Assignment 3 – COMP 135 – Spring 2016

Due: Feb 16 **Turn in hardcopy in class**

1. 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. What is the entropy of this distribution of colors?

2. Ideally a good learning algorithm will be robust against simple manipulation of features. In this question we consider the Naive Bayes algorithm. In particular consider a dataset D1 with discrete features (let's say binary for simplicity) and consider a variant D2 where one of the features has been duplicated 100 times. Is the classifier produced by Naive Bayes identical when it learns on D1 and D2? If you answer Yes please explain your answer clearly. If you answer No explain and give an example where the prediction of the two classifiers will be different.

3-5: Consider the dataset (https://github.com/kephale/TuftsCOMP135_Spring2016/blob/gh-pages/Lecture06/notebooks/SnowDays.csv):

Previous morning	Previous day	Previous night	Early morning	Closed?
Light	Light	Light	Heavy	TRUE
Light	Light	Heavy	Light	TRUE
Heavy	Heavy	Light	Light	FALSE
Heavy	Medium	Medium	Light	FALSE
Medium	Medium	Medium	Medium	TRUE
Light	Light	Heavy	Heavy	TRUE
Light	Heavy	Heavy	Medium	TRUE
Heavy	Medium	Medium	Light	FALSE
Medium	Medium	Light	Light	FALSE
Light	Light	Light	Light	FALSE
Light	Light	Medium	Light	FALSE
Light	Light	Light	Medium	TRUE

3a. How deep is the non-pruned ID3 tree for this data? (show your work)
3b. What is the non-pruned ID3 tree for this data? (show your work)

4. What are the priors for the class values?

5. What are the naïve Bayes probabilities for TRUE/FALSE on this day: (show your work)

Previous morning	Previous day	Previous night	Early morning	Closed?
Medium	Medium	Heavy	Heavy	???