BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES DIVISION
POST GRADUATE PROGRAMME IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
COMPREHENSIVE EXAM – CLASSIFICATION REGULAR

Course Code	**: PCAM ZC211**	**Course Title**	**: Classification**
Nature of Exam	**: Closed Book**	**Marks**	**: 30**
Duration	**: 2 Hours**		
Date of Exam	**: 09/08/2020 (FN)**		

Q1 **[5 Marks]**

a) Give any one example of probabilistic and non-probabilistic classification algorithm**[1 marks]**
Solution
Any one from Naïve Bayes or logistic for probabilistic and
Any one from SVM, decision tree, KNN for non-probabilistic

b) Consider the one-dimensional data set shown below, Y being the class label

X	0.6	3.2	4.5	4.6	4.9	5.2	5.6	5.8	7.1	9.5
Y	-	-	+	+	+	-	-	+	-	-

Classify the data point x=5.0 according to its 3, and 5-nearest neighbors. **[1 marks]**
Solution

NN(3-) of x=5.0 = 4.9 → +, 5.2 → - , 4.6 → + therefore +
NN(5-) of x=5.0 = 4.9 → +, 5.2 → -, 5.6 → - , 4.6 → +, 5.8 → + therefore +

c) In a manufacturing plant there are several conditions like abnormal temperature, pressure, humidity, electricity supply, machine failure, labor shortage etc which can impact the production. The management is interested to know what situation actually impact the production and hires data scientists to get a classifier prepared which predicts if the conditions will impact the production (Yes) or not (No). When the classifier was run on the test data, the following confusion matrix was obtained. Comment on the performance of the classifier using appropriate metric(s) to meet the management's objective. **[3 marks]**

Actual Class	**Predicted Class**	
	Yes	No
Yes	50	115
No	72	5000

Solution

Management's objective is to predict those conditions which impact the production. So YES is the class of interest.
The confusion matrix is given as:

Actual Class	**Predicted Class**	
	Yes	**No**
Yes	**50**	**115**
No	**72**	**5000**

There is class imbalance problem in the given data as the class of interest is rare.
Form the given data, TP = 50, FP = 72, FN = 115 and TN = 5000.
So, P = 50+115 = 165 and N = 72+5000 = 5072

A. The objective of this question is to check if a student is able to observe the class imbalance problem and contrast Accuracy/f-score metrics and finally comment about the classifier's relevance.

Q 2. **[5 Marks]**

a) Suppose that you have trained a logistic regression classifier, and it outputs on a new example "x", a prediction $h_\Theta(x) = 0.2$. What is the probability that "x" would belong to class y=0? **[1 mark]**
Answer: 0.8

b) You are working in google.com. As part of profile creation, the firm captures various important details about its members. Some of the samples are shown in the below table. Using this sample dataset, with the help of Naïve Bayesian classification technique, classify the following tuple either as "High", or "Middle" or "Low" income bracket member. You may ignore Laplacian correction.

{SrNo = "7", Age = "Above_40", Gender = "M", Occupation = "Software Engineer"} **[4 marks]**

Sr No	Age	Gender	Occupation	Income Bracket
1	Above_40	F	Software Engineer	High
2	Below_30	M	Marketing Executive	Middle
3	Between_31_to_40	M	Unemployed	Low
4	Below_30	M	Data Scientist	High
5	Between_31_to_40	F	Software Engineer	High

Solution:
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 | Age = 'Above_40' and IC = "High") = 1 /3 = 0.33
P(X | Age = 'Above_40' and IC = "Middle") = 0
P(X | Age = 'Above_40' 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

P(X | IC = "High") = 0.33 * 0.17 * 0.67 = 0.03
P(X | IC = "Middle") = 0 * 1 * 0 = 0
P(X | IC = "Low") = 0 * 0.5 * 0 = 0

P(X | IC = "High) * P(Income Bracket = "High") = 0.03 * 0.5 = 0.015
P(X | IC = "Middle) * P(Income Bracket = "High") = 0 * 0.17 = 0
P(X | IC = "Low) * P(Income Bracket = "High") = 0 * 0.33 = 0

As the P(X | IC = "High) * P(Income Bracket = "High") has the maximum probability value, the tuple will be classified as "High" Income category.

Q 3. **[5 Marks]**

After the parliament passed a bill on stringent traffic regulation, the following data was captured on a busy and representative traffic signal for a specific period. Consider "Crash Severity" as the class of interest and use multiway split for the discrete-valued attributes.

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, find out the attribute that should be considered at the root node.

Solution:

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:

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

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

	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	

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.

Q 4. **[10 Marks]**

a) What are the Pros (2 points) and Cons (2 points) of Support Vector Machines? (4 M)
Solution

Pros:

- **Works very well when there is a clear margin of separation**
- **Effective in High Dimensional Spaces**
- **Effective when the number of features (or dimensions) are greater than the number of rows (or samples)**

Cons:

- **Does not perform well on large dataset and training time is considerably high**
- **Computationally Extensive**
- **Does not perform well on overlapping classes**

b) What is Kernel trick in SVMs? Give an example of Kernel Trick. (2 M)
Solution
The Kernel trick is a method used by SVMs to project the input low dimensional space to a high dimensional space in order to make non-separable classes easily separable.

For instance, a 1-D feature with overlapping datapoints on one single line, when bent and converted to a 2-D polynomial feature, might have class A on the trough of the curve and the points of class B on the ends of the curve.

c) Define Support Vectors. (1 M)
Support Vectors are:

- **Data Points closest to the decision boundary**
- **Data Points Hardest to classify**
- **They directly affect the positioning of the decision boundary**

d) Coordinates of support vectors for class y=1 are (1,4) and (-1,4) and for class y=-1 are (0,-1). Lagrange multipliers for y=1 are 0.65 each, for y=-1 lagrange multiplier is 4.5. Find the weight vector w, and b for linear SVM problem.
Solution
-4.5[0 -1 1]+0.65[1 4 1]+0.65[-1 4 1]=[0 -4.5 -4.5]+[0.65 2.6 0.65]]+[-0.65 2.6 0.65] = [0 0.7 -3.2]
A=0 c=0.7
Ax+cy+b=0
Y=-b/c=3.2/0.7=4.57
Slope=0

Q 5. [5 Marks]

a) When do we use ensemble classifiers in machine learning (1 mark}
Ensemble classifier performs better than the base classifiers when each classifier error is smaller than 0.5
Necessary conditions for an ensemble classifier to perform better than a single classifier:

- Base classifiers should be independent of each other
- Base classifiers should do better than a classifier that performs random guessing

b) How are the weights of data points updated in Adaboost algorithm? How is this useful? (2 marks)

Weights can be determined using the error value. For instance, higher the error more is the weight assigned to the observation.

This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

Weight Update: $$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \begin{cases} \exp^{-\alpha_j} & \text{if } C_j(x_i) = y_i \\ \exp^{\alpha_j} & \text{if } C_j(x_i) \neq y_i \end{cases}$$ <- Eqn:5.88

where Z_j is the normalization factor

$$C^*(x) = \arg\max_y \sum_{j=1}^{T} \alpha_j \delta\left(C_j(x) = y\right)$$

- Reduce weight if correctly classified else increase
- If any intermediate rounds produce error rate higher than 50%, the weights are reverted back to 1/n and the resampling procedure is repeated

c) How is importance of each classifier decided in Adaboost algorithm? Explain mathematically (2 marks)

Base classifiers C_i: C_1, C_2, …, C_T

Error rate:

– N input samples

$$\varepsilon_i = \frac{1}{N} \sum_{j=1}^{N} w_j \delta\left(C_i(x_j) \neq y_j\right)$$

Importance of a classifier:

$$\alpha_i = \frac{1}{2} \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

https://en.wikipedia.org/wiki/AdaBoost#Choosing_at

****