Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
2020-2021
M.Tech. (Data Science and Engineering)
Comprehensive Examination (Makeup)

Course No. : DSECLZG565
Course Title : MACHINE LEARNING
Nature of Exam : Open Book
Weightage : 30%

No. of Pages = 3
No. of Questions = 5

Note: Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1. [3+2+2+3=10 marks]

A. Given a real world problem of organizing computing clusters, which type of algorithm is most suitable and why? If we also wanted to predict the computation time required for certain task on these clusters, given a 10 years data, can the same algorithm be used? If not which algorithm could be used for predicting the computation time. **[3 marks]**

B. Fruit basket contains half dozen oranges, and four apples. A girl eats three of them one after another. What is the probability of sequentially choosing two oranges and one apple?
[2 Marks]

Solution
Probability of choosing 1 orange= 6/ 10
After eating 1 orange, the total number of fruits left is 9.
Probability of choosing 2nd orange = 5/9
After eating 1 more orange, the total number of fruits left is 8.
Probability of choosing 1 apple out of a total of 8 fruits = 4/8 = 4/8
So the final probability of choosing 2 oranges and 1 apple = 6/10 * 5/9 * 4/8 =0.166

C. A woman was tested for diagnosis of heart attack with a 2% chance. The doctors examined ECG with a positive result. ECG accurately predicts about 80% of heart attacks and 96% of normal heart rhythm. The probability of heart attack was predicted to be about 30% by 90 out of a cardiac doctors. Do you agree?
[3 Marks]

Solution. Introduce the events: + =, B = n,
H = person having heart attack.
Hc = person not having heart attack/ Normal heart rhythm

We are given P(H) = .02, so P(Hc) = 1 − P(H) = .98.
- += ECG correctly predicts heart attack and - = ECG correctly predicts no heart attack

Given P(+ | H) = .80
and P(− | Hc) = .96, where the event − is the complement of +,
thus P(+ | Hc) = .04
Bayes' formula in this case is
P(H | +) = P(+|H)P(H) / P(+|H)P(H) + P(+|Hc)P(Hc))
= 0.80 × 0.02 / (0.80 × 0.02 + 0.04 × 0.98)

= 0.016/0.016+0.0392=0.016/0.0568=0.289
So the chance of heart attack would be 29%. Almost close to what had been predicted by most of the doctors.

D. The results of election are to be predicted for a candidate. There are four different hypothesis used to predict the result of candidate winning or losing an election. Four hypothesis are equally likely (15% probability) and predict that candidate will win the election. One of the hypothesis with 40% probability predicts that candidate will not win the election. Will the candidate win or not win the election?

[2 Marks]

Solution

$P(h_1|D) = .4$, $P(+|h_1) = 1$,
$P(+|D) = P(h_1|D) * P(+|h_1) = 0.2$
$P(h_2|D) = P(h_3|D) = P(h_4|D) = P(h_5|D) = .15$,
$P(-|h_2) = 1$, $P(-|h_3) = 1$, $P(-|h_4) = P(h_5|D) = 1$
$P(-|D) = P(h_2|D)*P(-|h_2) + P(h_3|D)*P(-|h_3) + P(h_4|D)* P(-|h_4) = 0.15+0.15+0.15+0.15=0.6$
The candidate will win the election

Question 2. [5+5=10 marks]

A. You are working in an insurance firm. Consider the firm would like to offer products and services based on the income bracket of the individuals. Following dataset depicts customer data with their age group, gender information and their occupation. Using this sample dataset, apply Naïve Bayes classification technique, to classify the following tuple either as "High", or "Middle" or "Low" income bracket member. **[5 Marks]**
{SrNo = "7", Age = "older", Gender = "M", Occupation = "Software Engineer"}

Sr No	Age	Gender	Occupation	Income Bracket
1	older	F	Software Engineer	High
2	young	M	Marketing Executive	Middle
3	Middle aged	M	Unemployed	Low
4	young	M	Data Scientist	High
5	Middle aged	F	Software Engineer	High
6	young	F	Unemployed	Low

Answer:
Need to find various conditional probabilities –
P(Income Bracket = "High) = 3 /6 = 0.5
P(Income Bracket = "Middle) = 1 /6 = 0.17

P(Income Bracket = “Low”) = 2 /6 = 0.33

P(X | Agegroup= ‘older’ and IC = “High”) = 1 /3 = 0.33
P(X | Agegroup= ‘older’ and IC = “Middle”) = 0
P(X | Agegroup = ‘older’ and IC = “Low”) = 0

P(X | Gender = ‘M’ and IC = “High”) = 1 /3 = 0.17
P(X | Gender = ‘M’ and IC = “Middle”) = 1/1 = 1
P(X | Gender = ‘M’ and IC = “Low”) = 1 / 2 = 0.5

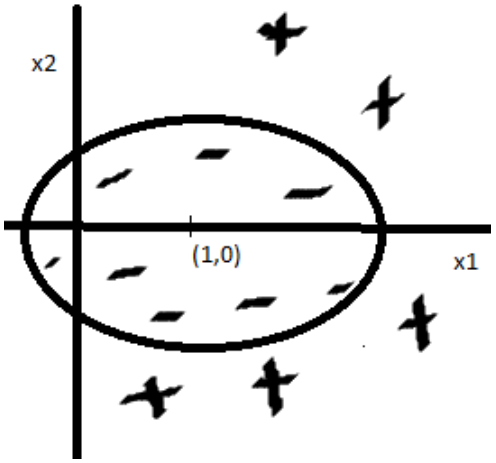
P(X | Occupation = ‘Soft Engg’ and IC = “High”) = 2 / 3 = 0.67
P(X | Occupation = ‘Soft Engg’ and IC = “Middle”) = 0
P(X | Occupation = ‘Soft Engg’ and IC = “Low”) = 0

B. Consider a binary classification problem and the hypothesis function $h(\mathrm{w},\mathrm{x}) = W^T X = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$ with $w^T =< -108, -36, 0, 9, 16 >$. Here x_1, x_2 represent the dataset features.
   a) Derive an equation for the decision boundary represented by this model, considering the sigmoid function and logistic regression. [3 M]
   b) Suggest a feature transformation that leads to linear decision boundary. [2 M]

Solution:

We have $h(\mathrm{w},\mathrm{x}) = W^T X$
Substituting the weight values leads to h(w,x)
$=-108 - 36x_1 + 9x_1^2 + 16x_2^2$

Sigmoid function is given by 1/ (1+e^{-z}) with z≥0 represents one class and z<0. We have $z = h(w,x) = -108 - 36x_1 + 9x_1^2 + 16x_2^2$ z≥0 represents one class and z<0 other.



Equation of decision boundary corresponds to z=0

$$=> -108 - 36x_1 + 9x_1^2 + 16x_2^2 = 0$$
$$=> 9(x_1^2 - 4x_1) + 16x_2^2 - 144 = 0 =$$
$$> \mathbf{9(x_1 - 1)^2 + 16x_2^2 = 144}$$

The decision boundary in this case is an ellipse with centre at 1, 0 and major axis 8, minor axis 6.

Feature transformation required: $\boldsymbol{X_1 = (x_1 - 1)^2/16}$, $\boldsymbol{X_2 = x_2^2/9}$

Question 3. [5+5=10 marks]

A. Consider the following training example dataset

Instance	a1	a2	a3	Class
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

What is the entropy of this collection of dataset with respect to the split class **[1 Mark]**

Find the Information gain of a1 and a2 **[3 Marks]**

What is the best split between a1 and a2 **[1 Mark]**

****

Solution:

What is the entropy of this collection of dataset with respect to the split class

Entropy = -(4/9 log (4/9) + 5/9log(5/9))

- (-0.51997-0.47111) = 0.99107

Find the Information gain of a1 and a2.

Entropy of a1 for T

= -(3/4log3/4)+1/4log(¼)

=-(-.3118-.5) = .81128

Entropy of a1 for F

=-(1/5log1/5+4/5log4/5)

=0.72193

Entropy of a2 for T = 0.97095

Entropy of a2 for F = 1

Information Gain a1 = 0.991 -(4/9)*.81128 -5/9(0.72193) = 0.229

Information Gain a2 = 0.991 - 5/9*o.971 -4/9*1 = 0.007

Best split is a1 due to higher

_____

B. What does the coefficient of correlation in a linear regression model signify? **[2 Marks]**

Solution:

The strength of the linear relationship between the two variables.

C. Explain how does bias-variance are related to model complexity? Explain with example whether increasing variance of a model is appropriate **[3 Marks]**