

Software Engineering (664)

Python Programming for Data Sciences

Department of Mathematics (Final Examination) University of Karachi

Lecturer: Engineer Syed Umaid Ahmed

Time Allowed: 3 hours

Note: Attempt any five questions. All carry equal marks

Total Marks: 50

Question 1:

(a) Compare the differences between Supervised Learning and Unsupervised Learning. Define Reinforcement Learning and real word examples of each of the following.

Table I shows a confusion matrix for a disease classification problem.

ID	Actual Sick?	Predicted Sick?	Outcome
1	1	1	TP
2	0	0	TN
3	0	0	TN
4	1	1	TP
5	0	0	TN
6	0	0	TN
7	1	0	FP
8	0	1	FN
9	0	0	TN
10	1	0	FP

- (b) Calculate Total True Positives and True Negatives?
- (c) Why confusion matrix is evaluated at the end of model development?
- (d) How many times a day do the minute and hour hands of a clock overlap?

Question 2:

(a) What level of measurement (nominal/ordinal/ratio/interval) do the following eight variables have? Discuss your findings in the form of a table.

Income, Year of Birth, Weight, Degree in Celsius, MSc Mathematics Degree, Skin Color, Hair Length, Gender

Which of the following variables can be used as appropriate features for building a model for the prediction of qualified person? Share your findings and ideas.

(b) Explain the working of the Decision Tree in case of the given training dataset of Fruits. Draw a neat and clean decision tree after necessary calculations.

Color	Diameter	Label
Hot	High	Apple
Hot	High	Lemon
Hot	High	Grapes
Warm	High	Apple
Cold	Medium	Lemon
Cold	Medium	Grapes

(c) At a party there were 66 handshakes. Everyone shook hands with everybody else. How many people were present at the party?

Question 3:

(a) Fill the applicable properties in each cell or the table below.

Algorithms	Supervised/Unsupervised	Numerical/Categorical/Ordinal	Class/Value
Linear Regression			
Logistic Regression			
K-Nearest Neighbor			
Naïve Bayes			
Decision Tree			
Support Vector Machine			

(b) Below are the eight actual values of the target/predictor variable in the train file. Calculate the actual entropy of the target variable

[0, 0, 0, 1, 1, 1, 1, 1]

(c) Explain the steps for making a decision tree. Make a clear decision tree for the conditions. If a job is accepted only there are three conditions.

- Salary is greater than 50,000
- Commute is less than an hour
- Incentives and Bonuses are offered

Question 4:

- (a) What is the difference between Random Sampling and Systematic Sampling? Discuss both of them with reference to real world examples.
- (b) What is Support Vector Machines Classifier? Discuss the use of line, plane and hyperplane in the SVM Classification Algorithm?
- (c) Discuss the problems with KNN algorithm. Mention the situation for biased results of prediction in KNN algorithm.

Table II shows a areas and price dataset for property in an area.

Area	Price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000

- (d) Which algorithm is suitable for prediction of unknown data points? Describe the reason for choosing the best algorithm.

Question 5:

- (a) Discuss what are outliers in dataset? Why it is necessary to remove them?
- (b) Describe how we can split the dataset in Python? Why we need to split the data before the preparation of the final machine learning model?
- (c) You have a data science project where you have to deal with 1000 columns and 1 million rows. The object of the problem is to carry out Classification. You are required to reduce dimensions in order to reduce computational time. Also, we have memory constraints. What will you do in this situation? Give three solutions.

OR

Four people A, B, C, D need to cross a bridge at night, and they have only one torch. The bridge is too dangerous to cross without a torch and is strong enough to support a maximum of two people at a time. They take 1, 2, 5 and 8 minutes respectively. What is the shortest time needed for all four of them to cross the bridge?


Question 6:

- (a) What are missing values in the dataset? Explain some different cases of filling the missing values by using mean or median in python. Also write python functions to find and full the null values in a dataset.
- (b) A robot is placed in the maze below at A1 position. It is programmed to move in four directions (up) ↑, (down) ↓, (right) →, (left) ←. If the robot can't move in a

programmed direction, then it stays in the same place to make the next move. Where will be the robot after executing the program?

↓ ← ↓ → ↓ → ↑ → → → ← ↑ ↓ ↓ ?

Give your answer as a pair of a letter & number e.g. B4

	1	2	3	4	5	6
A						
B						
C						
D						
E						
F						

Question 7:

(a) Here is a simple dataset of the training institution and the achieved results.

	Results	Training		
		Without Training	With Training	
Salary Packages Obtained by Students	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

- i. Find the Probability that a candidate has gone Training. Name this type of Probability.
- ii. Find the Probability that a candidate has gone Training and a Good Package. Name this type of Probability.
- iii. Find the Probability that a candidate has good package and not undergone University’s Training. Name this type as well.

(b) Explain Hypothesis Testing and Confidence Interval.

(c) We have categorized emails as ham and spam using Naïve Bayes Classifier. Discuss the complete process from the csv provided to building an accurate model.