CSC591: Foundations of Data Science
HW1: Exploratory Analysis, Basic Probability, Random Variables and Probability Distributions

Released: 9/05/18
Due: 9/17/18 (23:55pm). (One day late: -25%; -100% after that).

Instructor: Dr. Ranga Raju Vatsavai

Notes

- Submission filename for theory questions: Lastname_StudentID.pdf (only pdf).
- You can also submit scanned handwritten solution (should be legible, TA's interpretation is final).
- This homework contains 6 questions of which only a subset will be graded by the TAs. Questions to be graded will be decided by the instructor.
- This h/w is worth 4% of total grade.
- All submission must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- No makeups or bonus; for regarding policies, refer to syllabus and 1st day lecture slides.
- You are encouraged to do research, study online materials; discuss with fellow students; BUT ANSWERS SHOULD BE YOUR OWN. Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.

Q#	Max Points	Score
1	10	
2	10	
3	10	
4	10	
5	10	
6	10	

Q1. Describe concisely various forms of data preparation (or preprocessing). (10 points)

Q2. A company wants to know the customer satisfaction and conducts survey over 100 customers. The survey form includes the following questions. (10 points)

Survey Form:

a) Are you: Male Female (Circle one of the choice)
b) How old are you? _____ (in years)
c) How much do you spend on groceries? ____ (in $$$$.$$)
d) How much do you spend on soft drinks? ___ (in $$$$.$$)
e) Which soft beverage do you prefer? ____ (Coke, Pepsi, Dr. Pepper, …)
f) How satisfied are you with diet beverages? ____ (Very satisfied, Satisfied, Not Satisfied)
g) How likely are you to buy 6-pack diet coke?___ (Very likely, Likely, Not Likely, Very unlikely)

Assume that all surveyed customers returned survey forms with correct answers.

Answer the following questions:

Your objective is to:

1. Design the database (one table) and enter the data. Show the table with few sample data entries.

2. For each resulting column (attribute), list the type of attribute (in terms of Nominal, Ordinal, Interval, Ratio) . Please justify your answer.

3. For each attribute, what kind of summary (statics) make sense? Please justify your answer.

4. For each attribute, what kind of graphical representation makes most sense? (e.g., pie chart, bar chart, …) Please justify your answer.

Q3. (10 points) Define (precise; one or two sentences) the following:

(a) Statistics:
(b) Population:
(c) Sample:
(d) Event space:
(e) Event:
(f) Random variable:
(g) Experiment:
(h) Discrete data:

(i) Continuous data:
(j) Mean:
(k) Median:
(l) Variance:

Q4. (10 points) The following data shows body temperature readings in degree F of 10 patients.

98.8 98.4 98.2 98.1 99.0 98.9 99.2 98.3 98.2 98.8

(i) Based on these readings, is a body temperature of 104.2 degree F is unusual? [Hint: Anomaly, Outlier, or Unusual value is defined as Mean ± 2 Standard deviations].

(ii) Based on general medical knowledge, is 104.2 deg F is unusual? Also comment if the definition of anomaly or unusual value defined as above is useful in this analysis?

(iii) With respect to the body temperature data given above; what is the standardized value (also known as z score) for 104.2 deg F.

Q5. (10 points) Law of total probability: Suppose C_1, C_2, …, C_m are disjoint events such that C_1 U C_2 U … U C_m = Ω. The probability of an arbitrary event A can be expressed as: $P(A) = P(A| C_1)P(C_1) + P(A| C_2)P(C_2) + \ldots + P(A| C_m)P(C_m)$. Illustrate this law using Venn diagram (for m=5) and derive P(A) using this Venn diagram

Q6. (10 points) Answer each of the following questions.

(a) Let Ω = {a,b,c} be a sample space. Let P(a) = ½, P(b) = 1/3, and P(c) = 1/6. List all possible subsets of Ω and find probabilities for all these subsets.
(b) The following table summarizes the results of breathing test given to drivers suspected of driving under influence. Suppose if two persons included in the following table are randomly selected without replacement, the find the probability that the first person has positive test result and the second person has negative test result.

	Persons actually driving after consuming alcohol	Persons without alcohol consumption
Positive test result	90	10
Negative test result	5	95