Exercise0 Dataset description

e De Ronde Maria
e Andre Quentin
e11921655 Fabian Holzberger

March 14, 2021

Classification Dataset: Email-Spam

We have chosen a Emai-Spam Dataset for classification (link to dataset). The goal for this dataset is to distinguish by a machine learning algorithm between spam and non-spam emails. The datasets are given in `csv` format and the structure is shown in table 1. Here we see that the dataset has a Body-column that contains the text-body of an email and a Label-column that is either set to 1 for spam E-mails or 0 for non-spam E-mails.

Index	Body	Label
100	Subject: inexpensive online medication here pummel wah springtail cutler bodyguard we ship quality medications overnight to your door !...	1
6006	Subject: organizational changes we are pleased to announce the following organizational changes : enron global assets and services in order to increase senior management focus on our international businesses...	0

Figure 1: Structure of the Email-Spam Dataset

The dataset contains in total 10.000 samples where 50% of the samples are spam- and 50% are non-spam E-mails. Since there are no missing values the dataset is perfectly balanced with respect to the target attribute. We aim to apply the **Bag of Words** method to the dataset. This Method extracts the N most common Words from all E-mails and then maps an E-mail to a vector v such that the component m of v is an nonegative integer that counts the occureces of the mth most common word in the corresponding E-mail. From that we conclude that the dimension of our dataset is $N + 1$ since we also include the target attribute. Note that we apply the following cleanup steps to all Emails to remove data that we expect to not improve the classification: 1. remove links, 2. remove characters except alphabetical ones, 3.convert uppercase-chars into lowercase-chars, 3. lemmatize words 4. remove stopwords. By that we reduce the number of total different words in all e-mails from 86733 words to 74618 words by the cleanup step. This also illustrates that our classification algorithm has to work with data that is hight dimensional (dimension much larger than 1000). In figure **??** we show the most common words in spam and non-spam emails. One can see that the word count in spam e-mails is much higher than in the non-spam e-mails which means that overall spam mails have more words in common than non-spam e-mails.

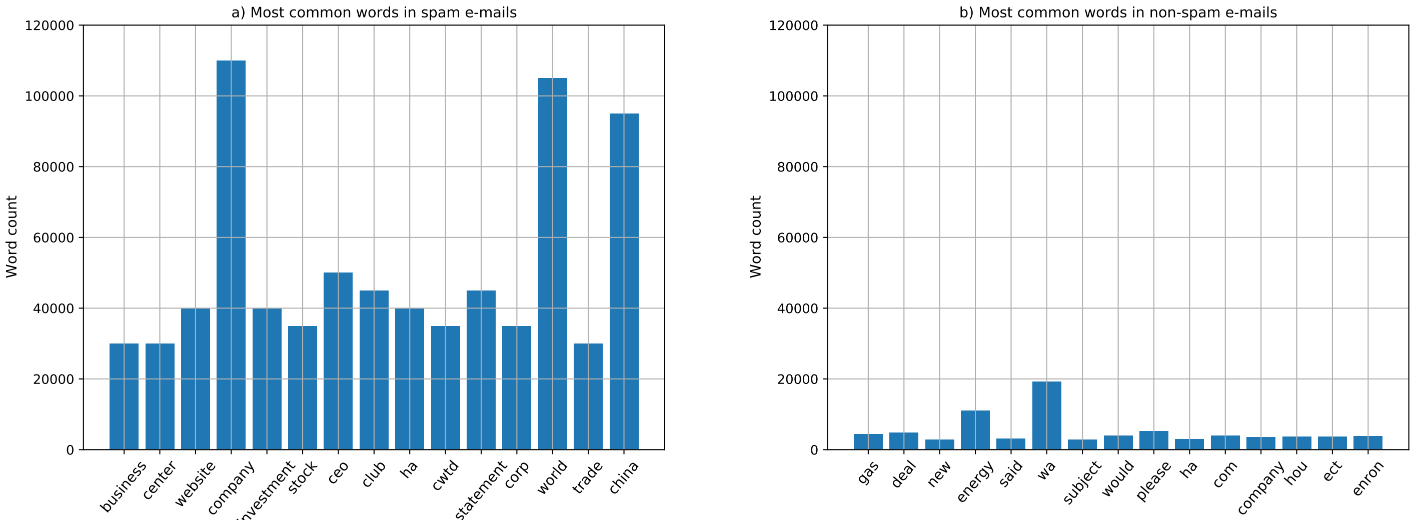

Figure 2: Figure *a*)15 most common words in spam e-mails *b*)15 most common words in non-spam e-mails

Regression Dataset: Automobile

The automobile dataset is used for regression. It consists of 26 attributes shown in Table 3.

1.symboling(O)	2.normalized-losses(I)	3.make(N)	4.fuel-type(N)	5.aspiration(N)	6.num-
7.body-style(N)	8.drive-wheels(N)	9.engine-location(N)	10.wheel-base(I)	11.length(I)	12.wid
13.height(I)	14.curb-weight(I)	15.engine-type(N)	16.num-of-cylinders(O)	17.engine-size	18.fuel
19.bore(I)	20.stroke(I)	21.compression-ratio(I)	22.horsepower(I)	23.peak-rpm(I)	24.city
25.highway-mpg(I)	26.price(I)				

Figure 3: Attributes of Automobile dataset

Comparison of the datasets

Dataset	samples	data-types	attributes	attribute format	missing Values
E-mail Spam	10.000	nominal	>1000	text, boolean	no
Automobile	205	nominal, ordinal,intervall	26	cathegorial, integer, real, text	yes

Figure 4: Characteristics of the chosen datasets