COMP 472/6721 **Assignment 3** **Fall 2012**

Due date: November 7, 2012
Late submission: 20% per day
Teams: You can do the assignment individually or in teams of 2.
Teams must submit only 1 copy of the assignment.
Purpose: The purpose of this assignment is to make you practice basic machine learning techniques.

Question 1 (10pts): Probabilistic Reasoning

Assume that a fancy food-store sells wild hand-picked mushrooms from a local farmer. In the store, the mushrooms are labelled as *gourmet*, *good* or *at-your-own-risk*. The store always keeps the following inventory: 25% of its mushrooms are labeled *gourmet*, 50% are labeled *good*, and 25% are labeled *at-your-own-risk*. Mushrooms labeled as *gourmet* have a 5% chance of being poisonous, a *good* mushroom has a 15% chance of poisoning someone, and a *at-your-own-risk* mushroom has a 25% chance.
If Jim bought a mushroom from the store and was poisoned,

A. What is the probability that the mushroom had been labeled as *gourmet*?

B. What is the probability that the mushroom had been labeled as *good*?

C. What is the probability that the mushroom had been labeled as *at-your-own-risk*?

Show all your work.

Question 2 (50pts): Naïve Bayes Classification

Assume that Jim and Julia meet at a medical clinic, and after a short conversation, Jim writes down his phone number on a piece of paper. Jim was still feeling sick, so his handwriting was not very clear. Once he left, Julia tried to figure out the first 3digits of his phone number, but in vain... so she immediately called you, because she knew you were taking an AI class this term. You reassured her and told her that you may be able to figure out these digits as long as you have a big enough training set. So Julia asks her 10 friends from Facebook to write the digits from 0 to 9 on a 4x4 grid, and label each grid with its correct classification. For example, her friend Gina wrote:

You immediately took the 100 data instances and converted them into a training set of 16 features with values B (for black) and W (for white). For example the digits above are represented with the vectors:

f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	digit
W	B	B	W	W	W	B	W	W	W	B	W	W	W	B	W	1
W	B	B	W	W	W	W	B	W	W	B	W	W	B	B	B	2
W	B	W	B	W	B	B	B	W	W	W	B	W	W	W	B	4

The entire dataset that Julia collected is available as a text file on Moodle as the file `4by4digits.txt`.

You digitise Jim's hand written note as a 4x4 picture, and come up with the following 16 features for the first 3 digits.

f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	digit
B	B	B	B	B	B	W	W	W	W	B	W	B	B	B	W	?
B	B	B	W	B	B	W	W	W	W	B	W	W	B	B	W	?
B	B	B	B	W	W	B	W	W	B	W	W	B	W	W	W	?

Use a Naive Bayes classifier and the add 0.5 smoothing technique to identify the most probable first 3 digits of Jim's phone number (i.e. classify the 3 data instances above).

You can assume that each digit of the phone number is independent of the others.

You are free to use a script or a spreadsheet to help you in this task, but then you must submit your script/spreadsheet and your formulas. In any case, show <u>all</u> your work. In particular, for each of Jim's digit, indicate the score of each output class and how you computed it.

Question 3 (40pts): Decision Trees

Assume that, thanks to you, Julia was able to reach Jim and they decided to go to the zoo together. Unfortunately, the zoo keeper forgot to lock the animal cages and they started to run wild – endangering the visitors. To help the zoo keeper, Jim & Julia decided to try and capture the wild animals but they know that some of their friends have been hurt by wild animals before... they remember the following 8 friends who tried to capture animals before. Some got hurt, and others didn't depending on the features of the animals:

Friend	animal isBig	animal isSmelly	animal is of color	animal isDangerous
1	no	no	white	no
2	no	no	white	no
3	yes	yes	brown	no
4	yes	no	brown	yes
5	no	yes	white	yes
6	no	no	brown	yes
7	no	no	brown	yes
8	yes	yes	white	yes

A- What is the entropy of the class is "animal isDangerous"?

B- Which attribute should you choose as the root of the decision tree? Justify your answer.

C- Show the decision tree that ID3 would build to classify the animals.

D- Jim & Julia just spotted to following 3 animals and are considering trying to capture them. Classify these 3 animals using your decision tree above.

animal isBig	animal isSmelly	animal is of color	animal isDangerous
yes	yes	brown	?
no	yes	brown	?
yes	yes	white	?

Submission:

The assignment **must be handed electronically** by midnight on the due date.

1. Make sure that you have signed the expectation of originality form (available on the Web page; or at: http://www.encs.concordia.ca/documents/expectations.pdf) and given it to me.

2. In addition, write one of the following statements on your assignment:
   - For individual work: *"I certify that this submission is my original work and meets the Faculty's Expectations of Originality***",** with your signature, I.D. #, and the date.
   - For group work***:*** *"We certify that this submission is the original work of members of the group and meets the Faculty's Expectations of Originality"*, with the signatures and I.D. #s of all the team members and the date.

3. To hand in your assignment electronically:
   - Create one zip file, containing all files for your assignment.
   - Name your zip file this way:
     - For individual work: name the zip file: *a2_studentID*, where *studentID* is your ID number.
     - For group work***:*** name the zip file: *a2_studentID1_studentID2*, where *studentID1* and *studentID2* are the ID numbers of each student.
   - Upload your zip file at: https://fis.encs.concordia.ca/eas/