QP Pattern:

- **Instance based learning: 4-8 marks**
- **SVM: 4 - 8 marks**
- **Bayesian Learning: 4 - 8 marks**
- **Ensemble Learning: 4 - 8 marks**
- **Unsupervised Learning: 4 - 8 marks**
- **Model Evaluation and Comparison: 4 - 8 marks**
- **Pre-midterm topics: 5-10 marks**

Course No.	: DSECLZG565/ AIMLCLZ565
Course Title	: Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	:

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.

Q.1 The results of the election are to be predicted for candidates based on dataset D. There are three different hypotheses h1, h2 and h3 are used to predict the result of candidates winning or losing an election. The probability of h1 given dataset D is 0.5, the probability of h2 given dataset D is 0.3 and the probability of h3 given dataset D is 0.2. Given a new candidate, h1 predicts that a candidate will win the election whereas h2 and h3 predict that candidate will lose the election. What's the most probable classification of a new candidate? [4 Marks]

Solution:

+ = win, - = lose

$P(h_1|D) = .5$, $P(-|h_1) = 0$, $P(+|h_1) = 1$

$P(h_2|D) = .3$, $P(-|h_2) = 1$, $P(+|h_2) = 0$

$P(h_3|D) = .2$, $P(-|h_3) = 1$, $P(+|h_3) = 0$

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D)$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D)$$

P(+|D) = 1*0.5+0*0.3+0*0.2=0.5 **[1.5M]**

P(-|D) = 0*0.5+1*0.3+1*0.2=0.5 **[1.5M]**

Winning and losing both are equiprobable [1M]

Q.2 Consider the problem

$$Minimize\ f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 3)^2$$
$$Subject\ to\ (x_1 - 1)^2 + x_2^2 \leq 5;\ x_1 \geq 0, x_2 \geq 0$$

a) Define the Lagrangian function for the above problem (1 Mark).

b) Establish necessary KKT conditions for the above problem. (2 Marks)

c) Are the Karush-Kuhn-Tucker conditions sufficient for a solution? That is, does a point satisfying the KKT conditions have to be a solution to the problem? Answer yes or no, then explain why. (2 Marks)

Solution

a) Defining the Lagrangian function

The problem can be restated as:

$$Minimize\ f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 3)^2$$
$$Subject\ to\ g(x_1, x_2): -(x_1 - 1)^2 - x_2^2 + 5 \geq 0 \longrightarrow (\lambda)$$

Define Lagrangian function

$$\mathcal{L}(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda\, g(x_1, x_2)$$
$$\mathcal{L}(x_1, x_2, \lambda) = (x_1 - 2)^2 + (x_2 - 3)^2 - \lambda\,(-(x_1 - 1)^2 - x_2^2 + 5)$$

b) Necessary KKT Conditions

$\nabla\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{\partial\mathcal{L}}{\partial\mathbf{x}} = 0$	⟹	$\frac{\partial\mathcal{L}}{\partial x_1} = 2(x_1 - 2) + 2\lambda = 0$ $\frac{\partial\mathcal{L}}{\partial x_2} = 2(x_2 - 3) + 2\lambda = 0$ $\frac{\partial\mathcal{L}}{\partial\lambda} = (x_1 - 1)^2 + x_2^2 - 5 = 0$
$\lambda_i^*(g_i(x^*) - b_i) = 0$	⟹	$\lambda((x_1 - 1)^2 + x_2^2 - 5) = 0$
$g_i(x^*) - b_i \geq 0\ \forall i$	⟹	$-(x_1 - 1)^2 - x_2^2 + 5 \geq 0$
$\lambda_i \geq 0$	⟹	$\lambda \geq 0$

c) Are the Karush-Kuhn-Tucker conditions sufficient for a solution? That is, does a point satisfying the KKT conditions have to be a solution to the problem? Answer yes or no, then explain why.

Yes. Since the objective function and constraints are convex, the KKT conditions are sufficient for optimality.

Q.3 Suppose we have the following one-dimensional data at -4.0, -.3.0, -2.0, -1.0, 0.0, 1.0, 2.0, 3.0, 4.0. Use the EM algorithm to find a Gaussian mixture model consisting of exactly one Gaussian that fits the data. Assume that the initial mean of the Gaussian is 10.0 and the initial variance is 1.0. [7 Marks]

Solution

First we note that $\pi_1 = 1$ since there is only one Gaussian in the mixture model.
Computing the posterior probabilities $P(z_{n1} = 1/x_n) = \gamma(z_{n1})$ we see that the posterior probabilities are all equal to 1 since both the numerator and denominator are equal to $\pi_1 N(x_n/\mu_1, \Sigma_1)$. **[1.5 M]**

Also $N_1 = \sum\gamma(z_{n1}) = N$, the number of data points. [**0.5M]**
This completes the E-step.

In the M-step, we see that

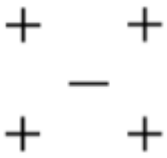
$\mu_1^{new} = \frac{1}{N_1}\sum_{n=1}^{N}\gamma(z_{n1})x_n = \frac{\sum_{n=1}^{n=N} x_n}{N} = \frac{-4.0 + -3.0 + -2.0 + -1.0 + 0.0 + 1.0 + 2.0 + 3.0 + 4.0}{9} = 0.0$

[2M]

and $\Sigma_k^{new} = \frac{1}{N_1}\sum_{n=1}^{n=N}(x_n - \mu_1^{new})(x_n - \mu_1^{new})^T$.
Here the x_n and μ_1^{new} are 1×1 matrices and the expression for Σ_k^{new} simplifies to $\frac{\sum_{n=1}^{N} x_n^2}{N}$ which is $\frac{2*(4.0^2+3.0^2+2.0^2+1.0^2)}{9}$=6.66. [2M]

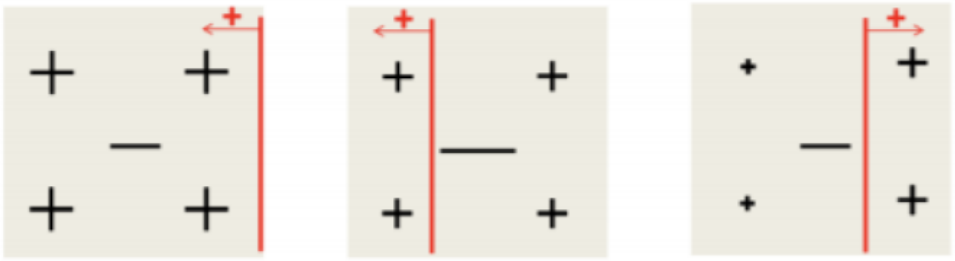
In the next iteration the E-step computes the posterior probabilities to be 1 and the M-step computes the same mean and covariance matrix as above, so the algorithm converges.[**1M]**

Q.4 Consider training a boosting classifier using decision stumps on the following data set. Circle the examples which will have their weights increased at the end of each iteration. Run the iteration till zero training error is achieved. [4 Marks]



Solution:

No of iteration - 3



Q5.a) In a clinical trial, height and weight of patients is recorded as shown below in the table. For incoming patient with weight = 58 Kg and Height = 180 cm, classify if patient is Under-weight or Normal using KNN algorithm assuming K = 3? **[3 marks]**

Weight (in Kg)	Height (in cm)	Class
61	190	Under-weight
62	182	Normal
57	185	Under-weight
51	167	Under-weight
69	176	Normal
56	174	Under-weight
60	173	Normal
55	172	Normal
65	172	Normal

Class – "Under-weight"
[2 marks for the calculation and correctly identifying 3 nearest neighbours]
[1 mark for the correct classification]

b) Suggest some of the ways to choose the optimal value of k in KNN. [2 marks]

- Using elbow curve: Derive a plot between error rate and K denoting values in a defined range. Then choose the K value as having a minimum error rate.
- Using methods like GridSearchCV and RandomisedSearchCV in python
- K should be the square root of n (number of data points in the training dataset).
  [2 marks for atleast 2 methods]

Q6: The following sample data is captured on a busy traffic signal for a certain period, after the parliament passed a bill on strict traffic regulations. Consider "Crash Severity" as the class of interest.

Weather Condition	**Driver Condition**	**Rule Violation**	**Seat Belt?**	**Crash Severity**
Good	Alcohol	Speed	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Red Signal	Yes	Minor
Good	Sober	Speed	Yes	Major
Bad	Sober	Other Rules	No	Major
Good	Alcohol	Red Signal	Yes	Minor
Bad	Alcohol	None	Yes	Major
Good	Sober	Other Rules	Yes	Major
Good	Alcohol	None	No	Major
Bad	Sober	Other Rules	No	Major
Good	Alcohol	Speed	Yes	Major
Bad	Sober	Red Signal	Yes	Minor

Using Information Gain, Identify the attribute that should be considered at the root node of decision tree model. [6]

Given that, 8 Major and 4 Minor classes in the data out of 12 records. The expected information needed to classify a tuple in the data: Information or Entropy = $- \sum p(i/t).\log_2 p(i/t)$

$= - \{ (8/12).\log_2 (8/12) + (4/12).\log_2 (4/12)\}$

$= - \{- 0.39 - 0.53\}$ **= 0.92**

Now the calculation of information (entropy) for all other attributes:

	Driver Condition	
	Alcohol	**Sober**
Major	4	4
Minor	1	3
Entropy	0.72	0.99
Weighted Avg Info		**0.88**

	Weather Condition	
	Good	**Bad**
Major	5	3
Minor	2	2
Entropy	0.86	0.97
Weighted Avg Info		**0.91**

	Seat Belt	
	Yes	**No**
Major	4	4
Minor	4	0
Entropy	1.0	0
Weighted Avg Info		**0.67**

	Rule Violation			
	Speed	**None**	**Red Signal**	**Other Rules**
Major	3	2	0	3
Minor	0	1	3	0
Entropy	0	0.92	0	0
Weighted Avg Info			**0.23**	

[1 mark for each attribute for entropy calculations]

Since the information is least for the attribute **Rule Violation**, gain will be maximum with it. So **Rule Violation** will be the first attribute selected for splitting. Students may (optionally) show information gain calculations also.

[1 mark for identifying the attribute for the root node]

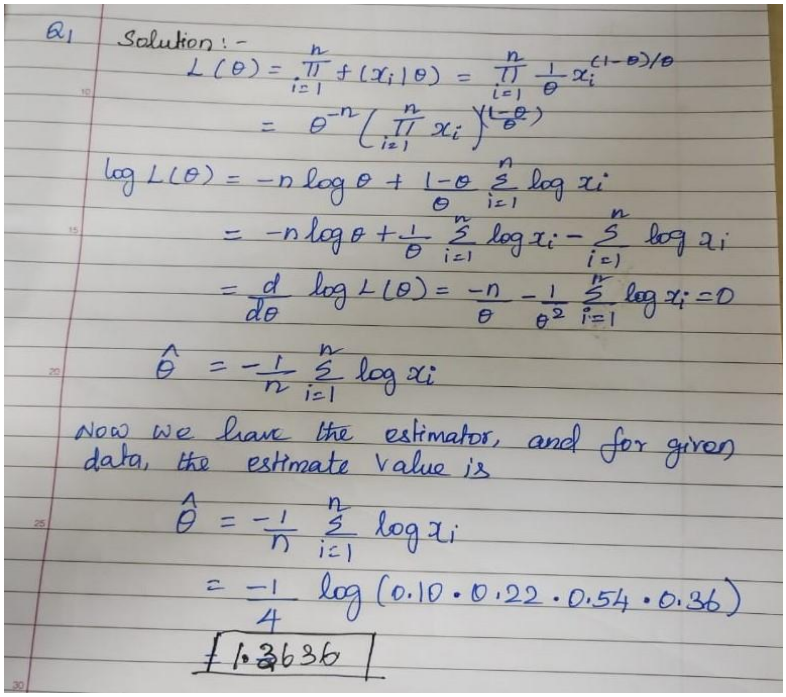
Question 7:

Let $T_1, T_2, \ldots T_n$ *be* a random sample of a population describing the website loading time on a mobile browser with probability density function given as:

$$f(t/\theta) = \frac{1}{\theta} t^{\frac{(1-\theta)}{\theta}} \quad where \;\; 0 < t < 1 \; and \; 0 < \theta < \infty$$

Find the maximum likelihood estimator of θ. What is the estimate of θ, if the website loading time from four samples are $t_1 = 0.10$, $t_2 = 0.22$, $t_3 = 0.54$, $t_4 = 0.36$. [5 Marks]

Solution:



Marking Scheme: Derivation of θ =3 marks (step wise marks)

θ Computation = 2 marls (wrong value = 0 marks)

Question 8: Given below is the confusion matrix for multi-class classifier when it was run on test data.

	Actual Class →	**Class A**	**Class B**	**Class C**
Predicted Class	**Class A**	23	34	22
	Class B	17	16	8
	Class C	20	10	10

Comment on the performance of this classifier by calculating precision and recall with respect to Class B. [4 marks]

Solution:

TP = 16, TN = 23+22+20+10

FP = (17+8)

FN = 34+10

Precision = TP/(TP+FP) Recall = Tp/(TP+FN)

[2.5 marks for calculation and formula]

[1.5 marks on observations]