CS373 Homework 1

Due date: Thursday October 1, 11:59pm (submit pdf to Gradescope)

Homework must be submitted as a PDF. Answers must 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. **The R code must be included in line with the questions and plots, not at the end of the document**. Please label all plots with the question number. Your homework must be typed and must contain your name and Purdue ID. If you are using a late day, please indicate it at the top of the document.

1 Basic Probability and Statistics

1. **(4 pts)** Consider an experiment where a coin is tossed repeatedly until the first time a head is observed.

   (a) What is the sample space for this experiment? What is the probability that the coin turns up heads after i tosses?

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

2. **(5 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. Compute the following:

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

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

   (c) $\mathrm{P}(F \cap G)$

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

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

3. **(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 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. **(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 4a 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.)

5. **(4 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, then prove that $\neg A$ 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.

6. **(4 pts)** Let X_n be the random variable that equals the number of heads minus the number of tails when n coins are flipped. Each flip has a probability of p of heads, $(1-p)$ probability of tails. (Do not assume $p = 1/2$.)

   (a) What is the expected value of X_n?

   (b) What is the variance of X_n?

   (c) Compute the expected value and variance of X_3.

7. **(2 pts)** The exam in a course consists of 25 true/false questions, each worth two points, 35 multiple-choice questions, each worth one point, and 5 fill-in-the-blank questions, each worth three points. The probability that a student answers a true/false question correctly is 0.8, the probability that they answer a multiple-choice question correctly is 0.75, and the probability that they answer a fill-in-the-blank question correctly is 0.65. What is the expected score on the exam? Now consider the case where an incorrect answer on a true/false question results in a penalty of $-\frac{1}{2}$ point. What is the resulting expected score?

8. **(2 pts)** 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 $E[X+Y] = E[X] + E[Y]$.

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:

1. Download and install R from: http://cran.r-project.org/
2. Download the Yelp dataset from Brightspace. This data set is part of the Yelp academic dataset and consists of data about 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 (4 pts)

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

(b) Print the names of the columns in the table using the names() function.

4 1D plots (15 pts)

A. (a) Plot a histogram of the *'checkins'* 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) Compute the logged values for *'checkins'* (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) Plot a density plot of the logged values of the *'checkins'* attribute using the `density()` function.

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

B. (a) Plot the logged values of the *'checkins'* attribute using `hist()` with `breaks=50`.

(b) Plot the logged values of the *'checkins'* attribute using `density()` with `adjust=0.5`.

(c) Discuss any differences from your previous plots and how the parameter settings change the way the distributions look.

(d) Plot `hist()` with `breaks=50`, `freq=FALSE` and compare with the density plot from this part.

C. (a) 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.)

D. Out of all the plots above for the *'checkins'* attribute above (A-B), which one gives you the most information? Why? What does it tell you?

5 Sampling and transforming data (15 pts)

A. (a) Transform the attributes: *'alcohol'* and *'noiseLevel'* into ordered, numeric features using the `factor()` function with the `levels` argument to specify a set of ordered levels. For *'alcohol'* use `levels=c("full_bar","beer_and_wine","none","")`; for *'noiseLevel'* use `levels=c("quiet","average","loud","very_loud","")`. Then transform the resulting levels into integers using `as.integer()`.

(b) Append the two new columns to the original data frame, using `cbind()` to increase the number of features to 30.

B. (a) Compute the quantiles (using `quantile()`) for the *'reviewCount'* attribute.

(b) Select a subset of the data with *'reviewCount'* value ≤ the 1st quartile (25th percentile). You can use `subset()` or select from the data frame with `[ ]` operations.

(c) Print a summary of the above subset and compare the results to those from Q1a. Discuss any differences in the distributions of the numerical attributes that you find.

6 2D plots and correlations (15 pts)

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

(b) 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. (b) Calculate the pairwise correlation among the above five attributes using the `cor()` function.

(b) 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 Q4a.

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

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

(c) For the attribute identified above, calculate its interquartile range for each value of *'goodForGroups'* (i.e., a separate IQR for the "TRUE" instances and the "FALSE" instances). You can do this with the `subset()` and `quantile()` functions. Calculate the overlap between the two 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.