Report 02 — Regression and classification

Markus s123692 & Jonathan s123094

November 8, 2016

Abstract

In this assignment we will explore 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.

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/signs, 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 different features and if its spam or not. It is hard to tell if there is a distinct relationship between the features. If any of them have a linear relationship they can be reduced to a single feature. (see figure1)

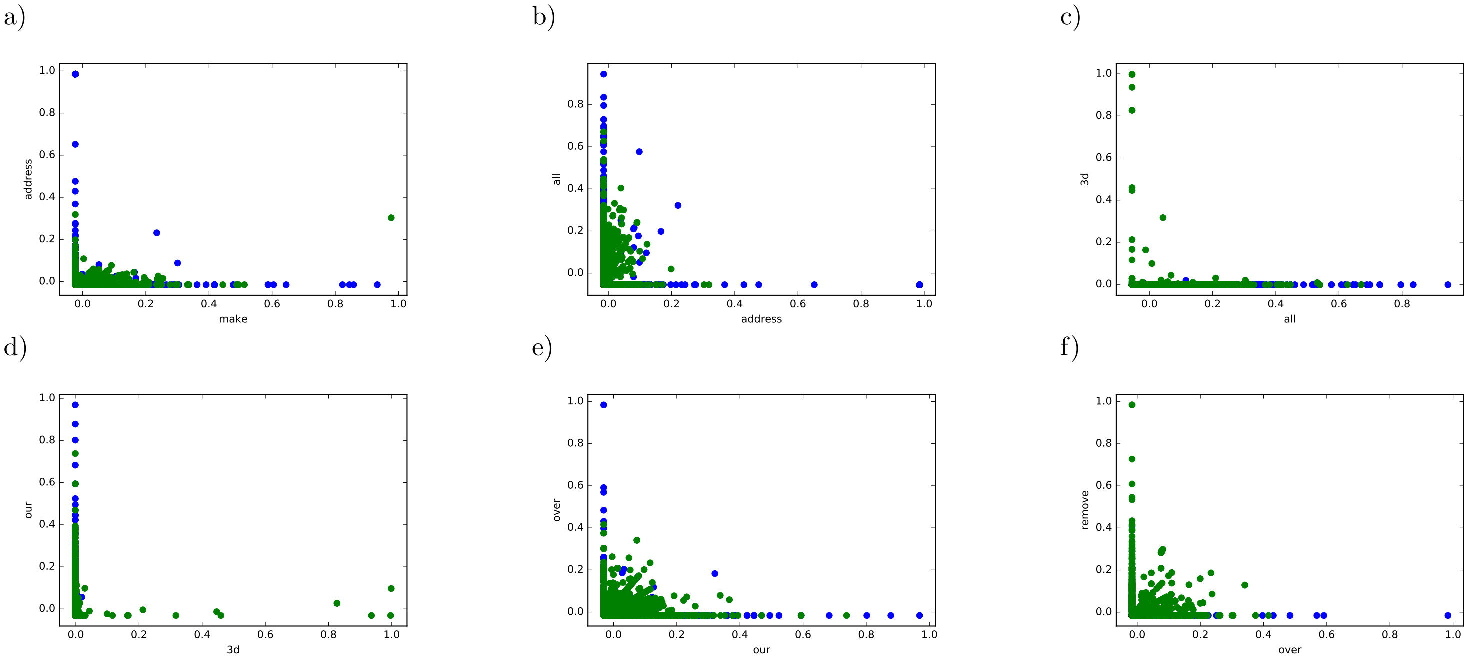

Figure 1. Green is spam and blue is not.

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 initial loss with zero features had a lower error than any of the features applied to it. Though to show that we forwar feature selection on our dataset the inital loss was hardcoded $loss\_record = [0.27]$.

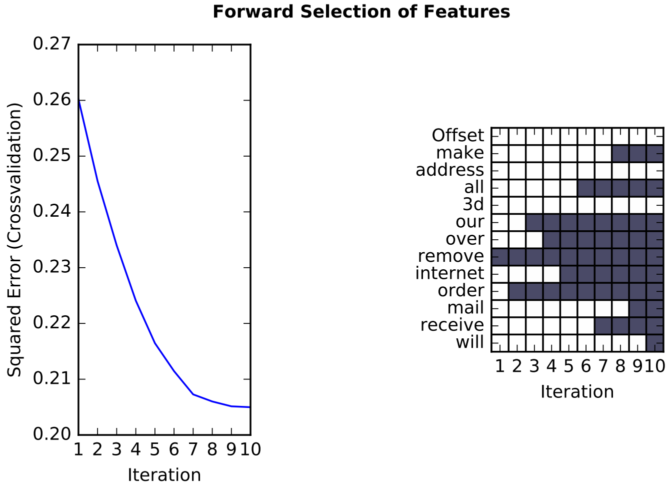

Figure 2. Forward selection form 12 features

Above is the best feature selection that LRF could create with 12 features. If we look at the error rates that are computed:

Linear regression without feature selection:

- Training error: 0.16731688100187755
- Test error: 0.1690774799900705
- R^2 train: 0.29925175624581624
- R^2 test: 0.2907848014802547

Linear regression without feature selection:

- Training error: 0.16731688100187755
- Test error: 0.1690774799900705
- R^2 train: 0.29925175624581624
- R^2 test: 0.2907848014802547

We can tell that without feature selection we get a smaller error. Which leads us to believe that Linear Regression will not give a great estimation of what is spam and what is not.

1.3 Data predicitons

The fitted data has now creatd a model that can be used to fit the email to spam or not. The following vector are the coeficcients of the trained model. These are the importance of each feature which in turn helps us estimate if an email is spam or not.

Attribute	make	all	our	over	remove	internet	order	mail	receive	will
Coefficient	0.490	0.685	1.158	1.631	2.510	2.079	1.237	0.958	0.660	-0.279

And below the residual error for the different words can be seen. From these we can conclued that something causes something but they do not show any distinct patterns.

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.

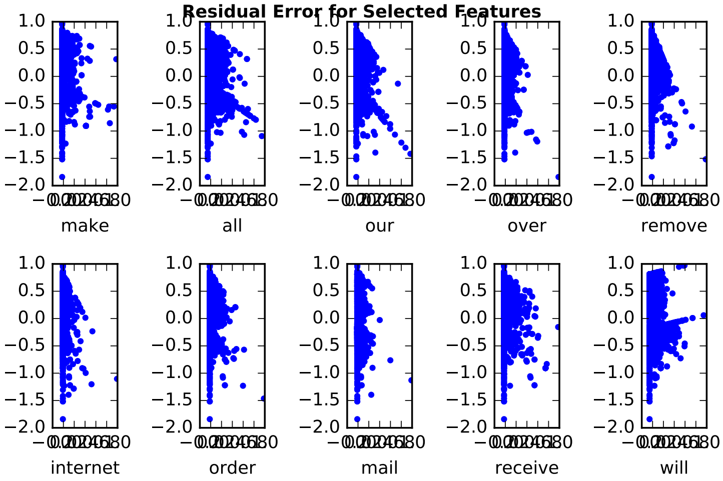

Figure 3. Text.