Project 2

Bilguun Chinzorig, Monika Avila, Lkham Nyambuu
EPFL

***Abstract*—**

I. Introduction

The aim of this project is to conduct a twitter sentiment analysis. Our goal is to classify untaged tweets into two groups: positive and negative. For this purpose, we begin our work by preprocessing the tweets. Following, we create the word space, i.e. matrix of embeddings using GloVe methodology. After, we span the tweet space by retrieving features from the embedding matrix. Finally, we used our futures in two types of classifiers: 1. Random Forests and 2. SVM. We conclude that random forests outperforms SVM considering both computing efficiency and classifying results.

II. Preprosecing

III. The word space: embedding matrix

In order to obtain a numerical representation of the words w_i contained in the data set as $\mathbf{w_i} \in \mathbb{R}^K$ we used GloVE model.

This model retrieves the numerical representation of the words that is closest to natural logarithm of the observed co-occurence value x_{ij}. Indeed, Pennington et al. (2014) propose to obtain the real valued vector by minimizing a weighted sum of squared errors. Where the error is the difference between the log of co-occurence value and the predicted one given by $\mathbf{w_i'w_j}$. Thus, the model minimizes the following cost function:

$$L(\mathbf{w_n}, \mathbf{w_c}) = \sum_{j=1}^{N} \sum_{i=1}^{D} f(x_{ij})(log(x_{ij}) - \mathbf{w_i'w_j})^2$$

with:

$$f(x_{ij}) = min(1, x_{ij}/x_{max})^{3/4}$$

IV. The tweet space: feature retrieval

Using the matrix of embeddings we retrieve the tweet space. For this, we generate features that

V. The Classification Model

After obtaining the tweet space i.e. the features that represent the data set of tweets, we trained the classification model. For this aim, we used two models: 1. Random Forest and 2. SVM. The first model is random forest....

The second classification algorithm used is SVM. In this method, we look to maximize the geometric margin between positive and negative tweets subject to the constraint that the classified vectors must lie outside this margin[1]. This optimization problem is equivalent to the following one for each training sample:

$$min_w ||w||^2$$

$$s.t. \quad y_i(x_i'w + b) \geq 1$$

Now, setting the dual problem we can obtain the support vectors which are a few training observations that lie on the margin.

VI. Methods

VII. The Data

The training data set is formed by 2 millions of tweets, half of them correspond to the positive emotions and the other half to negative ones. This data has been already tokenized, thus we start with the cleaning and the analysis of the data.

First, we remove the hashtags, URL directions and @...

Second, we counting the words and we check Zipf's law.

After this, we extract features: ... a. n-grams b.

VIII. Results

IX. Discussion

X. Summary

[1] AndrewNg, Support Vector Machines

References

Pennington, J., R. Socher, and C. Manning
2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Pp. 1532–1543.