AC209a Milestone 5: Data Science with User Ratings and Reviews

Andrew Ross, Sophie Hilgard, Reiko Nishihara, Nick Hoernle

November 28, 2016

Proposal of Future Work

In Milestone 4, we evaluated the performance of a number of baseline recommendation models (simple averaging, matrix factorization, popularity ranking, and PageRank) on restaurants and bars in the Yelp dataset. Going forward, and following some of the techniques used in [1], we want to see if we can surpass those benchmarks by blending models, computing time-sensitive user and business baseline ratings, and by exploring other models like collaborative filtering. We're also interested in exploring models that exploit more subtle time and location-based effects, as well as continuing to explore the data itself.

Blending

One important feature of the Netflix Grand Prize solution, and an increasingly talked about technique in general, is blending predictions from different models to achieve a lower overall error. Concretely, this usually means taking a linear combination of the individual model predictions, where the weights are computed per-example using another model (gradient-boosted decision trees are a popular option [1, 2]). It is also common to train separate blend weight models for different subsets of users [2]; we've also seen in homeworks that there are majority vote or selection-based methods for combining predictions, which may be more appropriate for classification tasks (depending on whether you treat predicting ratings as classification or regression – both perspectives seem to have merit). We plan to explore at least a few of these methods.

Time-sensitive baselines

Another important feature of [1] is the use of time-changing baseline predictors, as well as a frequency baseline that accounts for the number of ratings a user makes in a day, which was important in the Netflix dataset. As we saw in our data exploration, it is important in the Yelp dataset; 1- and 5-star reviews mean something different coming from a user rating many businesses per day than they do from a user rating just one. We will attempt to evaluate how including each particular term in our the baseline predictors affects our RMSE.

Even more time-sensitive baselines

In many ways, time is much more of a factor for Yelpers seeking food or drinks than it is for Netflix users picking their next show. Although public opinion about a show or movie sometimes changes over time, the actual quality of a restaurant's food can change much more rapidly. This can make past restaurant rating data less reliable for predicting future ratings. We're interested in seeing if we can reduce RMSE by weighting recent ratings more heavily when computing baselines, or perhaps trying to detect anomalies. To evaluate performance, however, we may have to test our performance only against recent ratings in our test set.

Another factor we noticed was that day of week had a strong correlation with ratings for certain types of businesses. This was especially true for bars and 'Nightlife' businesses (shown below). The plot shows the ratings that are above the mean in red and the ratings that are below the mean in blue. The average number of ratings per day of the week is shown in the right side vertical axis and the number of ratings on any one day of the year is shown in the axis above the plot. We clearly see that the reviews that are given on Sundays and Mondays tend to be more positive than the reviews given on the other days of the week.
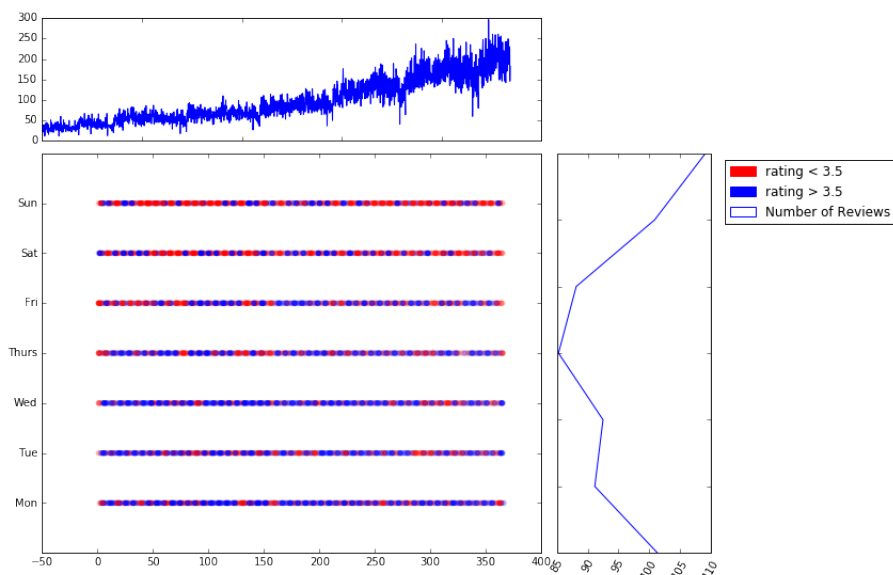

Figure 1: Plot ratings above and below the mean for days of the week vs days in a year for resturants of categories bar and nightlife in Phoenix

A similar analysis can be applied to various sub locations within the greater Phoenix area. Figure 2 shows the locations of businesses that are tagged to three cities within the Phoenix area. We can compare the reviews that these locations are receiving over time, as shown in Figures 3, 4 and 5.
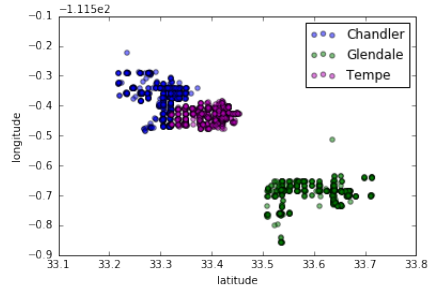
2

Figure 2: Plot of location of reviews tagged to three cities in Phoenix

For these different locations we see that to a greater or lesser extent, Sunday reviews receive the highest scores. We also see (focusing on the right hand line plot) that Chandler and Tempe receive roughly 5 and 10 reviews per day more than Glendale. The plot also uses the opacity of the dot to indicate numbers of reviews and there appears to be a trend that in the winter months (days of the year 200 through to 50), the Glendale area is particularly quiet (especially when comparing this to the other cities). Finally, we notice that Glendale appears to be a satellite city within Phoenix and this could be a possible reason for this.
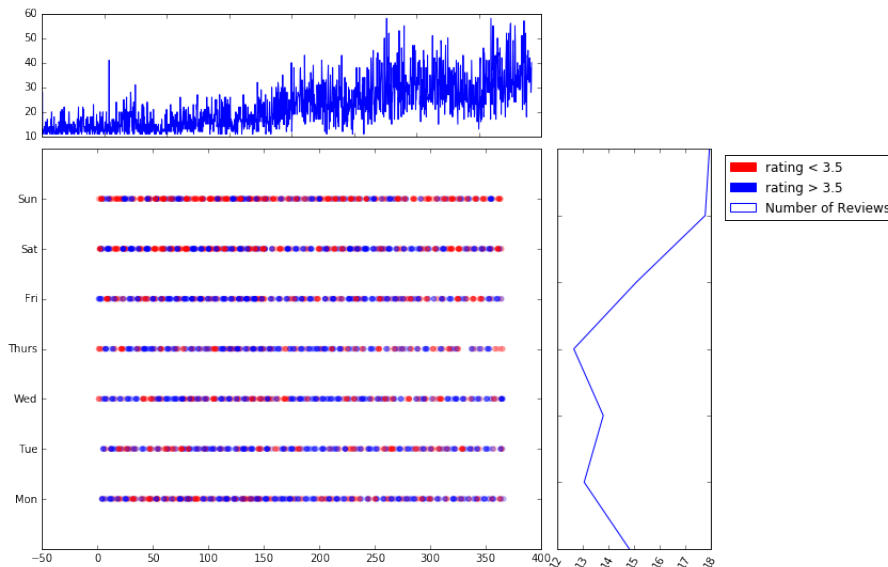

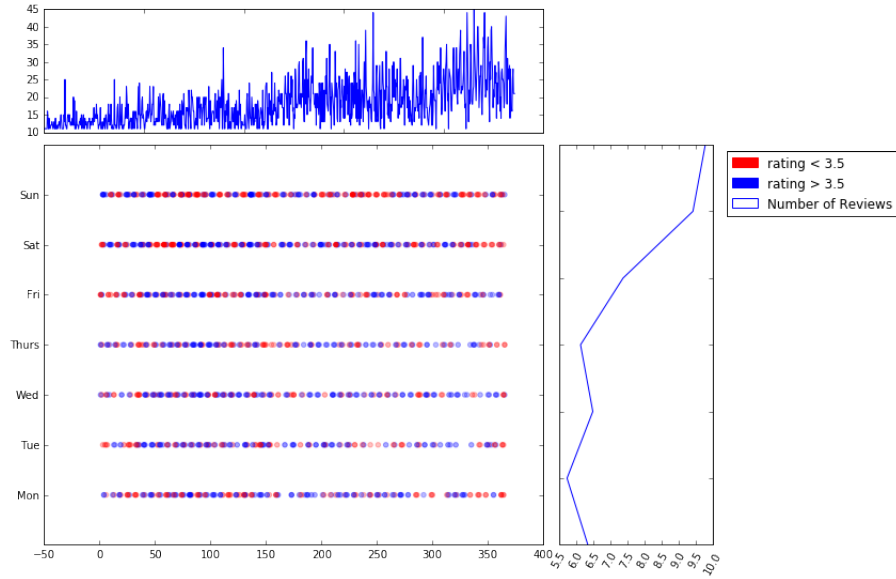Figure 3: Plot of reviews above and below the mean for Restaurants in Chandler

Figure 4: Plot of reviews above and below the mean for Restaurants in Glendale
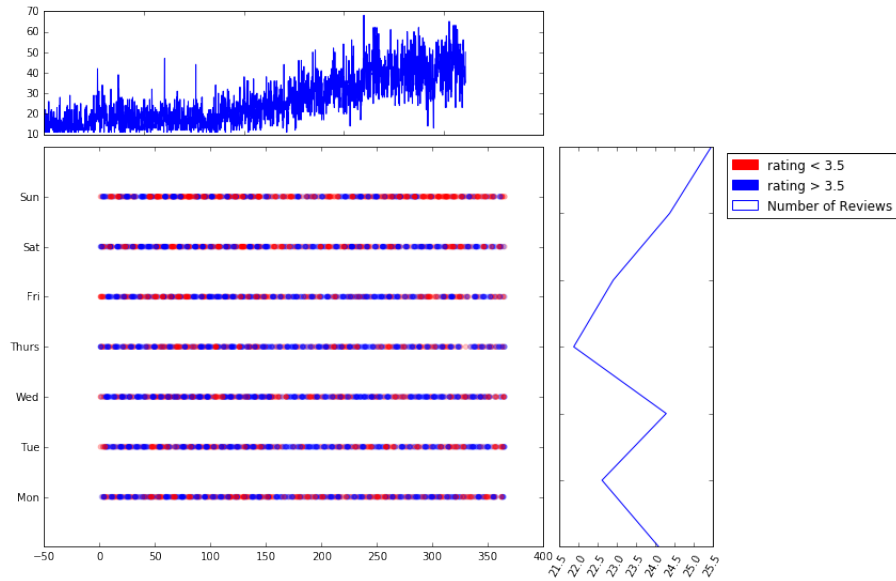

Figure 5: Plot of reviews above and below the mean for Restaurants in Tempe

Weekly or seasonal effects were not something discussed at all in [1], perhaps because they aren't as meaningful for movies (It's a Wonderful Life and Bad Santa aside), but we're interested in seeing if we can use them to improve our predictions for restaurants and bars a bit further.

Bayesian Baselines

Building on another avenue we went down during data exploration, we think it might make sense to assume a prior probability distribution (perhaps a scaled Beta) for each business and user baseline, initially centered at the global mean. For users and businesses with lots and lots of ratings, that prior will become irrelevant, and the baseline will converge to their average rating, but for users and businesses with few ratings, their baseline would be somewhere between the global-mean prior and their empirical average. It should be fairly straightforward to modify our simple averaging code to predict the posterior expectation of a Beta distribution rather than the raw average, and if we obtain promising results there, we may attempt to use it in our overall baseline computation.

Location-based analysis and going beyond rating RMSE

We have also explored a number of other models for predicting business success rather than focusing only on recommendation systems. For example, we can use the Yelp dataset to measure business popularity (using review counts or checkin volume as well as average rating), or even whether and when a business eventually closes. We've spent time investigating location-based, seasonal, and other temporal trends in these business metrics. Location based models have focused largely on clustering methods, as neighborhood is a better predictor than absolute latitude and longitude values. Trend analysis will use regression models over time to analyze if businesses are cyclically popular at any given lag or to find points in time when a given business was comparatively 'hot' within its lifetime.

We also built a network-based recommendation system based roughly on the PageRank algorithm. This model finds the most 'connected' business, where path lengths between businesses are weighted inversely to how much a user liked both of the restaurants. To find the next recommendation for any given user, we search the graph built off of similar users and return the 'closest' business the user has not yet been to. Our aim for the remainder of the project is to use these varying models and insights to build a combination model with reasonable priors that should outperform the baseline models that have been presented in milestone 4.

References

[1] Yehuda Koren, The BellKor Solution to the Netflix Grand Prize

[2] Michael Jahrer, Andreas Tscher, and Robert Legenstein, Combining Predictions for an accurate Recommender System