TRINITY COLLEGE DUBLIN

School of Computer Science and Statistics

Final Assignment 2019-20 STU33009: Statistical Methods for Computer Science

Submitting Your Report

- **Reports must be typed (no handwritten answers please) and submitted on Blackboard. Upload the dataset file to blackboard along with your report when submitting your assignment (please upload as separate files, not zipped).**
- **Reports should be no more than 5 pages in length including all plots (extra pages beyond this will be ignored when marking).**
- **You will need to use matlab to calculate values from the assignment dataset, or alternatively write a short program in python to do this. In either case give the code used as an appendix to the report (it doesn't count towards the page limit), but please keep the code short.**
- **In order to obtain full credit it is essential that you explain/justify how you obtained your results and, where appropriate, that you critically reflect upon them. Simply giving raw numbers as answers will receive few marks as will saying "see code for details" and the like, even if the code contains explanatory comments.**
- **It is mandatory to complete the declaration that the work is entirely your own and you have not collaborated with anyone - the declaration form is available on Blackboard.**

Downloading Dataset

- Download the assignment dataset from `https://www.scss.tcd.ie/doug.leith/ST3009/final2020.php`. Important: You must fetch your own copy of the dataset, do not use the dataset downloaded by someone else.
- The data file consists of columns of data. Each column corresponds to one question in an exam, each row to one student. The value in row i, column j is the mark awarded to the i'th student for their answer to question j.

Assignment

1. An exam consists of three questions. The three questions are constructed by selecting three topics (one per question) uniformly at random from a set of 10 possible topics, without replacement. Each question is worth 33% and the pass mark is 40%. I am planning my study strategy for the exam. Suppose I study only n out of the 10 possible topics.

(a) How many combinations (ignoring order) of three topics can an exam have ? Explain your answer. [5 marks]

aCb = combination -> unordered
aPb = permutation -> ordered

(b) Derive an expression for the probability that none of the n topics I studied come up in the exam. Explain your answer. [5 marks]

(c) Derive an expression for the probability that I fail the exam, i.e. fewer than 2 of the n topics I studied come up in the exam, and plot the value of the probability vs n. Explain your answer. [5 marks]

(d) Suppose the exam changes to consist of 4 questions, each worth 33% (so only 3 need to be answered correctly to get full marks, and 2 to pass). Derive an expression for the probability that I fail now, and plot vs n. Discuss with reference to part (c). [5 marks]

(e) Write a short stochastic simulation of the exam setup with three questions. The simulation takes n as input and should draw the three exam questions without replacement from 10 topics and output a random variable X which takes value 1 when the student passes the exam (i.e. when enough of the topics they have studied come up in the exam) and 0 otherwise. [5 marks]

(f) Extend your simulation so that it runs N times and outputs the empirical mean $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$ where X_i is the output of the simulation on the i'th run. This mean is an estimate of the probability the student passes the exam. Using the CLT derive an expression for a 95% confidence interval for Y in terms of $E[X_i]$, $var(X_i)$ and N. Your answer from part (c) gives the value of $1-E[X_i]$ vs n, using this calculate $E[X_i]$, $var(X_i)$ when $n = 7$ and so obtain 95% confidence intervals for Y for $N = 1000$ and $N = 10,000$. [5 marks]

(g) Now run your extended simulation multiple times with $N = 1000$. It is up to you to decide how many times to run it, but you should justify your choice. Using the sequence of Y values thus generated estimate how frequently these values lie within the confidence interval you calculated in part (f). Repeat for $N = 10,000$. Discuss. [5 marks]

(h) Suppose now that the topics that come up in an exam are more predictable i.e. if a topic came up in last years exam it is more likely to come up in this years exam. How might a student modify their study strategy to exploit this? Modify your simulation to model these changes. Using your simulation to evaluate the effect of these changes in the exam setup discuss how the probability of passing changes as the exam becomes more predictable and students adapt their study strategy accordingly. Similarly, discuss what happens when the probability that a topic comes up in this years exam *decreases* when it came up in last years exam and what happens to the probability of passing if a student applies a strategy that assumes topics are predictable but in fact they are chosen uniformly at random. [20 marks]

2. The assignment dataset contains the marks for student answers to a set of exam questions. Please cut and paste the first line of the data file (which begins with a #) into your report as it identifies your dataset. Your task is to use this data to investigate whether the level of difficulty of each of the questions is similar or not.

(a) Using the data for each question estimate the PMF of the mark and plot it. What does this data suggest regarding the relative difficulty of the questions? Discuss the

strengths and weaknesses of using this approach to evaluate relative difficulty of the questions. [5 marks]

Students can be expected to differ in ability and preparation. To control for this we can use one question as a baseline and compare the marks for the other questions relative to this.

(b) For each question estimate the mean mark conditioned on the student mark for question one, and also estimate the variance conditioned on the student mark for question one. If you judge it necessary you can lump the question one marks together into bins e.g. assign the mark to one of the ranges 0-10, 10-20, 20-30 etc and calculate the mean mark for each question conditioned on whether the question one mark falls into a particular range. If you decide to use such binning explain why and discuss the pros and cons of different choices of bin size. [10 marks]

(c) Using the values you calculated in part (b), for each question calculate a confidence interval for the mark conditioned on the mark for question one. Plot the confidence interval vs the mark for question one, e.g. as a series of vertical error bars. Using your results discuss the relative difficulty of the questions. [10 marks]

(d) Describe how you could use linear regression to predict the mark for a question given the student mark for question one. Hint: the mark for question one is the input feature and the dataset is the training data. Based on your calculations from parts (b) and (c) discuss whether the assumption in linear regression that the mean is a linear function of the features is appropriate. If not, discuss how the features used in the regression model might be modified to better fit the data. Discuss whether the type of randomness assumed in linear regression is appropriate or not for exam marks. [10 marks]

Another way to try to account for differences in student ability and preparation is to suppose student i has a "strength" S_i and question j has a "difficulty" D_j such that the mean mark X_{ij} for student i and question j is $S_i - D_j$ and marks are normally distributed about this mean.

(e) Write an expression for the PDF of X_{ij} conditioned on S_i and D_j. Using this, write an expression for the log-likelihood of the dataset assuming the marks of each student and each question are independent. [5 marks]

(f) Outline how gradient descent might be used to find the maximum likelihood estimates for the unknown parameters S_i and D_j. Note: there is no need to calculate the gradient of the log-likelihood. [5 marks]