COMP1433: Introduction to Data Analytics
COMP1003: Statistical Tools and Applications

Assignment

PolyU Spring 2022

Answers Submission Due: 23:59, Apr 7, 2022.

Important Notes.

- This is an *individual* assessment. So, no discussion (in any forms) is allowed among classmates.
- Following the syllabus, only the R language is allowed for answering the programming questions.
- Please submit the compressed folder (in zip or rar) with all the answers. Please also name the folder with your student ID, such as "21123456D.zip" or ""21123456D.rar".
- In the compressed folder, for non-coding questions, please put your answers and detailed problem solving steps into a report in PDF version (to avoid messy formula). For the coding questions with the index i below, you may create a folder named as "Qi" (e.g., Q1), which contains the codes for Q_i and the readme.txt file indicating how to run the code.
- It is highly recommended that the codes are well commented for easy reading and with clear indication in the comments which part is for which subquestion (if any). It is for the case that the implementation is imperfect (with bugs) and we need to somehow find scores from the codes to see if your algorithm is designed in a correct way.
- The compressed folder should be submitted to the *blackboard*.[1] The full mark is 100' and submission entry is: Assessments/Assignment. For Question 2 and 3 below, we have provided the input data in the form of the

[1] learn.polyu.edu.hk

compressed folder “Assignment_Data” available in the same entry. You’ll find two folders there, one named as Q2 and the other Q3, for the input of Question 2 and 3 respectively.

- No late submission is allowed and don’t forget to double check if the submission is saved successfully before leaving.
- When handling the paths for file loading and saving, please use relative paths for the TAs to run your codes easily in a different environment. It can be assumed that the input data file is put in the same folder as the codes used to read the data.
- It is assumed that the external libraries “tokenizers” and “ggplot2” have already been installed in the R system. For any other external libraries you need to use, please indicate them in the readme.txt file.
- Last but not least, best of the luck for this assignment! :)

Question 1. In the Harry Potter world, each Chocolate Frog has a card inside it for one to collect famous witches and wizards (a.k.a., *Chocolate Frog Card*). In Harry’s first journey on the Hogwarts Express, he got his very first Chocolate Frog Card, which is *Albus Dumbledore*. Assume that there are N possible wizards or witches to appear in the Chocolate Frog Cards and each unique card exhibits a equal chance ($\frac{1}{N}$) to be drawn in a Chocolate Frog.

(a) [**Non-coding Question**] Harry collected another X cards till he got another Dumbledore (i.e., Dumbledore is the X-th card while none of the previous $X-1$ cards is a Dumbledore). If there are 3 possible unique Chocolate Frog cards ($N=3$), what is the expected value of X (i.e., $\mathbb{E}[X]$)? (10’)

(b) [**Non-coding Question**] If we didn’t set N to an exact number, please generalize the solution of (a) to represent $\mathbb{E}[X]$ using a formula with N. (10’)

(c) [**Coding Question**] On the Hogwarts Express, Ron Weasley, one of Harry’s best friends, told Harry that he collected 500 cards while he has not yet got *Agrippa* or *Ptolemy*. Assume that $N=100$, write the R codes to run a Monte Carlo Simulation to estimate the probability that one has 500 cards (with possible replications) while neither *Agrippa* nor *Ptolemy* is included. (10’)

Question 2. [**Coding Questions**] Sentiment classification for tweets are very helpful for people to understand the public opinions from users. Here, we will implement a Naive Bayes classifier to classify the sentiment polarity of the tweets. In the class, we discussed the case with only positive and negative polarity, you can generalize that to the classification problem involving three classes — positive, negative, and neutral, where neutral means that the tweet does not exhibit subjective polarity of positive or negative, such as "*I did the laundry on the weekend.*", which tends to be objective. We've released the training data in "Q2/train_tweets.txt". Each tweet record is presented in a line, separated into "tweetID", "userID", "sentiment" (positive, neutral, or negative), and "text" by tab ($\backslash t$).

(a) Read in the training data and create a dataframe named as "tweets" with four columns to hold the data. Each row corresponds to a tweet record and each column indicates an attribute. Please also name the columns accordingly as "tweetID", "userID", "sentiment", and "text". (10')

(b) Employ an external R library "tokenizers" and tokenize the text in each tweet with the function "tokenize_words(·)".[2] Create the fifth column of the "tweets" dataframe and name the column as "tokens", which should carry the tokenized text for each tweet record. For the i-th row of the "tweets" dataframe, the "text" column presents the raw text of the i-th tweet while the "tokens" column presents its tokens obtained from the "tokenize_words(·)" function. (10')

(c) Create a new vector named as "vocab" to maintain the distinct tokens appearing in > 3 tweets from the training data. In other words, each token in the "vocab" vector should appear in > 3 tweets measured by the "tokens" column of "tweets" dataframe and there is no replicated tokens in the "vocab" vector. Print the count of tokens in the "vocab" vector on the screen. (10')

(d) Build a Naive Bayes classifier to measure the likelihood of observing each words conditioned on "positive", "neutral", and "negative" sentiment and the prior of observing each sentiment in the training data. Use your model to predict the sentiment label of the following three tweets and print the results on the screen:

$\langle i \rangle$ "I love the banner that was unfurled in the United end last night. It read: Chelsea - Standing up against racism since Sunday"

$\langle ii \rangle$ "So Clattenburg's alleged racism may mean end of his career; Terry, Suarez, Rio use it and can't play for a couple of weeks?"

$\langle iii \rangle$ "In our busy lives in Dubai could we just spare a moment of silence this Friday morning for the people who still wear crocs."

Please be reminded to tokenize the text of the above three tweets before predicting their sentiment. Use "add-1" smoothing for the likehood measure and map the probabilities into the log space when building and applying the model. (20')

Question 3. [**Coding Questions**] We learned the K-means clustering methods in the class. You can implement the algorithm from scratch (without using external

[2] Please refer to the page here for more details: `https://cran.r-project.org/web/packages/tokenizers/vignettes/introduction-to-tokenizers.html`

packages or libraries) and group the points into 3 clusters. The input points are put in a file named as “Q3/points.txt”: each line shows a 2D vector (point) where the first and second entry are separated by the comma “,”. After obtaining the clustering results, it is required to generate a figure to visualize the results. In the visualization, you can first draw a scatter plot with the input vectors (x-axis corresponding to the first entry while y-axis the second entry), and then color each point in red, green, and blue, indicating the cluster it is assigned to.

In the algorithm implementation, no external package or library should be used, while for visualization purpose, you can employ any graphics or visualization libraries you want (e.g., “ggplot2”). For the initialization, the representatives (centroids) of the three classes are $(40,40)$, $(100,0)$, and $(0,100)$, respectively. Please visualize the clustering results after 1,000 iterations. (20’)