CS373 HW1

January 29, 2018

Due date: 11:59pm Monday February 12, 2018.

Instructions for submission:

Create a single PDF with your answers. 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.

To submit your assignment, log into `data.cs.purdue.edu` (physically go to the lab or use ssh remotely) and follow these steps:

1. Make a directory named `yourusername-hw1` (all letters in lower case) and copy your PDF file inside it.
2. Go to the directory containing `yourusername-hw1` (e.g., if the files are in /homes/dan/dan-hw1, go to /homes/dan), and execute the following command:

   `turnin -c s373 -p hw1 yourusername-hw1`

   (e.g. Dan would use: `turnin -c s373 -p hw1 dan-hw1` to submit his work)

   Note that `s373` is the course name for turnin. It is not a typo.
3. To overwrite an old submission, simply execute this command again.
4. To verify the contents of your submission, execute the following command:

   `turnin -v -c s373 -p hw1`

1 Part I: 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) P(E $\cap$ F)
   b) P(E $\cup$ F)
   c) P(F $\cup$ G)
   d) P(E $\cup \neg$F)
   e) P(E $\cup$ F $\cup$ 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. **(6 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. **(6 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.

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 (`yelp.csv`) from the course page.

  This data set is part of the Yelp academic dataset and consists of data about 24,813 restaurants. The data file `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.csv()`. Use 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 the names of the columns in the table using `names()`.

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

(c) **(2 pts)** Print a summary of the *noiseLevel* attribute and the *stars* attribute.

4 1D plots

(a) **(4 pts)** 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.

Use the following R commands to plot the histogram and save it to `foo.jpg`:

```
jpeg('foo.jpg')
<insert your command here>
dev.off()
```

Add that image to your PDF. You do not need to submit it separately.

(b) **(4 pts)** Compute the logged values for *checkins* (you can use `log()` to compute the log of all the values in a vector). Plot a histogram of the logged values.

(c) **(4 pts)** Discuss the differences between the two plots and the information they convey about the distribution of *checkins* values in the data.

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: *isAmerican* and *goodForDinner* with a value of `TRUE` if the list contains "American" (in *categories*), "dinner" (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 to 32. Show the output of `summary()` for those two columns.

(b) **(4 pts)** Print the quantiles (using `quantile()`) for the `reviewCount` attribute.

(c) **(6 pts)** Select a subset of the data with `reviewCount` value $\leq$ 1st quartile (25th percentile). You can use `subset()` or select from the data frame with `[]` operations.

Print a summary of the above subset for the following attributes: `reviewCount`, `stars`, `attire`, `priceRange`, `delivery`, `goodForKids`, and compare them to their summary for the full dataset.

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

6 2D plots and correlations

(a) **(7 pts)** Plot a scatterplot matrix (using `pairs()`) 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) **(7 pts)** Calculate the pairwise correlation among the above five attributes using `cor()`.

- 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) **(7 pts)** Plot a boxplot (using `boxplot()`) for each of the following four attributes (`checkins`, `reviewCount`, `longitude`, `latitude`) vs. the `goodForGroups` attribute. Omit outliers using the `outline` argument.

Make sure to label both axes of the plot with the appropriate attribute names.

- 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.

- 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 `subset()` and `quantile()`. 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.