691: Data Analysis for Data Science

Optimization

Use one of the suggested analysis tools from class 1 to solve a k-means or hierarchical clustering problem and a linear programming optimization problem. You only need to do one of each.

1. K-means or Hierarchical Clustering

Use one of the following UCI machine learning datasets for a k-means or hierarchical clustering. Select an initial value for k based on a look at the descriptive statistics for the dataset.

A. http://archive.ics.uci.edu/ml/datasets/Auto+MPG (398 instances)

B. https://archive.ics.uci.edu/ml/datasets/Yeast (1484 instances)

Note: both of these examples have fixed widths, i.e. the file separators are some kind of blank character (spaces, tabs, etc.) but the fields are always exactly the same width. Use fixed width in the import dialogue for Excel/Calc. For PSPP select space and tab as your separators.

We can read these into R using read.fwf(). Open your fixed width file in a programmer's text editor like Notepad++ (do not use Word for this or any word processor unless you are using a monospace font). Count the number of characters in each segment (including white space).

```
ADT1_YEAST  0.58  0.61  0.47  0.13  0.50  0.00  0.48  0.22  MIT
ADT2_YEAST  0.43  0.67  0.48  0.27  0.50  0.00  0.53  0.22  MIT
```

```
123456789012345678901234567890123456789012345678901234567890123
```

Here is an example of counting the first column. We need the width of the column and width of the separator. Monospace font and adding numbers 1-(1)0 is an old trick for doing this manually. You can also get some editors to output column numbers or the width of selections.

```
ADT2_YEAST  |start of next column
```

```
123456789012
```

So the first segment is 12 characters wide. The separator in this case is 2 spaces, though there are a few instances in the file where it is 3. We can ignore the 3 spaces though as we care about the full width of each column and the width of the smallest separator. We can use strip.white=TRUE to remove excess white space as well.

Here is the command for reading the yeast dataset:

data <- read.fwf("https://archive.ics.uci.edu/ml/machine-learning-databases/yeast/yeast.data",header=FALSE,widths=c(12,6,6,6,6,6,6,6,6,3),strip.white=TRUE)

and the auto dataset:

data2 <- read.fwf("http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",header=FALSE,widths=c(7,4,11,11,11,7,4,2,40),strip.white=TRUE)

2. Optimization using Linear Programming

A. The Silly Nut Company makes two mixtures of nuts: Mixture A and Mixture B. A pound of

Mixture A contains 12 oz of peanuts, 3 oz of almonds and 1 oz of cashews and sells for $4. A pound of Mixture B contains 12 oz of peanuts, 2 oz of almonds and 2 oz of cashews and sells for $5. The company has 1080 lb. of peanuts, 240 lb. of almonds, 160 lb. of cashews. How many pounds of each of mixtures A and B should the company make to maximize profit? [Note: You need to check your units whenever you calculate an optimization problem. In this case you need to either use ounces for all values or pounds.]

B. Dr. Lum teaches part-time at two different community colleges, Hilltop College and Serra College. Dr. Lum can teach up to 5 classes per semester. For every class taught by him at Hilltop College, he needs to spend 3 hours per week preparing lessons and grading papers, and for each class at Serra College, he must do 4 hours of work per week. He has determined that he cannot spend more than 18 hours per week preparing lessons and grading papers. If he earns $4,000 per class at Hilltop College and $5,000 per class at Serra College, how many classes should he teach at each college to maximize his income, and what will be his income?

Note: You may choose to use a different data set than the ones provided above, but you should email me first to ensure that the data set is sufficient for this assignment. The UCI Machine Learning Repository is a good site for clustering datasets: http://archive.ics.uci.edu/ml/datasets.html

If using your own dataset, the dataset must be submitted with the assignment or a URL must be provided where the dataset can be downloaded.