CSC2506 Assignment 1

Matthew Scicluna

2017-01-13

Question 1 and 2

See paper attached at end of document for the answers to these questions.

Question 3: Spam classification using logistic regression

We used the glmnet package to get a l^2 regularized logistic regression function. We used the `train` function from the Caret package to train the regularization parameter (lambda) using 5 fold cross validation. The regularization parameter was selected from a simple grid search and we only considered values of lambda that were between 0 and 0.1, since practice runs indicated that values > 0.1 produced very innacurate classifiers. We transformed the data in one of four ways and trained the same regularized logistic model and compared the resultant optimal lambda along with its training and test error. The results are presented in table 1.

Table 1: The best Lambda and training and test errors of each model with different data transformations applied

	Lambda	Training Error	Test Error
No Transformation	0.020	0.086	0.074
Centered Data	0.021	0.074	0.081
Log Transformed Data	0.026	0.069	0.053
Binarized Data	0.027	0.076	0.064

In figure 1 we can see the effect of each data transformation of the selection of the optimal lambda. We can see that the accuracy is always highest at small values of lambda and gets progressively worse, justifying our decision to only consider lambda's ≤ 0.1

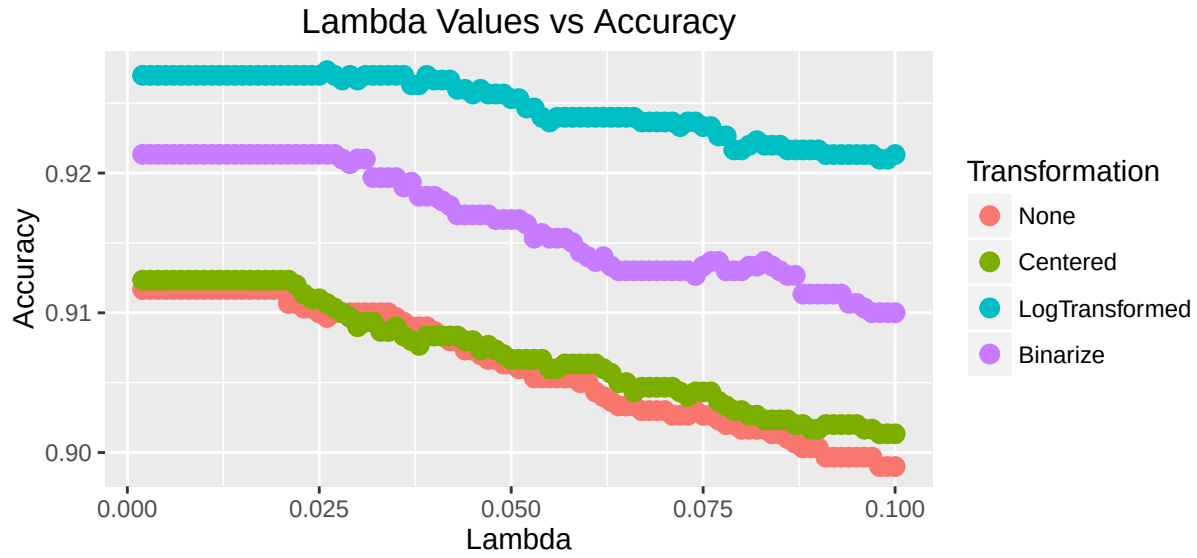

Figure 1: Comparing the accuracy of each Lambda between the models with different data transformations applied

We notice that by applying $\log(x_{ij}+1)$ to each $x_{ij} \in Dataset$ results in the model having the lowest test error. We look at absolute value of each of the weights of the logistic regression model trained on the logarithmically transformed data to see which 5 features are most informative in determining that an email should be marked as spam.

Table 2: The values of the 5 smallest and largest weights of the logistic model trained on log transformed data

V38	V12	V40	V2	V57	V7	V23	V24	V11	V53
0.002	0.003	0.0044	0.0047	0.005	0.0309	0.0319	0.0343	0.035	0.0479

We see that V38, V12, V40, V2, V57 are the weights with the smallest values (and thus they govern the features most likely to indicate non-spam) and V7, V23, V24, V11, V53 are the weights with the largest values, and so are most likely to indicate spam.

Question 4: Collaborative Filtering

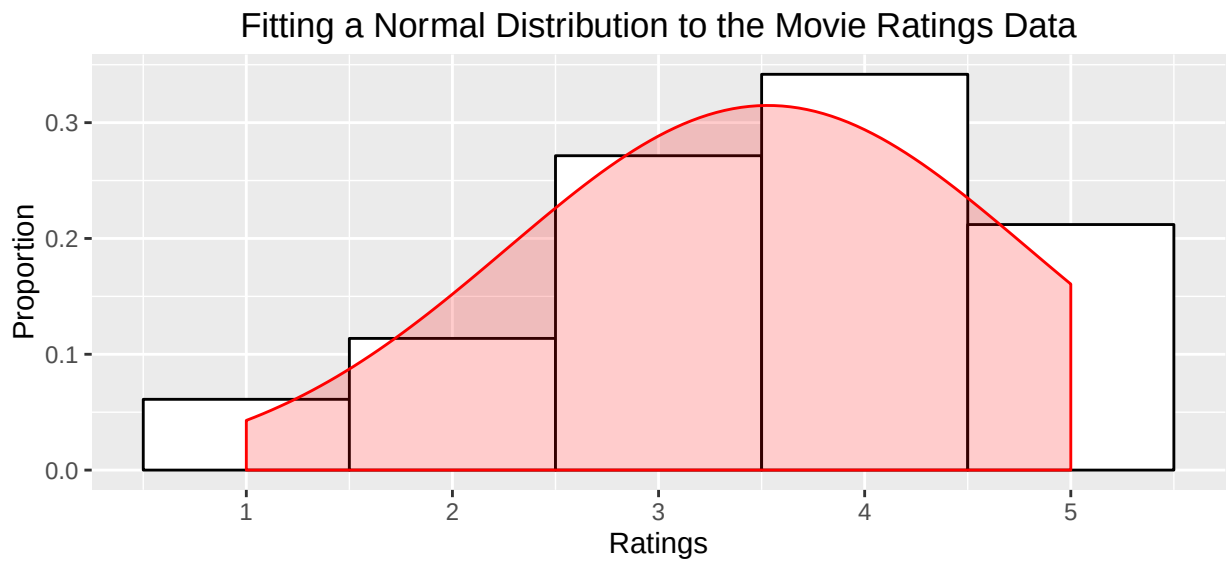

Figure 2: Assessing the fit of the normal distribution with its maximum likelihood parameters to the movie ratings data. The normal density is plotted in red.

The log likelihood for a normal distribution is as follows:

$$\ln \mathcal{L}(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Differentiating the log likelihood and setting this to zero yields the following maximum likelihood estimates: $\mu^{(mle)} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma^{2(mle)} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$.

We plotted the empirical distribution of the movie ratings along with the fitted normal distribution superimposed on it. We clearly see that the distribution has a heavy tail, meaning that there are a lot more ratings above 3 than below or equal to it. This is unsurprising since we are analyzing the distribution of user submitted movie ratings. People submitting rakings to these movies would have been willing to watch the movies in the first place– meaning that the movies must have had some sort of appeal to them apriori.

We can see that the normal distribution is not a good fit since it has symmetric tails and is unable to capture the skewness of the empirical distribution.

We next fit the data to the beta-binomial distribution. This distribution has the following likelihood:

$$\sum_{i=1}^{N} \ln \mathcal{L}(x_i \mid \theta) = \sum_{i=1}^{N} \ln \binom{N}{x_i} \frac{\mathrm{B}(x_i + \alpha, N - x_i + \beta)}{\mathrm{B}(\alpha, \beta)}$$

We can derive the method of moments estimates analytically. The estimators are as follows:

$$\alpha^{MoM} = \frac{Nm_1 - m_2}{N(\frac{m_2}{m_1} - m_1 - 1) + m_1}$$

$$\beta^{MoM} = \frac{(N - m_1)(N - \frac{m_2}{m_1})}{N(\frac{m_2}{m_1} - m_1 - 1) + m_1}$$

where $m_1 = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $m_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$

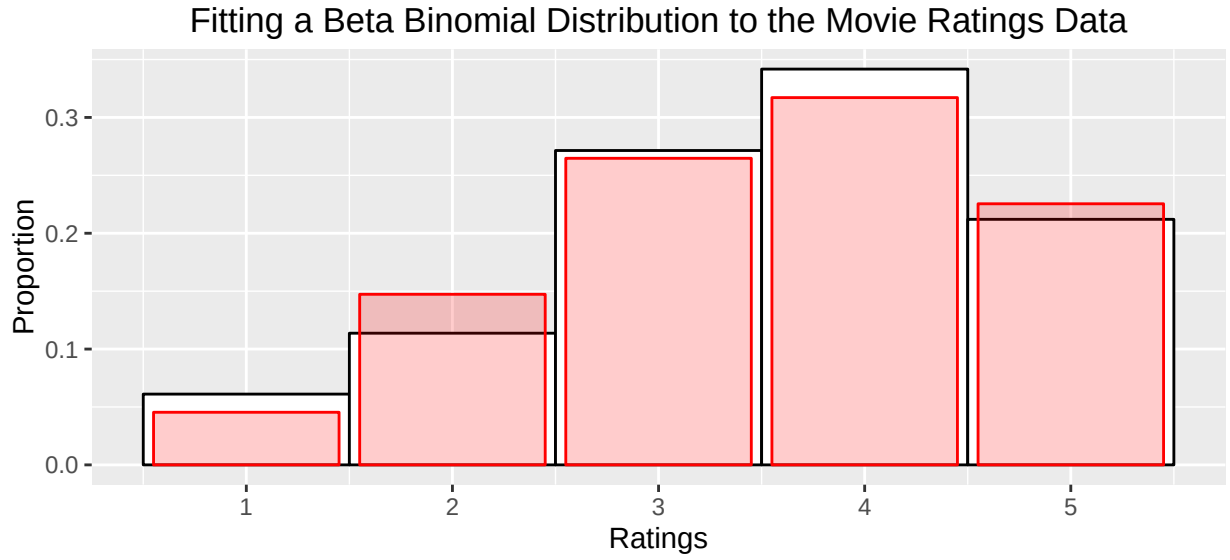

Figure 3: Fitting a beta binomial distribution to the movie ratings data

After some algebra, we get that $\alpha^{MoM} = 4.6$ and $\beta^{MoM} = 2.67$. The corresponding distribution is plotted in figure 3, superimposed to the empirical distribution. We can see that this distribution has a good fit, as it captures much more of the skewness compared to the previous normal fit.

Finally, We fit a normal distribution and a beta binomial distribution to a subset of the data, leaving 20% of the data out as a test set to compare the fit of each model.

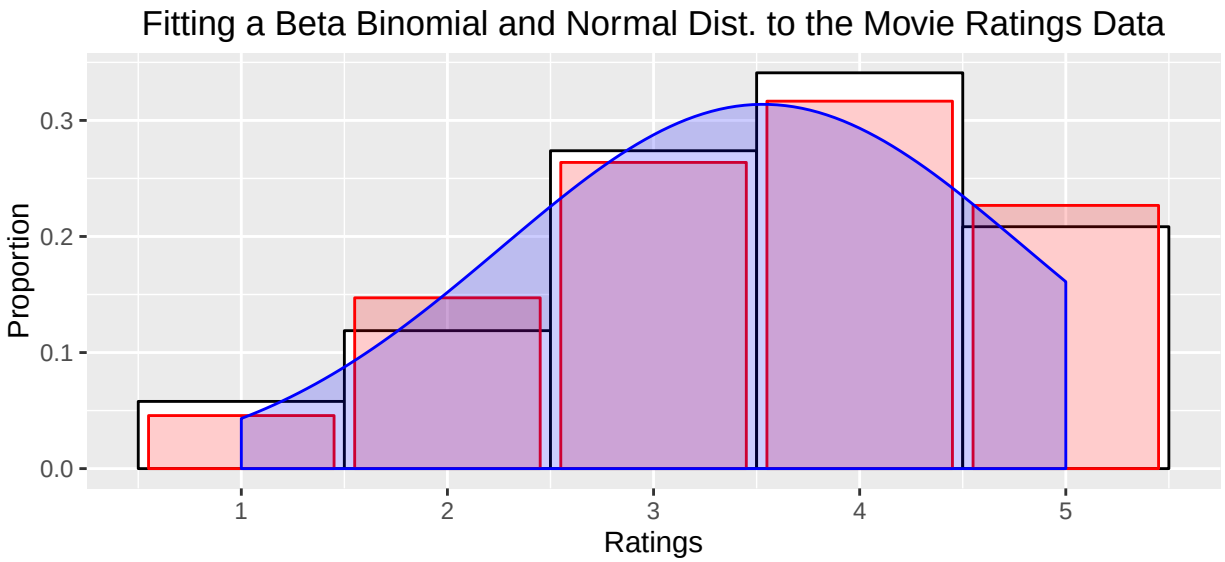

Figure 4: Assessing the fit of both Normal and Beta Binomial models fit to the same training data against the test data

The log likelihood for the Normally fitted model was -3.0922×10^4 and for the Beta Binomial model it was -2.9456×10^4 . Not surprisingly, we see that the data was more likely to come from the Beta Binomial distribution than from a normal one. This confirms our observation from figures 2 and 3 that the Beta Binomial was a better fit to the data.