CSC421 Assignment 2. Summer 2018 (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	5	
2	5	
Total	10	

1 Overview

The goal of this assignment is probabilistic modeling and statistical learning. Don't hesitate to contact the instructor via email or utilize the chat room of the ConneX course website for any questions/clarifications you might need.

The questions are marked as (*) necessary to pass, (**) expected and (***) exceptional.

IMPORTANT: YOU SHOULD ONLY SUBMIT A SINGLE PDF FILE WITH YOUR REPORT THROUGH CONNEX. ANY OTHER FORMAT WILL NOT BE ACCEPTED. THE ANSWERS SHOULD BE LABELED BY THE CORRESPONDING QUESTIONS NUMBERS

2 Probabilistic Simulation (5pts)

Consider simulating a game of Monopoly. Your goal is to experimentally determine the cummulative probability of landing on each square following the rules of the game after 100 moves. By cummulative I mean the total number of times you land on a square during a play. So for example if you land on the Free Parking square 5 times during the 100 moves in a particular run of the simulation then the probability of landing on that square is 5/100. To get accurate probabilities you will have to run 1000 simulated games.

You can assume that you are only a considering a single player and you ignore buying/selling property. You will need to simulate chance cards, rolling the dice including doubles, and going to jail. The following article has some information you can use in your simulation `http://www.businessinsider.com/math-monopoly-statistics-2013-6`.

Consider the problem of generating a random sample from a specified distribution on a single variable. You can assume that a random number generator is available that returns a random number uniformly distributed between 0 and 1. Let X be a discrete variable with P(X = xi) = pi for i 1, ..., k. The cummulative distribution of X gives the probability X x1, ..., xj for each possible j. Explain how to calculate the cummulative distribution in O(k) time and how to generate a single sample X from it.

Use this method to simulate playing the Monopoly game 1000 times and show on a table what is the probability of landing in each railway station, the GO square, Mediterenean Avenue and Boardwalk.

2.1 Questions

- Write code to generate random numbers corresponding to rolling a pair of dice and summing the ouput taking into account the doubling rule (when you roll doubles you get to roll again). Include all the code with comments describing what you are doing **(*) (1pt)**

- Write code for selecting a chance card at random based on the information provided in the article about the math of Monopoly **(*) (1pt)**

- Write code for simulating a game of Monopoly with a single player, go to jail, and chance cards. Record how many times you land on each square after playing a game consisting of 100 moves. Run the

simulation 1000 times and average the landing results. Show in a table the following probabilities: each railway station, the GO square, Mediterenean Avenue and Boardwalk. **(**) (2pt)**

For an A+ in this assignement undergraduate students can implement any of the following (implementing more than one is ok but will not give you additional points). Graduate students need to implement two out of four to get the A+ point. I am also open to other suggestions for extending the assignment **(***) (1pt)**

- Write code that visualizes the board and associated probabilities in color **(***) (1pt)**
- Implement code for handling the money aspect of monopoly and extend your simulation to handle multi-player playing by creating a very simple buying/selling AI for example randomly selecting one or the other action
- Implement a GUI for your game
- Implement a reasonably sophisticated Monopoly playing AI agent

3 Naive Bayes Text Classification

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 within a given category (condititional independence), with frequencies determined by document category. Download the following file: `http://www.cs.cornell.edu/People/pabo/movie-review-data/review_polarity.tar.gz` containing a dataset that has been used for text mining consisting of movie reviews classified into negative and positive. You will see that there are two folders for the positivie and negative category and they each contain multiple text files with the reviews. You can find more information about the dataset at: `http://www.cs.cornell.edu/People/pabo/movie-review-data/`.

Our goal will be to build a simple Naive Bayes classifier for this dataset. More complicated approaches using term frequency and inverse document frequency weighting and many more words are possible but the basic concepts are the same. The goal is to understand the whole process so **DO NOT** use existing machine learning packages but rather build the classifier from scratch.

Our feature vector representation for each text file will be simply a binary vector that shows which of the following words are present in the text file: **Awful Bad Boring Dull Effective Enjoyable Great Hilarious**. For example the text file cv996 11592.txt would be represented as (0, 0, 0, 0, 1, 0, 1, 0) because it contains Effective and Great but none of the other words.

3.1 Questions

- Write code that parses the text files and calculates the probabilities for each dictionary word given the review polarity **(*) (1pt)**
- Explain how these probability estimates can be combined to form a Naive Bayes classifier. You can look up Bernoulli Bayes model for this simple model where only presence/absence of a word is modeled. **(*) (1pt)**
- Calculate the classification accuracy and confusion matrix that you would obtain using the whole data set for both training and testing. **(**) (1pt)**
- Check the associated README file and see what convention is used for the 10-fold cross-validation. Calculate the classification accuracy and confusion matrix using the recommended 10-fold cross-validation. **(**) (1pt)**
- One can consider the Naive Bayes classifier a generative model that can generate binary feature vectors using the associated probabilities from the training data. The idea is similar to how we do direct sampling in Bayesian Networks and depends on generating random number from a discrete distribution. Describe how you would generate random movie reviews consisting solely of the words from the dictionary using your model. Show 5 examples of randomly generated positive reviews and 5 examples of randomly generated negative reviews. Each example

should consists of a subset of the words in the dictionary. Hint: use probabilities to generate both the presence and absence of a word **(***) (1pt)**

4 Grading

The submission should be a single PDF containing code snippets, text and figures with your answers in order. The questions are worth a total of 10 points. Grading will be based on the content of the answers as well as the quality of the report.