Lab 1

Thomas Haase

August 20, 2025

Table of contents

---

1 Part 1 - Bernie Sanders and Donald Trump tweets

1.1 Task 1.1

```
data <- fread(file = './trumpbernie.csv')

data[1:5,20:25]
```

```
     abe   abl abort absolut absurd  abus
   <int> <int> <int>   <int>  <int> <int>
1:     0     0     0       0      0     0
2:     0     0     0       0      0     0
3:     0     0     0       0      0     0
4:     0     0     0       0      0     0
5:     0     0     0       0      0     0
```

```
nrow(data)
```

```
[1] 1003
```

```
ncol(data)
```

```
[1] 1496
```

High-dimensionality describes a dataset where the number of variables is large relative to the number of observations. Since the columncount of the data is larger than the number of observations the dataset is highdimensional. A logistic regression would produce a perfect fit, which means that each observation can be completely explained. Since in this case also the "noise" () is part of the data modeling, the model becomes too flexible. This is also called overfitting.

1.2 Task 1.2

```
glm <- glm(trump_tweet ~ .,
           data = data,
           family = "binomial")

coef(glm)[1010:1050]
```

```
    prefer    premium     prepar       pres  prescript    present     presid
        NA         NA         NA         NA         NA         NA         NA
presidenti      press    pressur     pretti    prevent   previous      price
        NA         NA         NA         NA         NA         NA         NA
   primari      prime   prioriti     prison     privat   privileg        pro
        NA         NA         NA         NA         NA         NA         NA
   probabl    problem    process   proclaim     produc    product     profit
        NA         NA         NA         NA         NA         NA         NA
   program   progress    project     promis     proper     propos    protect
        NA         NA         NA         NA         NA         NA         NA
   protest      proud     provid     public     puerto       pull
        NA         NA         NA         NA         NA         NA
```