**CS 535/EE 514 – Machine Learning**

**Assignment 1**

Deadline: Sep. 22 (Friday) at 11:59 PM

**Objective: Basic understanding of machine learning, model selection, problem understanding and identification**

Question 1 (15 points)

1. (10 points) You are given a dataset for cancer detection having two classes (binary classification). 0 stands for “cancer not detected” and 1 for “cancer detected”. This dataset has train/test split. Training set has 10,000 instances/records where half of the instances belong to class 0 and remaining half belong to class 1. Testing set has 1,000 instances where 990 instances belong to class 0 and 10 instances belong to class 1. You create two models, model A and model B. Model A gives you training accuracy of 80% and testing accuracy of 75%. Model B gives you training accuracy of 50% but testing accuracy of 99%. Between these two, which model will you prefer and why? Discuss potential problems in both models and method how to rectify them.
2. (5 points) You create another model C which has low bias i.e is elastic enough to mimic the training data distribution. What potential problem might occur in such model and how would you tackle it?

Question 2 (10 points)

Manhattan Bites delivers pizza at LUMS. However, company’s riders are not able to deliver it on time hence customers of Manhattan Bites get unhappy. To keep customers happy, they have to give them free food items as compensation, which increases their cost. They hear you are studying machine learning at LUMS and contact you to come up with a solution for their problem of riders becoming late. Can machine learning algorithms solve this problem? If yes, which algorithm would you recommend and why? If not, give reasoning.

Question 3 (10 points)

1. (5 points) Can single decision tree learn logical operators “and” & “or”? If no, please provide reasoning. If yes, please create such a decision tree for following rules. It is binary classification problem and A,B,C can assume only two values, 0 or 0.5.

If A = 0 and B = 0 and C = 0 then class X.

If A = 0.5 or B = 0.5 or C = 0.5 then class Y.

1. (5 points) How can decision tree algorithm be used to reduce number of dimensions/attributes of a given dataset?

Question 4 (10 points)

1. (5 points) There is a cricket match between India and Pakistan. Your task is to predict who would win this match, based on all available past data of Pak Vs India matches. You want to use ML algorithm to make prediction. Would you approach this as classification problem or regression problem and why?
2. (5 points) Let’s suppose you got your hands on Bill Gate’s bank statements for all previous years. You want to predict his next year’s income using ML algorithms. Would you approach this as classification problem or regression problem and why?

Question 5 (5 points)

Suppose you are given a pure random number generating function. You have 1 million records of previously generated random numbers using this function. You task is to predict next random number that will be generated by this function. Can machine learning algorithms solve this problem? If yes, which algorithm would you recommend and why? If not, give reasoning.

Question 6 (20 points)

Find which of the following hypothesis can shatter i) 2 points in 2-D space ii) 3 points in 2-D space, and iii) 4 points in 2-D space (x,y are the dimensions and the remaining symbols are model parameters):

2. (a <= x <= b) AND (c <= y <= d)

Which hypothesis is more likely to overfit and why?

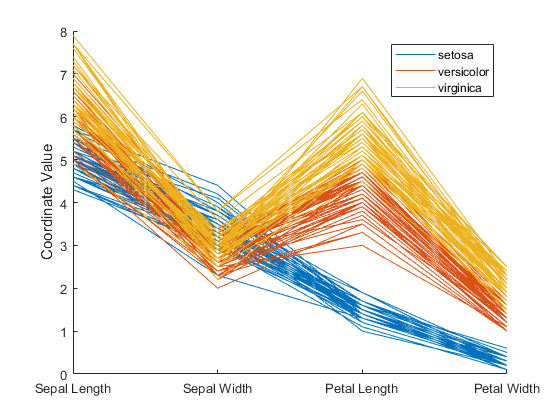
Question 7 (10 points)

1. (5 points) How would you distinguish between automation problems and learning problems? Give example of each.
2. (5 points) Explain bias-variance trade off with respect to a model.

Question 8 (5 points)

What is type-I and type-II error? Illustrate with the help of a confusion matrix of binary class problem let’s say having class labels “cancer = yes” and “cancer = no” having 100 examples for each class. Just a 5-10 lines explanation with the help of confusion matrix to support the answer would suffice.

Question 9 (10 points)



Consider above graph. It has three classes, each denoted by separate color. Number of lines in a particular color show number of instances/records of that particular class. It has 4 attributes plotted against X-axis, sepal length, sepal width, petal length and petal width. On Y-axis, we have coordinate values for all attributes. If you are asked to reduce number of attributes/dimensions of the dataset based on this graph and pick only top two attributes, which two attributes would you pick and why?

Question 10 (5 points)

What is difference between induction and deduction? Give an example for each.

**Instructions:**

1. **This is assignment is focused on understanding of basic concepts. So provide brief reasoning for your answers.**
2. **All answers must be hand-written.**
3. **Please write a** **brief answer. For some questions, even couple of lines would suffice.**
4. **All assignments must be submitted on LMS after scanning. You can use “camscanner” app for this purpose as well. Before uploading, please make sure it’s readable. Do not email your assignments to your instructors.**
5. **Plagiarism will be checked. Plagiarized assignments will either face points deduction or will be reported to disciplinary committee.**
6. **Late submissions will be penalized according to policy announced by the instructor.**