# Ridge and Lasso binary logistic regression in NLP using R

Eligijus Bujokas

Vilnius University

December 18, 2018

# Content

- What is natural language processing
- NLP definitions
- Logistic regression in NLP tasks
- Quora example

# Natural language processing

Natural language processing (NLP) is a sub field of computer science concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

- Classifying text
- Chat bots
- Word embedding

# NLP definitions

Corpus

A text corpus is a collection of texts of written (or spoken) language presented in electronic form.

A document is the row of a matrix

A token is single term from the corpus

# NLP definitions

Tokens

Eligijus is presenting now.

[Eligijus], [is], [presenting], [now]

[Eligijus is], [presenting now]

# NLP definitions

N - grams

An n-gram is a contiguous sequence of n items from a given sequence of text.

2 - gram (bigram):

[Eligijus is], [is presenting], [presenting now]

3 - gram:

[Eligijus is presenting], [is presenting now]

# Corpus in a matrix

A document is about cars:

| index | text | Y |
|---|---|---|
| 1 | The new car | 1 |
| 2 | fantastic car | 1 |
| 3 | the weather is dreadful | 0 |
| 4 | the weather is great | 0 |

The document term matrix would then look like:

| fantastic | new | great | dreadful | car | weather | is | the |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |

# Logistic regression in NLP context

Let us say we have n documents and k unique tokens. Each document is assigned a binary label - $\{0, 1\}$.

Then the general model is:

$$log \frac{P(Y = 1)}{P(Y = 0)} = \sum_{i=0}^{k} \beta_i X_i \, (1)$$

We start indexing from zero because of the intercept.

We get the coefficients $\widehat{\beta}$ by maximizing:

$$\sum_{i=1}^{n} \left( y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0 + \beta^T x_i}) \right)$$

in respect to $\beta_0$ and $\beta$

# Result from the example

| feature | coefficient |
|---|---|
| car | 0.4883108 |
| (Intercept) | 0.4681151 |
| new | 0.3547782 |
| fantastic | 0.2963947 |
| the | -0.2962816 |
| great | -0.3255586 |
| dreadful | -0.3255672 |
| is | -0.4882456 |
| weather | -0.4882754 |

# Lasso and Ridge logistic regression

Ridge (more uniform distribution of coefficient values):

$$\frac{1}{N}\sum_{i=1}^{n}\left(y_i(\beta_0+\beta^T x_i)-log(1+e^{\beta_0+\beta^T x_i})\right)-\alpha\sum_{j=1}^{k}\beta_j^2$$

Lasso (reduces number of features):

$$\frac{1}{N}\sum_{i=1}^{n}\left(y_i(\beta_0+\beta^T x_i)-log(1+e^{\beta_0+\beta^T x_i})\right)-\alpha\sum_{j=1}^{k}|\beta_j|$$

# Quora insincere question competition

Competition:

https://www.kaggle.com/c/quora-insincere-questions-classification

Data:

https://www.kaggle.com/c/quora-insincere-questions-classification/data

Y = 1 if the question is not sincere

Y = 0 if the question is sincere

# Example

Sincere questions:

[1] “How did Quebec nationalists see their province as a nation in the 1960s?”
[2] “Do you have an adopted dog, how would you encourage people to adopt and not shop?” [3] “Why does velocity affect time? Does velocity affect space geometry?”
[4] “How did Otto von Guericke used the Magdeburg hemispheres?”

Insincere questions:

[1] “Has the United States become the largest dictatorship in the world?”
[2] “Which babies are more sweeter to their parents? Dark skin babies or light skin babies?”
[3] “If blacks support school choice and mandatory sentencing for criminals why don’t they vote Republican?”
[4] “I am gay boy and I love my cousin (boy). He is sexy, but I dont know what to do. He is hot, and I want to see his di**. What should I do?"

# Ridge regression results

```
dim(dtm)
```

```
## [1] 161620  10956
```

```
model <- glmnet(x = dtm, y = train$target,
                family = 'binomial',
                alpha = 0,
                lambda = 0.01)
```

```
dim(coef_table)
```

```
## [1] 10957     2
```

# Ridge regression results (sincere sentiment)

| features | coefficient |
|---|---|
| pains | -3.130936 |
| immortality | -3.201486 |
| aftermath | -3.205285 |
| responded | -3.238110 |
| ipcc | -3.495939 |
| leap | -3.655830 |
| trades | -3.758036 |
| curved | -3.949973 |
| independently | -3.971127 |
| whey | -4.573653 |

# Ridge regression results (insincere sentiment)

| features | coefficient |
|---|---|
| mindless | 5.106554 |
| castration | 4.872542 |
| alabamians | 4.802497 |
| castrating | 4.697437 |
| tennesseans | 4.630287 |
| brandeis | 4.511478 |
| castrated | 4.327585 |
| nigerians | 4.296180 |
| satanism | 4.244813 |
| isaiah | 4.151869 |

# Lasso regression results

```
model <- glmnet(x = dtm, y = train$target,
                family = 'binomial',
                alpha = 1,
                lambda = 0.01)
```

```
dim(coef_table)
```

```
## [1] 266   2
```

# Lasso regression results (sincere sentiment)

| features | coefficient |
|---|---|
| tips | -0.1948865 |
| job | -0.2073240 |
| computer | -0.2271329 |
| app | -0.2314394 |
| online | -0.2642010 |
| company | -0.3068929 |
| study | -0.3218161 |
| difference | -0.3855465 |
| engineering | -0.5245851 |
| (Intercept) | -0.8000179 |

# Lasso regression results (insincere sentiment)

| features | coefficient |
|---|---|
| liberals | 1.931021 |
| indians | 1.856379 |
| trump | 1.818080 |
| muslims | 1.795633 |
| americans | 1.763183 |
| castrated | 1.608947 |
| democrats | 1.591572 |
| women | 1.552479 |
| girls | 1.519971 |
| jews | 1.385540 |