**CS 535/EE 514 – Machine Learning**

**Assignment 1**

Deadline: Sep. 18 (Monday) 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.

**Ans:** We will prefer Model A. That’s because if we look at the information of train test split, training set has 50% of instances belonging to class 0 and rest 50% belong to class1. Similarly, we can see that 99% of the instances belong to class 0 and only 1% instances belong to class 1. It means, if our model start predicting every instance/example as class 0, it will achieve training accuracy of 50% and testing accuracy of 99%. It’s not a good model because it has not learned the difference between class 0 and class 1.

**In other words:** Model A is preferred because it’s training accuracy is higher with testing accuracy following closely as compared to Model B. A bigger gap between training and testing accuracy means model is either underfitting or overfitting (underfitting in case of model B).

**In layman terms:** It’s highly unlikely to get 99% marks in exam when your performance in class is 50% (unless you cheat).

1. (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?

**Ans:** Low bias occurs when the model’s predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results. We can use random sampling again and again to get the split which does not mimic the training data exactly and use it to train the model.

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.

**Ans:** This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of three fundamental things:

1. There exist a pattern that you want your algorithm to learn.
2. You cannot solve it mathematically, even by writing exponential equations. Please refer to your quiz-1 question. If you have a mathematical equation to solve a problem, why would you need ML? Just put in the values and solve it.
3. You have data on it.

In this problem, what would be your data? What would be the features? What do you want to classify/predict? What would be the error/cost function that you want to minimize during training? One can argue that we want to predict optimal route to take to deliver pizza just like google maps does. But for that, you need live feed of current situation of a particular route and given conditions on that route, even rider can tell if he will be late or not.

In other words, you cannot identify what pattern you want your learning algorithm to learn and what do you want to predict. You wouldn’t be able to “Solve” this problem of riders becoming late. Try it with dummy data using decisions tree on paper!

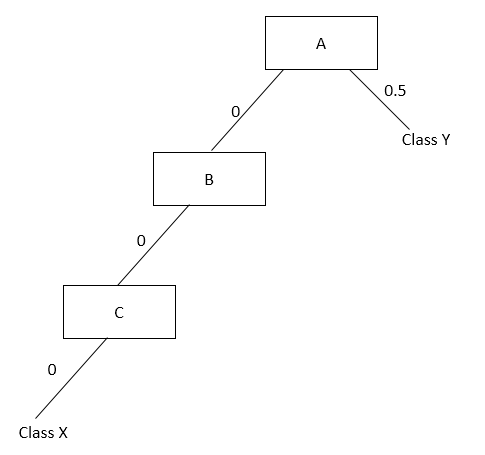
Question 3 (10 points)

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.

**Ans:** There can be multiple correct answers to this question and all will be considered correct. You were required to construct a tree that fulfills only these two rules. There are other combinations possible, for example, If A=0 and B=0.5 and C=0.5 then class X. But those were not required to consider because it would’ve been very complex tree to construct at beginner’s level. But if someone has done that, he/she will get full credit. A simple tree to incorporate two rules above would be:



Left part of the tree is for Rule 1 (and operator where values for every variable must satisfy) while right part is for or operator (where if one value satisfies, we don’t need to check any other value). Offcourse you can have nodes to the right of attribute B and C for class Y to accommodate more rules like if A = 0 and B = 0.5 then Class Y, but that was not required.

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

**Ans:** You can use entropy or information gain. Let’s suppose if any attribute gives you information gain less than a threshold, say 0.0001, you drop it because it’s not useful. You keep all other attributes above this threshold. Another way could be, calculate IG for all attributes and then sort them. Keep top p% attributes. Multiple correct answers are possible.

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?

**Ans:** Classification because output is discrete.

1. (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?

**Ans:** Regression because output is continuous.

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.

**Ans:** We cannot solve this problem using ML because it’s a “pure” random number generating function. If we are able to predict what this function is going to generate next, it would not be random now would it? In other words, every previously generated number is random and there exists no pattern to learn.

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.

**Ans:** Automation is basically making a hardware or software that is capable of doing things automatically. Consider the example of fire alarm systems deployed in buildings. As soon as smoke sensor is activated, water start pouring down the pipes. Any action is hardcoded in such process. It is not learned.

On the other hand purpose of ML is making intelligent machines. We try to make machines or software to mimic human behavior and intelligence.

Automation can be provided by placing some sensors and making something do corresponding to sensor readings. If you want to make the same thing using ML, then you might need to do extra bit of work. Like use of neural networks, graphs, and deep learning in your software. Your coding level will decide to how much extend you can make your system stimulate like human. In case of simple automatic you can easily predict the output, according to sensor readings and there is no learning involved. While in case of ML there is always bit of uncertainty, just like human brain because model learns from data make decisions based on that learning, not on hardcoded actions.

1. (5 points) Explain bias-variance trade off with respect to a model.

**Ans:** Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you’re using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is error due to too much complexity in the learning algorithm you’re using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You’ll be carrying too much noise from your training data for your model to be very useful for your test data.

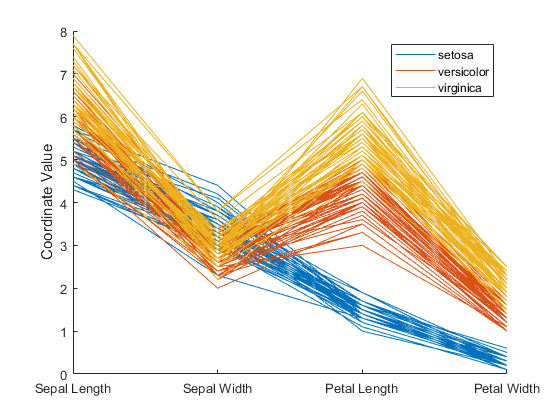
Essentially, if you make the model more complex and add more variables, you’ll lose bias but gain some variance — in order to get the optimally reduced amount of error, you’ll have to tradeoff bias and variance. You don’t want either high bias or high variance in your model.

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.

**Ans:** Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn’t, while Type II error means that you claim nothing is happening when in fact something is. Please refer to lecture 7 of ML for further clarification (Performance measures).

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?

**Ans:** We will pick Petal Length and Petal Width because they separate all classes more visibly than other two. In other words, we can put threshold that if Petal Length is greater than 5, it’s class virginica, if it’s value is between 3 and 5, it’s class versiclor and setosa otherwise.

Question 10 (5 points)

What is difference between inductive machine learning and deductive machine learning? Give an example for each.

**Ans**:  In inductive machine learning, the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn.  Let’s understand this with an example, for instance, if you have to explain to a kid that playing with fire can cause burns. There are two ways you can explain this to kids, you can show them training examples of various fire accidents or images with burnt people and label them as “Hazardous”. In this case the kid will learn with the help of examples and not play with fire. This is referred to as Inductive machine learning. The other way is to let your kid play with fire and wait to see what happens. If the kid gets a burn they will learn not to play with fire and whenever they come across fire, they will avoid going near it. This is referred to as deductive learning.

**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.**