Project 1 - Introduction to Machine Learning and Data Mining

Jens Collatz Laustrup (s204431) Gustav Morgensol (s205428)
Toshini Iyer (s236312)

03/10/2023

Constributions:

	Section 1	**Section 2**	**Section 3**	**Section 4**	**Exam quest.**
s204431	20%	20%	60%	40%	33.33%
s205428	20%	60%	20%	30%	33.33%
s236312	60%	20%	20%	30%	33.33%

1 A description of your data set

We have chosen to work with a data set called ”Spambase”. It can be found at:

[Hopkins,Mark, Reeber,Erik, Forman,George, and Suermondt,Jaap. (1999). Spambase. UCI Machine Learning Repository. https://doi.org/10.24432/C53G6X.]

The Spambase dataset is primarily designed to address the problem of email classification, specifically for distinguishing between ”spam” and ”non-spam” emails. The overall problem of interest lies with spam detection in emails.

The main goal is to develop a machine learning model that can classify incoming emails as ”spam” or ”non-spam” based on various attributes and features that are extracted from numerous email contents.

According to the authors, the spam e-mails came from individuals who had filed spam and from the author’s post master, and the non-spam emails came from work emails and personal emails.

The idea is that the classification model can be used to predict whether a new email is spam, and thus can be used to create a spam filter.

The dataset provides a collection of email messages in the form of a number of attributes and labels that make it possible to use supervised machine learning to train on the data. Here’s a breakdown of key components of the dataset:

- **Attributes and Features:** Each email message is represented using a set of features or attributes. These features include various statistics and characteristics derived from the email content, such as word frequencies, character frequencies, and the number of capital letters. These features can be used as input to machine learning models.
- **Labels:** The dataset includes labels for each email message, indicating whether it is *spam* (often labeled as "1") or *non-spam* (often labeled as "0"). These labels are used as the ground truth for training and evaluating the performance of the model.

We now explain a few findings from previous analysis of the data.
One article that has analysed the data is "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets". The conclusion of the article shows that the Naïve Bayes classifier performs well when used with the Spambase dataset. [Nurul Fitriah Rusland et al 2017 IOP Conf. Ser.: Mater. Sci. Eng. 226 012091]

An article published in the Global Journal of Computer Science in 2015 gives a table on page 26 stating the performance of various datasets with different machine learning algorithms, which also concludes that the Spambase dataset achieves a 94.28% accuracy when worked with various machine learning algorithms.

[Bhuiyan, H., Ashiquzzaman, A., Juthi, T.I., Biswas, S., & Ara, J. (2018). A Survey of Existing E-mail Spam Filtering Methods Considering Machine Learning Techniques. Global journal of computer science and technology.]

Thus, it seems feasible to predict whether emails are spam or not spam from the Spambase dataset based on previous research.

We will now explain how we plan to perform classification and regression to the data.
For the classification task, we plan to predict the spam/not spam class label based on all other attributes. This is the main machine learning aim, because predicting whether an email is spam or not is the overall problem of interest.
For the regression there are a few options. One option is to predict the number of capital letters or the frequency of a specific word or character based on the other attributes. Another possibility, which is likely more relevant is to predict the binary spam/not spam attribute based on the other attributes using logistic linear regression.
Either way, the main goal is to predict the spam/not spam binary attribute.
We do not plan to transform the data to carry out these tasks, except for potentially standardizing the data for some tasks, by subtracting the mean and dividing by the standard deviation of the attribute, for each value.

2 A detailed explanation of the attributes of the data

The data set contains 58 attributes in total. The first 48 attributes are word frequency attributes. That is, what percentage of words in the email consists of a specific word. This could for example

be word_freq_meeting for the word ”meeting”. If this attribute has the value 15, that would mean that 15% of words in the email is the word ”meeting”.
The next 6 attributes are character frequency attributes. These have the same meaning as the word frequency attributes, but for characters. This could for example be the attribute char_freq_! for the character !. If this attribute contains the value 15, this means that 15% of the characters in the email is the character !.
The data set then contains 3 attributes describing the number of capital letters in the email. These are capital_run_length_average, which is the average length of sequences of capital letters, capital_run_length_longest, which is the length of the longest sequence of capital letters, and capital_run_length_total, which is the total number of capital letters that the email contains.
Finally, the data set has a binary attribute that can take the value 0 or 1. This indicates whether the email is considered spam or not, where 0 means not spam and 1 means spam.
The word frequency and character frequency attributes are continuous ratio attributes. They possess a true zero, and multiplication and division are meaningful, I.E. you are able to say 10% of a given word occurrence is twice as much as 5%. The capital_run_length_average attribute is also a continuous ratio attribute.
The capital_run_length_longest and capital_run_length_total attributes are both discrete ratio attributes, since they can posses any non-negative integer. Again, they posses a true zero, and multiplication and division are meaningful.
The last attribute is a binary nominal variable, that can have the value, 0 or 1 dependant on the email being spam or not.
Note that none of the attributes can have negative values.
Additionally we have discovered that the data set contains no missing data or corrupted data.[1] This is also confirmed by the authors of the data set. Thus, there does not appear to be any data issues.

We will now show summary statistics of the attributes. Since there is a very large number of word frequency attributes we will show only five of them here. Summary statistics for all of the attributes can be seen in the appendix. We do not show the binary attribute. This will be explained later.

	Mean	**Std. dev.**	**Median**	**Min**	**Max**
word_freq_receive	0.05982	0.20154	0.0	0.0	2.61
word_freq_business	0.14259	0.44406	0.0	0.0	7.14
word_freq_money	0.09427	0.44264	0.0	0.0	12.5
word_freq_george	0.7673	3.36729	0.0	0.0	33.33
word_freq_telnet	0.06475	0.40339	0.0	0.0	12.5
char_freq_;	0.03857	0.24347	0.0	0.0	4.385
char_freq_(	0.13903	0.27036	0.065	0.0	9.752
char_freq_[	0.01698	0.10939	0.0	0.0	4.081

[1] Note that on the page that we have linked to it says that the data does contain missing data. However, in the documentation provided by the authors they specify that there is no missing data, which is what we have also found from our investigation.

char_freq_!	0.26907	0.81567	0.0	0.0	32.478
char_freq_$	0.07581	0.24588	0.0	0.0	6.003
char_freq_#	0.04424	0.42934	0.0	0.0	19.829
capital_run_length_average	5.19152	31.72945	2.276	1.0	1102.5
capital_run_length_longest	52.17279	194.89131	15	1	9989
capital_run_length_total	283.28928	606.34785	95	1	15841

From this, we see that the median of almost all of the word frequency and character frequency attributes is 0. This shows that at least half of the observations contain the value 0 for each of these attributes (this is the case where the email does not contain the given word/character). We also see that both the means and standard deviations are quite close to 0 for these attributes. The capital run length attributes appear to have much larger values in general than the other attributes, and a much larger standard deviation. This indicates that standardization of the data may be necessary. To summarise the binary spam/not spam attribute we mention that of the 4601 observations, 2788 of them are not spam (61%) and 1813 are spam (39%).

3 Data visualization(s) based on suitable visualization techniques in- cluding a principal component analysis (PCA)

First, we investigate whether there are outliers in the data set. Figure 1 shows box plots of the attributes (after standardizing the data). The idea is to identify possible outliers.

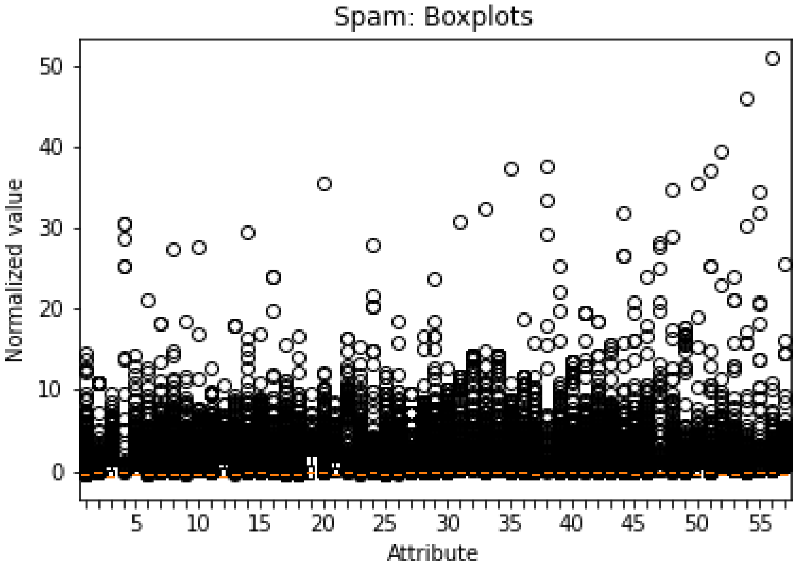


Figure 1: Histograms of 11 different attributes.

From the box plots, we do see some potential outliers for some of the attributes. However, we have examined these points closer and concluded that they do not appear to be outliers. Thus, we will not remove any points.

We will now investigate whether the data appears normal distributed. Since we have a large number of attributes and some are very similar we will only investigate some of them. In particular we will investigate four word frequency attributes (we arbitrarily choose ”make”, ”receive”, ”your”, and ”george”), four character frequency attributes (we arbitrarily choose ”;”, ”(”, ”[”, and ”!”), and the final three attributes (”capital_run_length_average”, ”capital_run_length_longest”, and ”capital_run_length_total”). Figure 2 shows a histogram for each of these 11 attributes.

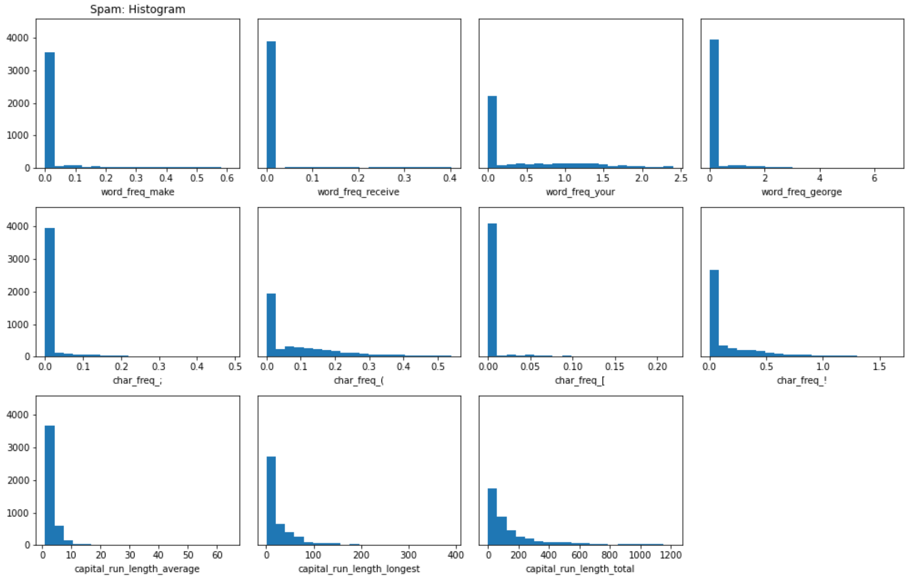


Figure 2: Histograms of 11 different attributes. Range shown is from the minimum value to the minimum value plus two times the standard deviation.

By looking at these histograms, the attributes do not appear to be normal distributed. We see for each of the attributes that a large number of points have the value (or close to) 0. This corresponds to the situation where the corresponding email does not contain the word, character, or any capital letters. Also, it should be noted that values cannot be negative. It may be possible that some of the word frequency and character frequency attributes follow a half-normal distribution if the zero values are not considered. The capital run length attributes also appear to potentially be half-normal distributed.

We will now investigate whether the attributes appear to be correlated. Figure 3 shows the correlation matrix of the attributes.

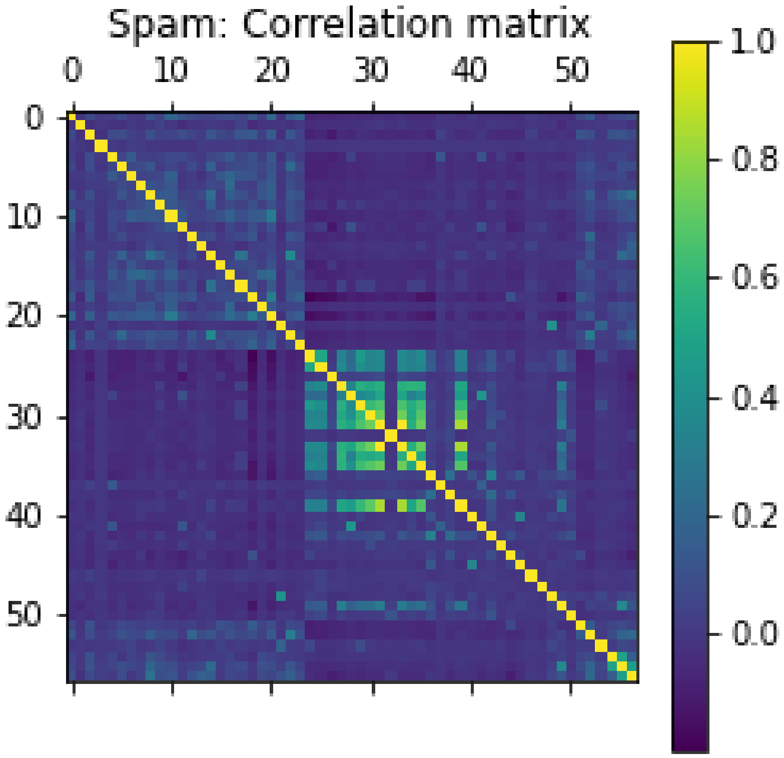


Figure 3: Correlation matrix of attributes.

As seen, some of the attributes do appear highly correlated (bright squares). Though, many attributes do not appear correlated (dark squares).

We will now carry out a PCA analysis of the data. The analysis will be performed on all attributes except the binary spam/not spam attribute. Thus, the analysis is performed on 57 ratio attributes in total.
The first 54 of these use the same scale (a percentage between 0 and 100), however, the final three attributes do not use this scale. These attributes can be arbitrarily large and by inspecting the data, we see that these attributes often have much greater values than the other attributes. Thus, we choose to standardize the data before performing the PCA analysis.

Figure 4 shows the percent variance explained for the individual principal components and the cumulative percent variance explained. Additionally, a 90% threshold for the percent variance explained is shown. All principal components are included.

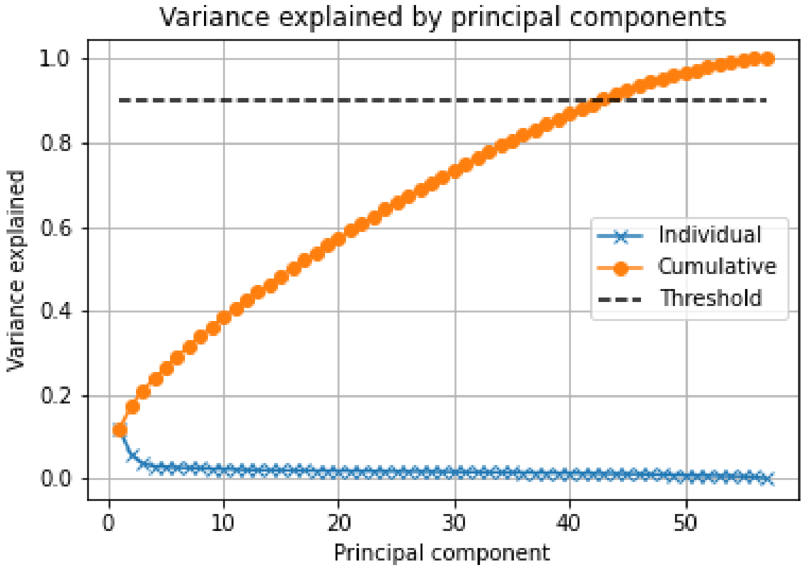
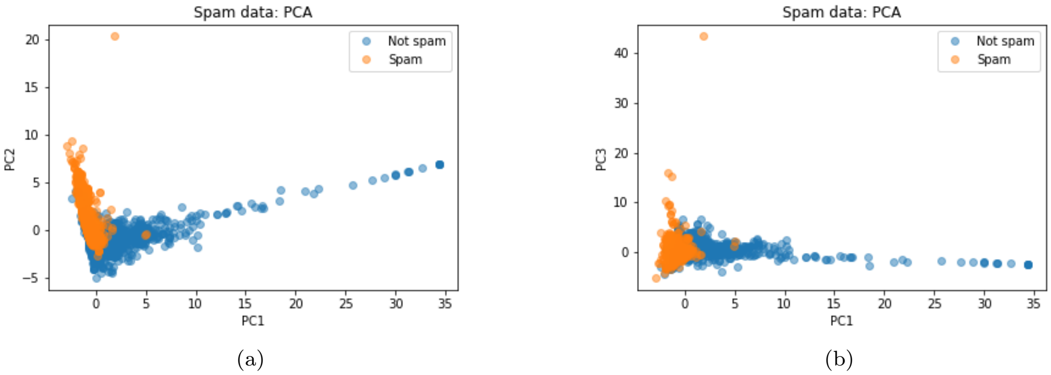


Figure 4: Percent variance explained for principal components.

As seen it requires a large number of principal components (42) to explain at least 90% of the variation. Additionally, the first principal component only explains around 12% of the variation. We will now inspect the first three of these principal components in more detail.

Figure 5 shows the data projected onto the first and second principal component (left figure), and the data projected onto the first and third principal component (right figure). Additionally, each data point has been labelled with its class (spam/not spam).

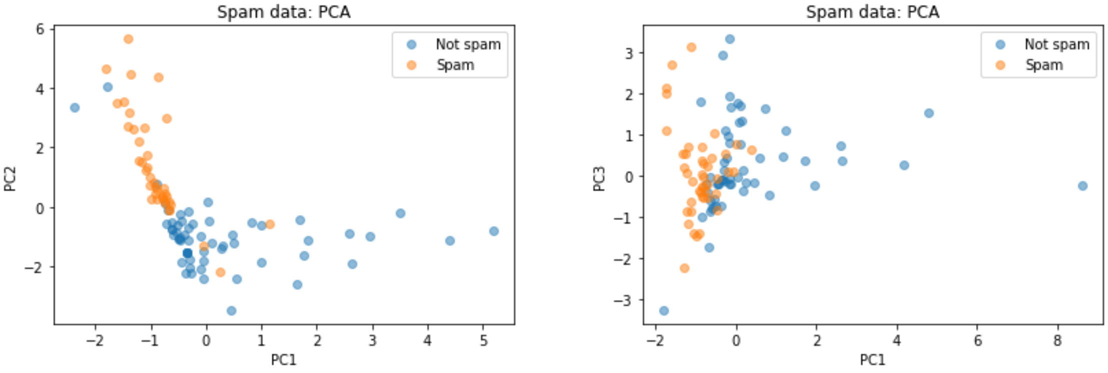


(a)

(b)

Figure 5: Projection on PC1 and PC2 (a), and on PC1 and PC3 (b).

Since the data set contains a large number of data points (4601) there are many overlapping points on the figures. Thus, we also show the same figures for 100 randomly selected data points on figure 6.



(a)

(b)

Figure 6: Projection on PC1 and PC2 (a), and on PC1 and PC3 (b) for 100 random points.

As seen on the figures, negative values of the first principal component and positive values of the second principal component generally seems to indicate that an email is spam. However, it is still difficult to separate the two classes based on just these components. For the third principal component it is difficult to tell from the figure what a positive or negative value indicates.

We now inspect the component coefficients for the first three principal components. These coefficients are shown in figure 7.

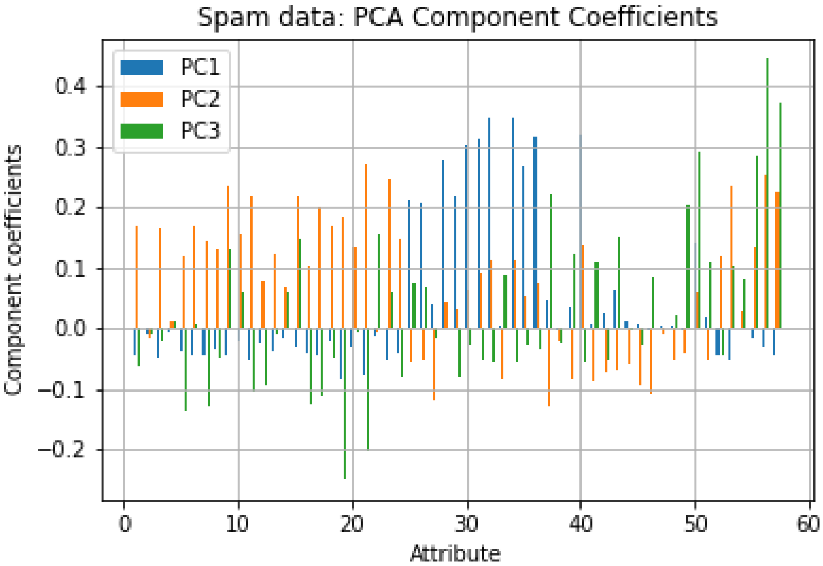


Figure 7: PCA component coefficients.

Let us first consider the first principal component. By observing the figure, we see that it has some attributes that are more important than the others. These are the attributes with a coefficient that is greater than 0.20. The names of these 11 attributes are:

word_freq_hp, word_freq_hpl, word_freq_650, word_freq_lab, word_freq_labs, word_freq_telnet,
word_freq_857, word_freq_415, word_freq_85, word_freq_technology, word_freq_direct

Since these have large positive coefficients this should indicate that emails with a large amount of the words hp, hpl, 650, lab, labs, telnet, 857, 415, 85, technology, and direct are unlikely to be spam emails. These words are also quite related.

We now consider the second principal component. This component also seems to have some large positive values. Again we consider the attributes with a coefficient greater than 0.20. The names of these 8 attributes are:

word_freq_order, word_freq_receive, word_freq_addresses, word_freq_your, word_freq_000,
char_freq_$, capital_run_length_longest, capital_run_length_total

As explained earlier a large positive value of the second principal component indicates that an email is spam. Thus, the words order, receive, addresses, your, and 000, as well as the character $, the longest string of capital letters and a large total number of capital letters seems to indicate that an email is spam. This is also what you would expect.

For the third principal component it becomes difficult to make sense of the coefficients. Thus, we will not discuss this further.

Overall, it seems that the first principal component generally identifies attributes, that often indicate that an email is not spam, while the second principal component identifies attributes, that often indicate that an email is spam.

Based on the analyses in this section we believe that the primary machine learning modeling aim (classification for spam/not spam) appears feasible. For example, it seems possible to distinguish between the classes from the principal components. However, it appears to be a relatively difficult task.

4 A discussion explaining what you have learned about the data

We have learned that the data set contains 55 continuous ratio attributes, 2 discrete ratio attributes and one binary nominal attribute. The main goal is to predict the binary attribute (spam/not spam) from the other attributes.

We have learned that the data set does not have any data issues (no missing values or corrupted data) and that by far the most common value for each of the ratio attributes appears to be 0. Additionally, none of the ratio attributes can have negative values.

We have concluded that the data set does not have issues with outliers. Thus, no observations should be removed. The attributes do not appear to be normal distributed, but some attributes may be

half-normal distributed. The attributes appear somewhat correlated. The principal component analysis showed, that it appears necessary to keep a quite large number of principal components in order to not lose too much information.
From this, we have concluded that the primary machine learning objective does not appear to be an easy task, but it does seem feasible. We believe that it is feasible to distinguish between the two classes (spam/not spam) using the given information, with a reasonable amount of error.

5 Exam Questions

1. Option C: Time of day is ordinal since the time intervals are ordered, but not interval since it does not make sense to add time intervals together. Traffic lights and running over are ratio since it makes sense to have 0 broken traffic lights or 0 run over accidents. y is ordinal since it contains ordered levels of congestion, but not interval since you cannot add these levels of congestion together.

2. Option A: We calculate the $p = \infty$ norm distance as:

$$d_{p=\infty}(x_{14}, x_{18}) = max\{|26-19|, |0-0|, |2-0|, |0-0|, |0-0|, |0-0|, |0-0|\} =$$

$$max\{7, 0, 2, 0, 0, 0, 0\} = 7$$

And thus, A is correct.

3. Option A: We calculate the explained variance of the first four principal components as:

$$\frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.87$$

Since this is greater than 0.8 option A is correct.

4. Option D: Time of day has a negative coefficient for principal component 2 and a low value. Broken truck, accident victim and defects all have positive coefficients for principal component 2 and a high value. The immobilized bus coefficient is small compared to the others and therefore only has a small effect. Thus, this will typically lead to a positive value of the projection.

5. Option A: We calculate the Jaccard similarity as:

$$J(s_1, s_2) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} = \frac{2}{2+6+5} = 0.153846$$

Thus, A is correct.

6. Option B: Since there are two possible values of x_7 where $x_2 = 0$ and $y = 2$, we calculate $P(x_2 = 0|y = 2)$ as:

$$P(x_2 = 0|y = 2) = P(x_2 = 0, x_7 = 0|y = 2) + P(x_2 = 0, x_7 = 1|y = 2) = 0.81 + 0.03 = 0.84$$

Thus, option B is correct.

6 Appendix

Full table of summary statistics:

	Mean	**Std. dev.**	**Median**	**Min**	**Max**
word_freq_make	0.10455	0.30536	0.0	0.0	4.54
word_freq_address	0.21301	1.29058	0.0	0.0	14.28
word_freq_all	0.28066	0.50414	0.0	0.0	5.1
word_freq_3d	0.06542	1.39515	0.0	0.0	42.81
word_freq_our	0.31222	0.67251	0.0	0.0	10.0
word_freq_over	0.0959	0.27382	0.0	0.0	5.88
word_freq_remove	0.11421	0.39144	0.0	0.0	7.27
word_freq_internet	0.10529	0.40107	0.0	0.0	11.11
word_freq_order	0.09007	0.27862	0.0	0.0	5.26
word_freq_mail	0.23941	0.64476	0.0	0.0	18.18
word_freq_receive	0.05982	0.20154	0.0	0.0	2.61
word_freq_will	0.5417	0.8617	0.1	0.0	9.67
word_freq_people	0.09393	0.30104	0.0	0.0	5.55
word_freq_report	0.05863	0.33518	0.0	0.0	10.0
word_freq_addresses	0.0492	0.25884	0.0	0.0	4.41
word_freq_free	0.24885	0.82579	0.0	0.0	20.0
word_freq_business	0.14259	0.44406	0.0	0.0	7.14
word_freq_email	0.18474	0.53112	0.0	0.0	9.09
word_freq_you	1.6621	1.77548	1.31	0.0	18.75
word_freq_credit	0.08558	0.50977	0.0	0.0	18.18
word_freq_your	0.80976	1.20081	0.22	0.0	11.11
word_freq_font	0.1212	1.02576	0.0	0.0	17.1
word_freq_000	0.10165	0.35029	0.0	0.0	5.45
word_freq_money	0.09427	0.44264	0.0	0.0	12.5
word_freq_hp	0.5495	1.67135	0.0	0.0	20.83
word_freq_hpl	0.26538	0.88696	0.0	0.0	16.66
word_freq_george	0.7673	3.36729	0.0	0.0	33.33
word_freq_650	0.12484	0.53858	0.0	0.0	9.09
word_freq_lab	0.09892	0.59333	0.0	0.0	14.28
word_freq_labs	0.10285	0.45668	0.0	0.0	5.88
word_freq_telnet	0.06475	0.40339	0.0	0.0	12.5
word_freq_857	0.04705	0.32856	0.0	0.0	4.76
word_freq_data	0.09723	0.55591	0.0	0.0	18.18
word_freq_415	0.04784	0.32945	0.0	0.0	4.76
word_freq_85	0.10541	0.53226	0.0	0.0	20.0

word_freq_technology	0.09748	0.40262	0.0	0.0	7.69
word_freq_1999	0.13695	0.42345	0.0	0.0	6.89
word_freq_parts	0.0132	0.22065	0.0	0.0	8.33
word_freq_pm	0.07863	0.43467	0.0	0.0	11.11
word_freq_direct	0.06483	0.34992	0.0	0.0	4.76
word_freq_cs	0.04367	0.3612	0.0	0.0	7.14
word_freq_meeting	0.13234	0.76682	0.0	0.0	14.28
word_freq_original	0.0461	0.22381	0.0	0.0	3.57
word_freq_project	0.0792	0.62198	0.0	0.0	20.0
word_freq_re	0.30122	1.01169	0.0	0.0	21.42
word_freq_edu	0.17982	0.91112	0.0	0.0	22.05
word_freq_table	0.00544	0.07627	0.0	0.0	2.17
word_freq_conference	0.03187	0.28573	0.0	0.0	10.0
char_freq_;	0.03857	0.24347	0.0	0.0	4.385
char_freq_(	0.13903	0.27036	0.065	0.0	9.752
char_freq_[	0.01698	0.10939	0.0	0.0	4.081
char_freq_!	0.26907	0.81567	0.0	0.0	32.478
char_freq_$	0.07581	0.24588	0.0	0.0	6.003
char_freq_#	0.04424	0.42934	0.0	0.0	19.829
capital_run_length_average	5.19152	31.72945	2.276	1.0	1102.5
capital_run_length_longest	52.17279	194.89131	15	1	9989
capital_run_length_total	283.28928	606.34785	95	1	15841