type has characteristic shown in the below figure?



   a) Partitional
   b) Hierarchical
   c) Naive Bayes
   d) None of the mentioned

Correct answer: b
Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

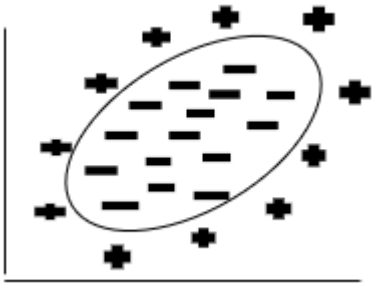
5. Point out the wrong statement.
   a) k-means clustering is a method of vector quantization
   b) k-means clustering aims to partition n observations into k clusters
   c) k-nearest neighbour is same as k-means
   d) none of the mentioned

Correct answer: c
Explanation: k-nearest neighbour has nothing to do with k-means.

6. Following measurements are known while testing a binary classifier with classes + and −:
   f_{++}, f_{+-}, f_{-+} and f_{--}, where f_{xy} denotes the number of instances that were with class "x" and classified as "y".
   Which of the following metrics represent "Recall"?
   a) $\frac{f_{++}}{f_{++}+f_{+-}}$
   b) $\frac{f_{++}}{f_{++}+f_{-+}}$
   c) $\frac{f_{-+}}{f_{-+}+f_{--}}$
   d) $\frac{f_{+-}}{f_{++}+f_{+-}}$

Correct answer: a
Explanation: Recall is the TPR (Tre positive rate).
$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{f_{++}}{f_{++}+f_{+-}}$

7. The distribution of data is given below?



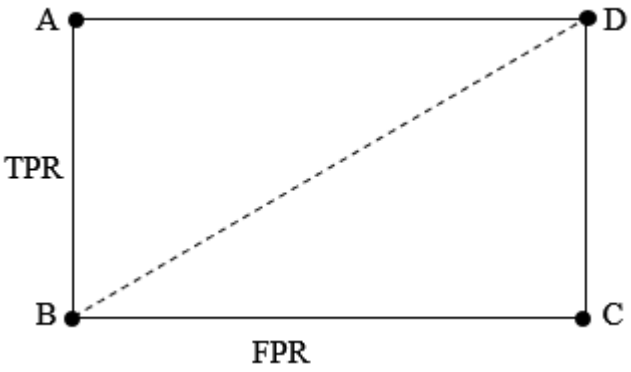
   The Kernel function, in this case, which should be chosen in building non-linear SVM is
   a) Laplacian Kernel.
   b) Polynomial Kernel.
   c) Gaussian RBF Kernel.
   d) Sigmoid Kernel.

Correct answer: c
Explanation: The Gaussian Radial Bias Function (RBF) kernel is the best kernel when the margin takes the shape of hypersurface.

8. Which of the following specifications is true for a worst classifier?
   a) TPR = 1, FPR = 0, Precision = 1, F_1 score = 1
   b) TPR = 0, FPR = 1, Precision = 0, Recall = 0
   c) TPR = 1, FPR = 1, Precision = 0, F_1 score = 1
   d) TPR = 1, FPR = 0, Precision = 0, Recall = 1

Correct answer: b
Explanation: For a worst classifier, TPR = 0 and FPR = 1 and hence both precision and recall are with 0 value.

9. In the ROC plot, among the 4 points A, B, C and D, which is (are) correct.



a) A = Ideal classifier
b) B = Worst classifier
c) C = Ultra-conservative classifier
d) D = Ultra-liberal classifier

Correct answer: a, d
Explanation: TPR=1 and FPR=0 means perfect or ideal classifier. TPR = 1, FPR = 1, the model predicts every instance to be a positive class, that is, it is an ultra-liberal classifier.

10. SVM computes the dot products of two vectors X_i. X_j. This implies that
a) SVM can be applied to the vectors with numerical attributes only.
b) SVM can be applied to the vectors with any type of attributes.
c) Computation time to build an SVM suffers from dimensionality problem as the cost of computation is influenced by dot products of vectors.
d) The cost of testing is not influenced by the number of support vectors.

Correct answer: a
Explanation: The dot product of vectors is computationally inexpensive and the testing only considers the dot product of each support vector with the test data.

11. Building an SVM is, in fact, solving an optimization problem. Which of the following statement is correct so far, the statement of optimal problem is concerned? All symbols bear their usual meaning:

a) Maximize $\frac{||w||^2}{2}$
b) Subject to $y_i(w.x_i + b) \geq 1$
c) Minimize $\frac{2}{||w||^2}$
d) Subject to $y_i(w.x_i + b) \leq 0$

Correct answer: d
Explanation: Options (a) to (c) are with wrong composition. Option (d) gives the dual form of Lagrangian.

12. Which of the following performance measure(s) consider(s) both Type-I and Type-II errors?
(a) Recall
(b) F_1 score
(c) Precision
(d) Specificity

Correct answer: b
Explanation: F_1 score is the harmonic mean of Recall and Precision, where Precision considers Type-I and Recall considers Type-II error.

13. Which of the following may be an accuracy of a random classifier?
(a) 0%
(b) 50%
(c) 100%
(d) All of the above.

Correct answer: d
Explanation: For a random classifier any accuracy estimation is possible. For example, if it is tested with all test data belong to only one class, then its accuracy may be with 50%, above 50%, or below 50%.

14. Suppose, H is a hyperplane to classify data. Which of the following statement(s) is (are) correct?
    a) Increasing the margin will increase the support vector count.
    b) Decreasing the margin will increase the support vector count.
    c) Increasing the margin will decrease the testing error.
    d) Decreasing the margin will decrease the testing error.

Correct answer: c
Explanation: If the distance between the margins of a hyperplane is increased, the. It will decrease the testing errors and vice-versa.

15. Which of the following is (are) unsupervised machine learning technique(s)?
    a) Agglomerative algorithm.
    b) Decision tree induction.
    c) Support vector machine.
    d) Auto regression analysis.

Correct answer: a
Explanation: Agglomerative algorithm is a kind of clustering technique and thus is an unsupervised technique.

16. Which of the following statement is NOT true?
    a) DIANA is based on splitting.
    b) AGNES is based on merging.
    c) BIRCH is an agglomerative clustering.
    d) PAM is a hierarchical clustering.

Correct answer: d
Explanation: PAM is a partitioning based algorithm.

17. "Bagging reduces the variance in the base classifier." – True or False?
    a) True
    b) False

Correct answer: a
Explanation: Bagging reduces the variance in the base classifier.

18. Optimal hyperplane is obtained in Support Vector Machines by solving a
    a) Linear Programming Problem
    b) Quadratic Programming Problem
    c) Semi-Definite Programming Problem
    d) Non-Linear Programming Problem

Correct answer: b

19. Support Vector Machines optimizes
    a) Empirical Risk
    b) Structural Risk
    c) Actual Risk
    d) None of the above

Correct answer: c

20. *"Random forest is less sensitive to the training data"* – True or False?
    a) True
    b) False

Correct answer: a

Part-II: Subjective type questions **6 X 10 = 60**

1. For a 5-class classification problem with 400 total data, 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

With reference to the data in the above table, 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?

Solution:

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

$p = t\text{otal number correct classification} = 78 + 54 + 44 + 105 + 19 = 300$

$N = \text{Total number of test data} = 400$

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

(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**

2. 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 O1, O2, and O16 are selected as shown int the table (in shaded row entries). Assume Euclidean 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) Obtained the contingency table.
b) List the objects which are under the cluster whose cluster centroid is O6.
c) List the objects which are under the cluster whose cluster centroid is O11.
d) List the objects which are under the cluster whose cluster centroid is O16.

Hint: *You are advised to obtain the contingency table storing d1, and d2 the two distances from two cluster centroids and then decides the assignment.*

Answer

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

Object	A_1	A_2	d1	d2	d3	Assignment
O1	1	3	24.75884	**2.236068**	7.615773	C2
O2	12	8	13.89244	12.20656	**4.472136**	C3
O3	2	4	23.34524	**3**	6.324555	C2
O4	10	5	17.49286	8.944272	**2.236068**	C3
O5	6	6	19.10497	6.403124	**2**	C3
O7	2	4	23.34524	**3**	6.324555	C2
O8	4	5	21.21320	4.472136	**4.123106**	C3
O9	5	5	20.51828	5	**3.162278**	C3
O10	10	10	13.45362	12.04159	**4.472136**	C3
O12	7	8	16.97056	8.602325	**2.236068**	C3
O13	3	1	24.83948	**1**	7.071068	C2
O14	12	10	12.20656	13.45362	**5.656854**	C3
O15	6	12	15.26434	11.70470	**6.324555**	C3

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

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

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

3. Consider the confusion matrix (see **Table Q6(b)**) showing the result on the performance of a classifier.

	Class A	**Class B**
Class A	80	25
Class B	15	70

Calculate the following:
a) Precision b) Recall c) Specificity d) F1 Score
You should clearly mention the formula for each metric for measuring the above.

Answer:

a) Precision $= \frac{TP}{TP+FP} = \frac{80}{80+25} = \mathbf{0.762}$

b) Recall $= \frac{TP}{TP+FN} = \frac{80}{80+15} = \mathbf{0.842}$

c) Specificity $= \frac{TN}{TN+FP} = \frac{70}{70+25} = \mathbf{0.737}$

d) F1 Score $= 2 \times \frac{Precision \times Recall}{Precision+Recall} = \frac{0.762 \times 0.842}{0.762+0.842} = \frac{0.642}{1.604} = \mathbf{0.4}$

4. A 2D training data set with attributes A and B is given as below. The vectors at #2 and #7 are the support vectors. Assume the Lagrange multipliers be λ_2 = 1.2 and λ_7 = 3.0

Vector#	A	B	Class
1	5.0	6.4	+
2	4.1	2.9	-
3	1.2	3.6	-
4	0.2	4.8	-
5	6.8	7.9	+
6	1.1	2.3	-
7	3.5	2.5	+
8	1.9	3.5	-
9	8.6	2.1	+
10	0.5	9.5	+

If W.X+b = 0 represents the equation of the maximum margin hyperplane, then obtain the following.

a) The weight matrix W
b) The value of b
c) The classification of a vector [4.5, 5.4]

Answer:

(a) Calculation of W

$$W1 = -4.1 \times 1.2 + 3.5 \times 3 = 5.58$$

$$W2 = -2.9 \times 1.2 + 2.5 \times 3 = 4.02$$

$$W = [5.58, 4.02]$$

(b) Calculation of b

$$b1 = -1 - 5.58 \times 4.1 - 4.02 \times 2.9 = -35.536$$

$$b2 = +1 - 3.5 \times 5.58 - 2.5 \times 4.02 = -28.58$$

$$b = (b1 + b2) \times 0.5 = -32.058$$

(c) $\delta(\mathbf{X})$ = The classification of a vector [4.5, 5.4] is

$$\delta(X) = +[4.5 \times 5.58 + 5.4 \times 4.02 - 32.058 > 0]$$

5. Consider two vectors, *X {3,2,0,5}* and *Y {1,0,2,0}*, calculate the cosine similarity and dissimilarity between X and Y. Calculate the angle (θ) between X and Y.

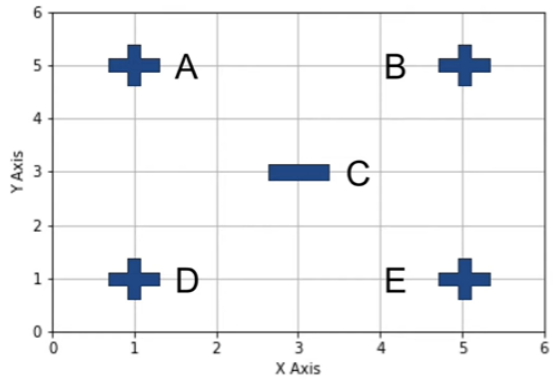
Answer:

Cosine similarity $= cos(x, y) = \frac{x.y}{\|x\| \times \|y\|} = \frac{3\times1+2\times0+0\times2+5\times0}{\sqrt{3^2+2^2+0^2+5^2} \times \sqrt{1^2+0^2+2^2+0^2}} = \frac{3}{6.16\times2.236} = \frac{3}{13.774} = \mathbf{0.22}$

Cosine dissimilarity $= 1 - cos(x, y) = 1 - 0.22 = \mathbf{0.78}$

$\theta = cos^{-1}cos(x, y) = cos^{-1}(0.22) = \mathbf{77.29°}$

6. Consider the 5 points as shown in the diagram given below.



Here, 4 points belong to **Class 1** and denoted by plus sign. Again, 1 point belongs to **Class 0** and denoted by minus sign. Consider the vertical decision boundary for $X < 2$, $X < 4$, $X < 6$, $X > 2$, $X > 4$, and $X > 6$, now perform **one (1) round** AdaBoost algorithm.

Answer:

Step-1: Initialize weights to all training points, as $w = \frac{1}{N}$

Here, $N = 5$

Points	Weights
W_a	$1/5$
W_b	$1/5$
W_c	$1/5$
W_d	$1/5$
W_e	$1/5$

Step-2: Calculate the error rate for each weak classifier, as $\in = \sum_{wrong} W_i$

C	Wrong	Error
$X < 2$	B and E	$2/5$
$X < 4$	B, C, and E	$3/5$
$X < 6$	C	$1/5$
$X > 2$	A, D, and C	$3/5$
$X > 4$	A and D	$2/5$
$X > 6$	A, B, D, and E	$4/5$

Step-3: Pick classifier with the lowest error rate.

C	Wrong	Error
$X < 2$	B and E	$^2/_5$
$X < 4$	B, C, and E	$^3/_5$
$\boldsymbol{X < 6}$	**C**	$\boldsymbol{^1/_5}$
$X > 2$	A, D, and C	$^3/_5$
$X > 4$	A and D	$^2/_5$
$X > 6$	A, B, D, and E	$^4/_5$

Step-4: Compute voting power for the classifier.
Here, $\in = \frac{1}{5}$ for $\boldsymbol{X < 6}$ classifier.

$$\alpha = \frac{1}{2} log_e \frac{1-\frac{1}{5}}{\frac{1}{5}} = \frac{1}{2}\log 4$$

Step-5: Append the classifier in the ensemble classifier and check if the classifier is good enough or not.

$$f(x_i) = \sum_{t=1}^{T} \alpha_t h_t(x_i)$$

$$\therefore h(x) = \frac{1}{2}\log 4 \times F(x < 6)$$

Step-6: Update weights of each point where the previous classifier went wrong.

$$W_{new} = \begin{cases} \frac{W_{old}}{2(1-\in)}, & if\ point\ classified\ correctly \\ \frac{W_{old}}{2\ \in}, & if\ points\ classified\ wrongly \end{cases}$$

Points	Weights
W_a	$^1/_8$
W_b	$^1/_8$
W_c	$^1/_2$
W_d	$^1/_8$
W_e	$^1/_8$