CS57300: Homework 1

Due date: Sunday February 1, midnight (submit pdf to Blackboard)

Submit both your answers to the questions and the R code that you used for analysis. Your homework must be typed and submitted as a PDF. Use of Latex is recommended, but not required.

1 Counting (2 pts)

Consider a simple password system. There are 26 lowercase letters, 26 uppercase letters, 10 digits, and 30 special characters on a keyboard.

(a) Our system accepts passwords of 6-10 characters, how many unique passwords are there containing only lowercase letters and digits? How many if the system requires at least 1 digit?

(b) Our system accepts passwords of 6-10 characters, how many unique passwords are there containing uppercase, lowercase, digits and special characters? How many if the system requires at least 1 digit, uppercase and special character?

2 Axioms of probability (2 pts)

(a) Prove that $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ using the axioms of probability.

(b) Prove the conditional version of Bayes rule:

$$P(B|A,C) = \frac{P(A|B,C)P(B|C)}{P(A|C)}$$

3 Probability and conditional probability (3 pts)

(a) Three men toss coins to see who pays for coffee. If all three match, they toss again. Otherwise the "odd man" pays for coffee.

(i) What is the probability that they will need to do this more than once?

(ii) What is the probability of tossing at most twice?

(b) Alice and Bob are playing a simple dice game. Each rolls one dice and the one with higher number wins. If the numbers are the same, they roll again. If Alice just won, what is the probability that she rolled a '5'?

4 Probability distributions (3 pts)

Let X be a discrete random variable such that $P(X = x) > 0$ if $x = 1, 2, 3, \text{or } 4$ and $P(X = x) = 0$ otherwise. Suppose the CDF is $F(x) = 0.05x(1 + x)$ at the values $x = 1, 2, 3, \text{or } 4$.

(a) Sketch the graph of the CDF.

(b) Sketch the graph of the discrete pdf $f(x)$.

(c) Find $E(X)$ and $Var(X)$.

5 Independence (3 pts)

A box contains four disks that have different colors on each side. Disk 1 is red and green, disk 2 is red and white, disk 3 is red and black, and disk 4 is green and white. One disk is selected at random from the box. Define the events as follows: A = one side is red, B = one side is green, C = one side is white, and D = one side is black.

(a) Are A and B independent events? Why or why not?

(b) Are B and C independent events? Why or why not?

(c) Are any pair of events mutually exclusive? Which ones?

6 Conditional Expectation (2 pts)

If X and Y are jointly distributed random variables, then the conditional expectation and conditional variance of Y given X are given by:

$$E(Y|x) = \sum_y y \cdot p(y|x)$$

$$Var(Y|x) = E(Y^2|x) - [E(Y|x)]^2$$

Let X and Y be discrete random variables with joint pdf $p(x, y) = 48/(45xy)$ if $x = 2, 4$ and $y = 1, 4$, and zero otherwise. Determine $E(Y|x)$ and $Var(Y|x)$.

7 Correlation (5 pts)

(a) Let X and Y be independent Bernoulli random variables with $p = \frac{1}{2}$. Show that $X + Y$ and $|X - Y|$ are dependent but uncorrelated.

(b) Discuss the differences between Correlation and Covariance. How does $Cov(X, Y) = 1$ differ from $Corr(X, Y) = 1$? Which statement is stronger?

(c) Show that $Corr(aX + b, cY + d) = -Corr(X, Y)$ when a and c have the opposite sign.
Make sure to include proofs for any identities or shortcuts that you use.
Discuss how this property would change if Covariance were used instead.

8 Exploratory Data Analysis (15 pts)

In this section, you will use the R statistical package to begin exploring, transforming, and analyzing data. To get started, do the following:

1. Download and install R from:
   `http://cran.r-project.org/`
   Links to a quick intro to the R programming language and a short reference card are below.
   `http://www.stat.cmu.edu/~larry/all-of-statistics/=R/Rintro.pdf`
   `http://cran.r-project.org/doc/contrib/Short-refcard.pdf`

2. Download the Yelp dataset from the course page.
   This data set is part of the Yelp academic dataset and consists of data about 14,192 restaurants. The datafile *yelp-data.csv* contains 44 attributes: 35 numeric and 9 discrete attributes.

   The first row of the data file is a header row with the names of the attributes, the values are separated by a ";" delimiter. The `categories` attribute is a list of local classification of the restaurants (e.g., *Pizza, Fast Food*).

3. Read the data into R using the `read.table()` function. Print a summary of the data using the `summary()` function. *Make sure to consider the arguments for the `read.table()` function. You will need to use the `comment.char=""` argument to avoid errors.*

R Questions

(a) Plot a histogram of the `tip_count` attribute. Use the `hist()` function with its default values and make sure to title the plot with the name of the attribute for clarity. Next plot a histogram using the log values of tip_count.

(b) Plot the `tip_count` attribute again but this time use the `density()` function in the plot, for both the original and the logged values.

(c) Find the continuous attribute with **largest** range and plot a histogram of the values. Make sure to title the plot with the name of the attribute for clarity.

(d) Find the discrete attribute (that is not a unique identifier) with the **maximum** number of values and plot a barplot to show the frequency of each value. Note that this will look like a histogram but for nominal values. Again, make sure to title the plot with the name of the attribute for clarity.

(e) Consider the four continuous attributes: `latitude, longitude, stars, likes`. Calculate the pairwise correlations among these four attributes. Plot scatterplots for the pair of attributes with largest positive correlation and the pair of attributes with largest negative correlation. Make sure to label both axis of the plot with the attribute names. Report the correlations and discuss whether the correlations are interesting or expected, given your domain knowledge.

(f) Choose a particular category (e.g., Nightlife) and create a new binary feature for each example that records whether the example contains the chosen category (e.g., Nightlife vs. not-Nightlife). You can use the `regexpr()` function to test whether the list contains a particular string. Plot a boxplots of your new binary feature vs. stars and likes (i.e., *feature vs. stars* and *feature vs. likes*). Make sure to label both axes of the plot with the attribute/feature names.

(g) Continue with the same approach you used above to explore several categories (e.g., *Bars, Diners*). Construct at least two new binary features from those categories that exhibit a difference in the star ratings (between categories). Plot the boxplots and discuss whether the relationship is interesting or expected, given your domain knowledge.