Summary of "Neighborhood and Price Prediction for San Francisco Airbnb Listings"

James Gearheart, Danny Zhuang, Bob Saludo, and Ryan Wallace

October 13, 2016

In "Neighborhood and Price Prediction for San Francisco Airbnb Listings", Tang and Sangani explore data from the Inside Airbnb project, containing a complete set of 7,029 listings of properties for rent on Airbnb in San Francisco as of November 2, 2015. For each listing, the data set contains the price per night, text information, such as the name of the listing and a description of the property and host, an image of the property, and a number fields describing quantifiable attributes of the property, such as amenities offered and square-footage. With these data, the authors seek to predict two attributes of a listing, the price per night, and the neighborhood in which the property is located. The authors choose these metrics for their potential usefulness to Airbnb's platform. In particular, predicting price may allow Airbnb to offer improved suggested pricing for hosts ready to list, while predicting neighborhood could provide a first step towards recommending new properties to clients based on where they have stayed in the past.

The authors make several interesting decisions about how to approach the tasks of predicting price and neighborhood. Rather than attempting to predict price as a continuous variable, a binary response variable for price is created that indicates whether the predicted price is above or below the median price in the data set. This method has the advantage of simplifying both the price prediction task and the evaluation of the goodness of the model. However, disadvantages of this approach include that simple linear regression is no longer a suitable option to model the price response variable, and that the results of the prediction are of limited usefulness, especially for the application of suggesting pricing to hosts. Additionally, in the data cleaning phase, all observations in neighborhoods containing 70 or fewer listings are removed. While this reduces the burden on the neighborhood classifier, it also restricts the models usefulness to neighborhoods with a relatively high number of listings.

The majority of the work is done in extracting features from the data set in order to prepare the data for modeling. A total of five features are chosen: listing information features, bag of words features, word class features, text sentiment features, and visual features. The authors employ a variety of techniques in order to extract these features. Interestingly, pre-developed packages are used to do the majority of processing for all but the listing information feature. In particular, packages from the Natural Language Toolkit such as Porter-Stemmer seem to be useful for the processing of hand-written data into quantifiable features. Similarly, OpenCV's libraries for extracting SURF features from images appear useful for the processing of the images associated with the listings.

After creating a feature vector for each listing, two support vector machine models are trained to predict each of the response variables. The choice of SVM classifier is not explained; although, given the large number of features, the choice seems reasonable. Nevertheless, alternative methods may be worth investigation. Two major adjustments are needed to improve the models. First, the optimal value of the regularization parameter is experimentally determined by gridsearch. Second, high variance in the initial model indicates that overfitting is likely. In order to combat this, the authors use the Recursive Feature Elimination functionality in sklearn to reduce the degrees of freedom in the SVM.

The final models for both price and neighborhood have impressive predictive power. In the test set, prices are categorized to the correct group (either above or below the median) with approximately 81% accuracy, while neighborhoods are categorized with 42% accuracy. While this performance seems good, the authors note that discretization of the predicted price into smaller bins is likely necessary for most applications, and is a promising direction for future work.