# COMP 135 – Machine Learning – Fall 2016

## Homework Assignment 1

**Due date:** Wednesday, September 28 (hardcopy in class)

This assignment includes review questions for material from recent lectures.

1. (1) Explain the difference between supervised and unsupervised learning and give an example from each type to illustrate the difference. (2) Explain the difference between Classification problems and Regression problems, give an example from each type to illustrate the difference, and explain whether these are supervised or unsupervised learning problems.

2. Consider a dataset with $N$ examples which is potentially perfectly classified by nearest neighbors but where a random subset of $N/100$ examples has their labels corrupted. What would be a good strategy to avoid being misled by the wrong labels? Draw a 2D example of a possible scenario to explain your answer.

3. The test time of the nearest neighbors algorithm is expensive as we have to search for the neighbors which means scanning all the training set. Explain how this search time can be reduced. You do not need to give full algorithmic details but do provide enough details to explain the general idea.

4. Following the example in class, we are expecting to encode a sequence of colored marbles, each having one of 6 possible colors (call the colors A, B, C, D, E, F). The frequency of the colors in the sequence is A:0.22, B:0.05, C:0.13, D:0.15, E:0.31, F:0.14. (1) What is the entropy of this distribution of colors? (2) Show how to construct a Huffman code for the sequence, and calculate its average code length? Recall that the entropy is the best asymptotic one can hope for. Does the Huffman code achieve the rate promised by the entropy?

5. In class we discussed several splitting criteria for decision trees. (1) Give an example of dataset and splits (in terms of +/- numbers) where Accuracy is less sensitive than Information Gain (i.e., it cannot distinguish splits where Gain does). (2) Give an example of dataset and splits (in terms of +/- numbers) that shows a difference between Information Gain and the Gain Ratio. Please make sure to explain how each criterion is calculated and explain your answer.