Data Mining: 36-462/36-662
Final Project

Crime Mining (or Data Criming?)

Thursday April 25 2013

Deliverables and deadlines

Here are some key deliverables and deadlines upfront:

- Your predictions and your slides are both due Thursday May 9 at 5:30pm, submitted by email to the professor (`ryantibs@cmu.edu`).
- Your write-up is due Friday May 10 at 5:30pm, submitted as a paper copy in Doherty Hall A302 (the room for our final exam period).

Each team only has to submit one of everything. Hence you can read "you" throughout as "your team". Read on to learn more...

Introduction

The data for your final project is collected over 79 neighborhoods from an East Coast city (name withheld on purpose). 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.

This is real data, and we've withheld the crime score for 24 of the 79 neighborhoods. So what's available to you looks like this:

```
   Crime % Pop. Change, 90-00 % Pop. Change, 00-10 . . .
1     NA                -4.13                -6.18
2      3               -28.11                33.37
3      3               -24.25               -20.35
4      2                10.78               -17.18
5      1                -4.78                -4.98
6      3               -14.67               -18.94
7      1               -12.30               -19.79
8     NA                -7.37                 6.85
9      3               -13.82               -14.58
10     1                -3.23               -10.05
.
.
.
```

Above, each row corresponds to a neighborhood, and the `...` denotes an additional 90 columns of measurements that are not shown. The `NA` values above represent the missing crime scores; for these neighborhoods, you have all of their other measurements. Your main goal by the final exam period is to predict the missing crime scores (i.e., fill in the `NA`s).

More details

Download the file "neighbor.Rdata" from the course website. 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 (snippet displayed above). 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. More precisely, there are 24 neighborhoods with missing scores, and they are numbered as follows:

```
> which(is.na(neighbor.dat[,1]))
 1  8 13 17 18 20 23 30 31 42 46 47 49 52 53 54 55 57 60 69 71 72 76 79
```

For the final exam period you will create a vector (of 1s, 2s, and 3s), corresponding to your predictions for the crime scores of the neighborhoods in this order. You should store this as an R object called `crime.pred`, and save this to an R data file (using, for example, `save(crime.pred, file="mypred.Rdata")`). You must send this file to the professor via email (`ryantibs@cmu.edu`) by **Thursday May 9 at 5:30pm**.

How would you make such predictions? Note that you have 92 feature measurements on each neighborhood (the last 92 columns of `neighbor.dat`) at your disposal—use them! 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.

Write-up

You will submit a write-up, as a hard copy, on **Friday May 10 at 5:30pm**. This write-up should be a polished report, with figures and snippets of R code as you deem helpful. You don't need to submit your R code in its entirety. Your report should have the following sections (you can of course add subsections if you want), and should be no more than 6 pages.

Introduction: Describe your data set. What is the problem you are trying to solve? This can be brief. (You don't need to do the typical "exploratory data analysis" that you learned in 36-401, but you should provide proper motivation for your work.)

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 in the first half of the course, 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. Data mining isn't about continuing to manipulate the data in some way until you get an answer.

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 in the second half of the course, 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?

Presentation

Your team will have a short period of time, during the final exam period (on Friday May 10 at 5:30pm), to show some slides summarizing your work. You will be given 4 minutes total, and can show a maximum of 5 slides. You can cover some of your unsupervised analysis if you have something interesting to share, you and should explain your predictive model and how you arrived at it. Your slides must be in PDF format, and must be emailed to the professor (`ryantibs@cmu.edu`) by **Thursday May 9 at 5:30pm**. Slides with mostly (color) pictures, and not much text, are encouraged.

Evaluation

We have the true crime scores, and we will compute each team's misclassification rate: if $y_1, \dots y_m$ are the true crime scores for the withheld neighborhoods, and $\hat{y}_1, \dots \hat{y}_m$ are your predictions, then this rate is

$$\frac{1}{m} \sum_{i=1}^{n} 1\{\hat{y}_i \neq y_i\}.$$

The misclassification rates will be revealed during the final exam period. The undergraduate and graduate teams will be considered separately; the top 5 undergraduate teams and top 3 graduate teams (with the lowest misclassification rates) will be given extra credit. The very top team will get a box of chocolate truffles.

For the grading of the write-up, the unsupervised analysis will be worth roughly 33%, and the supervised analysis worth roughly 67%. Your presentation will also contribute a small amount towards your final project grade.

TA mentors

Each group will be assigned a TA mentor, one of Cong, Jack, Li, and Michael. Ask them for help! They have valuable expertise and experience. And the professor has promised the TA of the top team (in terms of misclassification rate) a box of chocolate truffles too, for a little extra motivation.

Cheating

Don't cheat. We know that there are ways to cheat on this final project. If we suspect you of cheating (e.g., if you have a remarkably low misclassification rate, but your method is not really statistically motivated), then we reserve the right to give you a 0.

Suggestions

Here are some suggestions:

- Have fun! It's supposed to be a fun project.
- Give yourself a team name, and email this to the professor at the same time that you email your predictions and your slides. Otherwise, you will be assigned a name (sure to be less cool than one that you come up with).
- Talk things over with your group. Before you start programming, think about the data that you have. Think about how it fits in to things that we've learned in the course. Plan what you want to do. Then, delegate, and do it. You don't have to wait too long before you start programming, but a little thinking and planning might help you and save you from getting frustrated early on.