Problem Set 03
Data Mining and Machine Learning – Spring 2016
Due date: 2016-02-26 13:00

All assignments must be uploaded to the assignments tab in ClassesV2 (notice that this is **not** the dropbox) by the date and time specified. Make sure that you follow the instructions exactly as described. This is a large class with a limited number of TA's and we will be grading parts of the assignment using an automated grading engine; you will lose points for things such as not naming files correctly. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions.

You need to turn in all of your solutions as a zip compressed file, named `netid_pset03.zip`, with your actual netid filled in in all lower case letters. This archive should contain the following files:

- `pset03.csv`
- `pset03.pdf`

The pdf file should contain any written solutions; the contents of the other file is described below.

Chicago Crime Data

The data referenced in the rest of this problem set is a subset of the Chicago Crime Data, released by the City of Chicago. It lists (almost) all reported crimes that occur within the city limits. You can download the training and test data from here:

```
http://www.stat.yale.edu/~tba3/stat665/psets/pset03/data/chiCrimeTest.
psv
http://www.stat.yale.edu/~tba3/stat665/psets/pset03/data/chiCrimeTrain.
psv
```

These are pipe separated files with a single header row. The variables year, month, and hour should be self-explanatory. The variables arrest and domestic are binary flags for whether an arrest was made for the crime and whether the crime was a domestic dispute. Location is a string from a relatively large dictionary showing the type of location where the crime occurred. The variables beat, district, ward, and community area are integers indicating the region where the crime happened. These should be treated as categorical variables.

The variable **type** is a five level categorical response that you will want to build a prediction algorithm for. The categories correspond to:

1. THEFT OVER $500
2. SIMPLE BATTERY
3. DECEPTIVE PRACTICE
4. NARCOTICS
5. BURGLARY

I have down-sampled the data so that these all occur with roughly equal proportions in both the training and testing sets.

You could in theory merge external datasets with this (such as weather or meta-data about the various regions). However, to keep the scope the problem set contained, we will require that you work only with the data provided.

Prediction Task

Your task is to build a predictive model that estimates which of the 5 crime types is associated with a crime given the available covariates. You should ultimately make class predictions on the test set and save these as `pset03.csv`. Note that there will be only one column in the output, so even though it is called a `csv` file, there should be no commas. There should also be no row names or header rows. There should just be one of the numbers $1 - 5$ on each row.

To help you with the formatting, and to make sure you do not make any silly mistakes, you should use these two files:

```
http://www.stat.yale.edu/~tba3/stat665/psets/pset03/testScript03.R
http://www.stat.yale.edu/~tba3/stat665/psets/pset03/pset03_sample.csv
```

The first is a script to check your results and the second is a partial list of the results from the test set. About 80% of the rows are masked with NAs, but the other 20% have the true class labels. By running the test script you can see how well you are doing on this small portion of the data, and can see that the format of the results is correct.

Note: You are free to use the 20% sample of the test set for additional validation to which you can tune any hyper-parameters in your model. However, when we grade the set, only your results on truly hidden 80% of the data will be used, so there is no need to spend time overly tuning to these values.

Grading: We will assign 4 of the 10 points to your predictions. Extra credit is available for particularly predictive models, but full credit will be given to anyone performing reasonably well (Hint: You should be aiming to get at least below a mis-classification rate of 41%).

Write-Up

In the file `pset03.pdf`, write up a description of what you tried for building a prediction model and any thing of interest that you saw in the fitting procedure. In particular, you should at least address:

- How did you tune the various parameters in your models?
- How did you use hierarchical modeling (use the prediction in one model as an input to another) or stacking to combine multiple models together?
- How did you incorporate the categorical 'location' variable? Did it influence the model significantly
- How did you incorporate the categorical variables such as beat, district, ward, and community area (you can use only one if you would like; but don't ignore them entirely)
- Describe the confusion matrix. What is the easiest category to differentiate. Why? What categories are hard to tell apart? Does this make sense to you?
- What you expect your mis-classification rate on the test set will be?

The write-up does not need any figures or overly fancy analysis; just answer the questions above and describe the approach you used. I'll leave the length up to your choosing, but somewhere between 1-2 pages is about the right length.