Text Driven Forecasting

Pedro Lorga Ramos
pedrolorgaramos@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2013

Abstract

In this article, we address the problem of text-based forecasting. Two tasks are considered: predicting movie’s revenues and predicting the number of clicks on news from a website. We provide an empirical comparison among several regularization strategies (ridge regression, Lasso, and the elastic net), discussing each method’s strengths and weaknesses. The experimental results show that our chosen approach (SpaRSA) outperforms all the baselines, achieving promising results on the number of iterations needed to obtain the final trained model, and also, in the error obtained when comparing to the error using the baseline.

1. Introduction

1.1. Motivation

In the last decades, Internet became a phenomenon that completely changed the way society communicates and expresses itself. The massive amount of Internet users generate an enormous and continuous flow of information in the network. This presents an opportunity for statistical-based approaches to thrive in predicting and interpreting events (such as clickthrough rate, movie revenues, stock values, or opinion polls) based on consumer-generated data. Some services and sites alone are responsible for a large amount of the information flow. Those services and sites, like blogs and social networks, are used for interests, sentiments and opinions sharing. That amount of information created a great interest in scientific community because it could be used as a new source of data. One of the uses of this data is predicting something that is somehow related to information inputed by the users of those services. Although it is possible to get interesting information from these sources, it is hard to retrieve the information contained in these sources because it is coded in text.

The processing of natural language is difficult because there are too many languages and dialects, each one with an enormous and sparse vocabulary that changes daily, grammatical rules and exceptions to those rules and there are also ambiguous sentences. It gets even more complicated when the text describes opinions, sentiments or perceptions. Although these difficulties, some advances were recently made in order to solve those problems. Through statistical models it is possible to analyze the text, and connect it to measurable phenomena of the real world. This process is very useful, enabling predictions in several interesting areas, like economics and business. It is possible, for example, to predict the risk associated with an investment, the success of movies or the number of views of news articles. In this thesis some experiments are going to be made in order to predict the movies revenues and number of clicks in a news website.

1.2. State of the Art

1.2.1 Text driven forecasting

The use statistical models in order to made prediction using information contained in text is recent. It will be enumerated several experiments related to this line of investigation. Joshi et al. [8] compiled the data that we use in order to predict movies revenues. Their article also tries to predict the movies revenues using the text from reviews of the movies and some metadata also collected by them. Some other experiments were made, such as predicting financial risk from financial reports [9]. For that article support vector regression was used [6]. In that case, the financial risk is measured through stock return volatility. It consists in variation measure of the price of some financial instruments through time. Experiments were also made in order to predict the NFL (National Football League) match results and bet results using text contained in tweets and match statistics [11]. They used linear and logistic regression. Another example of work done in this area is predicting the age of the author of an article using the article’s text [10].

1.2.2 Statistical Models of Regression

Several models of statistical regression have been proposed by the scientific community of statistical
learning, being those the appropriate tools to build a prediction system from text. Examples of that are the ridge regression, which uses $\ell_2$ regularization [7], the Lasso [12] that uses $\ell_1$ regularization and Elastic Net [14] which uses a combination between both $\ell_1$ and $\ell_2$ regularization. Ridge regression is the simplest one and, like all the other regularizations, avoids overfitting. Overfitting is the lack of capacity of models to generalize and work properly in data sets that differs from the one in which the model was trained. The use of regularization also enables the model to compute only one solution. The Lasso was trained. The use of regularization also enables sets that differs form the one in which the model of models to generalize and work properly in data avoids overfitting. Overfitting is the lack of capacity

simplest one and, like all the other regularizations, to avoid it, regularization is used. With the use of regularization, the obtained model is more generalized and usable in data sets other than the training set. The regularization also reduces the number of solutions that can be obtained in the minimization problem to one even when the quantity of features is larger than the amount of documents. Adding the regularization creates the following general objective function:

$$F(w) = Q(w) + \lambda R(w)$$ (3)

Where $\lambda$ is the parameter that controls how much regularization is used, $R(w)$ is the regularization function and

$$Q(w) = \|Xw - y\|_2^2$$ (4)

The process of training and estimate the $\hat{w}$ is simply defined by problem

$$\hat{w} = \arg\min_w F(w)$$ (5)

There are several regularization functions that can be used, three of them are: the ridge regression, with $\ell_2$ regularization ($R_{\text{Ridge}}(w) = \frac{1}{2}\|w\|_2^2$), the Lasso with $\ell_1$ regularization ($R_{\text{Lasso}}(w) = \|w\|_1$) and the Elastic Net that uses a combination of the last tho norms ($R_{\text{Elastic Net}}(w) = \alpha \|w\|_1 + \alpha \frac{1}{2}\|w\|_2^2$).

1.2.3 Optimization Algorithms

Several algorithms can be used to minimize objective functions resulting from the models. The SpaRSA [13] is the chosen algorithm to be used in this article and it will be explained in the next section. Another example of an optimization algorithm is Least-Angle Regression [3] in which is used the correlation between the chosen direction for each iteration and the residue (distance to the final solution). That algorithm belongs to a group of algorithms called Active Set Methods. All the algorithms in this group divide the main problem in several smaller problems that are solved sequentially. Another algorithm that can be used is the Gradient Projection for Sparse Reconstruction [4]. This algorithm solves the problem by converting it to a quadratic problem, separating the weight vector in two, one containing all the positive weight and the other the negatives.

There is also another group of algorithms called Proximal Gradient, which can be used to solve the problem of minimizing the objective function. Some of the algorithms that belong to that group are Fast Iterative Shrinkage Thresholding (FIST) [1] and Two Steps Iterative Shrinkage Thresholding (TwIST) [2], both based on Iterative Shrinkage Thresholding (IST) algorithm which consists in alternating the gradient steps with a soft-thresholding function. Finally, other algorithm that can be used is Coordinate Descent [5]. This algorithm takes advantage of the sparsity of the data, searching the solution in the direction of only one axis at a time, alternating the axis cyclically.

2. Background

2.1. Linear Regression

To solve predicting problems, like those we propose in this thesis, it is often used linear regression. The linear regression has the form

$$y \approx Xw + \beta_0$$ (1)

Where $X$ is the matrix that contains the data related with the text, $y$ is the vector that contains the data related the values which it is intended to predict and $w$ is the vector that contains the weights of the features contained in $X$. The parameter $\beta_0$ is the offset. We consider that $\beta_0$ is zero, because instead of an offset we use a constant feature. In order to predict $y$ we need an estimate of $w$, this estimate is obtained in the process of training the model. If we use a simple linear regression, the process of training is reduced to minimizing the difference between the original $y$ and the predicted one, $Xw$. 

$$\hat{w} = \arg\min_w \{\|Xw - y\|_2^2\}$$ (2)

However this simple statistical model will produce a $w$ that is to much adapted to the training set (data used in the training process) causing a high error when it is used to make a prediction in a different data set. This effect is called overfitting and to avoid it, regularization is used. With the use of regularization, the obtained model is more generalized and usable in data sets other than the training set. The regularization also reduces the number of solutions that can be obtained in the minimization problem to one even when the quantity of features is larger than the amount of documents. Adding the regularization creates the following general objective function:

$$F(w) = Q(w) + \lambda R(w)$$ (3)

Where $\lambda$ is the parameter that controls how much regularization is used, $R(w)$ is the regularization function and

$$Q(w) = \|Xw - y\|_2^2$$ (4)

The process of training and estimate the $\hat{w}$ is simply defined by problem

$$\hat{w} = \arg\min_w F(w)$$ (5)

There are several regularization functions that can be used, three of them are: the ridge regression, with $\ell_2$ regularization ($R_{\text{Ridge}}(w) = \frac{1}{2}\|w\|_2^2$), the Lasso with $\ell_1$ regularization ($R_{\text{Lasso}}(w) = \|w\|_1$) and the Elastic Net that uses a combination of the last tho norms ($R_{\text{Elastic Net}}(w) = \alpha \|w\|_1 + \alpha \frac{1}{2}\|w\|_2^2$).
2.2. Optimization Algorithms

In order to minimize the objective function it can be used one of the several methods. For the ridge regression a solution can be found using its closed form:

\[
\hat{w} = (XX^T + \lambda I)^{-1}X^Ty
\]  

Using the closed form has some advantages, we get the exact solution and for a small amount of data it is easily computed. But in order to computed the result using this method it is needed to compute the inverse of a matrix, which is not easy to compute in a large matrix. In those cases, it is used an iterative method like SpaRSA. The SpaRSA is a gradient based method in which, for each iteration, a step is taken in the opposite direction of the gradient. In the SpaRSA (Sparse Reconstruction by Separable Approximation) the size of the step is variable and it is adapted in each iteration. When the algorithm is close to the solution, the step is reduced and when the solution is far, the step is increased in order to get to converge to it as quickly as possible. This approach is based in Barzilai-Borwein methods. In this method the step size (\(\eta\)) is chosen in such a way that \(\eta \times I\), where \(I\) is the identity matrix, is a approximation of the Hessian matrix \(\nabla^2 F(w)\). In order to satisfy this imposition, the step size \(\eta\) is obtained in the following way:

\[
\eta = \frac{n}{\alpha_t}
\]  

\[
\alpha_t = \frac{||Xs_t||^2}{||s_t||^2}
\]  

\[
s_t = w_t - w_{t-1}
\]  

Where \(n\) is a parameter used to adjust the step size if necessary but generally \(n = 1\) is used. It is important to note that the SpaRSA only estimates the step size, for the remaining calculations, the ridge regression, Lasso or Elastic Net are used. The ridge regression uses only a simple gradient step 2.1, the Lasso 2.2 and the Elastic Net 2.3 use a gradient step and a proximal step. The proximal step uses the soft-thresholding (10) function.

**Algorithm 2.1 Ridge Iterative**

\[
\hat{w}_t \leftarrow 0
\]  

while \(t < \text{max}_\text{iter}\) do

\[
s_t \leftarrow w_t - w_{t-1}
\]

\[
\eta \leftarrow \text{get_alpha}(X, s_t)
\]

\[
\hat{w}_{t+1} \leftarrow w_t - \eta \times \nabla F(w_t)
\]

\(t \leftarrow t + 1\)

end while

return \(\hat{w}_t\)

**Algorithm 2.2 Lasso - Proximal Gradient**

\[
\hat{w}_t \leftarrow 0
\]  

while \(t < \text{max}_\text{iter}\) do

\[
s_t \leftarrow w_t - w_{t-1}
\]

\[
\eta \leftarrow \text{get_alpha}(X, s_t)
\]

\[
u_t \leftarrow w_t - \eta \times \nabla Q(w_t) \quad \text{(Gradient Step)}
\]

\[
w_{i,t+1} = \frac{1}{2}(v_t - \eta) = \frac{1}{2}||v_t - \eta||^2_2 + \eta \lambda R(u_t) \quad \text{(Proximal Step)}
\]

\(t \leftarrow t + 1\)

end while

return \(\hat{w}_t\)

**Algorithm 2.3 Elastic Net - Proximal Gradient**

\[
\hat{w}_t \leftarrow 0
\]  

while \(t < \text{max}_\text{iter}\) do

\[
s_t \leftarrow w_t - w_{t-1}
\]

\[
\eta \leftarrow \text{get_alpha}(X, s_t)
\]

\[
u_t \leftarrow w_t - \eta \times \nabla Q(w_t)
\]

\[
\hat{w}_{i,t+1} \leftarrow \frac{1}{2}||v_t - \eta||^2_2 + KR(u_t)
\]

\(t \leftarrow t + 1\)

end while

return \(\hat{w}_t\)

The function \(\text{get}_\text{alpha}\) returns the value of the step size calculated with SpaRSA for the current iteration. It can also be used an acceptance criteria to avoid steps that lead the algorithm away from the solution. In (2.3) \(K\) and \(b_t\) are:

\[
K = \frac{\eta \lambda \alpha}{1 + \eta \lambda (1 - \alpha)}
\]  

\[
b_t = \frac{w_i}{1 + \eta \lambda (1 - \alpha)}
\]

3. Text Driven Forecasting

Before presenting our experiments, we provide a rigorous definition of the problem we are trying to solve. We define the problem as follows: given some document \(T\) associated to one social phenomena, a concrete prediction of a quantity \(M\) of that phenomenon that can only be observed in the future is made. That prediction is made using the information obtained from \(T\) which is accessible in the present. It is important to note the difficulty in accessing the information contained in \(T\). This information is generally coded in text, which is noisy, sparse and badly structured. It gets worse when
analyzing documents that describe sentiments or opinions where the text contains ambiguous sentences, ironies and metaphors. The language itself may be a problem. There are several languages, each one with its unique vocabulary and set of rules. Language is also constantly evolving, changing every day.

We intend to attack two specific cases of predicting using text as guideline: predicting movies revenues and predicting the number of clicks in news articles. For the movies case, we will use text from reviews published in newspapers and magazines and also metadata like present actors and directors. For the news articles case, the predictions are based on text obtained from the news themselves. In this case we are also going to use some metadata like a flag that indicates if the new was ever clicked while it was on the main page or the day of the week that the new was published.

4. Experimental Results

In this thesis several experiments were made. These experiments had the objective to test several models (ridge regression, Lasso and Elastic Net) in order to compare their behavior. Another goal of the experiments was to minimize the error in the prediction of movies revenues and forecasting the number of clicks in news articles in a newspaper site. The data used to predict movies revenues was collected by Joshi et al. from reviews made of the movies. The data used in the prediction of clicks comes from text contained in the news articles. Some of the results, marked with an *, were obtained by removing from the features some unwanted common words and with low value to the predicting problem. Removing these words in the marked cases increased the quality of the results.

4.1. Prediction of Movies Revenues

A baseline was implemented to serve as comparison with the proposed methods. The baseline consists on predicting that all movies would get the same revenue, namely the average of all movie’s revenues in the training set. Absolute error obtained in the test set for this baseline was 11.67 millions dollars and the mean was 7.05 millions dollars.

| Table 1: MAE in million dollars without source division |
|-----------|-----------|-----------|-----------|
| Metadata | Ridge     | Lasso     | EN        |
| 5.99     | 6.14      | 6.14      |           |
| Text     | 7.52      | 7.65      | 7.54      |
| Text + Metadata | 5.97*     | 5.97*     | 6.03*     |

4.2. Prediction of Number of clicks in news articles

For the news articles it also was considered a baseline to compare to the obtained results. The baseline is again based on the mean of the training set. The absolute error obtained in the test set for this baseline was 803.71 clicks and the mean was 775.75 clicks.

| Table 2: MAE in million dollars with source division |
|-----------|-----------|-----------|-----------|
| Text      | Ridge     | Lasso     | EN        |
| 7.54      | 7.51      | 7.54      |           |
| Text + Metadata | 5.99*     | 6.00*     | 6.03*     |

| Table 3: MAE for the news experiments |
|-----------|-----------|-----------|-----------|
| Text+Metadata | Ridge     | Lasso     | EN        |
| 659.31     | 676.42*   | 659.7*    |           |
| Text       | 665.03*   | 682.59*   | 663.71*   |

| Table 4: Words with highest weights |
|-----------|-----------|-----------|
| Top positive | Weight (# clicks) |
| 1 | [flag] | 322.48 |
| 2 | Interior[url] | 174.28 |
| 3 | pessoas[url] | 167.46 |
| 4 | inicio[url] | 144.31 |
| 5 | cafe | 144.09 |
| 6 | actriz | 137.21 |
| 7 | youtube | 135.34 |
| 8 | vdeo | 117.12 |
| 9 | tributaria | 110.07 |
| 10 | fatura | 107.35 |

| Table 5: Words with lowest weights |
|-----------|-----------|
| Top negative | Weight (# clicks) |
| 1 | lusa | -136.65 |
| 2 | publicado | -99.71 |
| 3 | artes[url] | -69.85 |
| 4 | porque | -65.68 |
| 5 | hoje | -62.55 |
| 6 | faz | -53.40 |
| 7 | album | -50.81 |
| 8 | onde | -49.75 |
| 9 | risco | -47.11 |
| 10 | pia | -46.75 |

The features that are marked with “[url]” are metadata. These features are obtained by analyzing the news article’s URL. The metadata indicates the category in which the news articles is inserted. There is also the “[flag]” feature, also a metadata, that indicates if the news articles were anytime accessed from the front page. In the most relevant
features there are several references to visual illustrations like videos and pictures. This reveals that people show more interest in news articles containing or referencing images or videos. There are also some features that indicate the category of the news article in the positive top 4. The most viewed sections are “interior”, “pessoas” and “nicio”. Those features were collected from the URL of the news article.

The feature with highest value is the feature “[flag]” which is a flag that indicates if the new article was opened in from the front page of the website. The lowest weight belongs to the feature “lusa”. Lusa is an news agency that collects news for newspapers. The reason for this feature have such a low weight might because of the news from Lusa are distributed to other newspapers, causing the views of the news article to be distributed for several newspapers in comparison to exclusive news which can only be found in a specific newspaper. The news article provided from this agency are also usually short, and in many cases the title contains almost all the information. In those cases the viewer do not click on the news article because he is satisfied with the information that he got from the title. It is also a very common feature and may be absolving the bias.

5. Conclusions

From the careful analysis of the results, several conclusions can be pointed out. In both experiments it is obvious the importance of the metadata, since in both the error decreased when it was used text and metadata in comparison to using only text. About the models, although the ridge regression is the simplest, it was also the one getting better results. The downside of using ridge regression is the feature selection characteristic of ℓ1 regularization that may be interesting to remove unnecessary features and speed up the process. An analysis should also be made to the resulting weight vector. Computed errors are smaller than the one’s obtained by the baseline’s method, indicating that the experiments had positive results.

References


