CS 662
Artificial Intelligence Programming
Homework #7
Various Topics
Due: Wednesday, 12/4, midnight

Summary
This assignment gives you the opportunity to practice working with the concepts we've covered since the second midterm.

Assignment:

1. **(8 points)** K-nearest neighbors problem:

   You are given a dataset, D = {(x1, 'yes'), (x2, 'no'), (x3, 'yes'), (x4, 'yes'), (x5, 'no'), (x6, 'yes')}

   The values of the similarity function, K, for a new point x0 are: K(x0,x1)=2, K(x0,x2)=1.5, K(x0,x3)=1.8, K(x0,x4)=2.3, K(x0,x5)=2.1, K(x0,x6)=1.7

   - What is the 1-NN classification of x0?
   - What is the 3-NN classification of x0?

2. **(12 points)** Naïve Bayes

   Imagine we are building a Naïve Bayes classifier to distinguish between two classes, A and B.

   We have 5 documents in class A, containing only the following words with frequencies as given: money (4 times), finance (1 time), stock (10 times), and market (6 times).

   We have 10 documents in class B, containing only the following words with frequencies as given: money (1 time), loss (20 times), finance (20 times), and gain (5 times)

   Using the Naïve Bayes assumption, determine whether it is more likely that the below document is of class A or class B. Assume that the conditional probability of a word given a category is its frequency within a category. If a word does not occur in a given category, assume its conditional probability is 1/10,000. Show and explain your work and any assumptions you make:

   Document: "money finance loss stock gain average"

3. **(49 points in all)** Sentiment analysis
   - Navigate to the following web page: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
   - Download the following dataset: **Pros and cons dataset** used in (Ganapathibhotla and Liu, Coling-2008)

- Download the Opinion lexicon: A list of positive and negative opinion words or sentiment words for English (around 6800 words).
- **(35 pts)** You will be writing a very simple sentiment classification system to run on the pros/cons dataset using the opinion lexicon. For each sentence in the text (delimited by <Pros></Pros> or <Cons></Cons>), count the number of positive and negative sentiment words, and classify the sentence as positive if there are more positive sentiment words, negative otherwise. What percentage of the Pros and Cons sentences are classified as positive and negative, respectively?
- **(14 pts)** Extend your system to handle negation in the manner discussed in class, using commas, periods, semicolons, and dashes as the punctuation delimiters between negated words and normal words. What is your sentiment classification accuracy now?
- For this part, turn in: your python program, named "sentiment.py", a README.txt for how to run your program, and the accuracies obtained for both Pros and Cons with and without negation handling (so four numbers). No need to do cross-validation since this isn't a learning algorithm. You can include your accuracies in the same document as the answers to all the other questions for this homework.

4. Utility & VPI
   - **(8 pts)** Assume we have two classifiers, C1, and C2 for filtering spam. C1 correctly classifies spam 85% of the time, but classifies ham as spam 8% of the time. C2 correctly classifies spam 70% of the time, but classifies ham as spam 2% of the time. Assume we have 1000 emails to process. If it costs us $0.05 each time a spam is misclassified as ham, and $1 each time a ham is misclassified as spam, what are the costs of using C1 and C2?
   - **(8 pts)** Given the following information, what is the value of knowing whether we are in good or bad economic conditions?
     - The value (utility) of buying an apartment building in good economic conditions is $50,000 and $30,000 in bad conditions
     - The value of buying an office building in good economic conditions is $100,000 and -$40,000 in bad conditions
     - The value of buying a warehouse in good economic conditions is $30,000 and $10,000 in bad conditions
     - The probability of being in good economic conditions is .6 and the probability of being in bad economic conditions is .4

5. MDP
   - **(6 pts)** For the 4x3 world shown in Fig 17.1 of R&N, calculate which squares can be reached by the action sequence [Up, Up, Right, Right, Right] and with what probabilities.
   - **(9 pts)** Below is a set of utilities for states in our simple grid world. Show the calculation that the value iteration algorithm would make for square 3. What is the updated utility? Show all your work.

1 -0.02	2 0.35	3 0.65	**+1**
4 -0.02		5 0.28	**-1**
6 -0.02	7 0.01	8 0.02	9 0.01

What to turn in:

- A text, Word, or PDF file with your answers to all 5 questions, “answers.xxx”; note that the program has associated questions that should be included in this document. Show all your work as relevant for full credit.
- The program and readme for question #3 (sentiment analysis)