CS40003
Data Analytics

End-Autumn Semester Test
(Session 2017-2018)

Full Marks: 100 Time: 180 minutes

Instructions

- *There are two parts in the question paper. Answer to both parts.*
- *You should write your answers in the same order as they are in the question paper. Please answer to all sub parts of a question together.*

Part A

All questions in this part are of multiple choice type questions. For a question, there may be one or more option(s) is (are) correct.
For question with more than one correct options, credit will be given on pro-rata basis.

No credit will be given, if wrong options(s) is (are) chosen.

There is NO NEGATIVE marking.

Each correct answer to a question carries 2 marks only.

1. Prior probability is P(Y='A') and posterior probability is P(Y='A'|X='x') on a given table, where X and Y are two attributes there.
   Which of the following is true always?
   a) P(Y='A') ≠ P(Y='A' | X= 'x')
   b) P(Y='A') > P(Y='A' | X= 'x')
   c) P(Y='A') < P(Y='A' | X= 'x')
   d) P(Y='A') ≥ P(Y='A' | X= 'x')

2. Classification and clustering are two different tasks followed in data analytics. They are different in the sense that
   a) Clustering predicts the class of a record.
   b) Classification defines a class to which a record should belong.
   c) Clustering is based on supervised training.
   d) Classification is based on non-supervised training.
   e) None of the above.

3. Which of the following is a statistical-based classification method?
   a) Bayesian classifier
   b) Support vector machine
   c) k-Nearest neighbor classifier
   d) CART

4. Mark the incorrect statement(s) in the following.

Bayesian classifier is called Naïve, if it

a) Assumes all classes are mutually exclusive and exhaustive.
b) The attributes are independent, given a class.
c) It classifies provided that all attributes are categorical only.
d) It predicts class membership probabilities only.

5. ID3 algorithm follows
   a) A greedy strategy.
   b) A top-down decomposition approach.
   c) A divide-and-conquer strategy.
   d) A recursive approach.
   e) None of the above.
   f) All of the above.

6. Which of the following decision tree induction algorithms always results a binary decision tree?
   a) ID3
   b) C4.5
   c) CART
   d) All of the above

7. What is/are true about entropy say E of a table containing m >0 number of a labeled records belonging to k distinct classes?
   a) E is always a non-zero positive quantity.
   b) The minimum value of E is zero.
   c) The maximum value of E is $\log_2 m$
   d) The maximum value of E is $\log_2 k$

8. The Gini Index G (D) on a table D with k classes is used to measure the "impurity" of data set D. Which of the following statement(s) is (are) not correct about G (D)?
   a) G (D) is maximum when all records in D belongs to one class only.
   b) G (D) is minimum when all record in D belongs to one class only.
   c) The maximum value of G (D) is $1-\frac{1}{k}$ when the frequency of each class is $\frac{1}{k}, k \geq 2$.
   d) The minimum value of G (D) is $1-\frac{1}{m}$, if all the classes are evenly distributed among m tuples.

9. In Table A (9), the left column represents a classifier and right column represents the heuristic for decision tree building.

Table A(9)

Classifier Algorithm	Heuristic
C_1. C4.5	H_1. Gini Index of Diversity
C_2. CART	H_2. Information Gain
C_3. ID3	H_3. Gain Ratio

Which of the following mapping is appropriate?

a) C_1 - H_1 C_2 - H_2 C_3 – H_3
b) C_1 - H_3 C_2 – H_1 C_3 – H_2
c) C_1 – H_1 C_2 – H_3 C_3 – H_2
d) C_1 – H_2 C_2 – H_3 C_3 – H_1

10. If the three decision tree induction algorithms namely ID3, CART and C4.5 are applied to a data D, then
   a) All of them yield a unique decision tree.
   b) All of them possibly result different decision trees.
   c) The decision trees from ID3 and C4.5 are with lesser heights than the decision tree with CART.
   d) C4.5 always yields better decision tree than ID3.

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?
   a) Maximize $\frac{||w||^2}{2}$
      Subject to y_{en} $(wux_{ia} + b) \geq 1$

   b) Minimize $\frac{2}{||w||^2}$
      Subject to y_{en} $(wux_{ia} + b) \leq 0$

   c) L = $\frac{||w||^2}{2} + \sum_{i=1}^{n} \lambda_i (y_i (w.X_i + b) - 1)$

   d) L = $\sum_{i=1}^{n} \lambda_i - 1/2 \sum_{i,j} \lambda_i . \lambda_j \, y_i . y_j . x_i . x_j$

12. SVM computes the dot products of two vectors X_i. X_u. 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 influenced by the number of support vectors.

13. Given a data set in 2D space as shown in Fig. A (13).

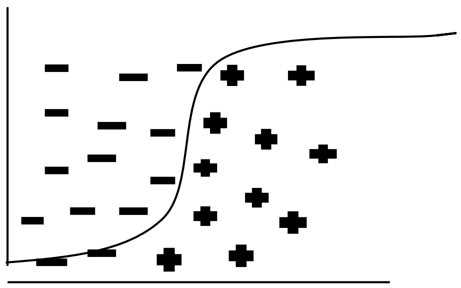


Fig. A (13)

Choose the most correct options, with reference to data in Fig. A (13).

a) Data is linearly separable and hence linear SVM should be used.
b) Data is linearly not separable and we can think for soft-margin SVM.
c) Data is linearly not separable and we can use linear SVM after transforming data into a higher dimensional space.
d) Sigmoid kernel can be applied to calculate gram matrix and then linear SVM.

14. The distribution of data is given below (see Fig. A (14)).

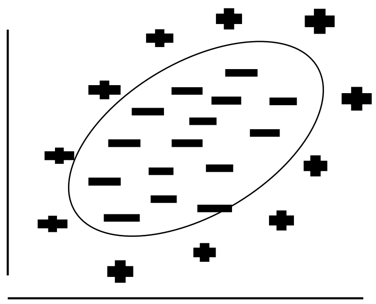


Fig. A(14)

The Kernel function, in this case, which should be chosen in building SVM is

a) Laplacian Kernel.
b) Polynomial Kernel.
c) Gaussian RBF Kernel.
d) Sigmoid Kernel.

15. 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 increase the error.
d) Decreasing the margin will increase the error.

16. Which of the following estimation strategy is called "Leave-one-out" validation strategy?

a) Hold-out method.
b) Random subsampling.
c) Cross validation.
d) Bootstrap validation.

17. A classifier model M when tested with training set T_1 and T_2 results the accuracy measure 95% and 75%, respectively. The size of T_1 and T_2 are 100 and 1000, respectively. Which of the following statements is/are not correct?

a) True accuracy is 75%
b) True accuracy is 85%
c) If the classifier is tested with bothT_1 and T_2 then the accuracy measure is closer to true accuracy.
d) Based on the above mentioned estimation none of the above estimations is acceptable.

18. Which of the following specifications is true for a perfect classifier?

a) TPP=1, FPR=0, precision=1, F_1 score=1
b) TPR=0, FPR=1, precision=0, F_1 score=1

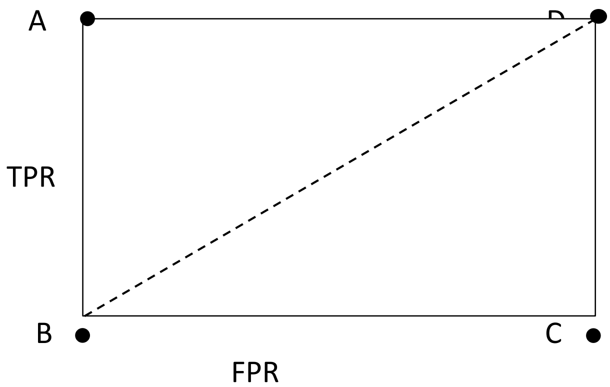
c) TPR=1, FPR=1, precision=0, F_1 score=1
d) TPR=0, FPR=0, precision=1, F_1 score=0

19. Which of the following metrics are for defining precision?

a) $\frac{f_{++}}{f_{++} + f_{--}}$
b) $\frac{f_{--}}{f_{++} + f_{+-}}$
c) $\frac{f_{++}}{f_{++} + f_{-+}}$
d) $\frac{f_{-+}}{f_{-+} + f_{--}}$

(All notations bear their usual meaning).

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



a) A = worst classifier
b) B = perfect classifier
c) C = ultra-conservative classifier
d) D = ultra-liberal classifier

Part B

This part includes 4 concept level or problem solving type questions.

You should write your answers in the same order as they are in the question papers.
Please answer to all sub parts of a question together.
Each part of the question carries 5 marks.

Problem 1.

(a) Consider the contingency table for a data of 20 observations (see Table B (1)).

Table B(1)

	Remedy		
Strategy		Yes	No
	Yes	4	9
	No	5	2

Calculate

i. P(Remedy = ‘yes’) Answer : 0.45
ii. P(Strategy = ‘No’) Answer : 0.35
iii. P(Strategy=’yes’| Remedy=’No’) Answer : 0.81
iv. P(Remedy= ‘yes’ | Strategy= ‘No’) Answer : 0.71

(b) For a class c_i, the posterior probability for attribute A_{ja} can be calculated using the following Gaussian normal distribution.

$$P(A_j = a_j|C_i) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{-\frac{(a_j - \mu_{ij})^2}{2\sigma ij}}$$

Give an idea how μ_{ij} and σ_{ij} can be calculated? Under what assumptions(s), the above calculations are possible?

Answer :

<u>Calculation of μ_{ij}</u>

μ_{ij} Can be calculated based on the mean of attribute values of A_j for the training records those belong to the class C_i.

<u>Calculation of σ_{ij}</u>

σ_{ij} Can be calculated from the variance of all the values in A_{ja} for the training records, which are labeled as class C_i.

<u>Assumption</u>

A_{ja} is numeric attribute.

(c) Consider the following data (See Table B(2))

Table B(2)

Age	Income	Married	Health	Class
Young	High	No	Fair	No
young	High	No	Good	No
Middle	High	No	Fair	Yes
Old	Medium	No	Fair	Yes
Old	Low	Yes	Fair	Yes
Old	Low	Yes	Good	No
Middle	Low	Yes	Good	Yes
Young	Medium	No	Fair	No
Young	Low	Yes	Fair	Yes
Old	Medium	Yes	Fair	Yes
Young	Medium	Yes	Good	Yes
Middle	Medium	No	Good	Yes
Middle	High	Yes	Fair	Yes
Old	Medium	No	Good	No

Using the data as shown in Table B(2), predict the record X = (Age= young, Income = Medium, Married = yes, Health = Fair”) belongs to a class?

Answer :

p_i = P(C_i) x $\prod_{j=1}^{n} P(A_j = a_j \mid C_i)$

Calculation of P(C_i)

P(Select = 'Yes') = 9/14 = 0.643

P(Select = 'No') = 5/14 = 0.357

Calculation of P(X | C_i) for each class C_i

P(Age='Young' | Select='Yes') = 2/9 = 0.222

P(Age='Young' | Select='No') = 3/5 = 0.6

P(Income='Medium' | Select='Yes') = 4/9 = 0.444

P(Income='Medium' | Select='No') = 2/5 = 0.4

P(Married='Yes' | Select='Yes') = 6/9 = 0.667

P(Married='Yes' | Select='No') = 1/5 = 0.2

P(Health='Fair' | Select='Yes') = 6/9 = 0.667

P(Health='Fair' | Select='No') = 2/5 = 0.4

Thus,

P(X | Select = 'Yes') = 0.222 x 0.444 x 0.667 x 0.667 = 0.044

P(X | Select = 'No') = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

P(C_i) x P(X | C_i) :

P(Select = 'Yes') x P(X | Select = 'Yes') = 0.643 x 0.044 = 0.028

P(Select = 'No') x P(X | Select = 'No') = 0.357 x 0.019 = 0.007

Problem 2.

(a) Define entropy of training set D.

Define information gain of a training set D while splitting on an attribute A. Assume that A has *m* distinct values in D.

Answer :

Entropy of training dataset D is given by

E(D) = -p_i $\log_2$ p_i

Where, $p_i = \frac{|C_i D|}{|D|}$, C_iD is the set of tuples of classC_i in D.

The expected information required to classify a tuple from D based on splitting A is also called weighted entropy and denoted as E_A(D) for all partitions of D with respect to A is given by

$E_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{|D|} \cdot E(D_j)$

Here, D_j denotes the j^{th} partition.

Information gain α(A,D) = E(D) – E_A(D)

(b) With reference to Table B(2), calculate the entropy of the data.

Answer :

$E = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

= (-0.643 * -0.643)-(0.35 * -1.51)
= 0.941

(c) With reference to Table B (2), obtain the Frequency table for the attribute Age. From the frequency table you have obtained, calculate the information gain of D while splitting on Age.

Answer :

Frequency Table for the attribute Age of D.

	Age = Young	Age = Middle	Age = Old	
Class Select = Yes	2	4	3	9
Class Select = No	3	0	2	5
	5	4	5	14

To calculate weighted entropy E_{Age} (D)

$V = f_{ij} \log f_{ij}$

Thus,

V = 2log2 + 4log4 + 3log3 +3log3 + 2log2
= 21.48

$S = S_i \log_2 S_i$ for all i=1,2,3.... M
= 5log5 + 4log4 + 5log5
= 31.2

Then,

E_{Age} (D) = (-V +S) / N here, N=14
= 0.69

Problem 3.

(a) Consider a training data with the attribute A, B and C and two classes + and – are given below. Here, λ_i denotes the Lagrangian multiplier.

Table B(3)

A	B	C	y_i	λ_i
1	3	5	-	0
2	4	6	-	0.5
8	9	7	+	0
6	5	4	+	0.3
2	2	4	-	0.6
3	1	2	-	0
10	11	10	+	0
7	8	9	+	0.2
9	8	7	+	0
10	10	10	+	0

How many support vectors are there? What are they?

Answer :

For the given table, support vectors are with non-zero Lagrange multiplier values. That is, number of Support Vector = 4.

(b) Obtain the support vector machine from the tabular data in Table B(3) .

Answer :

The SVM is

W X + b = 0

Here, $W=[W_1\ W_2\ W_3]$

$W_1 = \sum X_i\ (y_i\ .\ X_{ij})$

$= 0.5 \times -1 \times 2 + 0.3 \times 1 \times 6 + 0.6 \times -1 \times 2 + 0.2 \times 1 \times 7$

$= 1$

$W_2 = 0.5 \times -1 \times 4 + 0.3 \times 1 \times 5 + 0.6 \times -1 \times 2 + 0.2 \times 1 \times 8$

$= -0.1$

$W_3 = 0.5 \times -1 \times 6 + 0.3 \times 1 \times 4 + 0.6 \times -1 \times 4 + 0.2 \times 1 \times 9$

$= -2.4$

Thus, $W = [W_1\ W_2\ W_3] = [\ 1, -0.1, -2.4\ ]$

Next, We have to calculate b :

$b_1 = 1 - W\ X_1$ for Support Vector 1

$= 1 - (1.0 \times 2 - 0.1 \times 4 - 2.4 \times 6)$

$= 13.8$

$b_2 = 1 - W\ X_2$ for Support Vector 2

$= 1 - (1.0 \times 6 - 0.1 \times 5 - 2.4 \times 4)$

$= 5.1$

$b_3 = 1 - W\ X_3$ for Support Vector 3

$= 1 - (1.0 \times 2 - 0.1 \times 2 - 2.4 \times 4)$

$= 8.8$

$b_4 = 1 - W\ X_4$ for Support Vector 4

$= 1 - (1.0 \times 7 - 0.1 \times 8 - 2.4 \times 9)$

$= 16.4$

Averaging the above values, we get

$b = (b_1 + b_2 + b_3 + b_4) / 4$

$= (13.8 + 5.1 + 8.8 + 16.4) / 4$

$= 11.025$

Thus,

The support vector is

WX + b = 0

$W_1x_1+W_2x_2+W_3x_3+b = 0$

X1 -0.1x2 – 2.4x3 + 11.025 = 0

(c) How your support vector machine classify the following record?

X = [5, 6, 7]

Answer :

If δ(x) = W X + b

= 5 -0.1 x 6 – 2.4 x 7 + 11.025 = -1.375

= -ve sign, then it is in class -

Problem 4.

(a) With reference to k-fold cross validation, answer the following questions.

i. How many iteration(s) are there? What is the task in *i*-th iteration?

Answer :

There are k-iterations. In i-th iteration, D_i is used as test data whereas the other folds are used for training data.

ii. Can you claim that k-fold cross validation allows us to “trained by entire data as well as tested by entire data”?

Answer :

In an extreme case, if K=N (N is the size of the input dataset) we can say.

(b) A classifier is tested with a test set of size 520. Classifier predicts 480 test tuples correctly. With reference to this observation, answer the following questions.

i. What is observed accuracy?

Answer :

ϵ = 480 / 520 = .92307

ii. What is the standard error rate?

Answer :

Standard error rate σ = $\sqrt{\epsilon(1-\epsilon)/N)}$ here, N= 520

Then, σ =0.011

iii. What is the true accuracy? Assume that at confidence level α=0.99, the mean bound τ_a =2.58.

Answer :

If ϵ denotes the observed accuracy , then true accuracy

$\bar{\epsilon}$ = ϵ ± $T_\alpha\sqrt{\epsilon(1-\epsilon)/N)}$

=0.960 & 0.904

(c) The ROC plots for three classifier models are given below. (see Fig. B(4)).

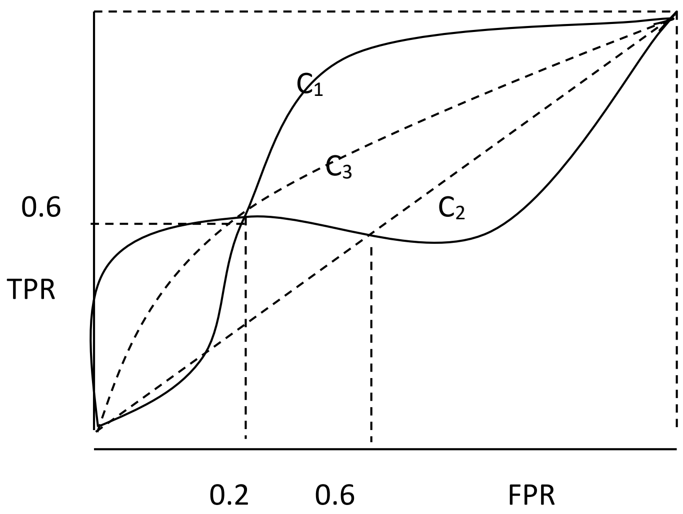


Fig. B(4)

Answer the following questions.

i. Which classifier is not acceptable? Why?

Answer :

Classifier C_2 is not acceptable as when FPR > 0.6, it work worse than the random classifier.

ii. How you compare C_1 and C_3?

Answer :

Area under C_3 is less than C_1. Hence, C_1 is better than C_3.

iii. How you quantitatively measure the performance of three classifiers at FPR=0.2

Answer :

$$\sigma = \sqrt{fpr^2 + (1 - tpr)^2}$$

For all three classifiers t_{pr} = 0.6 at fpr = 0.2

Hence, All C1, C2 and C3 have same performance and is

$$\sigma = \sqrt{(0.2)^2 + (0.4)^2}$$

$$= 0.4472$$

--- * ---