CSC421 Assignment 3. Spring 2015 (10pts)

Misunderstanding of probability may be the greatest of all impediments to scientific literacy.

Gould, Stephen Jay

`Student Name:` `Student Number:`

`Instructor George Tzanetakis`

Question	Value	Mark
1	2	
2	3	
3	2	
4	3	
Total	10	

1 Overview

The goal of this assignment is to familiarize you with probability theory, Baysian Networks, and Machine Learning. Be aware that some of the questions require much more work than others. Your deliverable will be a report documenting your answers. **Clarity and correct use of mathematical notation will count in grading**.

You can use any programming language to implement the programming parts of the assignment. Don't hesitate to contact the instructor via email form with any questions/clarifications you might need.

Your answer should be a single PDF file submitted through ConneX.

2 Probability Theory (2pts)

(*Variation I on cars and goats*). The master of ceremonies confronts you with three closed doors, one of which hides the car of your dreams, new and shinny and desirable (running on fuel cells for the ecologically sensitive). Behind each of the other two doors, however, is standing a pleasant but not so shiny and somewhat smelly goat. You will choose a door and win whatever is behind it. You decide on a door and announce your choise, whereupon the host opens one of the other two doors and reveals a goat. He then asks you if you would like to switch your choice to the unopened door that you did not at first choose. Being unsure about what to do you decide to toss a fair coin. If the coin falls head you switch, if it fall tails you don't switch. Suppose you win the car. What is the probability you switched doors. **(1pt)**

(*Variation II on cars and goats*) You play the car-goat game two times in independent sessions. The first time you don't switch and the second time you do switch. What is the probability that you win two goats ? Two cars ? **(1pt)**

3 Email categorization (2pts)

Text categorization is the task of assigning a given document to one of a fixed set of categories, on the basis of text it contains. Naive Bayes models are often used for this task. In these models, the query variable is the document category, and the "effect" variables are the presence/absence of each word in the language; the assumption is that words occur independently in documents, with frequencies determined by document category.

- Explain precisely how such a model can be constructed, given as "training data" a set of documents that have been assigned to categories. Explain precisely how to categorize a new document.

  (1pt)

- In a programming language of your choice, implement an email categorization system based on the model you designed. Create training data using your own email (for example spam, school, friends). Implement a naive bayes categorization system with three classes (you can choose what they will be). Document in detail what practical issues you had

to solve. Use a dictionary of relevant keywords of your choosing with 12 words (4 for each class) and simply consider whether the word is present as a feature or not.**(1pt but lots of work)**

- Using the trained models from the previous questions generate 5 "sample" emails for each class (basically each randomly generated email will consists of randomly generated keywords from your dictionary).

4 Baysian Networks (2pts)

Choose any application area that you like and model it using a Baysian Network formulation. Your network should have a minimum of 6 discrete random variables some of which (minimum 2) are not boolean. Write down the complete Baysian Network (topology + conditional probability tables). Provide reasonable estimates of the probabilities for your application domain and describe how you could obtain these numbers in practice.

Formulate 4 queries involving both evidence events and hidden variables that would be meaningful for the Baysian Network you defined in the previous question (the queries should be more complex than simple atomic events). Show the calculation of probabilities for two of the queries using exact inference using enumeration and two of the queries using variable elimination. You don't need to complete the calculation just show clearly the process and how it relates to the CPTs you have.

5 Learning and Decision Trees (3pts)

Choose any application area that you like and create a training set of of attribute-value entries with discrete values and a positive/negative label (similar to the restaurant example in your book). Your data should have a minimum of 10 positive entries, 10 negative entries and each entry should have a minimum of 4 attributes. You can either get the values using your imagination and reasonable assumptions or you can obtain them from actual examples (for example by asking your friends questions or by observing the weather etc).

- Draw the full trivial decision tree based on the truth table. **(1pt)**

- Using the information-gain heuristic and your training set show how a simpler decision tree can be learned. **(1pt)**
- Write your training set as a Weka attribute file (.arff) and load it into the Weka explorer interface. Run the ID3 classifier and include the decision tree and evaluation output given by Weka in your report. `http://www.cs.waikato.ac.nz/ml/weka/` **(1pt)**

6 Deliverables

Your deliverable is a report with your answers to the questions. For the questions requiring programming you must include the source code and test cases and provide enough documentation to make it easy to understand and read.