Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2023-2024

End-Semester Test
(EC-3 Regular)

Course No.	: DSE ZG565 / AIML ZG565
Course Title	: MACHINE LEARNING
Nature of Exam	: Open Book
Weightage	: 40%
Duration	:
Date of Exam	:

No. of Pages = 3

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1:

In a specific population, the probability of a person experiencing a particular symptom given they've had a Meningitis is 80%, whereas the chances of experiencing the same symptom without Meningitis is 10%. If the prevalence of Meningitis in the population is 5%, what is the probability that a person displaying the symptom indeed has Meningitis? [4]

Solution:
Probability of having the symptom given the disease (P(Symptom|Disease)) = 80%
Probability of having the symptom without the disease (P(Symptom|¬Disease)) = 10%
Prevalence of the disease (P(Disease)) = 5%
Goal:
We need to find the probability that a person who exhibits the symptom actually has the disease (P(Disease|Symptom)).

Step 1: Calculating Prior Probabilities:
P(Disease) = 5% P(¬Disease) = 100% - P(Disease) = 100% - 5% = 95% [1 mark]
Step 2: Calculating Likelihoods:
P(Symptom|Disease) = 80% P(Symptom|¬Disease) = 10% [1 mark]

Step 3: Applying Bayes' Theorem:

Using Bayes' theorem:

$$P(Disease|Symptom) = \frac{P(Symptom|Disease) \times P(Disease)}{P(Symptom)}$$

We need to calculate $P(Symptom)$, the probability that any given person exhibits the symptom. This can be calculated using the law of total probability:

$$P(Symptom) = P(Symptom|Disease) \times P(Disease) + P(Symptom|\neg Disease) \times P(\neg Disease)$$

Substituting the given values:

$$P(Symptom) = (0.80 \times 0.05) + (0.10 \times 0.95)$$
$$P(Symptom) = 0.04 + 0.095$$
$$P(Symptom) = 0.135$$

Now, we can calculate $P(Disease|Symptom)$:

$$P(Disease|Symptom) = \frac{0.80 \times 0.05}{0.135}$$
$$P(Disease|Symptom) \approx \frac{0.04}{0.135}$$
$$P(Disease|Symptom) \approx 0.296$$

[1mark]

Step 4: Interpretation:

The probability that a person who exhibits the symptom actually has the disease is approximately 29.6%.

[1 mark]

Question 2

a) How can Decision Tree models aid in enhancing the Interpretability of Machine Learning Systems, and What are their limitations? [1+1]

Solution:

Decision tree models inherently offer a high degree of interpretability because their decision-making process can be visualized and understood as a series of straightforward rules.

Aiding Interpretability: [1 mark]

Visual Representation: The tree structure can be easily visualized, showing exactly how decisions are made at each node based on specific feature values.

Simple Rules: Decision trees make decisions based on simple, if-then rules, which can be understood even by non-experts.

Limitations: [1 mark]

Overfitting: Decision trees can easily overfit the training data, especially with complex datasets, leading to a model that is too specific and not generalizable.

Complexity with Large Trees: While small trees are interpretable, very large trees can become unwieldy and difficult to interpret.

Biased with Imbalanced Data: Trees can become biased towards the majority class in imbalanced datasets, affecting their decisions.

b) You are a ML engineer working in a large organization. Your organization is very popular and hackers always want to target your organization to tarnish its image in the market. So, the employees in your organization receive a lot of spam emails. Your boss asks you to build a spam filter for distinguishing

between genuine e-mails and unwanted spam e-mails. Assuming spam to be the positive class, answer the following questions [5]

1) Describe Precision and Recall with respect to the given problem (1 mark)
2) Explain what happens if you optimize each of the above parameters (2 marks)
3) Which would be more important to optimize and why? (2 marks)

Answer

1) **0.5 mark for each only if Precision and Recall are explained as per the given problem, otherwise 0 marks should be given**

2) If we optimize recall, we may be able to capture more spam e-mails, but in the process, we may also increase the number of genuine mails being predicted as spam. [1 mark only if explained wrt the problem]

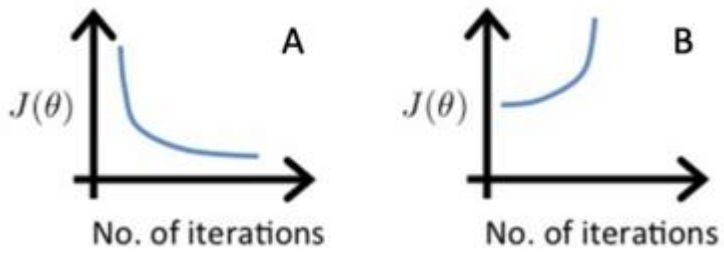
3) On the other hand, if we optimize precision, we may not be able to capture as many spam e-mails as in the previous approach, but the percentage of e-mails being classified as spam actually being spam will be high. [1 mark only if explained wrt the problem]

Increasing precision involves decreasing FP and increasing recall means decreasing FN.

4) Precision is more important to optimize in the given scenario.[0.5 marks]
It is okay if any spam email goes to INBOX, but it is not okay if any non-spam email goes to SPAM folder. We may miss an important email. [1.5 marks for justification]

Question 3

a) Suppose you tried logistic regression with 2 distinct values of learning rate and plotted the learning curve for each case where J(θ) represents the cost function. For which of the following cases (A or B), the learning rate possibly too large? Justify your answer. [2]

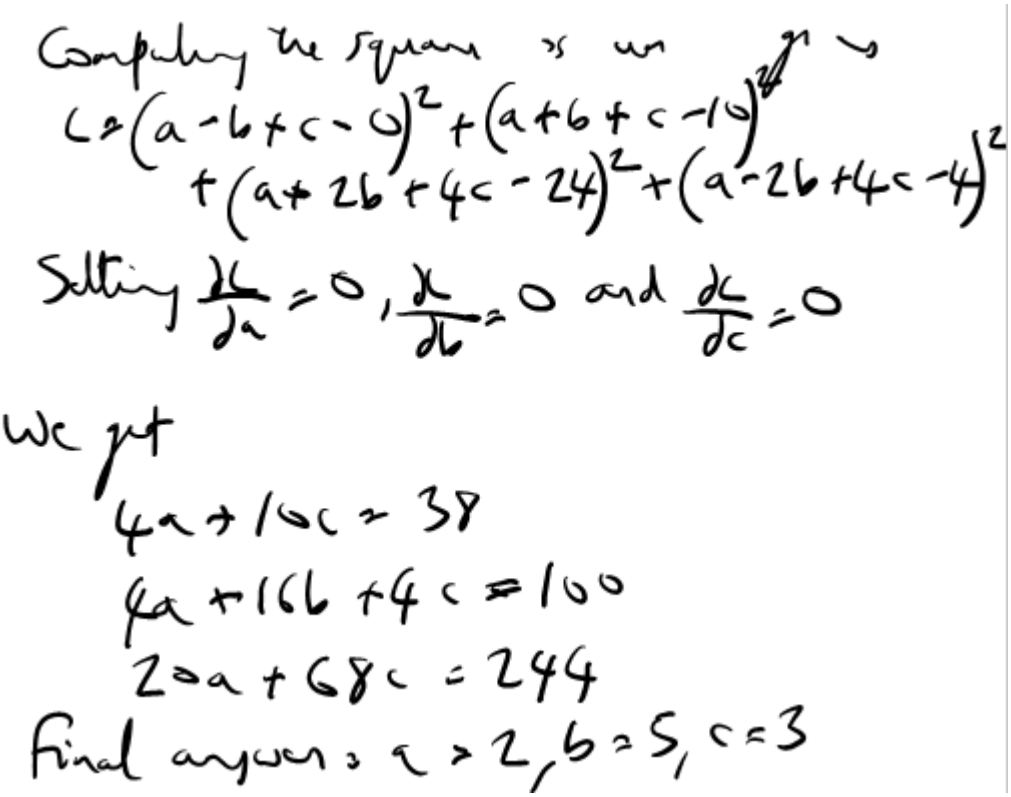


Solution

B is the right answer, instead of converging, it is diverging [0.5 marks for the correct answer 1.5 marks for the explanation]

b) What is the best curve of the form $y = a + bx + cx^2$ in terms of minimizing square error that fits the following training dataset (x, y): (-1,0), (1,10), (2,24), (-2,4)? [3]

Solution



Question 4

Use case: Committee of experts convene every year to nominate literary works to become eligible for the awarded of highest category by assessing the works on multiple parameters. Below is one subset of such features. Categorizing the works provides a transparent & streamlined way of nomination process. Quantified values of attributes are discretized in below data. Build a machine learning model to classify if an original literary work of writers has "High" or "Medium" or "Low" chances of nomination by the committee.

[3.5 + 1.5 + 1 = 6 Marks]

Readership Base	Writer's Reputation spread in other countries	Distinctive in Style	Chances of Nomination of Literary Works
Low	Low	High	High
High	High	Low	Medium
Low	Low	High	Low
High	Low	Low	Medium
Low	High	High	Medium
High	High	High	High
Low	Low	Low	Medium
Low	Low	Low	Medium

Test Instance: <Readership Base = High, Writer's Reputation = High, Distinctive in Style = High>

a) Use 6-NN, 3-NN & 1NN separately to classify the above test instance. Assume all the features are categorical in nature and use only the following measure of similarity for your calculation. Round-off all the proximity values to two decimal places.

$$Similarity\ (data1\ , data2) = \left(\frac{Number.of.matching.categorical.attributes}{Total.number.of.categorical.attributes}\right)$$

b) Which if the individual k-NN is more robust to outliers? Justify your answer with plagiarism free explanation in no more than 40 words.

c) Create an ensemble using all the above model with majority voting to predict the class for the given test instance

---

Answer Key:

a)

Similarity in the same order of input	Distance between query & instance
0.33	0.67
0.67	0.33
0.33	0.67
0.33	0.67
0.67	0.33
1	0
0	1
0	1
6NN-->	Medium
3NN-->	Medium
1NN-->	High

b) 6-NN is robust to outliers. Lesser the K more likely the outlier's effect is more provided they are the neighbors

c) From part a) majority voting = "Medium"

Marking Scheme:

a)

2 mark: Distance calculation between test and all other instances. Order the results of the distance .

1.5 mark: Results of each three model k-NN ie.., 0.5 marks for each model

b)

0.5 mark: Answer of correct K-NN

1 mark: plagiarism – free justification.

c) 1 mark: Majority voting among the result of part a)

Question 4

Use case: Committee of experts convene every year to nominate literary works to become eligible for the awarded of highest category by assessing the works on multiple parameters. Below is one subset of such features. Build a machine learning model to cluster the original literary work of writers to help committee with the nomination process.

[4.5 + 1.5 + 2 = 8 Marks]

Literary Work	Readership Base	Number. of. Translations
ID:1	70	5
ID:2	30	3
ID:3	50	7

Note: Wherever applicable round all the calculation to 4 decimal places. Use the given features as is without scaling.

a) Assume that the above literary works data follows Gaussian Distribution. Apply Gaussian Mixture Model based soft clustering only for one iteration to cluster the points among two clusters. Use initial values of (Mean of features with same order, Standard Deviation of features with same order from above tabulated attributes, weights) for cluster 1 = ((20,1), (10,1), 0.5) and cluster 2 = ((50, 7), (10,1), 0.5). Show all the step by step computations clearly. Show the final responsibility matric at the end of the first iteration.

b) Find only the new mixture weights and cluster means at the end of first iteration. **Note**: No need to calculate other prototypes standard deviations

c) Give a plagiarism free example & sample feature design from the domain of healthcare where soft clustering is best suited than hard clustering. Justify your choice in no more than 40 words

---

Answer Key:

New Mixture Weights = (0.33, 0.67)			Cluster - 1	Cluster - 2	Responsibility Matrix	
Literary Work	Readership Base	Number. Of. Translations	Weight * N(x|mean,SD)	Weight * N(x|mean,SD)	P(Zi|cluster 1)	P(Zi|cluster 2)
ID:1	70	5	0	0.0001	0	1
ID:2	30	3	0.0007	0	1	0
ID:3	50	7	0	0.008	0	1

Marking Scheme:

a)

1.5 mark: For Cluster 1 Weight * N(x|mean,SD) ie., 0.5 split up mark for each instance

1.5 mark: For Cluster 2 Weight * N(x|mean,SD) ie., 0.5 split up mark for each instance

0.5 mark: Responsibility matrix probability value CALCULATION for cluster 1

0.5 mark: Responsibility matrix probability value CALCULATION for cluster 2

b)

0.5 mark : New mixture weight calculation

0.5 mark : Mean of cluster 1

0.5 mark : Mean of cluster 2

c)

1 mark : Any copy – free Sample feature design of dataset for clustering in healthcare

1 mark : Justification of soft clustering requirement w.r.t. to student's data set

---

Question 5

Assume that in the AdaBoost algorithm, we are initially given a dataset of 6 points with classification (x,y = $class$):(1, +), (2, +), (3, —), (4, —), (5, +), (6, +) . The classifier is a decision tree stump choosing a constant c such that all points with $x > c$ are labeled one class and all points with $x \leq c$ are labeled the other class. Assume that the first classifier (i.e at the end of the first iteration) misclassifies only the points at $x = 1$ and $x = 2$. What are possible values of c for the first classifier? Find the importance of the first classifier, and values of instance weights at the end of the first iteration. [5]

If only the points $x=1$ and $x=2$ are misclassified, c must lie between 4 and 5, i.e $4 \le c < 5$

Classifier: $X \le c \rightarrow -$, $X > c \rightarrow +$

Error $\varepsilon = \frac{1}{n} \sum_j w_j \left(\delta(C(x_j) \neq y_j)\right) = \frac{1}{6} \times \frac{1}{6} \times 2$

$= \frac{1}{18}$

Importance $= \alpha = \frac{1}{2} \ln \frac{1-\varepsilon}{\varepsilon} = \frac{1}{2} \ln \frac{\frac{17}{18}}{\frac{1}{18}}$

Importance $= \frac{1}{2} \ln 17 = \alpha$

New weights: $w_1' = \frac{w_1 e^{\alpha}}{z}$, $w_2' = \frac{w_2 e^{\alpha}}{z}$, $w_3' = \frac{w_3 e^{-\alpha}}{z}$

$w_4' = \frac{w_4 e^{-\alpha}}{z}$, $w_5' = \frac{w_5 e^{-\alpha}}{z}$, $w_6' = \frac{w_6 e^{-\alpha}}{z}$

$z = w_1' + w_2' + w_3' + w_4' + w_5' + w_6'$

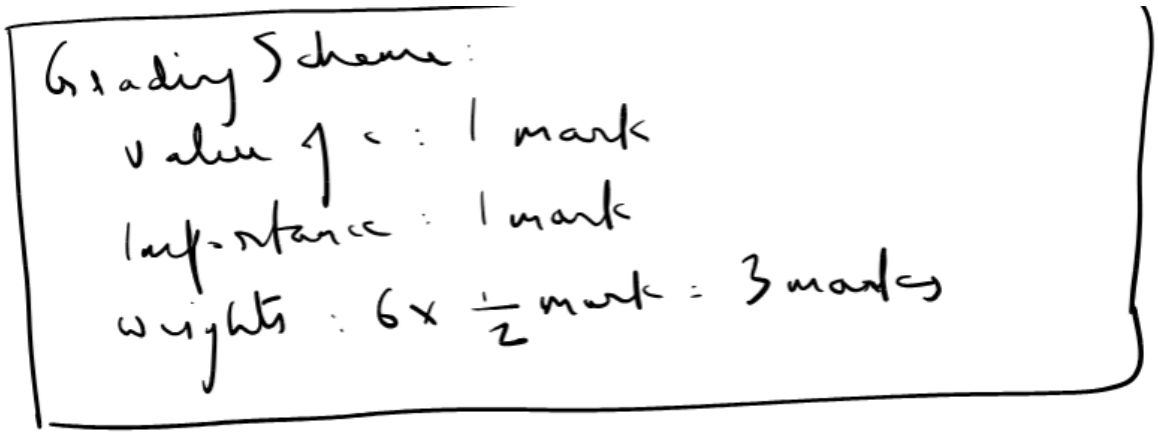
$e^{\alpha} = \sqrt{17}$, $e^{-\alpha} = \frac{1}{\sqrt{17}}$

$z = \frac{1}{6}\sqrt{17} + \frac{1}{6}\sqrt{17} + \frac{4}{6}\frac{1}{\sqrt{17}} = \frac{2}{6}\sqrt{17} + \frac{4}{6}\frac{1}{\sqrt{17}}$

$z = 1.53$

$w_1' = \frac{1}{6} \times \frac{\sqrt{17}}{1.53} = 0.449$ $\quad w_2' = 0.449$

$w_3' = w_4' = w_5' = w_6' = \frac{1}{6 \times 4.123} \times \frac{1}{z} = 0.026$



Question 6

- Suppose you have a dataset that has 10 features and 10,000 training instances. You have applied logistic regression with gradient descent on this dataset and trained a model. Unfortunately, this ML model exhibits poor performance on both the training data and test data. To address this issue, your team members have proposed several solutions, as mentioned below. Suggest which of the following looks promising in the given scenario and provide reasons for your choice? [5]

1. Use SVM with linear kernel without adding any new feature
2. Increase the regularization parameter λ, in logistic regression
3. Use SVM with RBF kernel
4. Transform the dataset using polynomial transformation and then use logistic regression

True/False	Answer	Explanation
False	Use an SVM with a linear kernel, without introducing new features.	An SVM with only the linear kernel is comparable to logistic regression, so it will likely underfit the data as well.
True	Use an SVM with a Gaussian Kernel	By using a Gaussian kernel, your model will have greater complexity and you can avoid underfitting the data.
False	Increase the regularization parameter λ	You are already underfitting the data and increasing the regularization parameter only makes underfitting stronger.
True	Create / add new polynomial features	When you add more features, you increase the variance of your model, reducing your chances of underfitting.

https://github.com/mGalarnyk/datasciencecoursera/blob/master/Stanford_Machine_Learning/Week7/SupportVectorMachinesQuiz.md

[0.25 marks for True/False, 1 mark for the justification]