MSD Homework 2

Jeff Hudson (jdh2182)

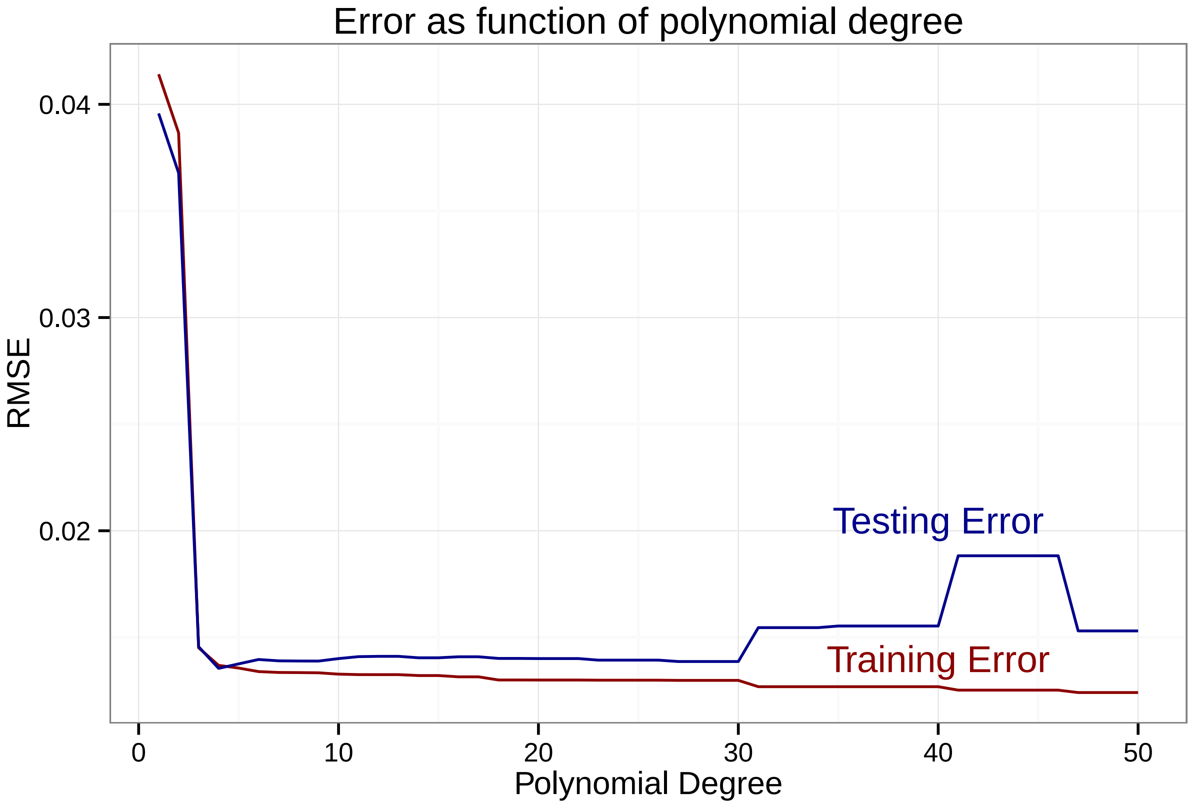
Monday, April 20, 2015

1. Cross-validation for polynomial regression

In this problem you will use cross-validation to determine a best-fit polynomial for the data provided in `polyfit.tsv`.

Use a 50% train / 50% test split to select the polynomial degree with the smallest test error, as measured by RMSE. You may use `lm()` to fit models along with the `poly()` function.

Provide a plot of the training and test error as a function of the polynomial degree, indicating the optimal degree. For this optimal degree, also provide a separate scatter plot of the data with the best-fit model overlayed. Report the coefficients for the best-fit model.

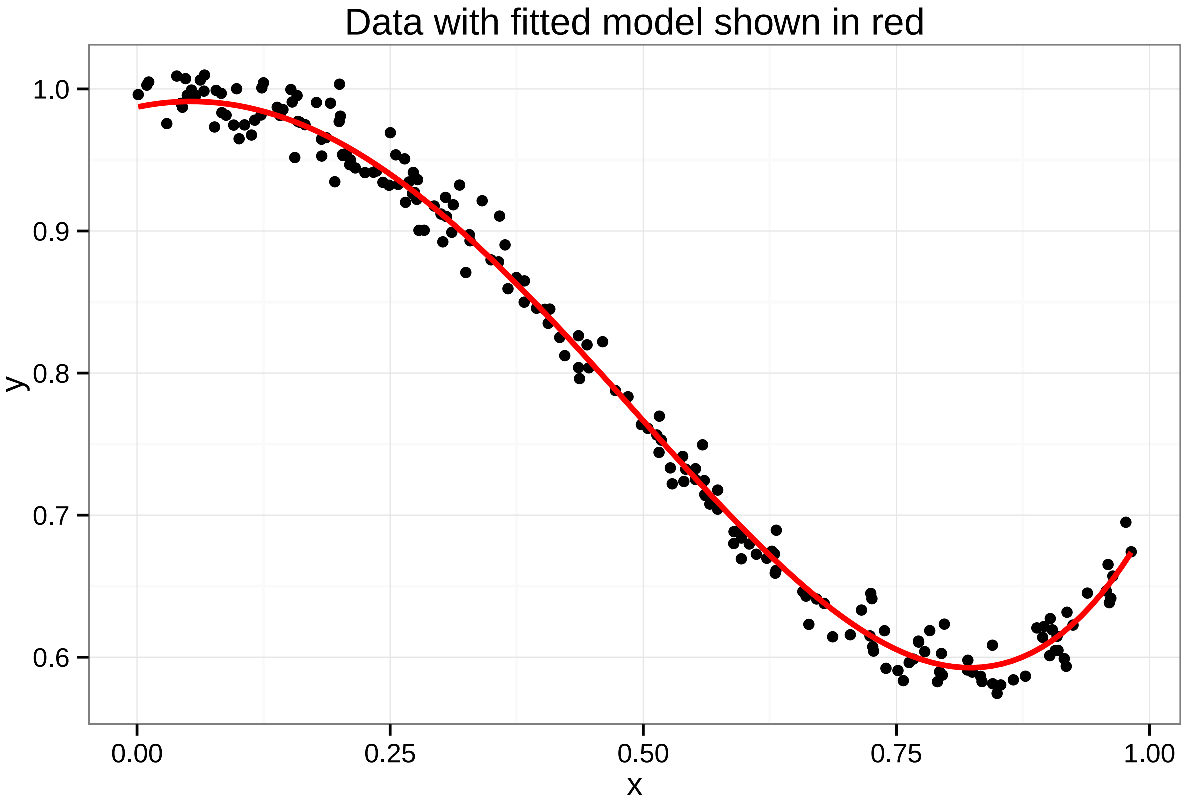


```
## [1] "Lowest RMSE: 0.0135"

## [1] "Best model is polynomial of degree: 4"

## [1] "Coefficients:"

## (Intercept)         `1`         `2`         `3`         `4`
##   0.9872302   0.1491746  -1.3904849  -0.1132470   1.0626240
```



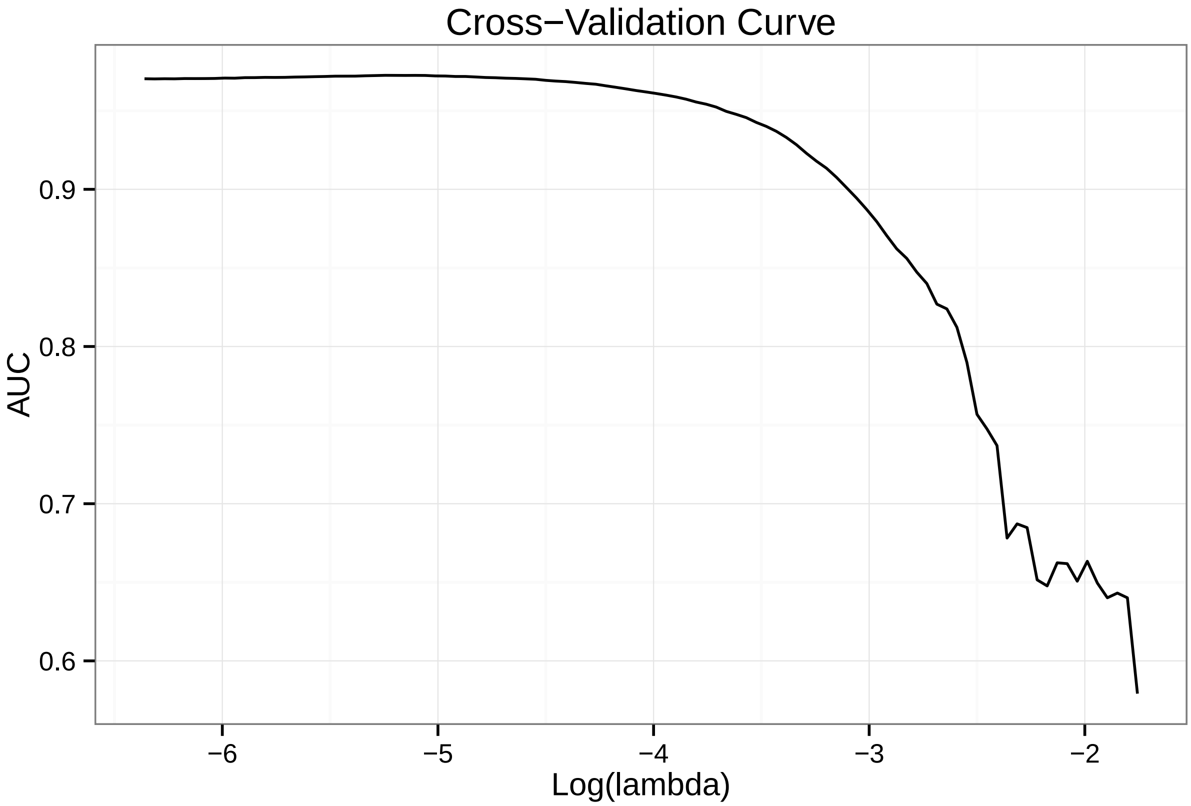
2. Logistic regression for article classification

In this problem you will use logistic regression to build a text classifier that predicts the section that an article from the New York Times (NYT) belongs to based on the words it contains.

`business.tsv` contains 1000 recent articles from the Business section of the NYT and `world.tsv` contains 1000 recent articles from the World section. `get_nyt_articles_by_section.R` was used to create these files, and is included for completeness, but does not need to be run.

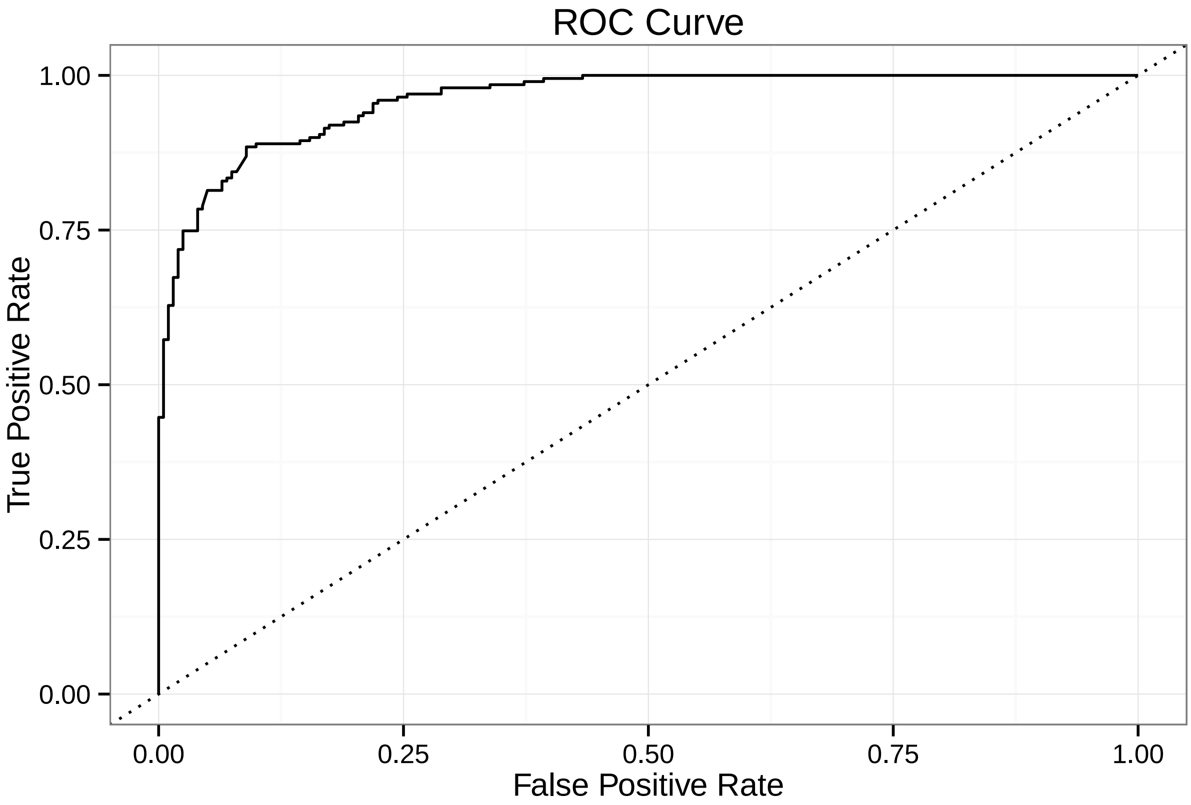
Read in each file and use tools from the `tm` package—specifically `VectorSource`, `Corpus`, and `DocumentTermMatrix`—to parse the article collection. Then convert it to a `sparseMatrix` (code provided) where each row corresponds to one article and each column to one word, and a non-zero entry indicates that an article contains that word.

Then create an 80% train / 20% test split of the data and use `cv.glmnet` to find a best-fit logistic regression model that maximizes area under the ROC curve (AUC) for the training data. Provide a plot of the cross-validation curve from `cv.glmnet`. Quote the accuracy and AUC on the test data and use the `ROCR` package to provide a plot of the ROC curve for the test data. Also show weights on words with top 10 weights for "business" and weights on words with the top 10 weights for "world".



```
## [1] "Best Lambda: 0.00528"

## [1] "Number of words: 666"
```



```
## [1] "Accuracy: 0.8975"
```

```
## [1] "Area Under Curve: 0.9611"
```

```
## [1] "Top words for 'Business' section:"
```

```
##      weight        word
## 1  2.000677     obama's
## 2  1.763163     updated
## 3  1.662234  publishing
## 4  1.384053  blackstone
## 5  1.379470      george
## 6  1.332936   executive
## 7  1.271848     company
## 8  1.230567    arbitron
## 9  1.136614     comment
## 10 1.111123         nbc
```

```
## [1] "Top words for 'World' section:"
```

```
##       weight       word
## 1  -4.220562      faced
## 2  -2.863906     pounds
## 3  -2.844702     donors
## 4  -2.753147       iran
## 5  -2.697001      fence
## 6  -2.685469       pope
## 7  -2.526128        war
## 8  -2.519375  explosion
## 9  -2.508869  combining
## 10 -2.466498 organizers
```