

WEEK 10 HOMEWORK

INSTRUCTIONS

- Every learner should submit his/her own homework solutions. However, you are allowed to discuss the homework with each other (in fact, I encourage you to form groups and/or use the forums) – but everyone must submit his/her own solution; you may not copy someone else's solution.
- The homework will be peer-graded. In analytics modeling, there are often lots of different approaches that work well, and I want you to see not just your own, but also others.
- The homework grading scale reflects the fact that the primary purpose of homework is learning:

Rating	Meaning	Point value (out of 100)
4	All correct (perhaps except a few details) <u>with</u> a deeper solution than expected	100
3	Most or all correct	90
2	Not correct, but a reasonable attempt	75
1	Not correct, insufficient effort	50
0	Not submitted	0

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.

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?

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!]

If you want to see what a more full-sized problem would look like, try solving your models for the file `diet_large.xls`, which is a low-cholesterol diet model (rather than minimizing cost, the goal is to minimize cholesterol intake). I don't know anyone who'd want to eat this diet – the optimal solution includes dried chrysanthemum garland, raw beluga whale flipper, freeze-dried parsley, etc. – which shows why it's necessary to add additional constraints beyond the basic ones we saw in the video!

[**Note:** there are many optimal solutions, all with zero cholesterol, so you might get a different one. It probably won't be much more appetizing than mine.]