IS 665 - Fall 2017 Final Questions

Required Datasets: You can find the following datasets on blackboard.

- data_babies.csv
- smokyph.csv
- churn.csv

1 (15pts) The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. `data_babies` introduces a data set on birth weight of babies. Another variable we consider is gestation, length of pregnancy in days. Is gestation a good predictor of birth weight of babies?

a. Study the relationship visually and with linear regression analysis.
b. Intrepret the regression analysis results.

2 (15pts) The babies data also include information about mothers' smoking habits. The variable smoke is coded 1 if the mother is a smoker, and 0 if not.

a. Like in question 1, study the relationship between the average birth weight of babies and the smoking status of the mother with linear regression model.
b. Interpret the results.

3 (20pts) We considered the variables smoke and gestation, one at a time, in modeling birth weights of babies in Question 1 and 2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of whether the child was 1st born (parity),0 if the child is the first born, and 1 otherwise; mother's age in years (age); mother's height in inches (height); and mother's pregnancy weight in pounds (weight).

a. Run regression analysis including all of the variables.
b. Interpret your results.
c. Based on your findings select the best variables to use in your model.

4 (20pts) Load data set `smokyph.csv` . This data set measures pH levels for water samples in the Great Smoky Mountains. There are claims that water ph level is not acceptable in this region (we expect average ph to be 7). Use the waterph column (waterph) to test if the water ph level is not acceptable (within 95% confidence level).

a. Construct your hypotheses.
b. Conduct the appropriate test
c. Interpret your results.

5 (20pts) `churn.csv` is provides historical information about randomly selected 3333 customers for a phone service provider. Among other things, the churn variable shows whether the customer stop subscribing to the server (yes= he/she stopoed). Using knn, develop a model that predicts the churn variable using (at least)4 other variables in the dataset.

a. Create a model with 80% of the data and test it with the rest.
b. Create a confusion matrix and briefly explain the accuracy of your model.

6 (20pts) For the same dataset create a decision tree again using 4 variables. (At least 2 of the variables in this model should be different than the ones you used in the 5th question)

a. Create a model and prun it.
b. Explain your findings in a sfew sentences.