Prediction of NYC Restaurant Health Inspection Results

Michael Walter

13 December 2013

Project Goal

Every year, the New York City Health Department inspects each of the 24,000 restaurants located in the city, and evaluates them on compliance in food handling, food temperature, personal hygiene, and vermin control. To help the public understand the results of these inspections, the Health Department developed a grading system in 2010, where a grade of 'A', 'B', or 'C' is assigned to each restaurant based upon the inspector's evaluation, with A being the best grade. Each restaurant is then required to post its grade at the restaurant's entrance, so potential patrons can have insight into the restaurant's health standards.

Since restaurants in New York City have a fairly high turnover rate, a restaurant will often have a sign saying 'Grade Pending' or 'Not Yet Graded', which means potential patrons are left to guess what the health grade would be. The potential patron would, however, know various facts about the restaurant, such as restaurant name, type of food, street address, borough, and zip code. The goal of this project is to take these basic facts and predict whether a restaurant will receive a grade of 'A' or a lower grade.

Data

New York City publicly posts the results of their restaurant inspections in an online database, and makes them downloadable as text files (https://nycopendata.socrata.com/). These text files contain information on every inspection the city has performed, and includes the inspected restaurant's name, address, zip code, borough, food type, inspection date, health violations found, inspection score, health grade, and any punitive actions taken by the health department. Since the goal of this project is to help a member of the general public predict the health grade, the input variables will be limited to only the information that a patron would know.

To make the information useable, the text files are first converted into a spreadsheet format for further processing in Matlab. Each restaurant inspection is then assessed and removed if the inspection did not produce a health grade, since many inspections are not graded. Next, those with 'Grade Pending' or 'Not Yet Graded' scores are removed. Additionally, since most restaurants have been inspected and graded multiple times, the grades for a given restaurant are averaged to eliminate counting a restaurant multiple times, which results in each restaurant being assigned its typical grade. Finally, the data is prepared for training and testing, by selecting a random distribution of 75% of the data as a training set, and 25% as a test set.

After processing the data and turning it into a useable format, 2,768 of the city's 24,000 unique graded restaurants are available for training and testing. Of these restaurants, 52.6% received a grade of 'A', while the rest received a lower grade.

Approach

The overall approach taken to classify a restaurant's health grade as an 'A' or lower is to first assess individual factors that a potential patron would know and determine the probability of the restaurant receiving a certain grade, given that individual factor. The probabilities for each factor are then collectively evaluated by performing logistic regression with each probability being treated as a feature. Finally, a classification prediction is made and evaluated for all values of a test set.

To help assist with feature selection, each feature is independently evaluated by holding out a test set when training, and then evaluating that features performance on that test set. This is relatively easy, since the output of each independent feature evaluation is the probability of an 'A' for each test case.

There are two issues worth noting when performing feature selection against the restaurant data set with features of borough, address, zip code, name, and food type:

1. The features can be grouped into three classes of data with similar characteristics: textual data (name and street address), features where each possible value is independent from the others possible values (borough and food type), and features where the possible values are related to each other (zip code).
2. There is correlation between certain features that occur in the data set, such as between borough and zip code, or name and food type.

While features of different classes can be evaluated together without a problem, using two correlated features can be problematic, and is therefore avoided.

Borough

To determine the probability of a restaurant receiving an 'A' given a specific borough, Bayes formula is easily implemented using indicator functions. The success of this factor is then tested by applying the Naïve Bayes classification algorithm, and running it against a test set. While one borough (Staten Island) results in a prediction accuracy of 74.2%, the overall prediction accuracy for the city is 51.3%, which is almost equal to random guessing and therefore implies that the borough is probably not a useful factor.

The poor performance of this factor is likely due to the fact that each borough is very large and consists of many diverse neighborhoods, which means there is too much variation within each borough. This point is supported by the fact that Staten Island, at 470,000 people, is one third the size of the next smallest borough, and therefore has the greatest probability for homogeneity throughout the entire borough.

Food Type

A Naïve Bayes classifier is also used to evaluate a restaurant by the feature of food type. Unlike borough, however, classifying a restaurant by food type results in a statistically significant accuracy of 61.4%, which implies that this feature is better for prediction. In particular, certain food types are more indicative of a certain health grade than others. For example, the following food types (each with more than 10 test restaurants) result in a generalization error of less than 30%:

Asian
Chinese/Japanese
Donuts
Greek
Ice Cream
Indian
Indonesian
Juice/Smoothie
Russian
Sandwiches
Steak
Turkish

This implies that these food types have a strong correlation between food type and health grade. On the other hand, the following food types (each with more than 10 test inspections) have a generalization error greater than 45%, and are therefore not as useful in predicting a grade:

African	Bagels
Bakery	Italian
Korean	Kosher
Middle Eastern	Pizza
Spanish	Vegetarian

Name

To evaluate a restaurant's name, a text classifier that determines the probability of a grade given a specific name needs to be used, such as a Naïve Bayes text classifier. While mathematically similar to the classifier used for borough and food type, this approach requires significantly more data processing. This is also similar to the spam filter discussed in the lecture notes. The accuracy of the independent name classifier is 58.3%, which is statistically significant.

The training set of 2,757 different restaurants results in 2,948 unique words with which to classify. The words that result in the strongest probability of accurately predicting a grade are:

Bombay	Chen
Donuts	Dunkin
Fusion	Ice
Pain	Quotidien
Shoppe	Town

It is important to notice the correlations between the name and food type. For example, the food type of 'donuts' is a strong predictor, as are the words 'Dunkin' and 'Donuts'. Other obviously correlated names and food types are: 'Indian' and 'Bombay', 'Ice Cream' and 'Ice', and 'Sandwich' and 'Pain'. This correlation demonstrates the futility in combining both name and food type when trying to make a final prediction, even though there is a significant accuracy for both name and food type.

Street

Classifying the restaurant by street address uses similar text classification algorithms as classifying the restaurant by its name. In this case, the training set of 2,757 different restaurants results in 656 different street words. While there are certain streets (such as Richmond, Flatlands, Pearl, and Knickerbocker) that are indicative of a specific restaurant health grade, the prediction accuracy against a test set is only 52.5%. The street evaluation factor is therefore not very useful.

Zip

The final feature that is easy for a restaurant patron to know is a restaurant's zip code. While the other factors are determined using a Naïve Bayes classification algorithm against a categorized set, the zip code affords other possibilities, which arise due to the fact that consecutive zip codes often share physical boundaries. Three approaches are tried to determine the best probability of a health grade: Naïve Bayes, k-means clustering, and fitting a polynomial to zip code groups.

Initially, a Naïve Bayes classification algorithm is implemented, since it provides a test case, and is easy to adapt from the previous factors. Naïve Bayes results in a test accuracy of 55.7% for 188 zip codes, which is statistically significant, albeit less accurate than the features of name and food type.

The next approach to predicting a health grade based upon zip code is to use a k-means clustering algorithm to attempt to group the zip codes into neighborhoods that might have homogeneous characteristics. 42 cluster centroids are chosen, since New York State defines 42 different neighborhoods in New York City, which results in an average of 4.8 zip codes per clustered neighborhood. After the clusters are formed, Bayes algorithm is applied to the clustered neighborhoods to determine the probability of each restaurant receiving an A. Unfortunately, at 53%, the test results are slightly worse than for the original Naïve Bayes approach.

The third approach is to attempt to fit a polynomial that inputs a zip code and outputs the probability of receiving a specific health grade, using the following steps:

1. The zip codes are divided into groups, because NYC's zip codes exist in seven groups that have roughly consecutive numbers within them, but fairly large gaps between the groups. For example, one group (Staten Island) has zip codes 10301-10314, and then the next group (the Bronx) has zip codes that jump over a hundred numbers to 10453-10473.
2. Values of 'A' are assigned a 1, and other values are assigned a 0. These values are plotted vs zip code for each group.
3. Holdout cross validation is used to determine the optimum dimension for a curve to fit through each of the seven groups. The optimal balance between bias and variance is achieved by cycling the hold out sample points, and averaging their test accuracies for various dimensions.
4. Finally, a new curve is fit for each group, given the optimum number of dimensions found in the previous step.

Using these steps, the probability of an 'A' grade for each zip code is extimated by each polynomial, since fitting the curve essentially averages the assigned 1's and 0's for that zip code, while also being influenced by the surrounding zip codes. Since adjacent zip codes will likely have some correlation, this is a promising technique. Unfortunately, this method independently results in a test accuracy of 53.1%.

Combination of Features

The final step to predicting a restaurant's health grade is to combine several of the features using Logistic Regression, and therefore the hypothesis is:

$$h_\theta(x) = \frac{1}{(1 - e^{-\theta^T x})}$$

Where [x] are the evaluation features for each restaurant, and [Θ] are the function's coefficients. The obvious evaluation features are probability give a food type and the Naïve Bayes output for zip code, since they are the most accurate uncorrelated features (Figure 1).

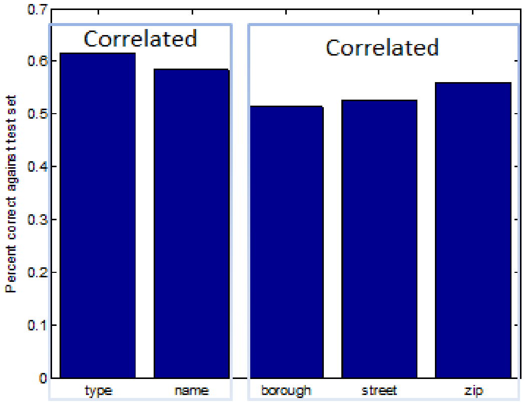

Figure 1. Test accuracy for each factor

Theta is found by running an update algorithm with a decreasing learning rate against the entire training set, until the theta values converge, which can be seen in Figure 2.

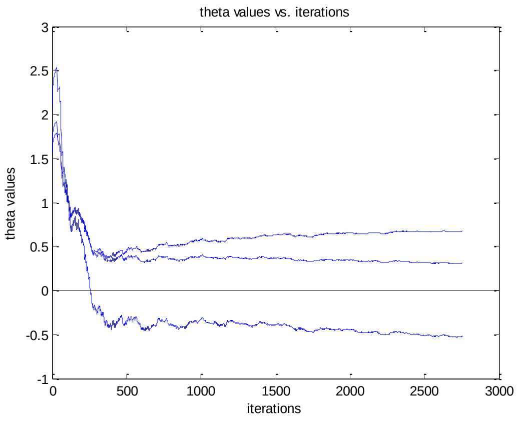

Figure 2. Theta values vs number of iterations

Finally, the logistic regression equation is run against the test set that had originally been left out, and the result is similar to the accuracy of food type, when run independently.

When logistic regression is instead run with the features of *food type* and the *polynomial fit zip code* the testing accuracy increases by approximately 1% to 62.4%. This result stays consistent when different mixes of testing and training sets are used, and therefore demonstrates a clear pattern. The reason for the polynomial fit performing worse than the Naïve Bayes during independent testing is likely due to the fact that the polynomial fit is an estimate attempt, rather than a calculated probability. Therefore applying Naïve Bayes probability logic is not quite appropriate. Once combined with food type in the logistic regression, the theta values compensate for any offset error in probability estimate and the polynomial fit works better.

In addition to zip code and food type, other feature combinations were tested to confirm that the assumptions about correlation and accuracy are valid. The results show that zip code and food type perform notably better than any other combinations (including combining all features). For example, food type and name are the largest single factors, however they were not chosen together due to their high correlation. The result of testing them together is an accuracy of 58.4%, which is 4% less than the chosen combination, and actually worse than food type by itself. This supports the decision to not include two highly correlated features.

Conclusion

The project shows that by properly applying machine learning techniques to historical New York City health grade information, a person can predict the grade of an ungraded restaurant with about 62% accuracy by only knowing the restaurant's zip code and food type.

References:

New York State Department of Health. "ZIP Code Definitions of New York City Neighborhoods". http://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

Ng. CS229 class notes. CS299.stanford.edu/

NYC Open Data. "Restaurant Inspection Results." https://nycopendata.socrata.com/