Bryson Cook
ISYE6501
Homework 10

Question 14.1
The breast cancer data set breast-cancer-wisconsin.data.txt from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ (description at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) has missing values.

1. Use the mean/mode imputation method to impute values for the missing data.
2. Use regression to impute values for the missing data.
3. Use regression with perturbation to impute values for the missing data.
4. (Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using
   (1) the data sets from questions 1,2,3;
   (2) the data that remains after data points with missing values are removed; and
   (3) the data set when a binary variable is introduced to indicate missing values.

For part 1, I used the mean method for any missing data in all columns but the Class attribute, where I used the mode to select between 2 and 4.

For part 2, I first removed the rows containing "?"'s from the data so that the model could be trained. I then created a regression model with this new data with the BareNuclei column as the prediction. However, the model is not very accurate with a adjusted R2 of only 0.2695. Then, using the factor p-values, I kept only the important factors and ended up with the following formula:

BareNuclei ~ MarginalAdhesion + NormalNucleoli + Class

This model did not have a much better R2, only 0.2719. More refinement such as PCA or stepwise regression could also be used to refine the model. I then applied the model to the rows missing data and inserted the results back into the data.

For part 3, I followed the same path as part 2, but created an array of normally distributed perturbations that were added to the results of the model. Since I was adding a perturbation, I added a check on the results to ensure the final value was between 1-10, as that is the scale of the original data.

Question 15.1
Describe a situation or problem from your job, everyday life, current events, etc., for which optimization would be appropriate. What data would you need?

I was at a hardware store and thought that optimization must be utilized all the time in supply chain decisions. Given an item's price from the manufacturers, those manufacturers' locations, your distribution centers or store locations, and the cost of transportation between them, you would need to optimize a model to have to lowest cost to get to item to the store. As an example, Company A may manufacture hammers at a lower price than Company B, but if the cost of transporting those hammers from Company B is cheaper, Company A would not necessarily be the best option. If more manufacturers or constraints such as minimum quantities or pallet size are added, it quickly becomes a complex problem that would require optimization.

Question 15.2

In the videos, we saw the “diet problem”. (The diet problem is one of the first large-scale optimization problems to be studied in practice. Back in the 1930’s and 40’s, the Army wanted to meet the nutritional requirements of its soldiers while minimizing the cost.) In this homework you get to solve a diet problem with real data. The data is given in the file diet.xls.

1. Formulate an optimization model (a linear program) to find the cheapest diet that satisfies the maximum and minimum daily nutrition constraints, and solve it using PuLP. Turn in your code and the solution. (The optimal solution should be a diet of air-popped popcorn, poached eggs, oranges, raw iceberg lettuce, raw celery, and frozen broccoli. UGH!)

2. Please add to your model the following constraints (which might require adding more variables) and solve the new model:

   a. If a food is selected, then a minimum of 1/10 serving must be chosen. (Hint: now you will need two variables for each food *i*: whether it is chosen, and how much is part of the diet. You’ll also need to write a constraint to link them.)

   b. Many people dislike celery and frozen broccoli. So at most one, but not both, can be selected.

   c. To get day-to-day variety in protein, at least 3 kinds of meat/poultry/fish/eggs must be selected. [If something is ambiguous (e.g., should bean-and-bacon soup be considered meat?), just call it whatever you think is appropriate – I want you to learn how to write this type of constraint, but I don’t really care whether we agree on how to classify foods!]

Using PuLP to minimize the cost of the daily food intake while adding the constraints located in the diet.xls file, the software gives the optimum cost of $4.34 per day with the following foods:

Food	Servings
Foods_Celery,_Raw	52.64
Foods_Frozen_Broccoli	0.26
Foods_Lettuce,Iceberg,Raw	63.99
Foods_Oranges	2.29
Foods_Poached_Eggs	0.14
Foods_Popcorn,Air_Popped	13.87

I find it interesting, but obvious once I think about it, that the results will consist of mostly the cheapest food available, as celery, lettuce, and popcorn are three of the four cheapest foods per serving. I assume that the optimizer chose the cheapest options to fill the largest requirements, such as calories, then added more foods as needed to meet the other constraints (such as the small amount of broccoli probably included to meet the Vitamin A constraint). I also think it’s worth noting that the above diet is horrible and would not actually suffice, as no person would probably eat 52 stalks of celery, 63 leaves of lettuce, and over three-quarters of a pound of popcorn in a day, which shows the need for more constraints to fit the real world.

Adding the constraints in part 2 yielded a total cost of $4.51 and the below components:

Food	Servings
Celery, Raw	42.40
Kielbasa,Prk	0.10
Lettuce,Iceberg,Raw	82.80
Oranges	3.08
Peanut Butter	1.94
Poached Eggs	0.10
Popcorn,Air Popped	13.22
Scrambled Eggs	0.10

Each of the constraints is obeyed, with three proteins (Kielbasa, Poached Eggs, and Scrambled Eggs), either celery or broccoli can be chosen, and a minimum of 0.1 serving must be served if that food is chosen. Once again, this is not a practical diet so further constraints would be needed to achieve a real world solution.