CS373 Homework 1

Due date: Monday February 11, 11:59pm (submit pdf on Gradescope)
Any use of late days must be explicitly mentioned at the top of your submission.

Homework must be submitted as a PDF; answers should be typed.

Instructions for submission
Submit a single PDF on Gradescope with all your answers. Make sure you select the page corresponding to the beginning of each answer, else points might be deducted. For part I, show the steps you took. For part II, include the R code you used for analysis, along with its output and any plots required by the question. Please label all plots with the question number. Your homework must be typed and must contain your name and Purdue ID.

1 Part I: Basic Probability and Statistics

A. **(6 pts)** Consider an experiment where a fair die is rolled repeatedly until the first time a 3 is observed.

   (a) What is the sample space for this experiment? What is the probability that the die turns up a 3 after i rolls?

   (b) What is the expected number of times we roll the die?

   (c) Let E be the event that the first time a 3 turns up is after an even number of rolls. What set of outcomes belong to this event? What is the probability that E occurs?

B. **(5 pts)** Two standard dice are rolled. Let E be the event that at least one of the dice lands on 5; let F be the event that the sum of the dice is even; and let G be the event that the sum is 7. Compute the following:

   (a) $\mathrm{P}(E \cap F)$

   (b) $\mathrm{P}(E \cup F)$

   (c) $\mathrm{P}(E \cup G)$

   (d) $\mathrm{P}(E \cap \neg G)$

   (e) $\mathrm{P}(E \cup F \cup G)$

C. **(4 pts)** We are given four coins C_1, C_2, C_3 and C_4. Coin C_i is chosen randomly with probability proportional to i for $i = 1, 2, 3, 4$. Let H_i represent the event that heads is observed when C_i is tossed; $P(H_1) = 1/4; P(H_2) = 1/3; P(H_3) = 1/2; P(H_4) = 2/3$.

   (a) What is the probability of selecting coin C_3?

   (b) If a coin is selected at random and tossed, find the conditional probability that the coin is C_3 given that a tail is observed. (State this as a conditional probability and show the calculation.)

D. **(4 pts)** 52% of the students at a particular college are female. 5% of the students in the college are majoring in computer science. 0.55% of the students are women majoring in computer science.

(a) If a student is selected at random, find the conditional probability that the student is female given that they are majoring in computer science. (State this as a conditional probability and show the calculation.)

(b) If a student is selected at random, find the conditional probability that the student is majoring in computer science given that they are female. (State this as a conditional probability and show the calculation.)

(c) Now suppose that the overall proportion of female students increases to 57% and that the conditional probability from Da changes (i.e., increases or decreases) to 15%. Compute the updated conditional probability that a student is majoring in computer science given that they are female. (Assume that the overall proportion of students majoring in CS stays the same.)

E. **(6 pts)** 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 W denote the event that the system will work, which happens if at least one of the disks works. Compute $P(W)$, the probability that the system will work.

(b) Let A denote the event that at least one of the following happens: (i) d_1 and d_2 work; (ii) d_3 works. If the system works when event A occurs, then compute the probability that the system will work.

(c) Considering the setting of Eb, given that d_1 works, what is the conditional probability that event A will occur and the system works?

F. **(6 pts)** Let an experiment consist of rolling three standard 6-sided dice.

(a) Compute the expected value of the sum of the rolls.

(b) Compute the variance of the sum of the rolls.

(c) If X represents the maximum value that appears in the two rolls, what is the expected value of X?

2 Part II: R

In this assignment, you will use the R statistical package to explore, transform, and analyze data. Based on your analysis you will formulate hypotheses about the data. To get started, do the following:

- Download and install R from: `http://cran.r-project.org/`
- Download the Yelp dataset from Piazza.
  This data set is part of the Yelp academic dataset and consists of data about 24,813 restaurants. The datafile *yelp.csv* contains 28 attributes: 6 numeric and 22 discrete. The first row of the data file is a header row with the names of the attributes where names are separated by a comma (,).

Use R to analyze the Yelp data and complete the questions below.

3 Data import and summarization

Read the data into R using `read.table()` function. Use the argument `sep=","` to specify the column delimiter, the argument `header=TRUE` to read in the column names, the argument `quote="\""` to read in the quoted fields, and the argument `comment.char=""` to treat the # characters as text rather than comments.

(a) **(2 pts)** Print a summary of the data using the `summary()` function.

(b) **(2 pts)** Print the names of the columns in the table using the `names()` function.

4 1D plots

A. (a) **(3 pts)** Plot a histogram of the *'reviewCount'* attribute. Use the `hist()` function with its default values and make sure to title the plot with the name of the attribute for clarity.

(b) **(3 pts)** Compute the logged values for *'reviewCount'* (you can use `log(d$column_name)` to compute the log of all the values in a column). Plot a histogram of the logged values.

(c) **(3 pts)** Plot a density plot of the logged values of the *'reviewCount'* attribute using the `density()` function.

(d) **(3 pts)** Discuss the similarities and differences between the three plots and the information they convey about the distribution of *'reviewCount'* values in the data.

B. **(2 pts)** Plot a barplot of the *'state'* attribute to show the frequency of each value. Use the `table()` function to get the counts for each value and the `names()` function to get the names of the values in the table. Use the `barplot()` function with the `names.arg` argument to label the bars with the appropriate value. Again, make sure to title the plot with the name of the attribute for clarity.

(Note that this will look like a histogram but for nominal values. In small renderings of this plot, you might not see all the state name labels, but if you stretch the window you will be able to see all the labels.)

5 Sampling and transforming data

A. **(4 pts)** The attributes *'categories'* and *'recommendedFor'* each contain a comma separated list of values associated with each restaurant. Compute two new boolean features: 'servesPizza' and 'goodForBreakfast' with a value of `TRUE` if the list contains *Pizza* (in *'categories'*), *breakfast* (in *'recommendedFor'*) respectively and `FALSE` otherwise. You can use the function `grepl(str, f$column_name)` to check whether the values in *column_name* contain the string *str*.

Append the two new columns to the original data frame, using `cbind()` to increase the number of features by 2.

B. (a) **(3 pts)** Compute the quantiles (using `quantile()`) for the *'checkins'* attribute.

(b) **(3 pts)** Select a subset of the data with *'checkins'* value $\leq$ the 1st quartile (25th percentile). You can use `subset()` or select from the data frame with `[ ]` operations.

(c) **(3 pts)** Print a summary of the above subset for the following attributes only: *'checkins'*, *'stars'*, *'noiseLevel*, *'priceRange'*, *'reviewCount"*, *'goodForGroups'*, and compare them to their summary for the full dataset.

Discuss any differences that you find in the distributions of these attributes.

6 2D plots and correlations

A. **(6 pts)** Plot a scatterplot matrix (using `plot()`) for the five attributes: *'stars'*, *'reviewCount'*, *'checkins'*, *'longitude'*, *'latitude'*.

Identify which pair of attributes exhibit the most association (as you can determine visually) and discuss if this is interesting or expected, given your domain knowledge.

B. **(6 pts)** Calculate the pairwise correlation among the above five attributes using the `cor()` function.

Identify the pair of attributes with largest positive correlation and the pair with largest negative correlation. Report the correlations and discuss how it matches with your visual assessment in part A.

C. **(6 pts)** Plot a boxplot (using `boxplot()`) for each of the following four attributes vs. the *'attire'* attribute: *'checkins'*, *'reviewCount'*, *'stars'*, *'latitude'*.
Make sure to label both axes of the plot with the appropriate attribute names.

(a) Identify the attribute that exhibits the most association with *'attire'* (as you can determine visually) and discuss whether this is interesting or expected, given your domain knowledge.

(b) For the attribute identified above, calculate its interquartile range for each value of *'attire'* (i.e., a separate IQR for the "casual" instances, the "dressy" instances, the "formal" instances and the instances with " " for *'attire'*). You can do this with the `subset()` and `quantile()` functions. Calculate the overlap between the four IQRs. Discuss whether these results support the conclusion you made based on visual inspection.

7 Identifying potential hypotheses (20 pts)

During your exploration above, investigate other aspects of the data. Explore relationships between variables by assessing plots, computing correlation, or other numerical analysis.

Identify TWO possible relationships in the data (other than the ones specified in earlier questions) and formulate hypotheses based on the observed data. For each of the two identified relationships:

(a) Include a plot illustrating the observed relationship (between at least two variables).

(b) State whether the variables are discrete or continuous and what type of plot is relevant for comparing these two types of variables.

(c) Formulate a hypothesis about the observed relationship as a function of two random variables (e.g., X is associated with Y).

(d) Write the hypothesis as a claim in English, relating it to the attributes in the data.

(e) Identify the type of hypothesis.