Problem Set 01
Data Mining and Machine Learning – Spring 2016
Due date: 2016-02-05 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_pset01.zip`, with your actual netid filled in in all lower case letters. This archive should contain the following files:

- `pset01.R` or `pset01.py`
- `pset01.csv`
- `pset01.pdf`

The pdf file should contain any written solutions; the contents of the script and csv file are described below.

Data files

The data referenced in this problem set is a subset of the feed from the NYC Taxi Commission `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`. The specific datasets that I have prepared for this problem can be downloaded from:

```
http://www.stat.yale.edu/~tba3/class_data/nyc_train.csv
http://www.stat.yale.edu/~tba3/class_data/nyc_test.csv
```

Implementation

Using the directions in the starter code in `pset01_starter.R` or `pset01_starter.py`, implement a simple version of k-nearest neighbors and kernel smoothing. You are able to use distance functions and density functions supplied by the language, as well as other basic control flow functions. However, you should (obviously) not call the versions of these functions supplied by default in the language. A script with these working functions should be uploaded as described above (i.e., named `pset01.R` or `pset01.py` and placed in a zip archive).

Prediction

Using the dataset `nyc_train.csv`, you will need to build three predictive models for whether a hailed taxicab will drop off its passengers outside of Manhattan (i.e., *dropoff_BoroCode* is not equal to one). These models will then be fit to the data in `nyc_test.csv` to give predictions. These should use predicted probabilities. That is, responses should be between 0 and 1, with 0 indicating zero probability the taxi will leave Manhattan. The models you will fit are as follows:

1. using knn with $k = 100$, with predictor variables of pickup latitude and longitude; use the standard euclidean distance function, even though this will not correspond directly to distances in real life
2. a linear regression that uses the pickup hour of the day (you will need to derive this from the raw data) and the pickup neighborhood coded as a factor for predictor variables
3. the same linear regression as above, but fit using ridge regression with a cross-validated shrinkage penalty

The first two predictions should be the same for everyone; the third will depend on how cross-validation is performed (the details are left to you).

The prediction results should be saved as `pset01.csv`. This should be a comma separated file with the same number of rows as the data in `nyc_test.csv` and three columns corresponding to the three models. Fill in the predicted probabilities for each observation in the test set. Make sure you do not include additional rows, row/column names, quotes, or other fluff.

For these predictions you do not need to write your own implementations from scratch, but may use whatever libraries you wish. We will not grade you based on whether your predictions are close to the actual values (finding the exact answers can be easily reverse engineered from the raw data anyway). We are just looking to see that you can implement the methods correctly at this point. Also, there is no need to attach your code for this task; we just want the prediction results.

Data analysis

Looking at the predictive models you built in the previous task, describe what factors seem to effect whether a hailed cab will leave Manhattan. What other models might help you understand this relationship better? Construct one or two visualizations that help to summarize your description. If it is helpful, you can look up the names of the neighborhood codes on this table:

`http://www.stat.yale.edu/~tba3/class_data/nyc_nta.csv`

Your answer should be roughly 150-300 words long (though that is a just a rough guideline so you know the level of detail we wont; we won't be counting). Save the visualizations and narrative as `pset01.pdf`.