MSAN 630: Advanced Machine Learning

Status Update: Yelp Recommendation Engine

Jeff Baker, Kailey Hoo, Griffin Okamoto

March 2, 2015

Project Objectives/Evaluation

Since our project proposal, we have updated the scope of our project in the following ways:

1. **Objective**: Our goal is to predict a user's rating of a business; this is updated from our original objective of generating a list of recommended businesses & their corresponding probabilities.
2. **Data**: We are now using the RecSys2013 Kaggle data set; this is different from our original plan to scrape data directly from Yelp's website and/or API.
3. **Evaluation**: Our minimum objective remains the same: using the readily available features, we would like to implement at least one machine learning model, which involves tuning hyper-parameters, calculating evaluation metrics, and final testing. We have also planned to do some additional feature engineering, but should these fail it will not mean failure for the project.

Data Status

Since we've changed the scope of our project, leveraging readily-available data has allowed us to pivot to exploratory data analysis and training data development much sooner than expected.

The RecSys2013 data consists of ~230K business reviews in the Phoenix, AZ metropolitan area. We've completed initial EDA on the data, consisting of:

- Univariate summary data
- Some bivariate EDA (limited by the nature of this data set)
- Text analysis (N-grams) of the business review text
- A slew of plots across various dimensions and fields, including histograms, barplots, and some impressive map plots (using 'RgoogleMaps' in R)

Having completed our EDA, we have begun algorithm exploration with our initial training data set; this data set consists of unmodified features directly from the source data itself. Our plan (as further explained below) is to do initial baseline models with this first training data set, and then add additional engineered features as we continue working through next week.

TECHNIQUES, EVALUATION RESULTS

We have begun trying out different machine learning algorithms in sklearn. We are dividing work in the following way:

Kailey	Jeff	Griffin
Logistic Regression	zeroR	Linear Regression
KNN	Random Forest	Naive Bayes
Boosting	Support Vector Machines	Neural Networks

As mentioned, for each of these initial models, we are using unedited features from the original data set. Once we have tried all of them (with minimal hyper-parameter tuning), we will narrow down the models based on their initial performance, but still keep zeroR for a baseline. We do not currently have any evaluation results or baselines. We expect to do a lot more model tuning, particularly with the models we narrow down to.

CODE STATUS

We are not using any new technology. We have written significant R code for exploratory data analysis of the user, business, and check-in data. We have written Python code for various data pre-processing and cleaning, as well as exploratory text analysis on the reviews. We have just begun writing Python code for the algorithms themselves.

OTHER STATUS

Given the scope changes to our project, we are slightly behind our original schedule. However, having completed our EDA step, we are on track to meeting our final completion deadline of Wednesday, March 11. The adjustments we've made to our objectives/evaluation should sufficiently accommodate this timeline. To ensure excellence in our presentation & report, we decided to work on our report as we go, providing a solid base off which to create and practice for our presentation.

During our meeting, we'd like to discuss a few concerns we have about data leakage. Otherwise, just general suggestions and guidance would be appreciated.