DEPARTMENT OF APPLIED MATHEMATICS AND COMPUTER SCIENCE

02450 Introduction to Machine Learning and Data Mining

Project 1

Feature Extraction and Visualization on a Data Set

Authors:
Jorge Sintes (s202581)
Georgios Tsimplis (s200166)
Klara Wesselkamp (S202382)

Course responsibles:
Jes Frellsen
Morten Mørup
Tue Herlau
Kristoffer Jon Albers

October 6, 2020

Contents

Introduction **1**

1 Description of the data set **1**

1.1 Main machine learning target 1

2 Data Attributes **1**

2.1 Attributes' Description 1

2.2 Basic Statistics 2

3 Data Visualization **3**

3.1 Distribution and Correlation 3

3.2 Principal Component Analysis (PCA) 5

4 Summary **9**

5 Appendix **10**

References **11**

The following contributions were made by the following students.

Student	**Exercise Contribution**
Jorge Sintes Fernandez	1, 2 (30%)
Georgios Tsimplis	2 (70%), 3 (70%)
Klara Wesselkamp	3 (30%), 4

Introduction

This is the first report for the group project in DTU's 02450 Introduction to Machine Learning and Data Mining course.

The purpose of this report is to apply extraction and visualization methods on a chosen data set, in order to get a basic understanding of it. This data set will be used in latter reports to perform further more complex analysis and implement supervised machine learning methods.

1 Description of the data set

The data set used in this project is the SpamBase Data Set, extracted from the UCI Machine Learning Repository [1]. This data base was created by workers at Hewlett-Packard Labs in California. They gathered a collection of 4600 different e-mails, of which almost 40% were filed as spam, and calculated a series of attributes, most of them counting the appearance of certain words or characters on each e-mail.

Previous academic uses of this data set can be found at C. Dimitrakakis and S. Bengio *Online Adaptive Policies for Ensemble Classifiers* [2], where they used it along other sets to test the performance of some improving algorithms for machine learning methods. Also, in Y. Wang and I. H. Witten *Modelling for Optimal Probability Prediction* [3], they created a new modelling method for optimal probability prediction over future observations, using the SpameBase Data Set, among others, for testing the performance.

1.1 Main machine learning target

As stated before, the main target of the project is to apply supervised learning methods using this data set. Specifically, to predict whether an e-mail is spam or not using the rest of the attributes, which is a typical classification problem. In addition to this main objective, it is also expected to observe what kind of common behaviours can be observed in the data set, what attributes are related, and try to propose and create other attributes that could help in the main task of the project.

2 Data Attributes

In this section, will be described the different attributes of our data-set and will be provided the basic statistics in order to have a more clear overview of the data. At this point, it is important to be noted that the specific data-set does not include any corrupted data or missing values.

2.1 Attributes' Description

As it was mentioned in the previous section, the data set consists of 4,601 rows which represent the observations of 4,601 different e-mails and 58 columns which represent the different attributes of each e-mail. The first 54 attributes are the relative frequencies of occurrence of some words/characters -which will be mentioned later- and are measured in proportions(%). The next three attributes are the average length of uninterrupted sequences of capital letters, the length of longest uninterrupted

sequence of capital letters and the total number of capital letters in the e-mail which are measured in units. Finally the last column is a binary attribute where "0" represents that the e-mail is "NOT SPAM" while "1" that is "SPAM". The next table illustrates the structure of the data set and divides the attributes into categories according to their type.

Data-Set		
54 columns	3 columns	1 column
Relative frequency of occurrence of words/characters	Capital-Run Average/Longest/Total	SPAM(Y/N)
Continuous Ratio	Discreet Ratio	Discreet-Binary Nominal
Range= [0-42.81]	Range=[1-15,841]	Range={0,1}

2.2 Basic Statistics

After the description of the attributes, it is useful to mention the basic statistics of the data set to have a deeper understanding of it. In order to understand the differences between the 2 classes of interest-"SPAM(Y/N)"-, the statistics were computed by dividing the data set into those classes. The *modes* of the last 3 attributes are given below.

Type - Attribute's Mode	Capital-Run Average	Capital-Run Longest	Capital-Run Total
Not Spam	1	1	5
Spam	1	12	139

Based on the need to present all the attributes we used dot plots to provide the mean and the standard deviation of each of them for both classes(SPAM-NOT SPAM).Here for space reasons we present only the plot for the means. At the Appendix of this report you can find the plot(Figure 9) which provides the Std's.

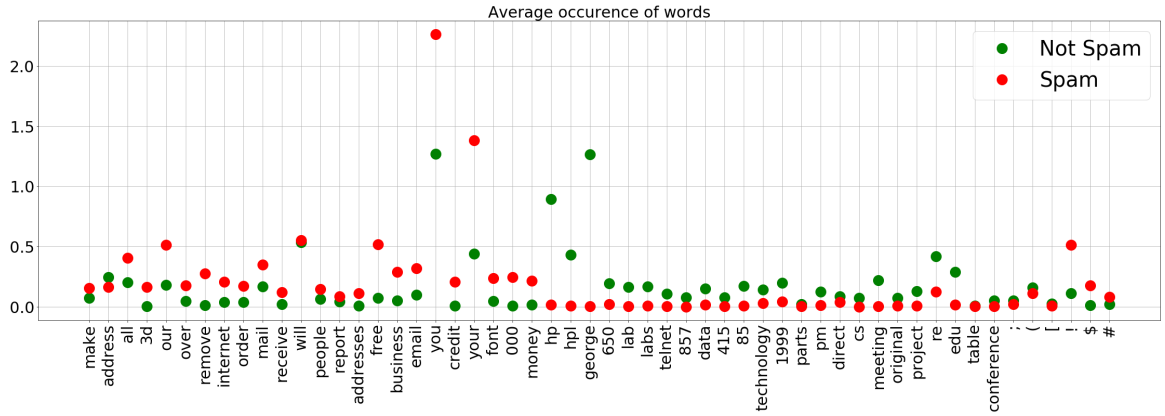
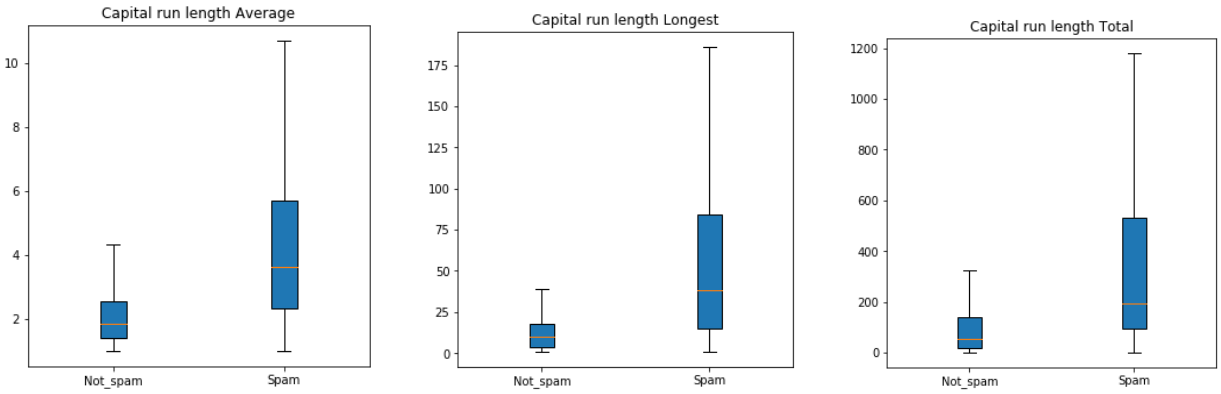


Figure 1

It is quite valuable to present the statistics of all the attributes because in this way we can understand that for the most of the attributes the means and the standard deviations are similar besides the means of the frequencies of some specific words(i.e: "you", "your", "hp", etc.). At the next table are given the means and the standard deviations for the last 3 attributes where there can be pointed out significant differences.

Type - Attribute's Mode	Capital-Run Average	Capital-Run Longest	Capital-Run Total
Means(Not Spam / Spam)	2.37 / 9.51	18.21 / 104.39	161.47 / 470.61
Std's(Not Spam / Spam)	5.11 / 49.83	39.08 / 299.20	355.68 / 824.85

For the last 3 attributes which have a different scale, the median and the quantiles are illustrated by the following box-plots.



(a) (b) (c)

3 Data Visualization

3.1 Distribution and Correlation

The data does not seem to be normally distributed as can bee seen the following example image. However, it could be said to be somewhat normally distributed if the number of 0-values was not so high. As the other distributions mostly look the same, they will be spared:

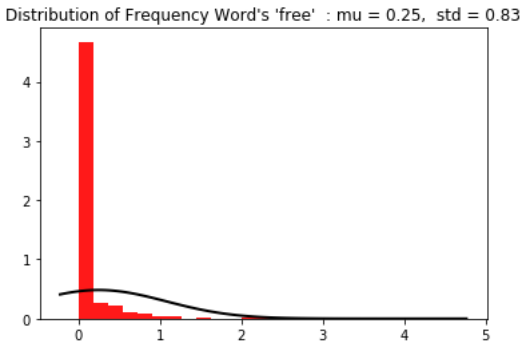


Figure 3: Distribution

In this case the correlation matrix does not contain a lot of new information, because the number of values renders it more confusing than meaningful:

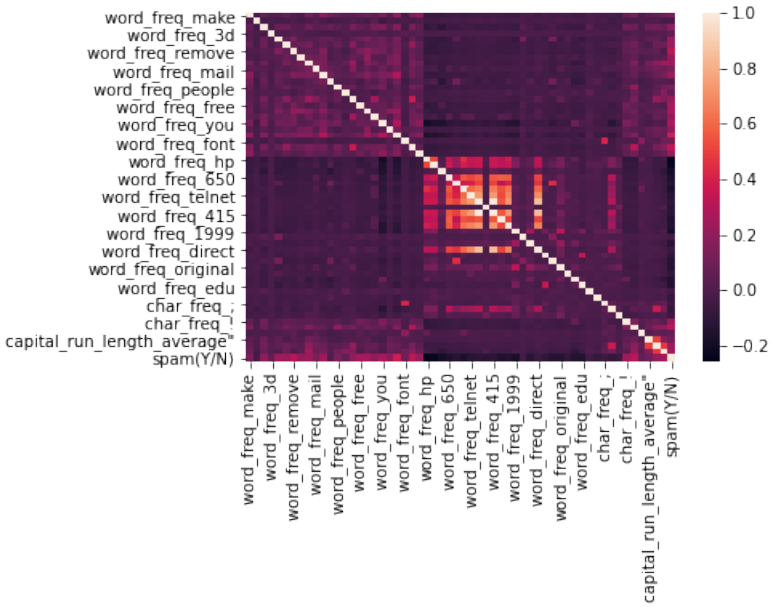
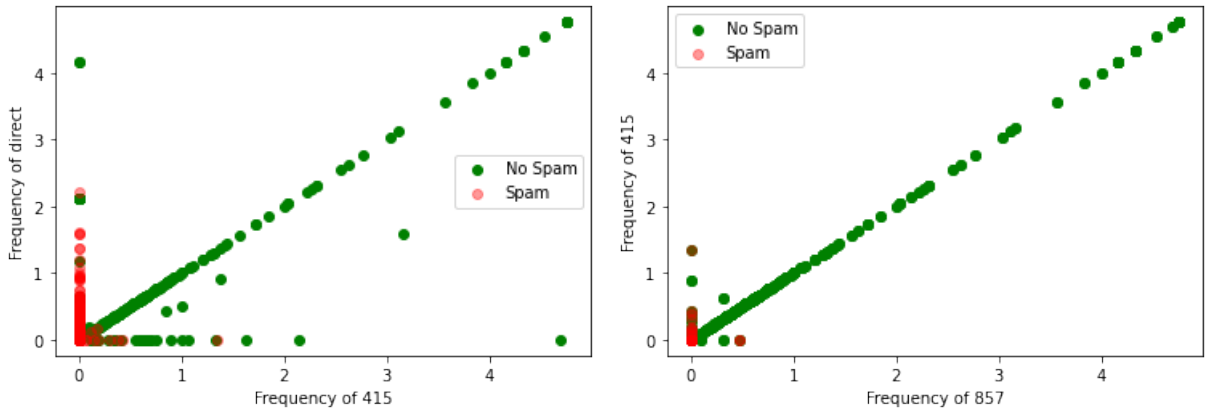


Figure 4: Heatmap of correlation matrix

Therefore, the data with the highest correlations will be examined, assuming that data with low correlation will not contribute to the understanding:

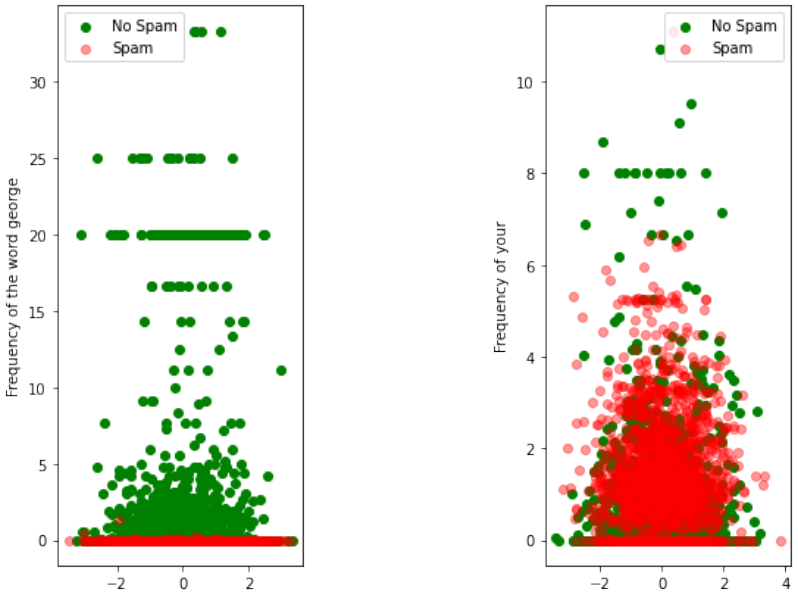
Variable 1	Variable 2	Correlations
word_freq_857	word_freq_415	0.996066
	word_freq_direct	0.848021
word_freq_415	word_freq_direct	0.845359
word_freq_telnet	word_freq_857	0.737555
	word_freq_415	0.735187

One correlation especially hits the eye, that is the one between the words 857 and 425. Below are plots of these variables as well as a comparison between the words 415 and direct.



In both cases, the shape is linear except for very low values and rare outliers. This is interesting, as it indicates, that the words in general appear together. One could take this as a starting point to see if certain words can be grouped together into one word. Furthermore the distribution of Spam/Not Spam Observations is meaningful, as the words mostly appear joined in a Not Spam-Context.

To get more insight into whether an e-mail is considered spam, the variable most correlated with that attribute is plotted, namely the word "your":



The suspicion arises, that the correlation might not be the best predictor of this. To bring home the point, there is a of the word "george", which is a strong indicator that an e-mail is not a Spam, which is also proven in the plot, whereas the correlation value is only -0.18:

An explication for this could be the sparseness of the matrix. With such a high number of 0-observations, the correlation and other values for the actually relevant values will be necessarily skewed. A more intricate evaluation of non-0-values might not only be interesting, but also necessary here.

3.2 Principal Component Analysis (PCA)

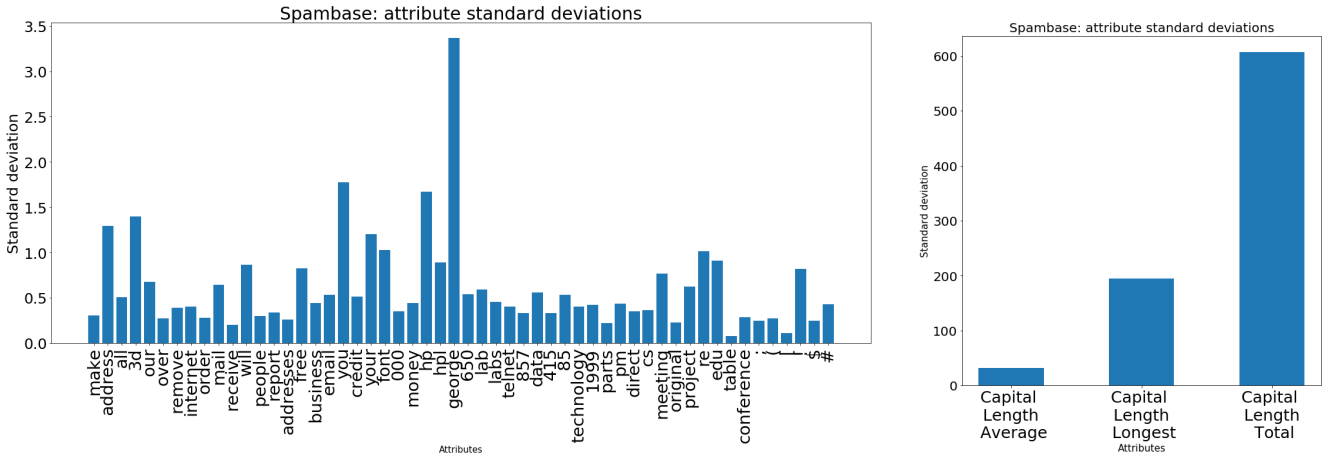
Principal Component Analysis is an effective technique which combines multivariate statistics and optimization methods to reduce the dimensions of a data set in order to use it for further analysis. The general idea is to extract the most "valuable" information from the data set and to express this information as a set of summary components (principal components). Specifically, this technique finds lower dimensional sub-spaces of an n-dimensional space and project the data so that to maintain and compress common information of different attributes to the new principal components. Let us consider X a matrix which represents the data set, -excluding the attribute that we want to predict in the future- m the mean of each attribute and s the standard deviation. The general steps of this technique are the following:

1. Standardize data set by computing: $\hat{x_i} = \frac{x_i - m}{s}$
2. Apply Singular Value Decomposition: $U\Sigma V^T = \hat{X}$

3. Project the data onto the subspace $V = [v_1, ..., v_k]$ which created by the first k principal components.

$$b_i^T = \hat{x_i}^T V_k$$

If the step of standardization is very important in data sets in which the attributes has different scales. If we do not apply this step it is quite possible that our analysis will be more influenced from variables of higher scales from others. Below are given the Standard Deviations of our attributes.



(a) (b)

Figure 5

After the standardization step we applied the SVD. From Figure 6, we can see that the first 3 Principal components explain only 20% of the total variance. This means that if we want to obtain as much information as possible from the original data set it is necessary to use more principal components in our analysis. A reasonable limit is the threshold of 90% so we can reduce the dimensionality from 57 to 43 in order to explain more than 90% of the total variance.

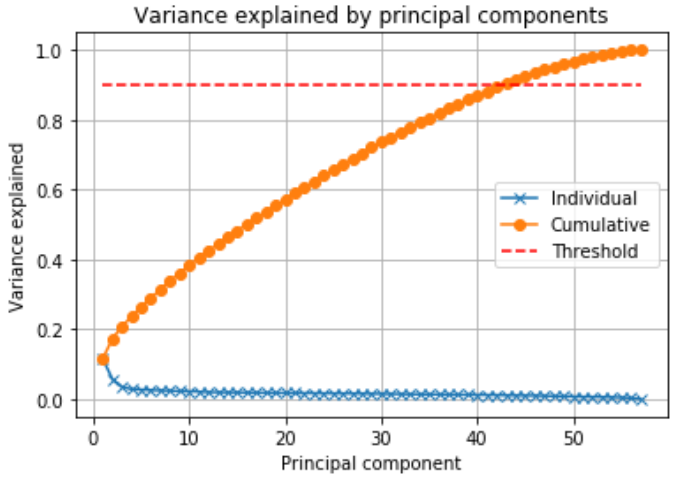
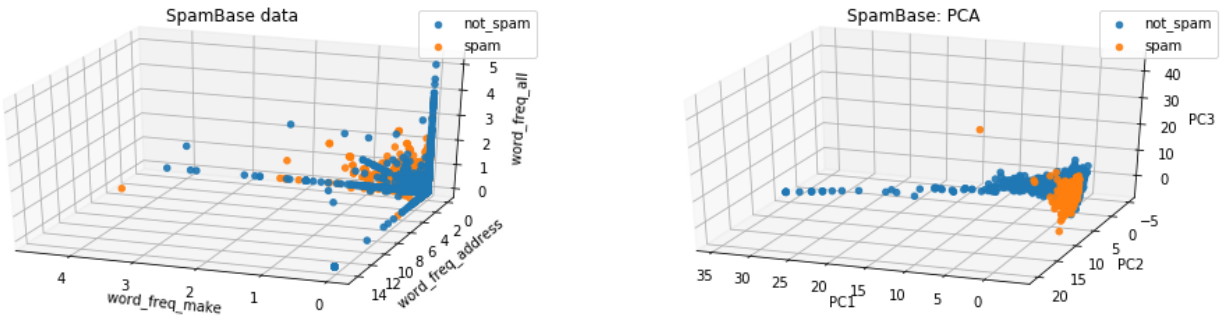


Figure 6

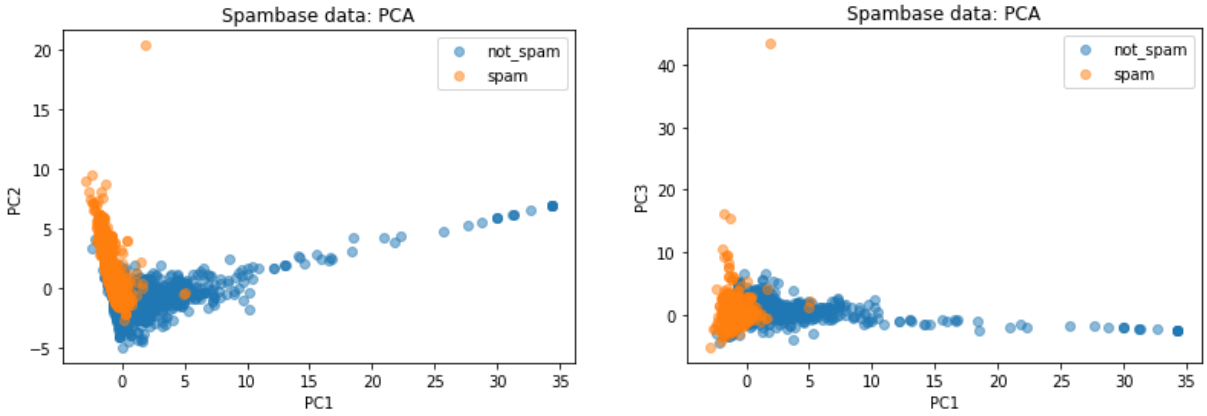
Below, we present firstly (Figure 5.a) a scatter plot which illustrates the correlation of the 3 first original attributes and then (Figure 5.b) the projection of the original data at the first 3 principal components to understand the way that the information of the dataset is compressed to the first components which explain the higher amount of the total variance compared to the next pc's.



(a) (b)

Figure 7

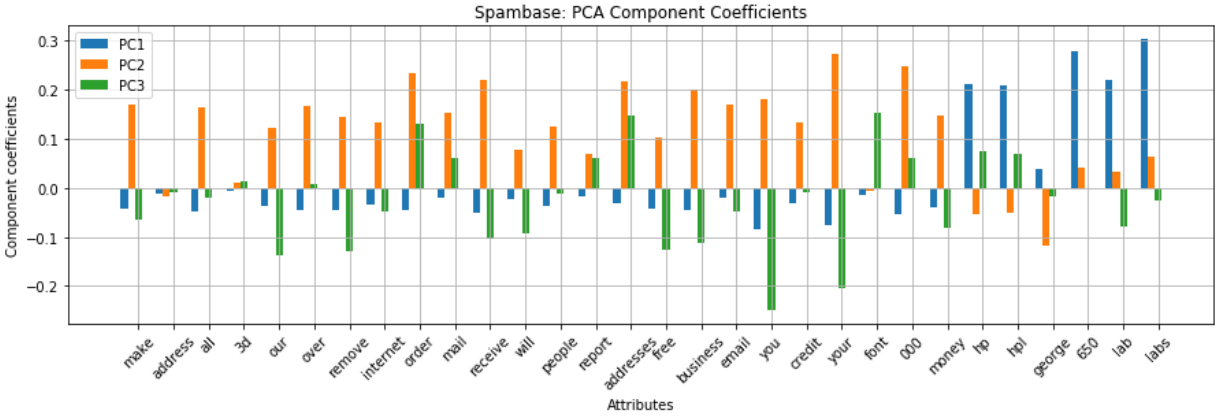
Moreover, below there are the projections of the data onto some combinations of the first principal directions. At the appendix you can find more projection combinations.



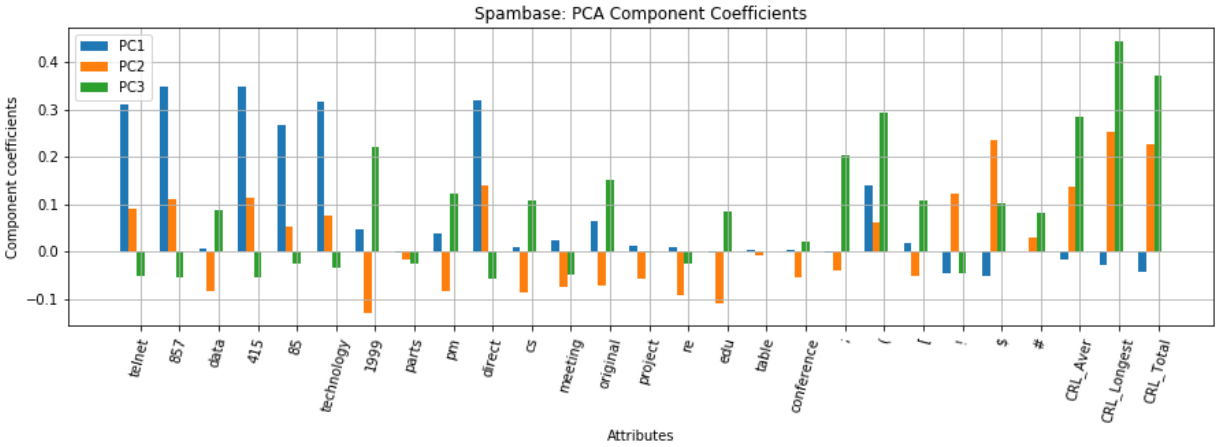
(a) (b)

Figure 8

Another valuable aspect of PCA is the interpretation of each principal component in terms of the original attributes. Below we present the PCA component coefficient plot for the first 3 PC's. From this plot we can understand that the larger absolute value of the coefficient of a corresponding attribute the more important is this attribute in calculating this specific component. For example, the original attribute that the first principal component mainly captures the variation of is the "frequency of word 857".



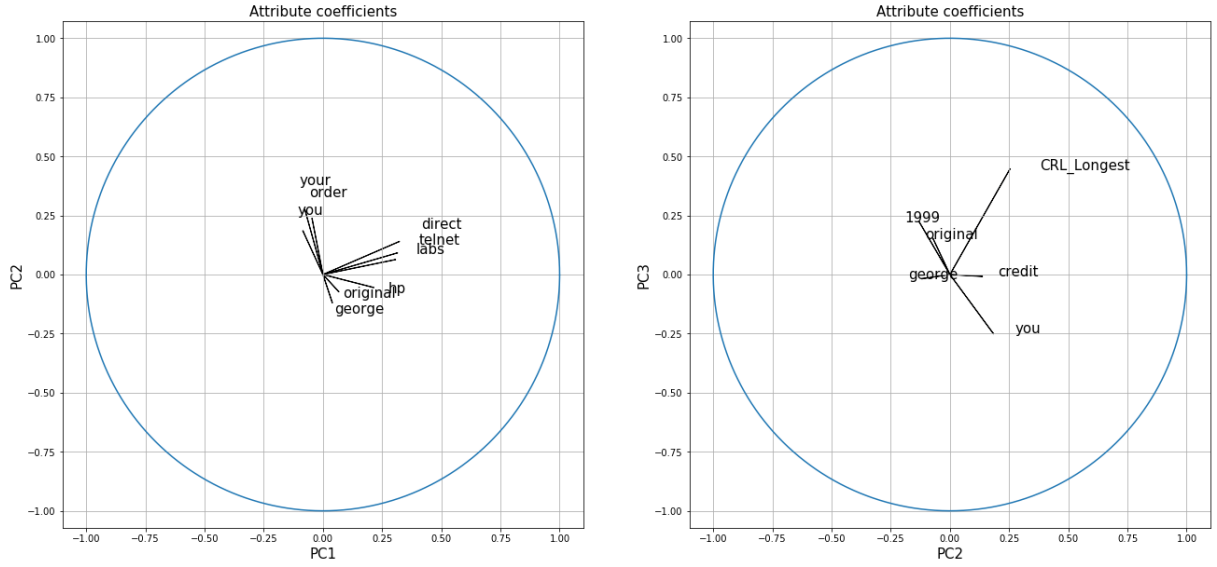
(a)



(b)

Figure 9

Finally, in order to have a more clear view not only for the magnitude of each coefficient but also for the direction we present the below plots. Due to the large number of the original attributes we included at this plot only the coefficients that have a significant magnitude according to the different directions that are projected onto the planes (PC1-PC2) and (PC2-PC3).



(a) (b)

Figure 10

4 Summary

It was possible to get quite a lot of information out of the data, only using very basic statistics and visualization techniques. One of the most important features of this data is its sparseness it is. This is not surprising, considering the content of the data, it should however constantly be respected.
One area in which this can be observes is talking about outliers. The statistics show, that the maximum of a certain variable is several magnitudes higher than it's mean. In many uncurated datasets this would immediately indicate an error, or, at least an outlier. However, in this case, these are the actually important variables. More specifically the examined E-Mail is defined by the words that occur in it often, not by the words that don't appear.

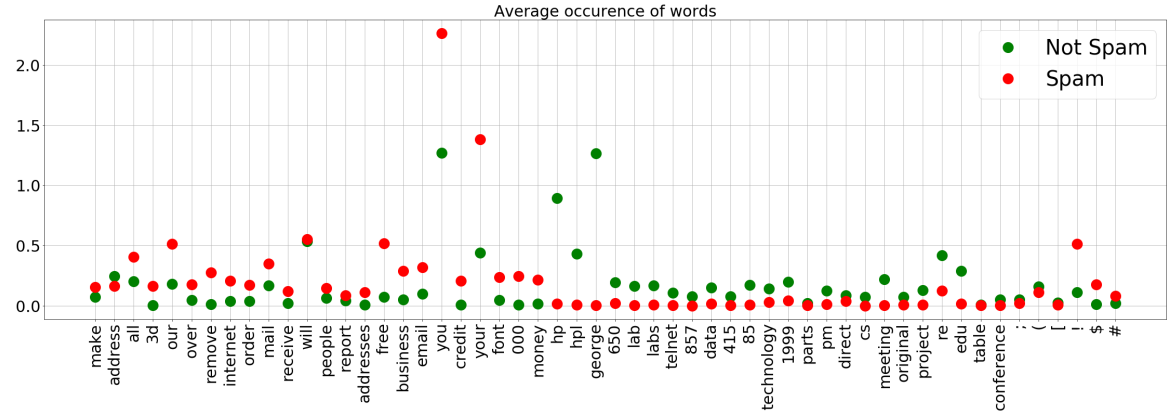
Another area, where the sparseness needs to be considered, is that the correlation of the data is strongly influenced by this feature of the matrix, so this data holds a variety of starting points for further examination. It is also important to remember that losses for falsely classified data are not evenly distributed: In general a "good" e-mail that was thrown away is going to be a way worse experience than a Spam that slid into the main mailbox and the algorithm should be considerate of this.

Now to the more substantial part: it is observed that some words almost exclusively appear paired or at very low rates. This might be a starting point to simplify the algorithm: having words that allow to be grouped together. In the basic statistics it could be observed that the mean of variables allowed, pretty quickly, to separate *Spam* from *Non-Spam* E-Mail. An idea could be to try to sort the variables according to their mean here, mostly disregard the variance.

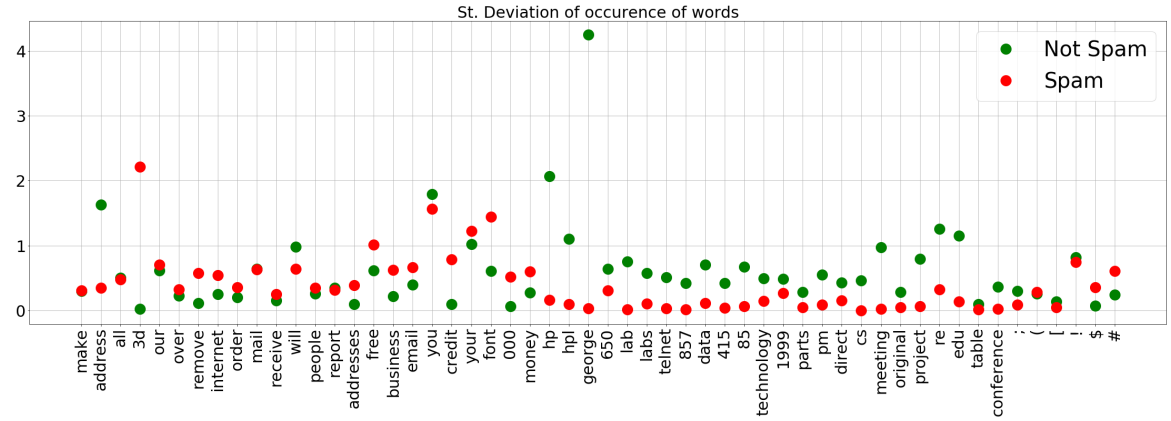
Finally, the most powerful statistical Instrument used here, Principal Component Analysis, did not yield very meaningful results. Apparently, an unreasonable high number of principle components would be needed to account for any meaningful proportion of the variance of the data. However, the question remains whether the variance here is a very useful measure. In the visualization, one can see that the observations in the PCA-Base are clearly splitted up into different directions, even though this separation remains quite blurry, and the overlap is quite big.

In the end, even though the dataset is highly curated, it remains complex data, not least because a clear definition of "Spam" is never given. This necessarily leads to some blurryness. However, this is also a very interesting dataset, giving machine-learning methods a very mundain and useful application.

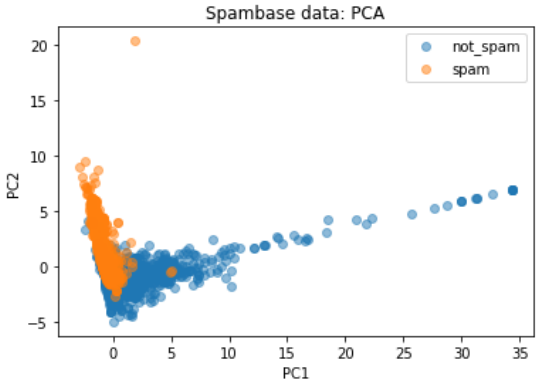
5 Appendix

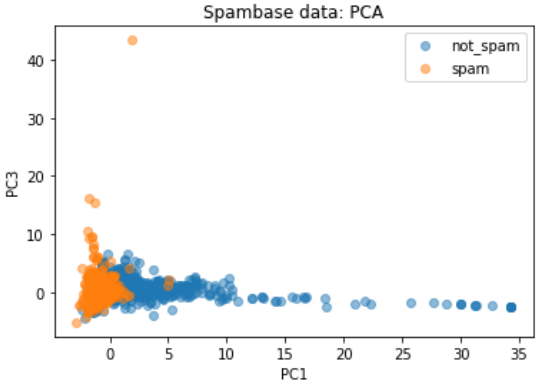


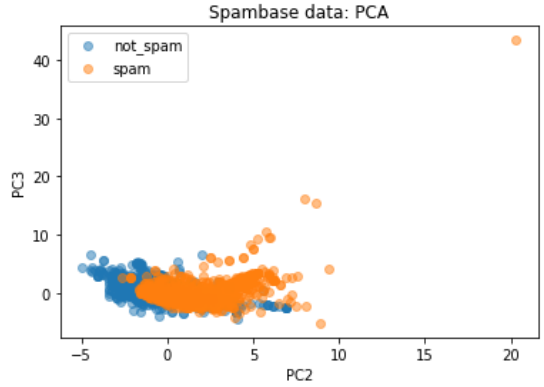
(a) Means

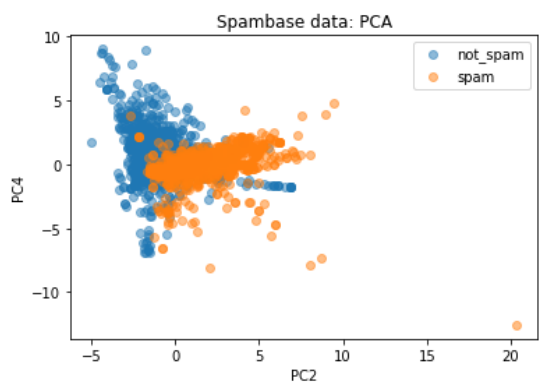


(b) Std's









References

[1] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[2] Christos Dimitrakakis and Samy Bengio. "Online Adaptive Policies for Ensemble Classifiers". In: *Neurocomputing* 64 (2005), pp. 211–221. DOI: https://doi.org/10.1016/j.neucom.2004.11.031.

[3] Yong Wang and Ian H. Witten. *Modelling for Optimal Probability Prediction*. en. Conference Contribution. July 2002. URL: https://hdl.handle.net/10289/2131.