Statistical Models for Frequent Terms in Text

Edoardo M. Airoldi, William W. Cohen, Stephen E. Fienberg

May 2004
CMU-CALD-04-106

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

In this paper we present statistical models for text which treat words with higher frequencies of occurrence in a sensible manner, and perform better than widely used models based on the multinomial distribution on a wide range of classification tasks, with two or more classes. Our models axe based on the Poisson and Negative-Binomial distributions, which keep desirable properties of simplicity and analytic tractability.

Carnegie MeJlon University
Pittsburgh, PA 15213-3890

Keywords: Bayesian Models, Multinomial, Binomial, Poisson, Negative-Binomial

Contents

1 Introduction

Almost 50 years ago, Herbert Simon (1955) argued that, "as a text progresses, it creates a meaningful context within which words that have been used already are more likely to appear than others." Such recurrences are not independent as the usual multivariate Bernoulli and multinomial models (e.g., naive Bayes) assume; they are captured by *contagious distributions,* one of which is the Negative-Binomial. In the same spirit Frederick Mosteller and David Wallace (1964, 1984) modeled the choices by the authors of the Federalist Papers about non-contextual words as expressive of their own personal writing styles. More recently, Church and Gale (1995) showed that Poisson mixtures fit the data better than standard Poissons, producing more accurate estimates of the variability of word occurrences over documents, entropy, inverse document frequency, and adaptation.

Contagious distributions naturally fit into language models; statistical assumptions underlying their genesis introduce a notion of context into the probability distributions for the occurrence of words, and they tend to better fit the observed word counts. From a practical perspective, the cross-validated classification accuracy achieved by multivariate Bernoulli or multinomial models can be improved on many problems, both standard and novel, by simply using contagious distributions as we show in this paper. Further, the naive Bayes approach can be seen as a combination of independent models for the frequencies of different words, and a natural extension of it would be to use a mixture of models, depending on observed characteristics of the data at hand, like average occurrence and variability of the counts of words. We believe these findings can be easily extended to LDA, or author-topic type of models using the general model presented in Erosheva, Fienberg, and Lafferty (2004), by simply plugging in these more realistic distributions and then updating the formulas—with some necessary approximations.

2 Related Work

The naive Bayes approach is usually associated with multivariate Bernoulli and multinomial models, but it is in fact more general. It consists of a simple application of Bayes' theorem to solve a classification problem: given a generative model, *p(x\0{),* for the presence (or the frequency) of a word, x, in each class ż, possibly given certain relevant parameters, 0$, and a prior probability distribution, *p(i),* over the classes, *i* G /, then Bayes' theorem prescribes the predicted class to be:

$$\hat{i} = \arg\max_{i \in I} \frac{p(x|0i)p(i)}{\sum_i p(x|\theta_i)\, p(i)}.$$

This approach is "naive" in the sense that different words are considered to be pairwise independent, and that the model, *p(x\0i),* is not specific as to the position of the words in the text (Mitchell 1997). Domingos and Pazzani (1997) give a more complete characterization of naive Bayes types of models, and study conditions for their optimality from a decision theoretic perspective.

Over the years, much research has been focused on Bernoulli and multinomial models for word counts, possibly because they are easy to use and allow for closed formula solutions in many cases, thanks to their strong mathematical relationship, and thanks to the multinomial's conjugate prior density; the Dirichlet distribution. The naive Bayes multinomial and Bernoulli models are often regarded as the baseline models to outperform, when testing a new algorithm.

Several works focused on the analysis of the limitations of naive Bayes Bernoulli and multinomial models (McCallum and Nigam 1998, Rennie et al. 2003); in particular the assumption that

occurrences of a same word *happen* independently of one another has seriously been challenged and strong evidence, both theoretical and empirical, has been produced against it in extensive studies of text and speech data. Ad hoc models have been proposed to go beyond the independence assumption (Church 1995, Church and Gale 1995, Beefennan et al. 1997, Teevan and Karger 2003).

Recently, a number of extensions have been proposed to the naive Bayes approach, which prescribe hierarchical (graphical) models, with both observed and hidden variables, in order to describe, cluster, and classify documents (Minka and Lafferty 2002, Blei et al. 2003, Griffiths and Steyvers 2004, Erosheva et al. 2004). In order to perform inferences in these models, the constants underlying the distributions of the variables in the top layer of the hierarchy have to be fixed. The empirical Bayes approach (Efron and Morris, 1972) is used here, often in combination with methods to approximate certain intractable (marginal) distributions, for example, MCMC (Robert and Casella 1999), variational methods (Jordan et al. 1999), and expectation propagation (Minka 2001). Again, the main ingredients of such extensions are the multinomial model for word counts and its conjugate Dirichlet prior.

3 Contagious Distributions for Text

Our data consists of the numbers of times words appeared in a set of texts. Specifically, for each of / categories (t' = 1,..., /) we have a collection of texts *(j* = 1,..., J»), and we represent each as a bag of words (a random vector *Xij := [XUJ, X2ij, ..., Xvij*]), where the words indexed by *n* = 1,..., *V* belong to a certain vocabulary. In the following discussion we denote the observed word counts, instances of the corresponding random numbers, with lowercase x's. We will also generally assume that *1 = 2.*

3.1 The Naïve Likelihood

In our experiments we considered both Poisson and Negative-Binomial models for the word counts. According to the naïve Bayes assumptions, documents are independent of one another, and words within a document are independent of one another. The likelihood of the set of all texts, denoted as ^({^ti}|{^m})j is written as

$$*(\{*v\}\backslash\{0m\}) = \prod_{n=1}^{v} \prod_{i=1}^{2} \prod_{j=1}^{Ji} \wedge\text{-}i\,i\wedge) > \qquad c^{3}\text{-}^{1})$$

where $\{0_n i\}$ denotes the entire set of parameters for $p(x_n ij \backslash 0_n i)$. Below, we briefly discuss three alternative models for word counts.

3.2 Baseline Model

The model that is most often associated with the naive Bayes classifier prescribes that, in general,

$$p\{Xnij\backslash 0_{ni}) \quad \text{OC} \quad (\dot{0}_{ni})^{X}\wedge,$$

and the model can be formalized in terms of Bernoulli or Binomial models. The Bernoulli random variable has one parameter, the mean (p), but has support on [0,1]; that is, it can be used to assign probabilities to presence or absence of words only. A Binomial random variable has support

on [0,7V] and can be used to assign probabilities to word counts, but it needs two parameters; the probability of occurrence *(p)* and the maximum number of occurrences *(N).*[1] The Binomial distribution arises under the assumptions that: (1) a same word occurs at each position in the text with a fixed probability; and (2) its occurrences are independent of one another.

3.3 The Poisson Model

If a random variable *X* has a Poisson distribution with expected value *E(X) = 0,* then

$$Pr(X = x\backslash 0) = \frac{e^{-0} 0^{x}}{x!}, \quad z \geq 0. \tag{3.2}$$

Consider a Binomial distribution with mean *E(X) = Np* and support [0, *N].* The Poisson distribution can be seen as a limit of this distribution for *N* —> oo, keeping the mean constant at *Np = 0.* The Poisson has only one parameter, the rate of occurrence *0,* instead of the two parameters *p, N* of the Binomial distribution, a convenient fact since choosing *N* in the Binomial scheme is non-trivial (recall that *Pr(X = N\p,N)>0).*

For text data, using the Poisson model implicitly assumes that words or terms occur randomly and independently, but with some mean frequency. Stated differently, suppose the usage of word each word *w* is modeled as random variable *T* denoting the expected 'time till usage" of *w.* The Poisson distribution gives a particular form for the density of T, since one may interpret a Poisson distribution with parameter j^y as the probability of *w* being used *x* times in a time interval of length r. More discussion is given in Johnson, Kotz, and Kemp (1992).

For our analysis, we rewrite *0* = u/i, where *u* is the size of a document in thousands of words and *fi* is the rate of occurrence of a word per thousand words, so that

$$Pois(x_{nij}|\omega_{ij}\mu_{ni}) = \frac{(\omega_{ij}\mu_{ni})^{x_{nij}} e^{-\omega_{ij}\mu_{ni}}}{x_{nij}!}.$$

The maximum likelihood estimator for $fi_n i$ is

$$\hat{\mu}_{ni} = \frac{\sum_j x_{nij}}{\sum_j \omega_{ij}},$$

which takes into account the variable length of the texts (ω_{ij}).

3.4 The Negative-Binomial Model

The Negative-Binomial model is as follows:

$$Neg\text{-}Bin(x_{nij}|\omega_{ij}\mu_{ni}, \kappa_{ni}, \omega_{ij}\delta_{ni}) =$$

$$= \frac{\Gamma(x_{nij}+\kappa_{ni})}{x_{nij}!\Gamma(\kappa_n 0} (\omega_{ij}\delta_{ni})^{x_{nij}} (1 + \omega_{ij}\delta_{ni})^{-(x_{nij}+\kappa_{ni})},$$

for $x \geq 0$, and such that u^ > 0, /x^ > 0, k™ > 0, S_{ni} > 0, and ^5^ = /x^, for any set of indices. We index the parameters consistently, so that: /i$_{m}$ is the *normalized* rate of occurrence

[1] **Also note that *Pr(X = N* |p, *N)* > 0, so that choosing *N* as the total number of words in the document is an extremely unrealistic assumption!**

for the n^{th} word in the i^{th} class, that is, the number of such words we would expect to see in any thousand consecutive words of text; *u>ij* is the word-length of a document expressed in thousands of words as above; $6_n\{$ is the *non-Poissonness* rate, that is, a parameter that controls how far the Negative-Binomial distribution is from its corresponding Poisson limit; and *Kni* := f^{21}_{ni} is a redundant parameter useful for some derivations.

Note that if <Jm = 0, the Negative-Binomial distribution with becomes the Poisson distribution. The extra parameter allows us to model heavy tails, relative to the Poisson.

Two prior studies of authorship attribution, Mosteller and Wallace (1964, 1984) and Airoldi et al. (2004) used this parameterization for the Negative-Binomial (in terms of (/im> ^m, *ux)) and observed that values of *6* were relatively stable across words and authors; for most words, S_{ni} € [0, 0.75].

	Poisson Model		Negative-Binomial Model	
Pool of words	Reagan— (38 texts)	Reagan+ (75 texts)	Reagan— (38 texts)	Reagan4- (75 texts)
50 highest frequency words	12 (50)	3(50)	31 (50)	49 (50)
21 semantic features	3(21)	1(21)	21 (21)	20 (21)
27 words by information gain	0(7)	0(8)	7(7)	8(8)

Table 1: Goodness of fit of Poisson and Negative-Binomial models for various pools of words. The pools are selected from positive (written by Reagan) and negative (written by Hannaford) examples of Reagan's radio addresses. Unbracketed counts are predicted number of words; in brackets we give the actual number of words. Predictions were made using p-values from a two-sample Kolmogorov Smirnov test. Low-frequency words (less than 8 per 10,000 words) were discarded. Source: Airoldi et al. (2004).

Following Mosteller and Wallace (1964, 1984) we used method of moment estimators that take into account the different lengths of the texts *(wij)*, and are "optimal" at the Poisson limit:

$$\begin{cases} \hat{\mu}_{ni} &= m_{ni}, \\ \hat{\delta}_{ni} &= d_{ni} = \max\left\{0, \frac{v_{ni}-m_{ni}}{m_{ni}r_i}\right\}, \end{cases}$$

where,

$$\begin{aligned} rrini &= \frac{\sum_j x_{nij}}{\sum_j \omega_{ij}}, \\ v_{ni} &= \frac{1}{J_i-1}\sum_j \omega_{ij}\left(\frac{x_{nij}}{\omega_{ij}} - m_{ni}\right)^2, \\ r_i &= \frac{1}{J_i-1}\left(\sum_j \omega_{ij} - \frac{\sum_j \omega_{ij}^2}{\sum_j \omega_{ij}}\right). \end{aligned}$$

The Negative-Binomial model often captures the variability in observed word-frequency data better than the Poisson model. In Table 1 we show some examples from Airoldi et al. (2004). The results in the table demonstrate that the flexibility we gain by introducing two parameters allows us to capture the way words with different frequency/topicality profiles occur in the texts.

Dataset	Categories	Examples	Features	Binomial	Poisson	**Neg-Bin**
20 Newsgroups	5	5000	43486	4.680%	**3.440%**	**3.400%**
Reuters-21578	3	11367	30765	13.830%	**7.346%**	**8.164%**
Fraud detection	3	3591	181307	0.501%	**0.390%**	**0.201%**
Opinions: Finance	3	600	11220	32.500%	**32.500%**	**30.833%**
Opinions: Mixed	3	600	15685	29.667%	**29.500%**	**28.667%**
Spam-Assassin Corpus	3	3302	118175	3.634%	**3.604%**	**3.523%**
Web-Master	3	582	1406	23.883%	**22.852%**	**22.337%**
Reagan's Addresses	2	748	19243	8.970%	**8.290%**	**7.716%**
Movie Reviews	2	1400	34944	**31.714%**	**24.571%**	**23.857%**
Prostate Cancer	2	326	779	11.656%	**12.270%**	**11.043%**
Dealtime	2	32349	230	12.118%	**9.840%**	**9.830%**

Table 2: Estimated prediction errors using five-fold cross-validation.

4 Experiments

4.1 Datasets and Methods

We explored the performances of the Poisson and Negative-Binomial models relative to the baseline on ten different problems. In Table 2 we describe our data in terms of number of documents, number of classes, and number of features. Features were selected according to information gain in these experiments.

In the "naïve Bayes" model of Equation 3.1, it is possible to use different models $p(x_n ij \backslash 0_n i)$ for different words *n*. If we use relatively powerful generative models like the Negative-Binomial and Poisson feature selection may help us to reduce over-fitting.

In the experiments described below, we used the following simple "back-off" scheme to select a model for each word *n*. If the word counts are always zero or one in the training data, we used a Bernoulli model. If counts were ever greater than one, we computed the mean and variance of $x_{ni}j$ on the training data, and if the mean was less than the variance (i.e., the sample is under dispersed) we used the Poisson model. Otherwise, we used the Negative-Binomial model.

Below is a short summary of the 11 datasets we analyzed along with the relevant tasks. Nine of them are text, with word-frequencies as features. The remaining two are non-textual datasets with discrete-valued features.

- *20 Newsgroups:*

  The 20 newsgroups data is described in Mitchell (1997). Here we want to classify newsgroups posts according to their topic.

- *Reuters-21578:*

  We considered a different classification task than Lewis (1998) and others. We consider three broader taxonomy nodes, namely *Money, Crops,* and *Natural Resources.* We adopted such *high level* categories in the belief that weakly topical words, with possibly medium to high frequencies, would be more informative.

- *Fraud detection:*

  The task is that of discriminating between messages which contain patterns of fraudulent intent for the type of scam and other e-mail, namely,* regular e-mail and spam. The scam corpus consists of 534 distinct messages posted to the Nigerian Fraud E-mail Gallery from April 2000 to April 2004 . Each message was previously been classified as the Nigerian 4-1-9 scam by the proprietor of the website. See Airoldi and Malin (2004).

- *Opinion Extraction:*

  The task is to categorize the main opinion expressed in online news articles as *Positive, Neutral,* or *Negative.* The ground truth consists of labels assigned by three independent reviewers, courtesy of Infonic.com. Bai et al. (2005) provide further details.

- *Spam Assassin:*

  The task here is to classify email as *Easy Ham, Hard Ham,* and *Spam* in the SpamAssassin corpus, available online at http://www.spamassassin.org/.

- *Web-Master:*

  The task is to classify web site update requests into three categories; *Add, Change,* or *Delete.* The corpus we use was introduced and analyzed in Cohen et al. (2004b).

- *Ronald Reagan's Radio Addresses:*

  The problem is that of attributing authorship to text for Ronald Reagan's 1975-1979 radio addresses of unknown authorship. The electronic texts of these addresses belong to a private collection, recently analyzed and discussed in Airoldi et al. (2004).

- *Movie Reviews:*

  The data consists of reviews from IMDB, collected and analyzed in Pang et al. (2002). The task is to associate a positive or negative sentiment with movie reviews.

- *Prostate Cancer:*

  The task is to classify whether a patient has prostate cancer given outcomes of different tests. Test outcomes are real-valued, and were discretized in a sensible manner.

- *Dealtime:*

  The task is to classify whether a customer will buy a certain book or not given integer-valued features that encode his or her browsing history, as captured by Dealtime sessions.

4.2 Discussion

Using the results of the experiments above we performed sign tests to weight the evidence in support the following claims:

- the Poisson model dominates the Bernoulli model (p-value ~ 0.0117),
- the Negative-Binomial model dominates the Bernoulli model (p-value ~ 0.000977), and
- the Negative-Binomial model dominates the Poisson model (p-value ~ 0.0117).

The sign test shows that the improvements are consistent, but does not indicate their magnitude. In fact, the magnitude of improvement varies widely from dataset to dataset. Often it is small; however, in some cases, the errors rate are reduced by a substantial factor over the simpler multinomial model (e.g., Dealtime, Movie Reviews, Nigerian Scam, 20 Newsgroups).

A consideration of the underlying geometry of these models sheds some light on why this happens; there is an increasing complexity as we go from the multinomial to the Poisson and then from the Poisson to the Negative-Binomial. In fact, the multinomial classifier can be represented by one hyperplane in the feature space, the Poisson classifier by two, and the Negative-Binomial by three hyperplanes, thus allowing for more complex partitions.

5 Concluding Remarks

We have described a simple, principled extension to the widely-used multinomial model for text, which allows better modelling of frequent words. The extension is based on replacing the widely-used multinomial distribution with a simple "contagious" distribution, by allowing word frequencies to be modelled with distributions that do not assume independence of different occurrences of the same word in a document.

A sign test over a collection of eleven datasets shows that the model generally leads to better classification accuracy, and sometimes leads to substantially better classification accuracy. Our experiments have been with simple "naive Bayes" classification models; however, an important advantage of the proposed extension is that is easy to integrate with more complex models of text, for instance, mixtures of multinomials.

In the current paper, we used maximum likelihood estimators for each word-frequency model. In future work, we also hope to develop tractable non-informative priors for the models, so that the extension can be used in settings for which a fully Bayesian or empirical Bayesian approache is appropriate.

Acknowledgements

The authors wish to thank Dr. Latanya Sweeney, and Dr. Tom Mitchell in the School of Computer Science at Carnegie Mellon University for helpful discussions and for their support throughout this project.

References

[1] E. M. Airoldi, A. G. Anderson, S. E. Fienberg, and K. K. Skinner (2005). Who wrote Ronald Reagan radio addresses? *Journal of Bayesian Analysis. Forthcoming.*

[2] E. M. Airoldi, and B. Malin (2004). Data mining challenges for electronic safety: the case of fraudulent intent detection in e-mails. In *Workshop on Privacy and Security Aspects of Data Mining* Brighton, England: IEEE Computer Society.

[3] X. Bai, R. Padman, and E. M. Airoldi (2005). On learning parsimonious models for extracting consumer opinions. In *Hawaii International Conference on System Sciences* Hawaii: IEEE Computer Society. *Forthcoming.*

[4] D. Beeferman, A. Berger, and J. Lafferty (1997). A model of lexical attraction and repulsion. In P. R. Cohen and W. Wahlster (eds.) *Annual Meeting of the Association for Computational Linguistics* Madrid, Spain: ACL.

[5] D. Blei, A. Ng, and M. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** 993-1022.

[6] K. Church (1995). One term or two? In E. Fox, P. Ingwersen and R. Fidel (eds.) *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Seattle: ACM Press.

[7] K. Church, and W. Gale (1995). Poisson mixtures. *Natural Language Engineering* **1** (2):163-190.

[8] W. W. Cohen (2004). Minorthird: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. http://minorthird.sourceforge.net

[9] W. W. Cohen, E. Minkov, and A. Tomasic (2004b). Learning to understand web site update requests. *Manuscript.*

[10] P. Domingos, and M. Pazzani (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning* **29** 103-130.

[11] B. Efron, and C. Morris (1972). Limiting the risk of Bayes and empirical Bayes estimators-Part II: the empirical Bayes case. *Journal of the American Statistical Association* **67** 130-139.

[12] E. A. Erosheva, and S. E. Fienberg (2004). Bayesian mixed-membership models for soft clustering and classification. *Manuscript.*

[13] E. A. Erosheva, S. E. Fienberg, and J. Lafferty (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* **97** (22):11885-11892.

[14] T. L. Griffths, and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* **101** (1):5228-5235.

[15] N. L. Johnson, S. Kotz, and A. W. Kemp (1992). *Univariate Discrete Distributions.* John Wiley.

[16] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning* **37** (2):183-233.

[17] D. Lewis (1998). The Reuters-21578 text categorization test collection. Available online at http://www.daviddlewis.com/resources/testcollections

[18] A. McCallum, and K. Nigam (1998). A comparison of event models for naïve Bayes text classification. *AAAI Workshop on Learning for Text Categorization.*

[19] T. Minka, and J. Lafferty (2002). Expectation-propogation for the generative aspect model. *Annual Conference on Uncertainty in Artificial Intelligence* San Francisco, CA: Morgan Kaufmann.

[20] T. Minka (2001). A family of algorithms for approximate Bayesian inference. *PhD Thesis*, MIT.

[21] T. Mitchell (1997). *Machine Learning.* Me Grow-Hill.

[22] F. Mosteller, and D. L. Wallace (1964). *Inference and Disputed Authorship: The Federalist* Addison-Wesley.

[23] F. Mosteller, and D. L. Wallace (1984). *Applied Bayesian and Classical Inference: The Case of "The Federalist" Papers.* Springer-Verlag.

[24] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell (2000). Text Classification from labeled and unlabeled documents using EM. *Machine Learning* 39 (2-3):103-134.

[25] B. Pang, L. Lee, and S. Vaithyanathan (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Conference on Empirical Methods in Natural Language Processing*

[26] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger (2003). Tackling the poor assumptions of naive Bayes text classifiers. In T. Fawcett and N. Mishra (eds.), *International Conference on Machine Learning* Washington D.C.: Morgan Kaufmann.

[27] C. Robert, and G. Casella (1999). *Monte Carlo Statistical Methods.* Springer-Verlag.

[28] H. A. Simon (1955). On a class of skew distribution functions. *Biometrika* 42 425-440.

[29] J. Teevan, and D. Karger (2003). Empirical development of an exponential probabilistic model for text retrieval: using textual analysis to build a better model. *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Toronto, Canada: ACM Press.