CS373 Homework 1

Due date: Thursday, September 19, 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)** Suppose there is a basket containing one apple and two oranges. A student randomly pick one fruit from the basket untile the first time the apple is picked. (Sampling with replacement)

   (a) What is the sample space for this experiment? What is the probability that the student pick the apple after i tosses?

   (b) What is the expected number of times the students need to pick the apple?

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

B. **(5 pts)** Two standard dice are rolled. Let A be the event that at least one of the dice lands on 5; let B be the event that the sum of the dice is even; and let C be the event that the sum is greater than or equal to than 9. Compute the following:

   (a) $\mathrm{P}(A \cap B)$

   (b) $\mathrm{P}(A \cup \neg B)$

   (c) $\mathrm{P}(A \cap C)$

   (d) $\mathrm{P}(A \cup \neg C)$

   (e) $\mathrm{P}(A \cup B \cup C)$

C. **(4 pts)** There are six dice in the bag: one red die, two yellow dice and three blue dice. A student randomly pick one dice from the bag (with replacement):

   (a) What is the probability of selecting a die in yellow color?

   (b) If a die is selected at random and tossed, find the conditional probability that the die is yellow given that 6 is observed.

D. **(4 pts)** Suppose in a high school, 3% of the students come to Purdue University, 45% of the students in a high school are female, 0.55% of the students are girls coming to

Purdue University. Let A be the event that a student from this high school is male, B be the event that a student from this high school comes to Purdue. Compute the following:

(a) $P(A|B)$

(b) $P(\neg B|\neg A)$

(c) Now suppose that the overall proportion of female students increases to 57% and that the conditional probability from D(a) changes to 15%. Compute the updated probability on $P(B|A)$

E. **(6 pts)** CS37300 has 4 exams d_1, d_2, d_3, d_4. The possibility of passing them would be 0.98, 0.96, 0.95 and 0.93 respectively. Suppose they are independent.

(a) Let W denote the event that you pass at least one of the exams. Compute $P(W)$.

(b) Let A denote the event that at least one of the following happens: (i) you pass d_4 but fail d_3; (ii) you pass both d_1 and d_2. If you would get an A when event A occurs, then compute the probability that you get an A.

(c) Considering the setting of question Eb, given that d_4 works, what is the conditional probability that event A will occur?

F. **(6 pts)** Suppose we would roll two 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? What's the probability of $sum = 7$?

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 *'stars'* 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 *'stars'* (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 *'stars'* 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 *'stars'* 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 *'reviewCount'* and *'stars'* each contain a comma separated list of values associated with each restaurant. Compute two new boolean features: 'reliableReview' with a value of `TRUE` if the *'reviewCount'* > 10 and `FALSE` otherwise.'highStar' with a value of `TRUE` if the *'reviewCount'* > 10 and *'stars'* > 4, `FALSE` otherwise.

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 three attributes: *'stars'*, *'reviewCount'*, *'checkins'*.

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 three 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 *'alcohol'* attribute: *'stars'*, *'reviewCount'*, *'checkins'*, *'longitude'*, *'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 *'alcohol'* (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 *'alcohol'* (i.e., a separate IQR for the "full_bar" instances, the "beer_and_wine" instances, the "none" instances and the instances with " " for *'alcohol'*). 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.