- Academic honesty is important. Strict plagiarism policy will be enforced.
- You can work in a group of 2–3 people. (You can group with previous partners.)
- Attach your R code after your typed-homework.
- Submit directly to Randy or put it into Randy's mailbox located in Statistics department office.
- Send also your R code to sta208.spring2015.ucd@gmail.com in a zipped file.

Introduction

The data for your final project is collected over 79 neighborhoods from an East Coast city. You have 92 measurements on each neighborhood, such as the percentage of population change from 1990 to 2000, the land area in square miles, the percentage of the population between ages 20 and 34, the percentage of residents who ride their bicycle to work, etc. Each neighborhood also corresponds to a score of 1, 2, or 3, reflecting its level of crime, with 1 being low, 2 being medium, and 3 being high.

Each row of the data corresponds to a neighborhood. The NA values represent the missing crime scores; for these neighborhoods, you have all of their other measurements. Your main goals to predict the missing crime scores (i.e., fill in the NAs).

Details

Download the file "neighbor.Rdata" from the smartsite. Load it into your R session (for example, using `load("neighbor.Rdata")`, provided that it is in your working directory). Now you should have two objects: `neighbor.dat` and `pair.dist`.

The object `neighbor.dat` is a matrix of dimension 79×93. Each row corresponds to a particular neighborhood. The first column represents the crime score, on a scale of 1 to 3; recall that 1 = low, 2 = medium, and 3 = high. Some of the scores here are missing, i.e., they are given NA values. There are 24 neighborhoods with missing scores, you will create a vector (of 1s, 2s, and 3s), corresponding to your predictions for the crime scores of those neighborhoods.

How would you make such predictions? Note that you have 92 feature measurements on each neighborhood (the last 92 columns of `neighbor.dat`). You also have spatial information about the neighborhoods, in the `pair.dist` object: this is a 79×79 dimensional matrix containing the pairwise distances between the neighborhoods, i.e., `pair.dist[i,j]` contains the (Euclidean) distance between neighborhoods i and j as measured on a map. It would be a good idea to investigate the utility of this information as well.

Report

Your report should have the following sections (you can of course add subsections if you want), and should be no more than 6 pages (do not include R code).

Introduction Describe the problem and possibly do some exploratory data analysis.

Unsupervised analysis Is there any interesting structure present in the data? Describe what this means in the context of the neighborhoods and their relationships to each other. Here you can use some of the techniques that we learned or any other techniques as long as they are adequately explained. Note that this question is intentionally vague. If you don't find anything interesting, then describe what you tried, and show that there isn't much visible structure.

Supervised analysis How did you make your predictions? Describe this process in detail. Again, you can use any of the classification techniques that we learned or any other techniques as long as they are adequately described. What predictor variables did you include? What technique did you use, and why did you choose it? What assumptions, if any, are being made by using this technique? If there were tuning parameters, how did you pick their values? Can you explain anything about the nature of the relationship between the predictors in your model and the predictions themselves?