How well do we know Bernoulli?

Giorgio Maria Di Nunzio[1] and Alessandro Sordoni[2]

[1] Dept. of Information Engineering – University of Padua
dinunzio@dei.unipd.it
[2] Dept. of Computer Science and Operations Research – University of Montreal
sordonia@iro.umontreal.ca

Abstract. Naïve Bayes probabilistic models are widely used in text categorization because of their efficient model training and good empirical results. Bayesian classifiers face a common issue called data sparsity problem which makes an adequate estimation of probabilities a difficult task. Therefore, smoothing techniques are needed in order to adjust the maximum likelihood estimators. In this preliminary paper we make use of a visualization technique to further investigate the expressiveness of the well known Bernoulli Naïve Bayes classifier. Various smoothing methods are tested by means of a visual analysis which makes the estimation of optimal parameters straightforward. Experimental results demonstrated that: (1) visual analysis is a valuable tool for understanding the behaviour of smoothing methods and their limits (2) the Bernoulli multivariate model performance can increase significantly with a suitable setting of smoothing parameters.

1 Introduction

A large number of studies have shown that Support Vector Machines (SVM) can outperform other approaches in many categorization applications [1], but Naïve Bayes (NB) is still widely used in practice mostly likely due to its tradeoff between very efficient model training and good empirical results. NB classifiers are sensitive to the data sparsity problem which is particularly evident when the size of training data is small. Due to data sparseness, the maximum likelihood estimation of the probability of unseen features (terms in the case of text classification) tend to be zero. To prevent this undesirable behaviour, smoothing techniques are a possible solution. Smoothing a probability actually means assigning a non-zero probability to the features that describe the object we want to classify. Several smoothing methods have been proposed [2]: additive, or *Laplacian* smoothing, Jelinek-Mercer, Dirichlet, absolute discount and two-stage smoothing. Some of these approaches operate an interpolation with a background collection model, some others simply add extra counts to the observed frequency of each feature.

In this preliminary work, we are interested in studying smoothing methods for the multi-variate Bernoulli classifier. Most research so far has shown that the multinomial Naïve Bayes generally outperforms the Bernoulli classifier both in text categorization [3] and information retrieval [4]. From a probabilistic point

of view, the latter model makes a weaker independence assumption about word occurrences at the price of not being able to model multiple word occurrences. Even if there has been some empirical evidence that multinomial outperforms multi-variate Bernoulli, the need for a more systematic comparison between these model is needed [5]. Therefore, we put forward the following research question: how far can we improve the performance of the Bernoulli classifier by setting optimal Beta prior smoothing parameters? The objective of our experimental evaluation (inspired by the work of [2]) is to compare three well-established smoothing methods against a manual optimization of the Beta parameters by means of the two-dimensional visual approach [6].

2 Bayesian and Jelinek-Mercer Smoothing

Given a set C of categories, the bayesian approach to categorization consists by estimating $\mathrm{P}(d|c_i)$ and calculating the posterior $\mathrm{P}(c_i|d)$ via Bayes rule[3]. The multi-variate Bernoulli model represents a document as a binary vector over the space of terms in which each dimension indicates whether the term occurs in the document. The occurrence of each term is governed by a Bernoulli distribution. Learning the parameters of this model corresponds to estimating class-conditional Bernoulli parameters $\theta_{t_k|c_i} \equiv \mathrm{P}(t_k|c_i;\theta)$, where t_k is a term of the vocabulary. The maximum likelihood (ML) estimators of this parameters are of the form:

$$\hat{\theta}^{ML}_{t_k|c_i} = \frac{\tau_{k,i}}{m_i} \tag{1}$$

where $\tau_{k,i}$ is the number of documents belonging to c_i in which term t_k appears and m_i is the total number of documents in c_i. The ML is zero for terms that never occur in documents in c_i. To prevent this undesirable behavior, the choice of a suitable prior to smooth probabilities is a possible solution. The conjugate prior of the Bernoulli distribution is the beta-distribution $beta(\theta;\alpha,\beta)$, where α and β are hyper-parameters. Assuming this prior, the smoothed estimate of the probability of a term t_k given a category c_i is given by the posterior mean [7]:

$$\hat{\theta}^{B}_{t_k|c_i} = \frac{\tau_{k,i} + \alpha}{m_i + \alpha + \beta} \ , \tag{2}$$

Setting $\alpha = 1$, $\beta = 1$ is called Laplace smoothing. Using the Jelinek-Mercer (JM) method, this parameter is computed by interpolating the maximum likelihood estimate with a collection language model $\theta_{t_k|C} \equiv \mathrm{P}(t_k|C;\theta)$:

$$\hat{\theta}^{ML}_{t_k|C} = \frac{\tau_k}{m} \ , \tag{3}$$

where τ_k is the number of documents in which term t_k appears and m the number of documents in the collection. Using λ as the interpolation parameter, the Jelinek-Mercer can be written as:

$$\hat{\theta}^{JM}_{t_k|c_i} = (1-\lambda)\hat{\theta}^{ML}_{t_k|c_i} + \lambda\hat{\theta}^{ML}_{t_k|C} \ , \tag{4}$$

[3] $\mathrm{P}(c_i|d) = \mathrm{P}(d|c_i)\mathrm{P}(c_i)/\mathrm{P}(d)$, where $c_i \in C$ and d is a document.

with $0 \leq \lambda \leq 1$. For $\lambda = 0$, we obtain the maximum likelihood estimator, while for $\lambda = 1$ we completely rely on the collection language model. Indeed, opposite to Beta smoothing, the Jelinek-Mercer smooths each parameter $\hat{\theta}^{ML}_{t_k|c_i}$ by a different amount depending on the probability of the term with respect to the entire collection. Nevertheless, looking closer at Eq. (2), we can write:

$$\hat{\theta}^{B}_{t_k|c_i} = \frac{m_i}{m_i + \alpha + \beta} \frac{\tau_{k,i}}{m_i} + \frac{\alpha + \beta}{m_i + \alpha + \beta} \frac{\alpha}{\alpha + \beta} \ , \tag{5}$$

which means that the probability of a term is obtained by interpolating the maximum likelihood estimator with the prior mean $\alpha/(\alpha + \beta)$. Setting $\alpha = \beta\, \tau_k/m - \tau_k$, such that $\alpha/(\alpha + \beta) = (\tau_k/m)$, we recover the JM except that the interpolation slope is flatter: we must allow β to vary through a bigger interval in order to recover JM estimations[4]. In our experiments, and for the purpose of visual analysis, we limit ourselves to use the same α and β for each smoothed estimate of term t_k. As we will see in the next sections, this will represent a lack of expressiveness of the Beta prior smoothing and opens a path for the continuation of this work.

3 Visualization of Priors' Effects

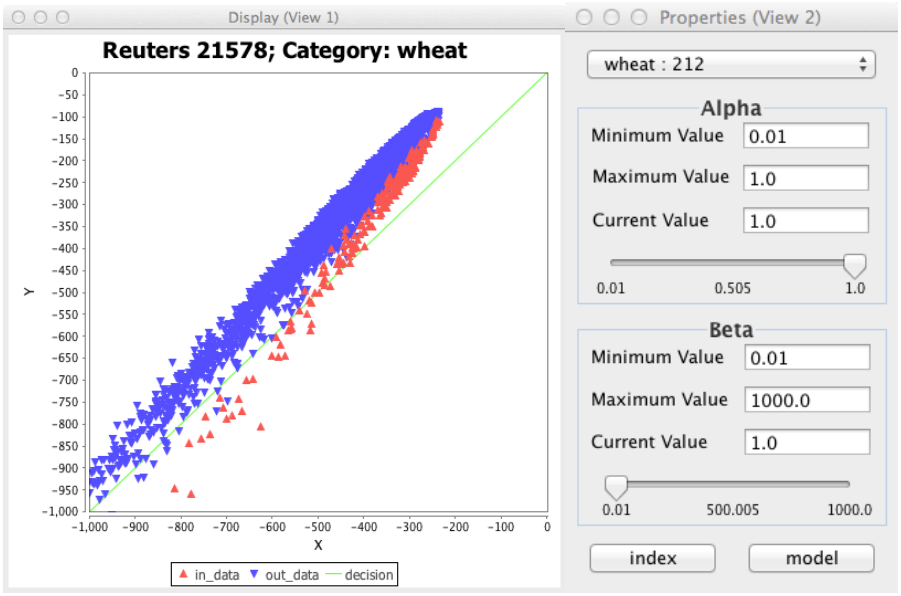
In this work, we make use of a visual analysis tool, namely the two-dimensional visualization of probabilistic models [6], for understanding the behaviour of smoothing methods and their limits. In the two-dimensional visualization, two coordinates are calculated for each document d and for each category c_i. These two coordinates correspond to the two posterior probabilities $\mathrm{P}(c_i|d;\hat{\theta})$ and $\mathrm{P}(\bar{c}_i|d;\hat{\theta})$ governed by the estimated parameter $\hat{\theta}$. We compare these two probabilities to decide whether the document belongs to c_i or not. By applying Bayes rule and taking the logs in order to avoid arithmetical anomalies (products of very small numbers tend to zero very quickly) we obtain:

$$\log\Big(\mathrm{P}(d|c_i;\hat{\theta}_{c_i})\Big) + \log\Big(\mathrm{P}(c_i;\hat{\theta})\Big) > \log\Big(\mathrm{P}(d|\bar{c}_i;\hat{\theta}_{\bar{c}_i})\Big) + \log\Big(\mathrm{P}(\bar{c}_i;\hat{\theta})\Big) \tag{6}$$

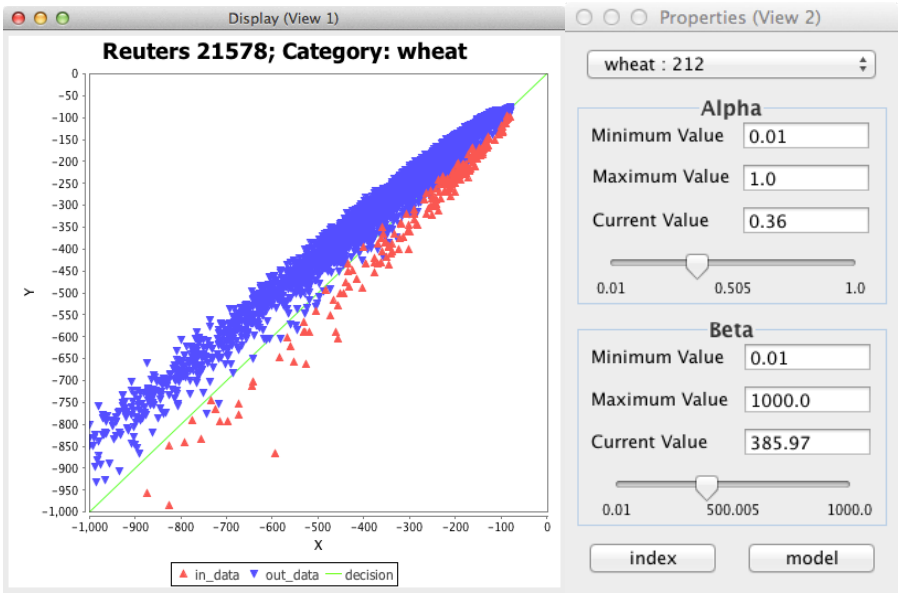
Given a category c_i, each coordinate of a document is the sum of two addends: a variable component which depends on the terms that appear in the document, and a constant component related to probability of the category itself. The probability $\mathrm{P}(d|c_i;\hat{\theta}_{c_i})$ is in turn estimated by combining the estimates $\hat{\theta}_{t_k|c_i}$ for each term in the document. We can therefore determine and change the position of the document in the two-dimensional space by adjusting the hyper-parameters α and β.

An example of this visualization is shown in Figure 1. The decision boundary is represented by the green line: below the line, the document is assigned to the category c_i, above the line, the document is assigned to $\bar{c}_i$. The influence of a change in the values α and β is visualized with an animation of the documents in the space.

[4] A similar derivation is done in [8] with a Dirichlet prior.



(a) Display window. (b) Properties window



(c) Display window. (d) Properties window

Fig. 1: Two-dimensional tool display for category "wheat" of REUTERS-21578 collection. Figure 1a and 1b show the distribution of documents for $\alpha = 1$, $\beta = 1$, Laplace smoothing. Figure 1c and 1d show the distribution of documents for a different setting of the parameters. The red triangles $\triangle$ are the documents of the category to classify, the blue diamonds $\bigtriangledown$ are all the other documents of the collection. The decision frontier is drawn in green.

		Reuters-21578			20-newsgroups			Ohsumed		
Average		Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Macro	(la)	0.341	0.418	0.350	0.047	0.269	0.076	0.105	0.531	0.138
	(je)	0.745	0.669	0.701	**0.749**	0.731	**0.727**	**0.636**	0.549	**0.577**
	(ey)	0.751	0.542	0.622	0.707	0.612	0.610	0.475	**0.557**	0.494
	(vi)	**0.798**	**0.717***	**0.749***	0.708	**0.755**	0.723	0.480	0.550	0.500
micro	(la)	0.672	0.661	0.666	0.047	0.292	0.047	0.235	**0.699**	0.351
	(je)	0.857	0.785	0.820	**0.752**	0.717	**0.736**	**0.659**	0.553	**0.601**
	(ey)	0.869	0.644	0.740	0.713	0.517	0.600	0.578	0.581	0.579
	(vi)	**0.879**	**0.841**	**0.860**	0.715	**0.755**	0.734	0.579	0.581	0.580

Table 1: Comparison of micro and macro average Precision, Recall and F1 measure for three of the four smoothing methods tested on the considered collections. The best performance is highlighted in bold. The star denotes a statistical significant improvement of the measure according to the Wilcoxon test applied to the vectors of scores on each category with the alpha value of 5%

4 Experiments

We tested Jelinek-Mercer (je) against three different parametrization of the beta prior: (la) a uniform (Laplace smoothing) beta prior, $beta(\theta; 1, 1)$; (ey) a beta distribution was set as found by Eyheramendi et al. [9], $beta(\theta; 0.1, 0.3)$; (vi) a beta distribution with optimal parameters α^* and β^*, $beta(\theta; \alpha^*, \beta^*)$. We found λ^* and α^*, β^* for each category by optimizing the F1-score (F1) on the training set: the Jelinek-Mercer λ^* was selected by iterative searching over the interval $[0, 1]$; for α^*, β^*, we exploited the document visualization technique. As overall quality measures, we used standard ATC micro- and macro-averged Recall, Precision, and F1 measures [1].

We selected three of the most widely used collections in literature. We tested Reuters-21578 using the 10 most frequent categories following the "ModAptè" split (9,603 training and 3,299 test documents); 20 Newsgroups, 20 categories with 18,846 stories, divided in 60%-40% training/test; Ohsumed, 6,286 training and 7,643 for test documents classified into 23 Medical Subject Headings (MeSH). These subsets of the collections were chosen accordingly to most of the literature in Automated Text Categorization (ATC) [3, 1, 10]. Default English stopwords were removed and all letters have been converted to lowercase. The two-dimensional interface was implemented in Java using Java Swing technologies.

The baseline obtained by (la) performed statistically worse than any other approach upon the considered datasets: Church and Gale presented strong arguments against the effectiveness of *add-one* smoothing for language data in [11]. As we started the visual search from the parameters set by (ey), (vi) cannot be worse than (ey). Nevertheless, since the parameters found with (vi) were optimized by monitoring the F1 measure, it may happen that with a higher F1, either the value of Recall or Precision are less than (ey). The averaged results on the three datasets are reported in Table 1.

Visual parameter optimization significantly improves categorization performances over the three methods in REUTERS. Fig. 1 illustrates how visual optimization operates for the category "wheat". Applying the same amount of smoothing to each term reveals to be effective in this collection: almost all categories are well represented and using the collection language model as an evidence source for smoothing is not of much interest. Nevertheless, by taking a closer look to performances on each category (not reported in this paper), we found indeed that JM performs best on difficult categories (SHIP, WHEAT). This tendency is clearly emerging on the other two collections. On 20 NEWSGROUPS, visual optimization greatly increases Precision performances over static (EY) parameters. Despite this fact, Beta prior smoothing with optimal parameters reaches the same expressiveness as Jelinek-Mercer (JM) smoothing. On the OHSUMED collection, visual optimization confirms that Beta prior smoothing is lacking expressiveness for this dataset. Computing the mean and the variance of the optimal λ^* parameter found for each category we obtained $\mu_{\lambda^*} = 0.82$, $\sigma^2_{\lambda^*} = 0.02$ thus confirming that taking evidence at a collection level is relevant when dealing with noisy documents and semantically overlapping categories.

5 Conclusions

In this preliminary work, we have studied the effects of smoothing methods for the NB classifier by means of visualization analysis. In the initial phase of this research, we have focused our analysis on the simplest NB model: the multi-variate Bernoulli model. We put forward the following research question: how far can we improve the performance of the Bernoulli classifier by setting optimal Beta prior smoothing parameters? The objective of our experimental evaluation was to compare three well-established smoothing methods against a manual optimization of the Beta parameters (which govern the smoothing of the probabilities) by means of the two-dimensional visual approach.

Experiments have shown that it is possible to find hyper-parameters of the Beta prior that improve the classification significantly. However, in this first set of experiments we limited ourselves to the use the same α and β for each term. A natural continuation of this research will be to find an automatic way to estimate different α and β parameters for each term and to understand if this actually improves performance measures. This problem will consist in characterizing the first and second order moment of each Beta prior distribution based on some relevant empirical evidence of term occurrence in the collection.

This initial set of experiments will lead to the second phase of the study: the analysis of the smoothing methods for the multinomial NB model. Most research so far has shown that the multinomial Naïve Bayes generally outperforms the Bernoulli classifier both in text categorization and information retrieval. From a probabilistic point of view, the Bernoulli model makes a weaker independence assumption on word occurences. This is why we believe that a more systematic comparison between these model is still needed [5]. Another thread of research will be to apply the visualization analysis to more complex NB models, such

as the Chain Augmented NB models (also known as CAN models) which allow a straightforward the application of sophisticated smoothing techniques from statistical language modeling [10].

Acknowledgments. This work has been partially supported by the QONTEXT project under grant agreement N. 247590 (FP7/2007-2013).

References

1. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34** (2002) 1–47
2. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. **22** (2004) 179–214
3. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. Volume 752. (1998) 41–48
4. Metzler, D., Lavrenko, V., Croft, W.B.: Formal multiple-bernoulli models for language modeling. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. (2004) 540–541
5. Zhai, C.: Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies **1** (2008) 1–141
6. Di Nunzio, G.: Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. Int. J. Approx. Reasoning **50** (2009) 945–956
7. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science). 2 edn. Chapman and Hall/CRC (2003)
8. Smucker, M.D., Allan, J.: An investigation of dirichlet prior smoothing's performance advantage. Technical Report Technical Report IR-391, The University of Massachusetts, The Center for Intelligent Information Retrieval (2005)
9. Eyheramendy, S., Lewis, D.D., Madigan, D.: On the Naive Bayes Model for Text Categorization. In Bishop, C., Frey, B., eds.: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. (2003)
10. Peng, F., Schuurmans, D., Wang, S.: Augmenting naive bayes classifiers with statistical language models. Inf. Retr. **7** (2004) 317–345
11. Gale, W.A., Church, K.W.: What's wrong with adding one? In: Corpus-Based Research into Language. Rodolpi. (1994)