CS37300 Homework — Homework 1— Last update: August 28, 2017

https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2017 Due: Sept 6, 2017 (11:59pm) [Due date changed]

Instructions and Policy: Each student should write up their own solutions independently. You need to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.

YOU MUST INCLUDE YOUR NAME IN THE HOMEWORK

The answers (without the python scripts) must be in submitted in a PDF via Blackboard. The python scripts will be submitted separately (CODE SUBMISSION INSTRUCTIONS WILL BE GIVEN SOON). Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.

There are FOUR (4) questions in this homework.

Python code guidelines

Turn in each question of your homework in a separate python file named `hw1-X.py`, where X is the question number. Please use `print("QY (Z) (K)")` to MARK the beginning of the answer to question QY, item Z, subitem K in file `hw1-X.py`.

Example:

```
print("Q2 (1) (a)")
# code for Question 2 item (1) (a)
...
print("Q2 (1) (b)")
# code for Question 2 item (1) (b)
...
print("Q2 (2) (a)")
# code for Question 2 item (1) (b)
...
```

Your code is REQUIRED to run on either Python 2 or Python 3 at scholar.rcac.purdue.edu. Preferably use Python 3 (Python 2 will also be accepted). The TA's will help you with the use of the scholar cluster. If the name of the executable is incorrect, it won't be graded. Please make sure you didn't use any library/source explicitly forbiden to use. If such library/source code is used, you will get 0 pt for the coding part of the assignment. If your code doesn't run on scholar.rcac.purdue.edu, then even if it compiles in another computer, your code will still be considered not-running and the respective part of the assignment will receive 0 pt.

Q1 (3 pts): Random Variables and Probability

1. **(6 pts)** Two standard dice are rolled. Let E be the event that the sum of the dice is odd; let F be the event that at least one of the dice lands on 1; and let G be the event that the sum is 5. And let Ω be all possible events. Compute the following:

   (a) P($E \cap F$)

   (b) P($E \cup F$)

   (c) P($F \cap G$)

   (d) P($E \cap (\Omega \backslash F)$)

   (e) P($E \cap F \cap G$)

2. **(6 pts)** (Bayes' rule) Peter is worried about a dangerous disease he saw on the news. Only 5 people in every 1000 will have this disease. Peter takes a test to tell if he has the illness or not. This test correctly gives a positive for 95% of cases, if the person has the disease. If the person doesn't have the disease, the test correctly gives a negative result with probability of 95%. What is the probability that Peter has the disease, given that he tested positive for it?

3. **(6 pts)** (Joint and Conditional Probabilities) A system is built using 3 disks d_1, d_2, d_3 having probabilities of failure 0.01, 0.03 and 0.05 respectively. Suppose the disks fail independently.

   (a) Let E denote the event of loss of data, which occurs only if two or more disks fail. Compute $P(E)$, the probability of loss of data.

   (b) Instead, let F denote the event that at least one of the following happens: (i) d_1 fails; (ii) d_2 and d_3 both fail. If loss of data only occurs when event F occurs, then what is the probability that there is loss of data?

   (c) Considering the setting of 3b, given that d_3 has failed, what is the conditional probability that event F will occur and there will be loss of data?

4. **(6 pts)** (Independence)

   (a) Suppose that E, F and G are independent events. Prove that

   $$P[E \cap (F \cup G)] = P(E)P(F \cup G)$$

   (b) Let A and B be independent events, and Ω the set of all events, then prove that $\Omega \backslash A$ and B are also independent.

   (c) Let X be the random variable that counts the number of heads when two fair coins are flipped. Let Y be the random variable that records whether both coin flips are the same (i.e., 1 if both heads or both tails, 0 otherwise). Are X and Y independent? Show why or why not.

5. **(6 pts)** (Linearity of Expectation) Show that the expected value of the sum of two random variables is the sum of their expected values. That is, prove that if X and Y are random variables defined on sample space S, then $E[X + Y] = E[X] + E[Y]$.

Q2 (2.5 pts): Working with Python.

File Description: For this assignment, your are provided with comma separated values file `https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2017/data/retail.csv`. It has the following attributes:

1. InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.
2. StockCode: Product (item) code. Uniquely assigned to each distinct product.
3. Description: Product (item) name.
4. Quantity: The quantities of each product (item) per transaction. Numeric.
5. InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. UnitPrice: Unit price. Numeric, Product price per unit in dollars.
7. CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. Country: Country name. Nominal, the name of the country where each customer resides.

1. **I/O: (0.5pt)** Write a python script that loads the file into memory and answers the following questions. Keep in mind to take care of the data types. We suggest to open the file once and just have a quick look at it, so you know what you are dealing with.
   (a) How many rows does the dataset have (not including the header line)? (write the code, report result in PDF)
   Hint: Use dictionaries to store each product by their unique ids. This dictionary can be used to answer the remaining questions without having to open the file again.
   (b) What is the number of unique items being sold by the retail company? Each item has a unique stock code. (write the code, report result in PDF)
2. **Data Processing: (1.5pt)** Using the data from the previous question (1), answer the following:
   (a) What is the average unit cost for the product with stock code **20685**. (write the code, report result in PDF)
   (b) Which hour in the day are most items sold in the given data set? Only consider the quantity, and not the amount spent. Use the invoice date to answer this. (write the code, report result in PDF)
   (c) Create a bar graph showing the amount spent by residents of each country(as long as the amount spent is more than $50,000). To do this, we suggest creating a python dictionary with keys as the name of the country and the values be the total amount spent. You can then filter this to only select countries whose residents spent more than $50,000. (write the code, show the plot in the PDF)
3. **Writing: (0.5pt)** Split the dataset randomly into two equal parts. Write the resulting into two csv files, namely output-1.csv and output-2.csv. (write the code, nothing goes in the PDF)
   Hint: Use `np.random.shuffle` to shuffle numpy arrays (remember to shuffle only rows, not columns) before splitting the data. Use `random.shuffle` if shuffling lists (requires `import random` at the beginning of the file). Dictionaries cannot be shuffled, but you can create a list of the dictionary keys and shuffle the list.

Submission: You're required to create a single python script, `hw1-2.py`. DO NOT include the source code in your submitted PDF. You also need to answer the questions asked above (except for the graph) in the PDF you submit on Blackboard.
You're only allowed to use the following libraries: csv, numpy, matplotlib, random, datetime

Q3 (2.5 pts): (Working with Matrices in Python) Download the Marvel comic book data[1] `https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2017/data/Marvel-data.tar.gz` containing the following files:

- marvel.txt: characterID (1st column) and comicID (2nd column).
- marvelCharacters.txt: charcterID and the name of the character.
- marvelComicBooks.txt: comicID and the name of the comic book.

All items of this question should be written in a single python script named `hw1-3.py`

(a) (1pt) Plot the empirical complementary cumulative distribution (ECCDF) of comic characters appearances in comic books. The ECCDF $P[X > x]$ is defined as the fraction of characters with more than x comic book appearances. For instance, if there are 6486 characters but only 10 of these appear in a 1001 or more comic books, then $P[X > 1000] = 10/6486$. (write the code, report the plot in the PDF)

IMPORTANT: Your plot should be in log-log scale (see matlibplot_example.py for an example).
Include the plot in your PDF (remember to label your axis).

Example of a bare-bones python code to plot in log-log scale (CCDF of geometric distribution):
`https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2017/hw/hw1/matlibplot_example.py`

(b) (1.5pt)

(i) In how many distinct comic books does the character "QUILL" appear? (write the code, report result in the PDF)

(ii) Find a comic book (comic book name) that has the most number of characters (if there is more than one, just give one of them). (write the code, report result in the PDF)

(iii) Let A be a numpy 2D array (matrix), where $A[i, j] = 1$ if character i appears on comic book j. $A[i, j] = 0$ otherwise (you can use np.zeroes to create this matrix). Let A^T be the transpose of matrix A. In your python script, compute and print $W = AA^T$. (write the code that prints W. Nothing goes on the PDF)

[1] Courtesy of Jay-Yoon Lee, Cesc Rossello, Ricardo Alberich, and Joe Miro. Reference `http://bioinfo.uib.es/~joemiro/marvel.html`

Q4 (2 pts): Write a python script hw1-4.py that creates a 100 by 20 matrix, X as follows

$$X_{i,j} = \begin{cases} 2 * i + j^2 + 1 & \text{if } i \leq j, \\ i^2 - 2 * j & \text{if } i > j \end{cases}$$

It also creates a 100-dimensional vector y, with

$$y_i = i^2 - 1$$

(a) (1pt) In this python script, compute the vector

$$b = (X^T X)^{-1} X^T y$$

and provide the value of b in the PDF.

(b) (1pt) In the same python script, also compute the inner product between the first row of X and the vector b created in the previous question. Report the result in the PDF.

Submit hw1-4.py for correction as a separate file (NOT in the PDF).