Report 02 — Regression and classification

Jonathan

Abstract

We consider linear regression (LR), linear regression with forward selection (FLR), decision trees (DT), k nearest neighbors (KNN), naive bayes (NB), artificial neural networks (ANN), and multinomial regression (MNMR) methods. For regression problem, we found that the ANN model performs best, and the FLR and the average age (AVE) models are indistinguishable from each other. For classification problem, we found the ANN and MNMR to be the best performing models, and showed them to perform better than the largest class (LCl) classifier and the linear regression models (LR) using the paired t-test.

1 Regression

1.1 Problem description

We have selected the regession problem detect spam emails from regular emails given 57 features. The features used are wordfrequencies of 54 different words/sings, the amount of capital letters, the longest sequence of capital letters and the average and the average sequence of capital letters.

1.2 Linear Regression with forward selection

As our dataset has 57 features it is infeasable to compare them all. Therefore for the for this section only the 12 first features are used. The 12 features.

Firstly we check if there are any obvious relationships between the features and if its spam or not. As can be seen there is a difference between the spread for spam and non spam.

a) b) c)

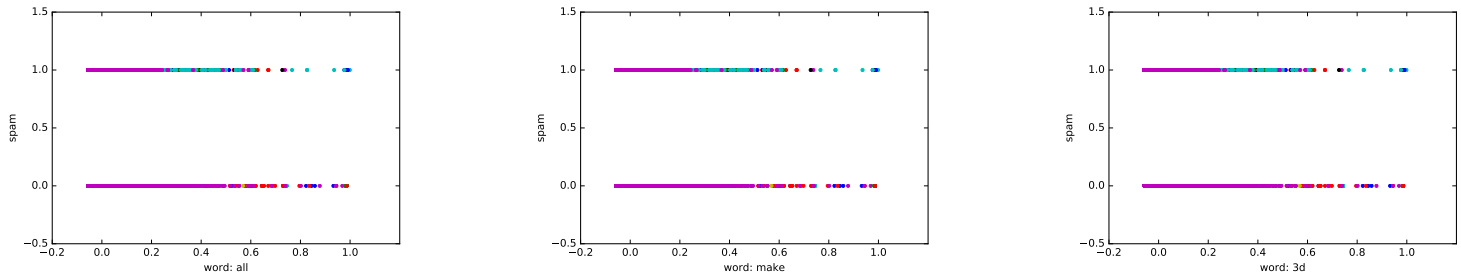

Figure 1. Spam with some features

From this we can see that the spam and nonspam is hard to distinguish just using one feature. But it is possible to see some patter as the spam is slighly less spread than nonspam.

To figure out what features to use with LR we do forward selection with a 10-fold cross-validation. Intially this would not run as the 0 feature error is less than any of the features applied to it. To show that we did it though a inital loss was hardcoded $loss\_record = [0.27]$.

This was the best feature selection that LRF could create with 12 features. If we look at the error rates that are computed
2 enumerate

Training error: 0.16731688100187755

Test error: 0.1690774799900705

R^2 train: 0.29925175624581624

R^2 test: 0.2907848014802547
enumerate

Training error: 0.1681589246441137

Test error: 0.1699619894406495

R^2 train: 0.2957251509210689

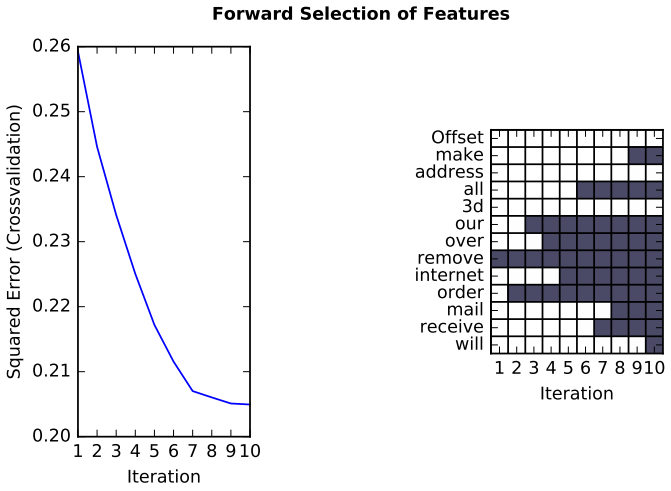

Figure 2. Forward selection form 12 features

R^2 test: 0.28707462348598956

We can see that without feature selection we get a smaller error. From this we can then conclude that Linear regression will not give us a very good estimator.

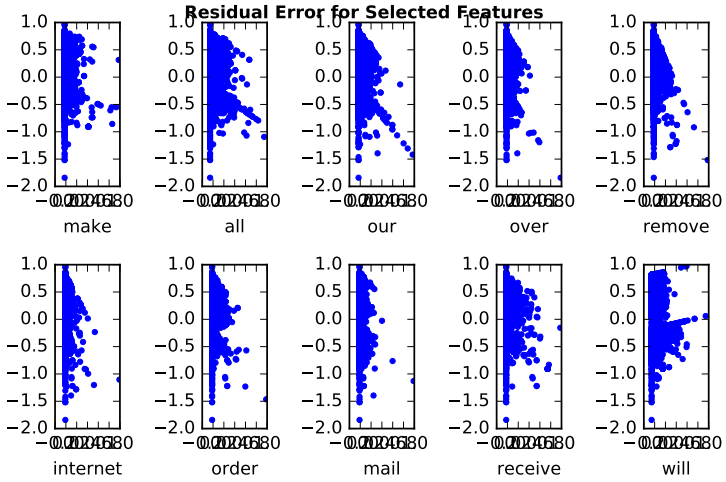

Figure 3. Text.

References

[1] http://archive.ics.uci.edu/ml/datasets/Abalone

[2] https://en.wikipedia.org/wiki/Generalized_additive_model

[3] S. Waugh,"Extending and Benchmark Cascade-Correlation," Thesis, 1997.

[4] H. Mayukh, "Age of Abalones using Physical Characteristics: A Classification Problem," ECE 539 Fall 2010 Project Report, Department of Electrical and Computer Engineering University of Wisconsin-Madison, 2010.