Capstone Report

Robert McKay Lothian

Sunday, November 15, 2015

1. Title

Predicting ratings from review text - can we transfer learning from restaurant reviews to a different city?

2. Introduction

A great deal of effort and expense goes into the machine learning process. Furthermore, labelled data may be expensive or difficult to obtain. In view of these issues, the ability to reuse learning in a new context would be very useful.

Using the Yelp review data, a small experiment was carried out to investigate the transferability of learning. Restaurant reviews was selected as the domain and a classifier was trained to predict the rating from the review text. The classifier was trained on the reviews of Edinburgh restaurants and then the resulting model was applied to the reviews from Scottsdale. Edinburgh was selected in order to compare a UK city with a US city. Scottsdale was chosen as the US city, because it has a similar number of restaurants listed to Edinburgh.

The model should transfer well if two conditions hold. Firstly, reviewers use the same words to describe the quality of their restaurant experience. Secondly, the mapping between those words and the rating is consistent between the two cities.

3. Methods

Data files and R scripts can be found at https://github.com/giraffebob/datasciencecapstone

Initial Data Selection

All businesses with "Restaurant" in their list of types were selected. (Alternatively, a more homogenous set of businesses could have been obtained by rejecting businesses with additional types.) After counting restaurants by city, restaurants in Edinburgh and Scottsdale were selected for study. The experimental data set consisted of texts and corresponding star ratings for every review of these restaurants.

Exploratory Analysis

It turned out that Scottsdale had many more reviews than Edinburgh, despite the similar number of restaurants. The distribution of star ratings for the two cities is shown in Figure 1.

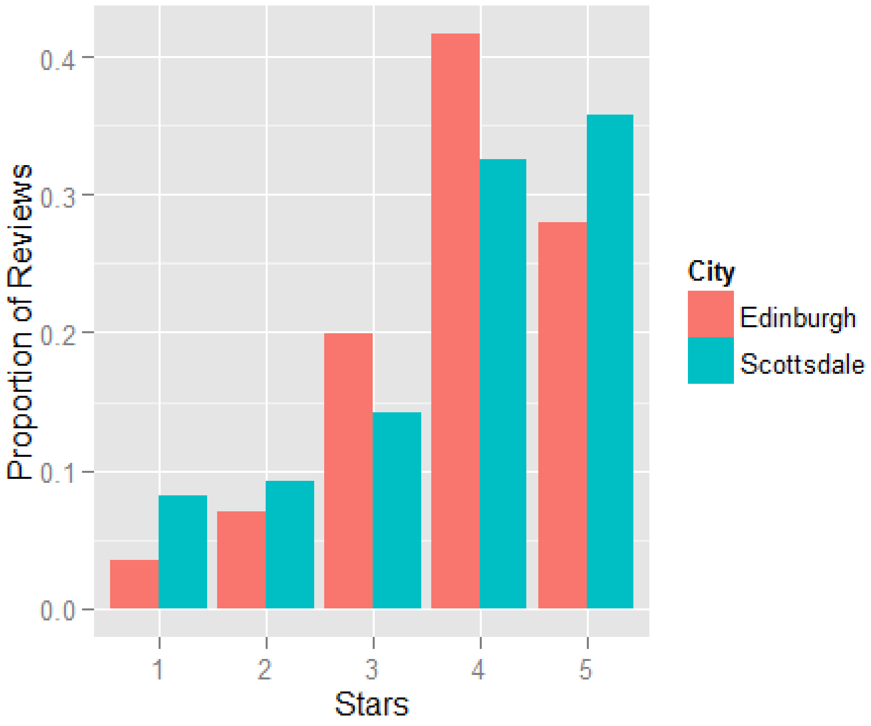

Figure 1: The distribution of star ratings for reviews of restaurants in Edinburgh and Scottsdale. Both distributions have been normalised for ease of comparison.

The Scottsdale reviews are more widely spread, giving one, two and five stars more often. This may mean that restaurant quality has greater variability in Scottsdale or simply reflect the attitude of reviewers. In both cities, only about a third of ratings were less than four stars, so it was decided to label all 1-3* ratings as "poor". Four and five star ratings were labelled as "good" and "excellent" respectively. This re-labelling led to a classification problem with three classes of roughly equal prevalence. The new class distributions are shown in Figure 2.

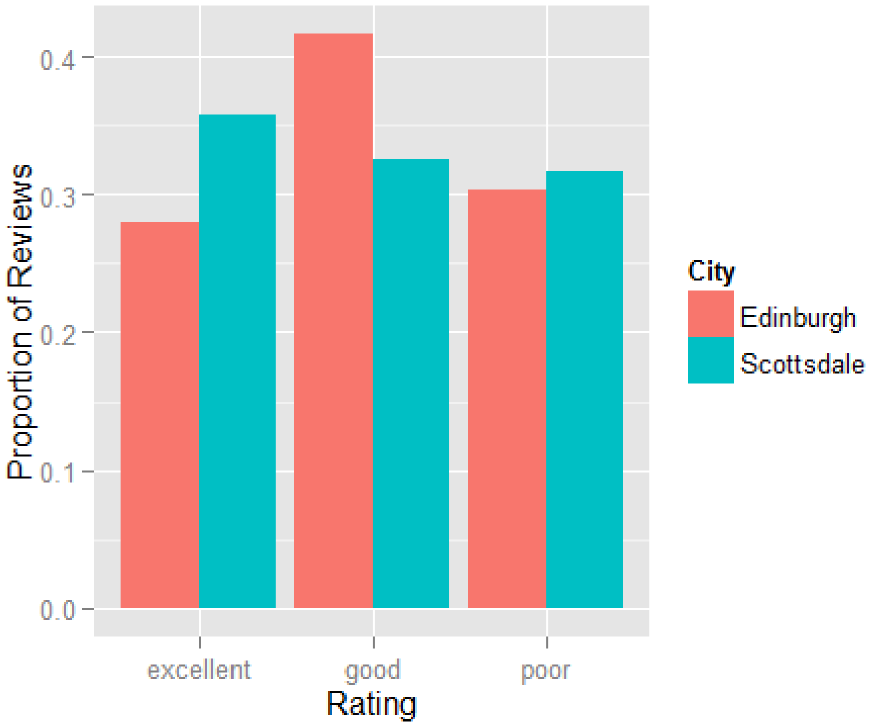

Figure 2: The distribution of rating classes for restaurants in Edinburgh and Scottsdale. Both distributions have been normalised for ease of comparison.

Preprocessing

The Edinburgh reviews were split into a training set (70%) and a test set (30%). The training set review texts were loaded into a corpus structure. Punctuation, numbers and surplus whitespace were removed and all words were converted to lower case. Stopwords were removed and stemming was applied. The document-term matrix was formed, excluding those words with fewer than 100 appearances in the corpus. This step was intended to reduce noise and allow faster training on the smaller feature set.

The same preprocessing steps were applied to the Edinburgh test data and to the Scottsdale data, except that the document-term matrices were formed using the set of terms derived from the Edinburgh training data. This allowed a model built on the Edinburgh corpus to be applied directly to the Scottsdale corpus. The performance on the Edinburgh test set could then be compared with the performance on the Scottsdale data.

Training the Classifier

The Naive Bayes classifier is simple and robust, and often gives comparable accuracy to more sophisticated methods such as the support vector machine. The Naive Bayes classifier with Laplace smoothing was applied to the training data. The model was then applied to the held-out test data and the Scottsdale data.

Evaluation Methodology

To establish that learning had occurred, the classifier accuracy was compared with the no-information rate (NIR). The NIR is just the prevalence of the most common class, which is the highest prediction accuracy possible without examining the reviews.

The metrics used to compare performance on the different data sets were accuracy and the mean of the three class sensitivities. The latter is the accuracy that would be achieved if the three classes were completely balanced in each test set. A more detailed picture was obtained by examining the individual class sensitivities.

The metrics discussed above can be derived from the confusion matrices, which are given in full for completeness.

4. Results

Confusion Matrices

Table 1 shows the confusion matrices for the training set and the two test sets.

Edinburgh Training Data

```
##            Reference
## Prediction  excellent good poor
##   excellent      1641 1215  439
##   good            442 1554  495
##   poor            202  563 1510
```

Edinburgh Test Data

```
##            Reference
## Prediction  excellent good poor
##   excellent       593  611  184
##   good            236  571  276
##   poor             98  267  577
```

Scottsdale Test Data

```
##            Reference
## Prediction  excellent  good  poor
##   excellent     22142 14726  7039
##   good           3706  6197  3680
##   poor           2226  4603 14142
```

Table 1. Confusion matrices for Edinburgh training and test and Scottsdale test data.

In all cases, the most common error is to misclassify a good review as an excellent one. This effect is most pronounced in the Scottsdale reviews. The confusion matrices are the raw data for performance evaluation. The key metrics derived from these matrices are presented and discussed below.

Overall Performance

The performance metrics for all three data sets are summarised in Table 2.

```
##                  Accuracy       NIR Mean_Sensitivity
## Edinburgh-train 0.5836745 0.4133482        0.6007960
## Edinburgh-test  0.5101084 0.4245532        0.5300585
## Scottsdale      0.5414282 0.3578083        0.5334387
```

Table 2. Performance metrics for Edinburgh training and test data and Scottsdale data.

The Naive Bayes classifier achieves 58.4% accuracy on the training set against the NIR baseline of 41.3%. We get a true estimate of the classifier accuracy from the test set, where the Naive Bayes achieves 51.0% against a NIR of 42.4%. This may not seem impressive, but some learning has been observed, so its transferability can be tested.

Surprisingly, the classifier achieves higher accuracy on the Scottsdale data: 54.1% against the NIR baseline of 35.8%. The class prevalences are somewhat different, so it may be better to compare the mean sensitivities. These are very close, at 53.3% for the Scottsdale data and 53.0% for the Edinburgh test data.

Performance by Class

Table 3 shows the sensitivity achieved on each data set for each class. We can see that the mean sensitivities conceal a large difference between classes. The poor performance on the "good" class is again evident. The similar overall performance on the two test sets conceals substantial differences in performance on individual classes.

```
##                 excellent      good      poor
## Edinburgh-train 0.7181619 0.4663866 0.6178396
## Edinburgh-test  0.6396980 0.3940649 0.5564127
## Scottsdale      0.7887013 0.2427721 0.5688428
```

Table 3. Sensitivity for each class for Edinburgh training and test data and Scottsdale data.

5. Discussion

Main Conclusions

The classifier achieves a modest, but highly statistically significant, learning gain. The most common misclassification is to label good reviews as excellent. Performance is comparable, or even slightly better, on the Scottsdale data, despite the classifier having been trained on the Edinburgh data. For this one small experiment, the learning appears to have been completely transferable.

Further Research Directions

This result suggests further interesting experiments. Firstly, note that there are many alternative classifiers, tuning parameters and data preparation choices available. It is likely that improved performance can be achieved on the Edinburgh data. Would this higher performance transfer as fully as the modest learning seen here? Would optimum performance on the current domain come at the expense of transferability? Secondly, transferability between cities is not a very ambitious goal. How much learning could be transferred to reviews of different business types? How much learning could be transferred to reviews from a different source?