# COMP 4745: Machine Learning

Instructor: Deepak Venugopal

Fall 2019: Homework 1

Due Dates: September 23, 2019 in dropbox on elearn

1. Can the following functions be represented using decision trees? If your answer is yes, draw the corresponding tree, if your answer is no, briefly state why. (10 points)
   - $A \wedge \neg B$
   - $A$ XOR $B$ (the XOR truth table is here `https://en.wikipedia.org/wiki/XOR_gate`)
2. What is the tree that a decision tree algorithm will learn for the following dataset to predict whether students like a restaurant or not. (10 points) (Hint: you don't need a calculator to do this)

| **Price** | **Fast** | **Distance** | **Like** |
|---|---|---|---|
| Low | No | Near | Yes |
| Low | Yes | Far | Yes |
| High | No | Near | No |

3. A decision tree learned from a dataset with $n$ features can have at-most $n$ nodes. Is this statement true or false? Briefly explain your reasoning. (10 points)
4. Suppose we have a classifier that classifies if an image contains a Human face or not. Suppose we have 100 images, 50 of which contain human faces. If our classifier accurately classifies that 30 images contains human faces, but at the same time wrongly classifies that 30 images contains human faces. What is the precision and recall. (10 points)
5. Suppose we use k-fold cross validation, how many times should we train the classifier? (10 points)
6. I have two machine learning algorithms, $A_1$ and $A_2$. $A_1$ has a training error of which is smaller than $A_2$'s training error? Can we conclude that $A_1$ is the better algorithm? Briefly explain your reasoning. (10 points)
7. Can the following functions be represented using perceptrons. If your answer is yes, compute the weights for the perceptron such that it can classify all instances of the functions correctly (you can do this without any complex math). Give a brief justification if your answer is No. (10 points)
   - $A \vee B$
   - $\neg A \vee B$
   - $A$ XOR $B$

8. If you have perceptron that is trained on a training dataset that is linearly separable (the positive and negative examples can be separated by a linear boundary), can we claim that the perceptron will give us 100% accuracy on any test dataset (not used in training) that is also linearly separable? Briefly explain your reasoning. (10 points)

9. Suppose your boss asks you design a ML algorithm for real-time prediction. Specifically, the requirement is that the ML algorithm needs to preform predictions very quickly. Can decision trees be used for such an application? Briefly explain your reasoning. (10 points)

10. (10 points) In this week's assignment, you will predict survivors of the Titanic disaster using the decision trees. The dataset is included as titanic-1.csv. The features are the $passenger-ticket-class$, $sex$, $age$, $number-of-siblings-in-ship$, $number-of-parents-in-ship$, $embarking-port$ and the label indicates whether they survived or not. Which was the most important feature for survival? (Run the decision tree and check which feature ends up on top more often). Draw a graph that shows avg. precision vs avg. recall (for 5-fold cross validation), for varying values of the pruning confidence parameter (e.g. 0.1,0.25,0.5 and 1). Did pruning greatly affect performance? (You can use either Weka J48 algorithm or Scikit-learn decisiontreeclassifier for this)

    With Weka, we can run ML algorithms from the GUI or through java/python programs. For Weka follow the below steps.

    (a) Download the software from `www.cs.waikato.ac.nz/ml/weka/`

    (b) Start the GUI by double-clicking on the Weka Icon

    (c) The GUI has many tabs. The first tab is used to load the data and pre-process it. The second tab runs the classification algorithms.

    (d) Load the data given as a .csv file. You should be able to see each feature in the GUI along with its summary statistics, etc.

    (e) Once you load the data, you will be able to go to the next tab and run classification algorithms. The algorithms are divided into groups such as trees (for decision trees type of algorithms), function (for algorithms such as logistic regression), etc.

    (f) Hit Start to run a specific ML method

    (g) The output will contain the error-metrics.

    If you want to use Scikit-learn, please install the anaconda environment (it will include all relevant packages needed for ML). `https://www.anaconda.com/distribution/` and write python code using the appropriate APIs.