UCM VI Modelling Week
Supervised information retrieval

Problem description

High performance. Delivered.

accenture neometrics

consulting | technology | outsourcing

01 Problem description

An important part of Text Mining is document categorization and classification. It consists in automatically classify documents, into one or more categories, based on their content.

Text mining can make our life easier helping us:

- Identifying document's class
- Filtering spam
- Selecting fundamental information to spread
- Etc

Applied to social networks it can help us to understand how information spreads and also to control critical information underneath the network to handle reputational risk.

01 Problem description

We provide a set of documents (corpus) in which each document has been manually classified into a polarity target. Hence every document may be marked as positive, negative or neutral.

The objective is to build a model capable to/of automatically classify new documents according to their polarity.

There are several ways to do this task. For example, it is possible to build a polarity classifier based on a term value dictionary, where every term is valued with a high score in case the term does express positivity and with a low score in case it expresses negativity. The final classification is realized employing term scores that appear on each text. This technique offers a limited precision.

The most common procedure used to apply analytic techniques to text data consists on transforming raw text in quantitative information. For this aim NLP (Natural Language Processing) techniques are applied.

02 Mathematical treatment – data transformation

Textual content treatment using NLP may be very complex. Due to the lack of time, you will be provided with a second dataset, in addition to the corpus, that contains the tf-idf measure for each term at each document.

This dataset will be in sparse format and it will allow you to gap using the NLP techniques and to concentrate on the mathematical polarity modelization. The second dataset has been obtained this the following data treatment:

- Normalization of the documents (orthographical error's correction)
- Tokenization and Sentence detection
- Part of Speech tagging
- Lemmatization
- Word selection (stop-word dictionary).
- Document-term matrix with term frequency per document
- tf-idf transformation

Neither the corpus or the tf-idf frequency or both datasets may be used to solved the problem.

In case raw text from the corpus is used, you will need to tokenize the documents and extract your own relevant information.

In case second tf-idf dataset is used, this would provide a quantity of information difficult to work with. It is necessary to select word or/and reduce dimension.

Another difficulty is that this type of data does not work well with linear classifiers, however we will have the opportunity to work with other classifiers (knn, SVM).

02 Mathematical treatment - modelizacion

This is a supervised problem, due to the fact that we have information of the real target value of the document. To adjust the model you will need to separate the data into two different groups:

- Train set: this observations will give information to adjust the model and to predict new observations. We recommend you to divide this set into a training set and a validation set in order to avoid overfitting.
- Test set: this observations will help us to obtain a real measure of the model's goodness of fit. It will not be used to fit the model.

Because of the nature of factors (explicative variables) obtained by dimension reduction it is almost impossible to find linear relations between the explicative variables and the target. Here we show a couple methods which could be applied for this aim:

- K nearest neighbors
- Support Vector Machines
- Dictionaries of words with their polarity joint
- Boosting and others

It could work well also to combine several methods.

The aim is to develop a classifier capable of distinguish between categories with high performance. Be aware of the skewness of the target's distribution!

Several measures could be used:

- Percentage of well-classified (distribution dependent)
- Precision and recall of every value of the target
- Confusion matrix
- Area under the ROC curve (distribution independent)

03 Guide-lines to solve the problem

The corpus is a set of documents extracted from social media, all written in the same language, with a three-value polarity target. Their length may be variable.

You will need to find the corpus' descriptives, such us: number of documents, number of factors/terms, distribution of the target and any other descriptive which could be useful.

You will have freedom to choose the model and to use the dataset you consider more convenient.

You will be provided with the raw text: this could be used to develop a model based on word with high polarity, helped with a polarity dictionary.

- You will be provided with the document-factor matrix: this ables you to build an agnostic model without looking at the content of the texts.

Once the model is adjusted, you will run the final model into the test set in order to view if the model is over-adjusted or not.

For the two previous steps you will need a measure to optimize. The ROC curve must be included in the list of measures used to fit the model, in summation of the typical measures used. The ROC curve is independent from the target's distribution and you will have to compute one ROC curve per target value.

To conclude prepare several reports to show how your model works.