

# 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 Group issue

Initially we were three people in our group. But due to our last group member not finishing her part, and not letting us know that she would not finish her part of the assignment, some of the answers will be missing/lackluster. This is the second time she does not do anything and now she no longer replies to our message. We have therefore decided to remove her from our group.

### 1.2 Problem description

We have selected the regression 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.

### 1.3 Linear Regression with forward selection

As our dataset has 57 features it is infeasible 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)

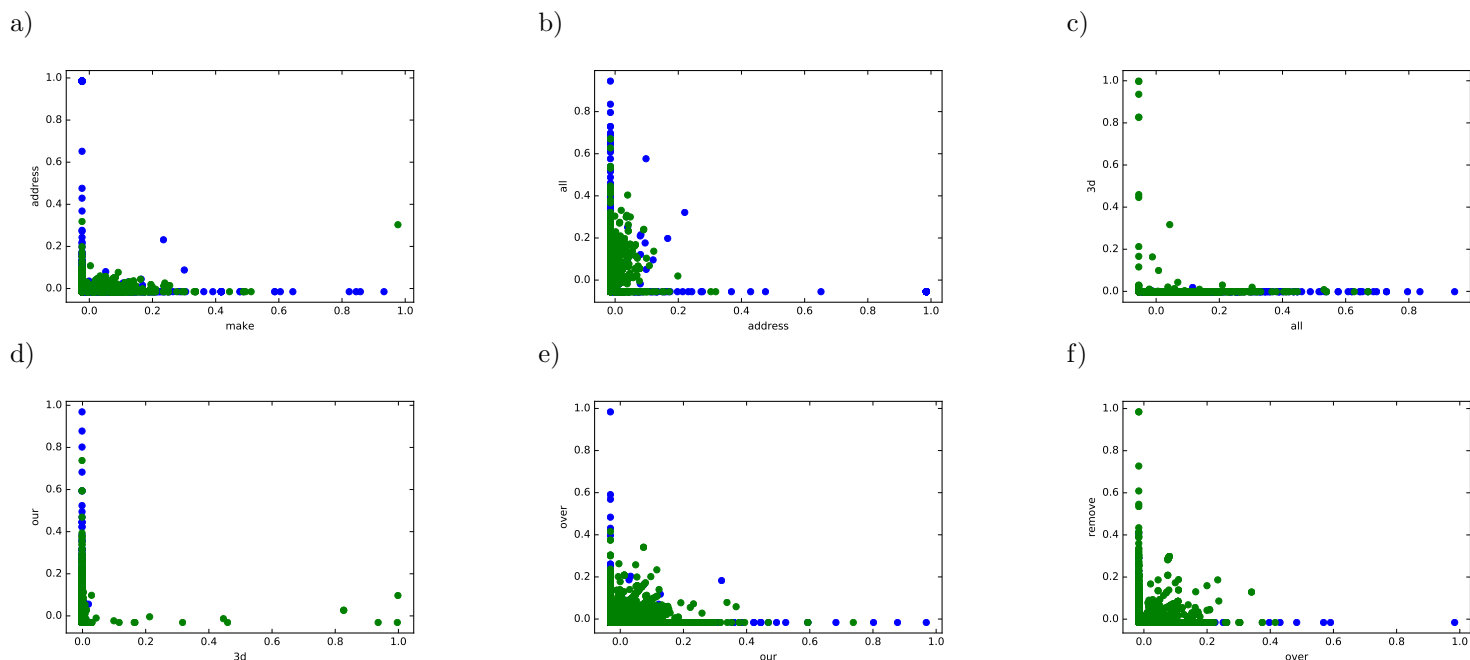


Figure 1. Green is spam and blue is not.

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 patten as the spam is slightly less spread than nonspam.

To figure out what features to use with LR we do forward selection with a 10-fold cross-validation. Initially 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 forward feature selection on our dataset the initial loss was hardcoded  $loss\_record = [0.27]$ .

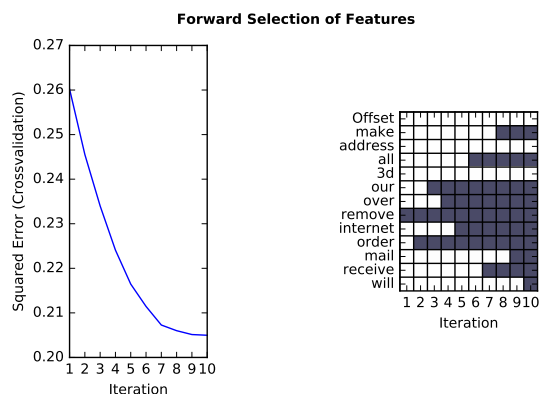


Figure 2. Forward selection form 12 features

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.

## 1.4 Data predicitions

The fitted data has now created a model that can be used to fit the email to spam or not. The following vector are the coefficients 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

Below (figure3) the residual error for the different features can be seen. From these we can conclude that something causes something but they do not show any distinct patterns.

## 1.5 Artificial Neural Network

We implement an ANN model to try to solve the regression problem. Spam is a binary variable as the mails can either be spam or not. The ANN is implemented with  $k = 5$  cross-validation folds and 1 – 4 hidden units. Below we have error rates for the 5 fitted ANN's.

## 1.6 Model Comparison

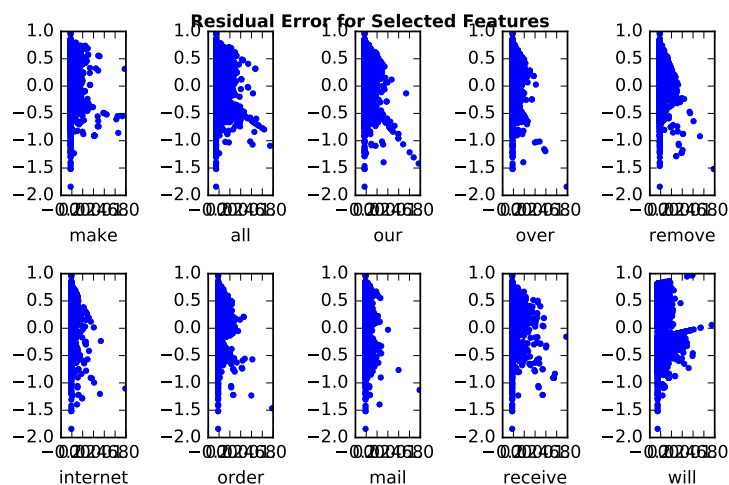


Figure 3. Text.



Figure 4. Text.

#### Linear regression without feature selection:

- Training error: 0.053436522970900534
- Test error: 0.05737122447956839

#### Linear regression with feature selection:

- Training error: 0.16731688100187755
- Test error: 0.1690774799900705