Capstone Proposal
Udacity Machine Learning Nanodegree

Chiel Peters

December 23, 2016

Domain Background

In order to keep our email inbox clean and maintainable we classify incoming e-mails according to their content. A well known example is the spam filter, this algorithm distinguishes spam from ham and keeps the inbox clean. However there are many more classifications possible. In this capstone project I would like to build a email classifier based on content. This algorithm would benefit users in automatically categorizing their emails instead of doing it manually.

I am personally interested in this topic because of an acquaintance of mine. She works at a customer care department for a large dutch energy company. Each day they receive uncategorized emails which have to be manually forwarded to the correct department based on the question of the customer. This takes time and money. One other side goal is that automatically labelling the emails allows for good reports. By labelling the email according to the content senior management can observe which common problems customers are facing and spend their attention accordingly.

Problem Statement

In order to formally define the email categorizing problem I use the notation found in [1]. An inbox is a set of emails each belonging to a certain category[1]: $E_{train} = \{(e_1, c_1), ..., (e_n, c_n)\}$ where each email e_i is accompanied by a category c_i. The set of categories is predefined: $C = \{c_1, ..., c_m\}$ The goal of the machine learning algorithm is to learn the mapping $E \rightarrow L$ using the training set such that new documents can be correctly categorized.

Datasets and Inputs

For traceability and benchmark purposes I prefer a (big) public dataset over a private/new one[2]. As dataset I will use the well known Enron email corpus

[1] I will assume that an email can only belong to one category

[2] For instance, my own inbox

([2]). This corpus consists of over 200.000 email messages belonging to 158 users. Although not every user might have good classified inboxes, it is shown by [1] that at least seven users have valid inbox foldering with over 25.000 messages.

Solution Statement

In order to classify the emails I will use the support vector machine (SVM) algorithm. This approach is described in detail in [3]. For my purposes there might be more than two categories per user. The SVM algorithm is however a binary classification algorithm. To overcome this problem the one-vs-all approach is applied. Here separate models are trained for each of the categories. The algorithm tries to separate the emails of a certain category by a hyperplane. This hyperplane exists in the feature space.

The features are obtained from the emails content. The content of the email is represented as a bag-of-words. That is messages are represented as vectors where elements of the vectors contain the count of a certain word. In order to combine similar words the content of the email is first stemmed and also the stopwords are removed. The 1000 [3] top frequent words will be used as features.

Benchmark Model

The solution is compared to the Naive-Bayes classifier. This classifier works by calculating the probability that a word belongs to a certain category. These probabilities are aggregated to calculate the probability that an email belongs to a category.

Evaluation Metrics

The simplest measure I will use is accuracy. That is the fraction of emails that is correctly classified. Other alternatives that could be used are the precision,recall and F1-Score. The precision measures the fraction of correct predictions for a certain category while the recall measures the fraction of instances of a category that were correctly predicted. The F1-score combines the two evaluation metrics.

Project Design

The project is divided into four separate phases: proposal, data analysis, modelling and evaluation & interpretation. This document is contained within the first phase and describes the problem domain.

In the second phase the Enron data is researched. While the corpus itself is huge only a portion of it might be viable for this type of research. The model will

[3] hyperparameter which will be optimized in the project

work a per user basis, but conditions might exclude certain users. For instance, if a user has only one folder or all its messages are saved in automatic folders (Inbox, Spam). Another important concept to look at is time. The emails are received on a timeline which might influence the word usage (e.g. new words) and categorization (e.g. new categories are made later on).

After the data is carefully cleansed the third phase is started. In this phase the SVM and Naive Bayes model are built. Pyhton and sci-kit will be used as a programming platform.

The fourth phase consists of evaluation & interpretation. After the model is built the results are interpreted according to the evaluation metrics described earlier. It is likely that some results might warrent further investigation into the data. Also there might exist differences in performance between Enron employees.

References

[1] R. Bekkerman, A. McCallum, and G. Huang, "Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora," *Center for Intelligent Information Retrieval, Technical Report IR*, vol. 418, 2004.

[2] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *European Conference on Machine Learning*, pp. 217–226, Springer, 2004.

[3] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms.* Norwell, MA, USA: Kluwer Academic Publishers, 2002.