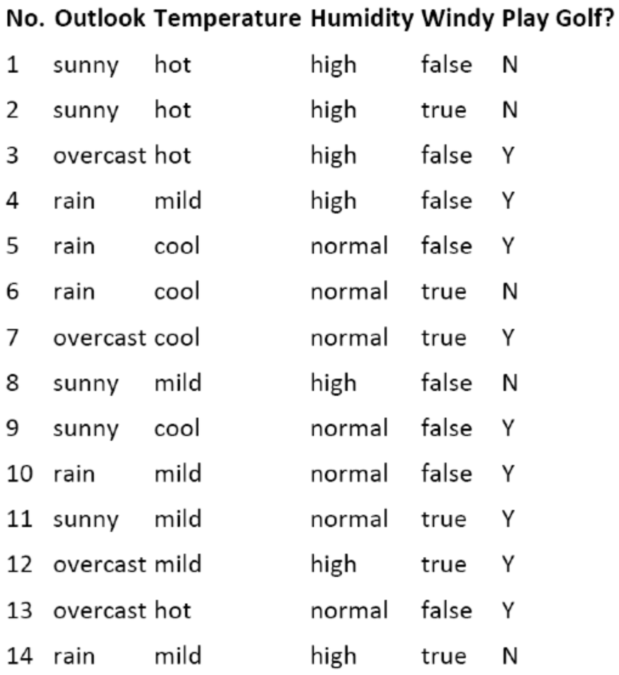
Fall 2022 CS165B: Introduction to Machine Learning – Homework 4

Due: **Wednesday, Nov 23rd, 11:59 pm PST**

Note: Please upload your written part as a PDF, and please upload your code as a separate file in (".py" or ".ipynb") form. Both files should be uploaded to Gauchospace.
NOTE: this homework will be shorter than previous ones and will be due on **WEDNESDAY**

1. (50 pts) Consider the following table of observations:



No.	Outlook	Temperature	Humidity	Windy	Play Golf?
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	Y
4	rain	mild	high	false	Y
5	rain	cool	normal	false	Y
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	Y
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	Y
10	rain	mild	normal	false	Y
11	sunny	mild	normal	true	Y
12	overcast	mild	high	true	Y
13	overcast	hot	normal	false	Y
14	rain	mild	high	true	N

From the classified examples in the above table, construct two decision trees (by hand) for the classification "Play Golf."

(a) (20 pts) For the first tree, use Temperature as the root node (this is a really bad choice.) Continue the construction of tree for the subsequent nodes using information gain. Remember that different attributes can be used in different branches on a given level of the tree.

(b) (30 pts) For the second tree, decide which attribute to use as the root node using the Decision Tree Learning algorithm described in class. It means that you need to calculate the information gains for splitting the data using each of the attributes you have and decide which one should be the root based on the discussion we have in the class. After you decide the root, rest will be similar to (a), where you choose the attribute with the highest information gain at each step.

Show your computations of information gain and draw the decision trees.

2. (20 pts) Please answer the question about Naive Bayes classifier below.

- State what is the simplifying assumption made by Naive Bayes classifier.

- Given a binary-class classification problem in which the class labels are binary, the dimension of feature is d, and each attribute can take k different values. Please provide the numbers of parameters to be estimated with AND without the simplifying assumption. Please explain your answer. Briefly justify why the simplifying assumption is necessary.

3. (30 pts) Assume we want to classify science texts into three categories - physics, biology and chemistry. The following probabilities have been estimated from analyzing a corpus of pre-classified web-pages gathered from Yahoo.

c	Physics	Biology	Chemistry
$p(c)$	0.35	0.40	0.25
$p(\text{atom}\|c)$	0.1	0.01	0.2
$p(\text{carbon}\|c)$	0.005	0.03	0.05
$p(\text{proton}\|c)$	0.05	0.001	0.05
$p(\text{life}\|c)$	0.001	0.1	0.008
$p(\text{earth}\|c)$	0.005	0.006	0.003

Assuming that the probability of each evidence word is independent of other word occurrences given the category of the text (Naive Bayes assumption), compute the (posterior) probability for each of the possible categories each of the following short texts; and based on that, their most likely classification. Assume that the categories are disjoint and exhaustive (i.e., every text is either physics, or biology or chemistry and no text can be more than one).

Please classify the documents A and document B given below. Assume that words are first stemmed to reduce them to their base form (e.g., atoms → atom) and ignore any words that are not in the table.

A : the carbon atom is the foundation of life on earth

B : the carbon atom contains 12 protons