THE LNM INSTITUTE OF INFORMATION TECHNOLOGY, JAIPUR

Introduction to Data Science

End Term Examination (29th Nov 2018)

Duration: 3 Hours **Total Marks: 80**

Instructions: The question paper consists of two parts (PART A and PART B) and each part must be answered on a separate answer sheet. Each question should be answered in a new page. Answer the questions in the same order as it appears in the question paper.

PART A (20 Marks)

Answer all questions. In MCQs correct choice(s) will get only 25% marks, correct justifications get 75% marks. If choices are wrong no marks will be given.

1. Suppose a company wants to revise the salary of a section of employees in a specific level using the salary of employees at the same level in other companies as reference. What measure(s) is (are) more appropriate to use? [2 M]
   a. Mean of salaries
   b. Mode of salaries
   c. Median of salaries
   d. None it will find the minimum of the lot and use that

   Justify your choice. Also justify why other choices won't work.

2. *Suppose there is a conversation between a son and a father like below:*
   Son: Do you have office?
   Father: Yes

   If you want to interpret this conversation using a NLP system at what level you would have to interpret it correctly? [2 M]
   a. Lexical Analysis
   b. Syntactic Analysis
   c. Semantic Analysis
   d. Pragmatics

   Justify your choice. Also justify why other choices won't work.

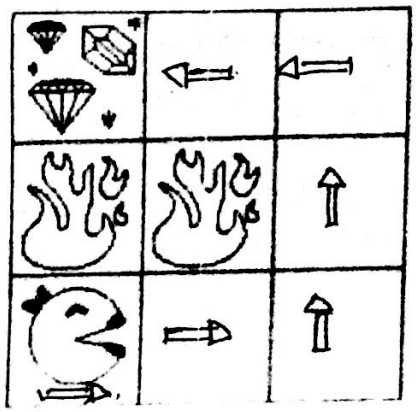
3. Consider the normal usage of the sentence: "The dog brings me the newspaper every morning". What kind of ambiguity is more appropriate here? Justify your selection. [2 M]
   a. Syntactic ambiguity
   b. Semantic ambiguity
   c. Lexical ambiguity
   d. Scope ambiguity

4. The mayor claims that blacks account for 25% of all city employees. A civil rights group disputes this claim, and argues that the city discriminates against blacks. A random sample of 120 city employees has 18 blacks. Test the mayor's claim at the .05 level of significance. (Note: Critical value of Z is -1.65 for 0.05 level of significance, i.e., if $Z > -1.65$ accept H_0, otherwise reject H_0). [3 M]

5. We have learned three kinds of semantic relationship in WordNet - hyponymy, antonymy and meronymy. What kind of semantic relationship exist between the following pairs of words: [3 M]

   a. Brother and Sister
   b. Relative and Person
   c. Relative and Family
   d. Arm and Bone
   e. Men and Women
   f. Family and Group

6. Pacman is out on a treasure hunting in the Gridworld island (see the Picture). She has no idea of the grid so she has to learn by episodes / observations. From any unmarked square, Ms. Pacman can take the standard actions (N, S, E, W), but she is sure-footed enough that her actions always succeeds (i.e. there is no non-deterministic movements). If she lands in a hazard square (marked as fire) or a treasure square (marked with diamonds), her only action is to call for an airlift (X), which takes her to the terminal 'Done / Exit' state, receiving a reward of -64 but +128 if she's running off with the treasure. There is no living expense, i.e., for movement from one square to another square there is no penalty.



Assume that the learning rate $\alpha = 0.5$. Answer the following questions: [4+2+2 = 8 M]

a. Suppose we are also given a policy (see the arrows) that from the lower leftmost square (marked with Pacman) and then go around the hazards to the treasure square and go to exit what will be utility value of the states after you learn / iterate for 7 times using Temporal Difference Learning Method. Illustrate each iteration.
b. Is this Temporal Difference Learning, Model-free or Model-Based approach? Why?
c. Is the above problem Passive Learning or Active Learning? Why?

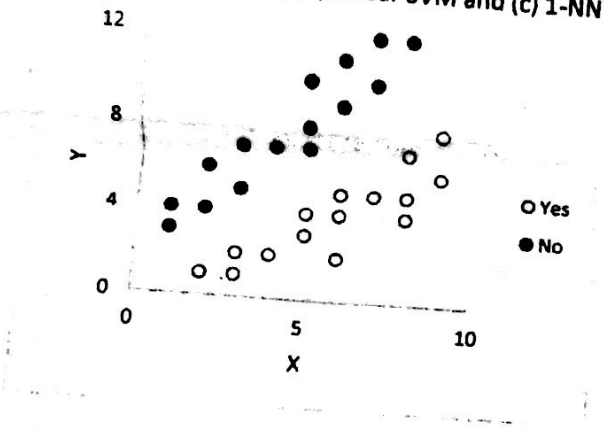
PART B (60 Marks)

Answer all questions. Show your calculations wherever required.

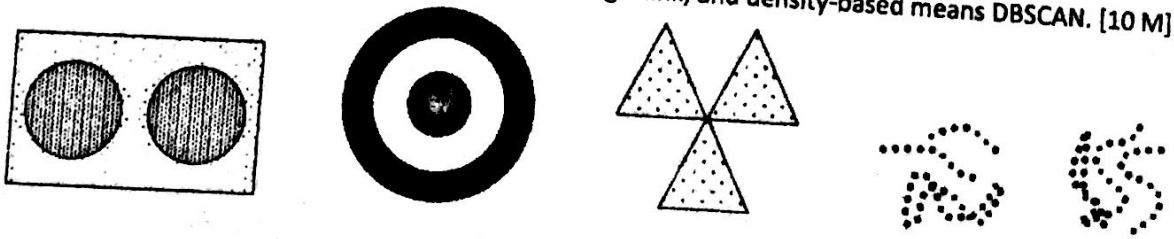
1. Draw full decision tree using gini index for the dataset given in Table 1 (without any pruning). Show all necessary splitting conditions as part of your calculation. [10 M]

Temperature	Outlook	Windy	Play
Hot	Rainy	Low	No
Cold	Overcast	High	No
Cold	Overcast	High	No
Cold	Sunny	Low	Yes
Hot	Sunny	Low	Yes
Hot	Rainy	High	No
Hot	Sunny	High	Yes
Cold	Overcast	Low	Yes
Cold	Overcast	Low	Yes
Hot	Rainy	High	No

2. Consider the binary class dataset shown in following figure. What do you think would be the nature of the decision boundaries if we apply (a) Decision Tree, (b) Linear SVM and (c) 1-NN? Explain. [10 M]



3. Identify the clusters in the following four datasets (shown from (a) to (d)) using the center-, contiguity-, and density based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN. [10 M]



(a) (b) (c) (d)

4. Use the **similarity matrix** given in the following table to perform single and complete link hierarchical clustering. Show the updated proximity matrix after each step. Also show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged along with the similarity values. [10 M]

	P1	P2	P3	P4	P5
P1	1	0.1	0.35	0.55	0.41
P2	0.1	1	0.98	0.47	0.64
P3	0.35	0.98	1	0.76	0.85
P4	0.55	0.47	0.76	1	0.44
P5	0.41	0.64	0.85	0.44	1

5. *Asterix and Obelix are confused about the result of a clustering method. They have the similarity matrix as shown below and the clustering result as: Cluster1 = {P1,P2} and Cluster2 = {P3,P4}. Help them understand whether the clustering result that they have obtained is good or not by quantifying it. Use correlation measure and check cluster validity. [10 M]*

	P1	P2	P3	P4
P1	1	0.9	0.7	0.5
P2	0.9	1	0.6	0.4
P3	0.7	0.6	1	0.8
P4	0.5	0.4	0.8	1

6. ***Answer the following questions: [4+3+3 = 10M]***
   a. Explain with example, why 'Accuracy' measure of a classifier does not give better idea about the performance of the classifier in a class imbalance problem.
   b. Explain the similarities and dissimilarities between bagging and boosting techniques.
   c. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on PCA.