Homework No.2 based on Lectures 3, 4, and 5

Due Date: March 6th 2016 11:59pm EST!

You will submit your answers in Canvas under Homework 2. There should be two files based on the questions shown on the next page.

Homework 2 doc
Homework.R

- Graduate student must complete all questions 1 through 5 .
- Undergraduate students may skip Question 5.

This assessment is based on the required readings, lectures, labs, and videos.

Homework 2 is worth 10 points total:

Rubric grading:

- R Coding assessment is 4 points
- Reading assessment is 2 points
- Hypothesis testing is worth 4 points

Please make sure you provide us clear instructions on how to run your code through a step method. Your answers should have comments as it is a standard of good coding principles.

Homework No.2 based on Lectures 3, 4, and 5

1) In **lecture No. 3** (slide 9) we did a simple exercise on decision trees and drew a tree with Outlook as the root. **Using this same tree**, find the entropy and information gain for this tree. You can follow the model of the lecture or the textbook. Show all your work. Using R and the three of packages shown in the lecture, draw the derivation tree.

If one of the packages gives you better results, explain why this is so. A picture of the tree and the code you used to produce will suffice along with a short explanation of why this model is better than the other two.

***Comments** to this question*: The entropy and information gain must be calculated by hand. The packages will not provide that information.

2) The table shown below indicate a sample data relating the number of hours spent by individual students outside of class on a course in information science for the first 4 weeks of classes. The table shows the student's scores on an examination at the end of these 4 weeks. Using the method of the Least Squares answer the following questions:

a) Can linear regression be used given the data shown in the table? If so, how can you justify your decision of using linear regression?

b) Assuming that the answer to the previous question is positive, find the line that best fit this data by following the model of the problem explained in the lecture. Make sure that you include all the relevant values in a table similar to the one shown in the lecture?

c) If there is a line that fits the data the best, use R and the packages used during the lecture to draw the line and obtain its slope and Y-intercept. If there is any discrepancy between the values that you obtained and those indicated by R, how can you justify these discrepancies?

d) Use the regression line obtained in part 3 of this question and estimate the the examination grade of a student who devoted 30 hours of study to the course material.

e) Can you use the data to justify the grade that student would get if the course lasted 8 (eight) weeks instead of four? Justify your answer.

Hint: as part of this question you may have to create a data frame. You can do it interactively or not, but in any case, it is always wise to show the structure of the data frame to make sure you got the data in correctly. *Notice that even though the questions have a linear order, that is, they are numbered, this does not imply that to answer them you have to follow that order. Therefore, our advice is that you read all questions first. Once you have a clear idea of what is needed, then proceed to answer them. You may have to do some previous work before even answering the first question*.

The table showing the population for this exercise is shown next.

Student No.	1	2	3	4	5	6	7	8
Hours of Study (X)	20	16	34	23	27	32	18	22
Examination Grade (Y)	64	61	84	70	88	92	72	77

Comments to this question: The table associated with determining the linear regression, following the model of slide 38 of Lecture 4, and all its entries need to filled in by hand; then, using R you can then verify if you calculations were correct. All entries in the table should be the same for all students. If you need to round, round even to two decimals. R may not round even but the values should be close. *You need to show in addition to the associated table, the scatter plot and the line that best fit the model, the R instructions to create them.*

3) In Chapter 3 of Lynda.com R Statistics Essential Training we review Hypothesis testing for 2012 Major League Baseball Season, the Washington Nationals had the best record at the end of the season: 98 wins out of 162 games (.605).

But what if I change the record does the testing hold true for 92 wins out of 162?

Show your R Code based on this change and write out your answers and code for each test. This is a one sample proportions test with continuity correction.

- Calculate P value?
- Calculate Chi Square value?
- What is the Confidence Interval?
- Calculate the mean, average, and standard deviation?

4) State the Hypotheses Test in Statistical form similar to what was discussed in Lab for Q3.

P value is .05

State the Null and Alternative for both Type I and II Errors.

- Resources: https://en.wikipedia.org/wiki/List_of_statistics_articles#F

5) Create a Histogram for Question 3 with R Code to show the win, lose, and breakeven point.