Homework Assignment 1

Due date: Monday, September 25 (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. 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:0.04, B:0.1, C:0.1, D:0.16, E:0.25, F:0.35. (1) What is the entropy of this distribution of colors? note: Remember to use log base 2. (2) Show how to construct a Huffman code for the sequence (please show the tree and code), and calculate its average code length. (3) Recall that the entropy is the best asymptotic one can hope for. Does the Huffman code achieve the rate promised by the entropy?

4. In class we discussed several splitting criteria for decision trees. (1) Give an example of dataset and splits (in terms of +/- numbers before and after the split) where Accuracy cannot decide between two potential splits (because they have the same score) but Information Gain does have a clear winner. (2) Give an example of dataset and splits that shows a difference between Information Gain and the Gain Ratio. In both cases, please make sure to explain how each criterion is calculated and explain your answer.

5. In class we discussed two pruning algorithms for decision trees: REP and the heuristic used by J48. Briefly explain the idea behind each algorithm (2-3 sentences each) and then discuss advantages and disadvantages of each algorithm.