Parametric Models of Linguistic Count Data

Martin Jansche
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
jansche@acm.org

Abstract

It is well known that occurrence counts of words in documents are often modeled poorly by standard distributions like the binomial or Poisson. Observed counts vary more than simple models predict, prompting the use of overdispersed models like Gamma-Poisson or Beta-binomial mixtures as robust alternatives. Another deficiency of standard models is due to the fact that most words never occur in a given document, resulting in large amounts of zero counts. We propose using zero-inflated models for dealing with this, and evaluate competing models on a Naive Bayes text classification task. Simple zero-inflated models can account for practically relevant variation, and can be easier to work with than overdispersed models.

1 Introduction

Linguistic count data often violate the simplistic assumptions of standard probability models like the binomial or Poisson distribution. In particular, the inadequacy of the Poisson distribution for modeling word (token) frequency is well known, and robust alternatives have been proposed (Mosteller and Wallace, 1984; Church and Gale, 1995). In the case of the Poisson, a commonly used robust alternative is the negative binomial distribution (Pawitan, 2001, §4.5), which has the ability to capture extra-Poisson variation in the data, in other words, it is *overdispersed* compared with the Poisson. When a small set of parameters controls all properties of the distribution it is important to have enough parameters to model the relevant aspects of one's data. Simple models like the Poisson or binomial do not have enough parameters for many realistic applications, and we suspect that the same might be true of log-linear models. When applying robust models like the negative binomial to linguistic count data like word occurrences in documents, it is natural to ask to what extent the extra-Poisson variation has been captured by the model. Answering that question is our main goal, and we begin by reviewing some of the classic results of Mosteller and Wallace (1984).

2 Word Frequency in Fixed-Length Texts

In preparation of their authorship study of *The Federalist*, Mosteller and Wallace (1984, §2.3) investigated the variation of word frequency across contiguous passages of similar length, drawn from papers of known authorship. The occurrence frequencies of *any* in papers by Hamilton (*op. cit.*, Table 2.3–3) are repeated here in Figure 1: out of a total of 247 passages there are 125 in which the word *any* does not occur; it occurs once in 88 passages, twice in 26 passages, etc. Figure 1 also shows the counts predicted by a Poisson distribution with mean 0.67. Visual inspection ("chi by eye") indicates an acceptable fit between the model and the data, which is confirmed by a χ^2 goodness-of-fit test. This demonstrates that certain words seem to be adequately modeled by a Poisson distribution, whose probability mass function is shown in (1):

$$\text{Poisson}(\lambda)(x) = \frac{\lambda^x}{x!}\,\frac{1}{\exp\lambda} \qquad (1)$$

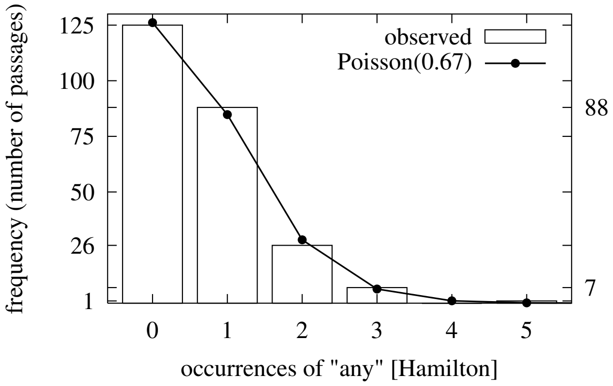

Figure 1: Occurrence counts of *any* in Hamilton passages: raw counts and counts predicted under a Poisson model.

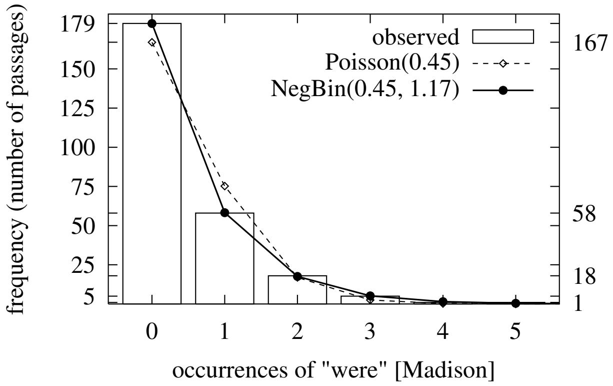

Figure 2: Occurrence counts of *were* in Madison passages: raw counts and counts predicted under Poisson and negative binomial models.

For other words the Poisson distribution gives a much worse fit. Take the occurrences of *were* in papers by Madison, as shown in Figure 2 (*ibid.*). We calculate the χ^2 statistic for the counts expected under a Poisson model for three bins (0, 1, and 2–5, to ensure that the expected counts are greater than 5) and obtain 6.17 at one degree of freedom (number of bins minus number of parameters minus one), which is enough to reject the null hypothesis that the data arose from a $\text{Poisson}(0.45)$ distribution. On the other hand, the χ^2 statistic for a negative binomial distribution $\text{NegBin}(0.45, 1.17)$ is only 0.013 for four bins (0, 1, 2, and 3–5), i. e., again 1 degree of freedom, as two parameters were estimated from the data. Now we are very far from rejecting the null hypothesis. This provides some quantitative backing for Mosteller and Wallace's statement that 'even the most motherly eye can scarcely make twins of the [Poisson vs. empirical] distributions' for certain words (*op. cit.*, 31).

The probability mass function of the negative binomial distribution, using Mosteller and Wallace's parameterization, is shown in (2):

$$\text{NegBin}(\lambda, \kappa)(x) = \frac{\lambda^x}{x!} \frac{\Gamma(\kappa+x)}{(\lambda+\kappa)^{\kappa+x}} \frac{\kappa^\kappa}{\Gamma(\kappa)} \tag{2}$$

If one recalls that the Gamma function is well behaved and that

$$\exp \lambda = \lim_{\kappa \to \infty} \left(1 + \frac{\lambda}{\kappa}\right)^\kappa = \lim_{\kappa \to \infty} \frac{(\lambda+\kappa)^\kappa}{\kappa^\kappa},$$

it is easy to see that $\text{NegBin}(\lambda, \kappa)$ converges to $\text{Poisson}(\lambda)$ for λ constant and $\kappa \to \infty$. On the other hand, small values of κ drag the mode of the negative binomial distribution towards zero and increase its variance, compared with the Poisson.

As more and more probability mass is concentrated at 0, the negative binomial distribution starts to depart from the empirical distribution. One can already see this tendency in Mosteller and Wallace's data, although they themselves never comment on it. The problem with a huge chunk of the probability mass at 0 is that one is forced to say that the outcome 1 is still fairly likely and that the probability should drop rapidly from 2 onwards as the term $1/x!$ starts to exert its influence. This is often at odds with actual data.

Take the word *his* in papers by Hamilton and Madison (*ibid.*, pooled from individual sections of Table 2.3–3). It is intuitively clear that *his* may not occur at all in texts that deal with certain aspects of the US Constitution, since many aspects of constitutional law are not concerned with any single (male) person. For example, Federalist No. 23 (*The Necessity of a Government as Energetic as the One Proposed to the Preservation of the Union*, approx. 1800 words, by Hamilton) does not contain a single occurrence of *his*, whereas Federalist No. 72 (approx. 2000 words, a continuation of No. 71 *The Duration in Office of the Executive*, also by Hamilton) contains 35 occurrences. The difference is that No. 23 is about the role of a federal government in the abstract, and Nos. 71/72 are about term limits for offices filled by (male) individuals. We might therefore expect the occurrences of *his* to vary more, de-

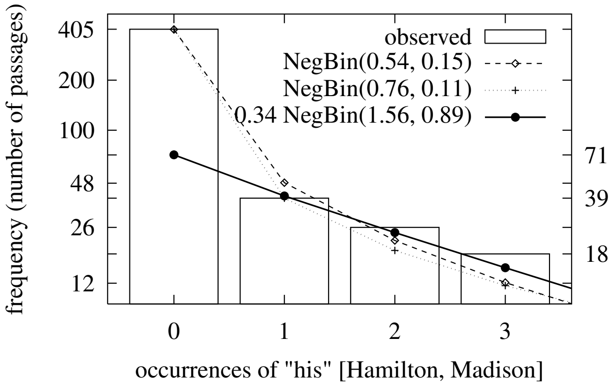

Figure 3: Occurrence counts of *his* in Hamilton and Madison passages (NB: y-axis is logarithmic).

	obsrvd	NegBin expctd	ZINB expctd
0	405	403.853	405.000
1	39	48.333	40.207
2	26	21.686	24.206
3	18	12.108	14.868
4	5	7.424	9.223
5–6	9	8.001	9.361
7–14	7	6.996	5.977
χ^2 statistic		6.447	2.952
df		4	3
χ^2 cumul. prob		0.832	0.601
$-\log L(\hat{\theta})$		441.585	439.596

Table 1: Occurrence counts of *his* in Hamilton and Madison passages.

pending on topic, than *any* or *were*.

The overall distribution of *his* is summarized in Figure 3; full details can be found in Table 1. Observe the huge number of passages with zero occurrences of *his*, which is ten times the number of passages with exactly one occurrence. Also notice how the negative binomial distribution fitted using the Method of Maximum Likelihood (MLE model, first line in Figure 3, third column in Table 1) overshoots at 1, but underestimates the number of passages with 2 and 3 occurrences.

The problem cannot be solved by trying to fit the two parameters of the negative binomial based on the observed counts of two points. The second line in Figure 3 is from a distribution fitted to match the observed counts at 0 and 1. Although it fits those two points perfectly, the overall fit is worse than that of the MLE model, since it underestimates the observed counts at 2 and 3 more heavily.

The solution we propose is illustrated by the third line in Figure 3. It accounts for only about a third of the data, but covers all passages with one or more occurrences of *his*. Visual inspection suggests that it provides a much better fit than the other two models, if we ignore the outcome 0; a quantitative comparison will follow below. This last model has relaxed the relationship between the probability of the outcome 0 and the probabilities of the other outcomes. In particular, we obtain appropriate counts for the outcome 1 by pretending that the outcome 0 occurs only about 71 times, compared with an actual 405 observed occurrences. Recall that the model accounts for only 34% of the data; the remaining counts for the outcome 0 are supplied entirely by a second component whose probability mass is concentrated at zero. The expected counts under the full model are found in the rightmost column of Table 1.

The general recipe for models with large counts for the zero outcome is to construe them as two-component mixtures, where one component is a degenerate distribution whose entire probability mass is assigned to the outcome 0, and the other component is a standard distribution, call it $\mathscr{F}(\theta)$. Such a nonstandard mixture model is sometimes known as a *'modified' distribution* (Johnson and Kotz, 1969, §8.4) or, more perspicuously, as a *zero-inflated distribution*. The probability mass function of a zero-inflated $\mathscr{F}$ distribution is given by equation (3), where $0 \leq z \leq 1$ ($z < 0$ may be allowable subject to additional constraints) and $x \equiv 0$ is the Kronecker delta $\delta_{x,0}$.

$$\mathrm{ZI}\mathscr{F}(z,\theta)(x) = z\,(x \equiv 0) + (1-z)\,\mathscr{F}(\theta)(x) \quad (3)$$

It corresponds to the following generative process: toss a z-biased coin; if it comes up heads, generate 0; if it comes up tails, generate according to $\mathscr{F}(\theta)$. If we apply this to word frequency in documents, what this is saying is, informally: whether a given word appears at all in a document is one thing; how often it appears, if it does, is another thing.

This is reminiscent of Church's statement that '[t]he first mention of a word obviously depends on frequency, but surprisingly, the second does not.' (Church, 2000) However, Church was concerned with language modeling, and in particular cache-based models that overcome some of the limitations introduced by a Markov assumption. In such a setting it is natural to make a distinction between the first occurrence of a word and subsequent occurrences, which according to Church are influenced by *adaptation* (Church and Gale, 1995), referring to an increase in a word's chance of re-occurrence after it has been spotted for the first time. For empirically demonstrating the effects of adaptation, Church (2000) worked with nonparametric methods. By contrast, our focus is on parametric methods, and unlike in language modeling, we are also interested in words that fail to occur in a document, so it is natural for us to distinguish between zero and nonzero occurrences.

In Table 1, ZINB refers to the zero-inflated negative binomial distribution, which takes a parameter z in addition to the two parameters of its negative binomial component. Since the negative binomial itself can already accommodate large fractions of the probability mass at 0, we must ask whether the ZINB model fits the data better than a simple negative binomial. The bottom row of Table 1 shows the negative log likelihood of the maximum likelihood estimate $\hat{\theta}$ for each model. Log odds of 2 in favor of ZINB are indeed sufficient (on Akaike's likelihood-based information criterion; see e. g. Pawitan 2001, §13.5) to justify the introduction of the additional parameter. Also note that the cumulative χ^2 probability of the χ^2 statistic at the appropriate degrees of freedom is lower for the zero-inflated distribution.

It is clear that a large amount of the observed variation of word occurrences is due to zero inflation, because virtually all words are rare and many words are simply not "on topic" for a given document. Even a seemingly innocent word like *his* turns out to be "loaded" (and we are not referring to gender issues), since it is not on topic for certain discussions of constitutional law. One can imagine that this effect is even more pronounced for taboo words, proper names, or technical jargon (cf. Church 2000). Our next question is whether the observed variation is best accounted for in terms of zero-inflation or overdispersion. We phrase the discussion in terms of a practical task for which it matters whether a word is on topic for a document.

3 Word Frequency Conditional on Document Length

Word occurrence counts play an important role in document classification under an independent feature model (commonly known as "Naive Bayes"). This is not entirely uncontroversial, as many approaches to document classification use binary indicators for the presence and absence of each word, instead of full-fledged occurrence counts (see Lewis 1998 for an overview). In fact, McCallum and Nigam (1998) claim that for small vocabulary sizes one is generally better off using Bernoulli indicator variables; however, for a sufficiently large vocabulary, classification accuracy is higher if one takes word frequency into account.

Comparing different probability models in terms of their effects on classification under a Naive Bayes assumption is likely to yield very conservative results, since the Naive Bayes classifier can perform accurate classifications under many kinds of adverse conditions and even when highly inaccurate probability estimates are used (Domingos and Pazzani, 1996; Garg and Roth, 2001). On the other hand, an evaluation in terms of document classification has the advantages, compared with language modeling, of computational simplicity and the ability to benefit from information about non-occurrences of words.

Making a direct comparison of overdispersed and zero-inflated models with those used by McCallum and Nigam (1998) is difficult, since McCallum and Nigam use multivariate models – for which the "naive" independence assumption is different (Lewis, 1998) – that are not as easily extended to the cases we are concerned about. For example, the natural overdispersed variant of the multinomial model is the Dirichlet-multinomial mixture, which adds just a single parameter that globally controls the overall variation of the entire vocabulary. However, Church, Gale and other have demonstrated repeatedly (Church and Gale, 1995; Church, 2000) that adaptation or "burstiness" are clearly properties of individual words (word types). Using joint independent models (one model per word) brings us back into the realm of standard independence assumptions, makes it easy to add parameters that control overdispersion and/or zero-inflation for each word individually, and simplifies parameter estimation.

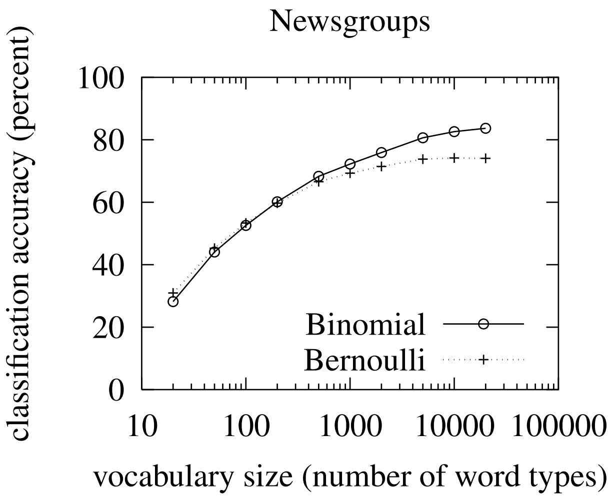

Figure 4: A comparison of event models for different vocabulary sizes on the Newsgroup data set.

So instead of a single multinomial distribution we use independent binomials, and instead of a multivariate Bernoulli model we use independent Bernoulli models for each word. The overall joint model is clearly wrong since it wastes probability mass on events that are known a priori to be impossible, like observing documents for which the sum of the occurrences of each word is greater than the document length. On the other hand, it allows us to take the true document length into account while using only a subset of the vocabulary, whereas on McCallum and Nigam's approach one has to either completely eliminate all out-of-vocabulary words and adjust the document length accordingly, or else map out-of-vocabulary words to an unknown-word token whose observed counts could then easily dominate.

In practice, using joint independent models does not cause problems. We replicated McCallum and Nigam's Newsgroup experiment[1] and did not find any major discrepancies. The reader is encouraged to compare our Figure 4 with McCallum and Nigam's Figure 3. Not only are the accuracy figures comparable, we also obtained the same critical vocabulary size of 200 words below which the Bernoulli model results in higher classification accuracy.

The Newsgroup data set (Lang, 1995) is a stratified sample of approximately 20,000 messages total, drawn from 20 Usenet newsgroups. The fact that 20 newsgroups are represented in equal proportions makes this data set well suited for comparing different classifiers, as class priors are uniform and baseline accuracy is low at 5%. Like McCallum and Nigam (1998) we used (Rain)bow (McCallum, 1996) for tokenization and to obtain the word/document count matrix. Even though we followed McCallum and Nigam's tokenization recipe (skipping message headers, forming words from contiguous alphabetic characters, not using a stemmer), our total vocabulary size of 62,264 does not match McCallum and Nigam's figure of 62,258, but does come reasonably close. Also following McCallum and Nigam (1998) we performed a 4:1 random split into training and test data. The reported results were obtained by training classification models on the training data and evaluating on the unseen test data.

We compared four models of token frequency. Each model is conditional on the document length n (but assumes that the parameters of the distribution do not depend on document length), and is derived from the binomial distribution

$$\mathrm{Binom}(p)(x \mid n) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad (4)$$

which we view as a one-parameter conditional model, our **first** model: x represents the token counts ($0 \le x \le n$); and n is the length of the document measured as the total number of token counts, including out-of-vocabulary items.

The **second** model is the Bernoulli model, which is derived from the binomial distribution by replacing all non-zero counts with 1:

$$\begin{aligned}&\mathrm{Bernoulli}(p)(x \mid n)\\ &\quad = \mathrm{Binom}(p)\left(\left\lceil \frac{x}{x+1} \right\rceil \,\middle|\, \left\lceil \frac{n}{n+1} \right\rceil\right) \qquad (5)\end{aligned}$$

Our **third** model is an overdispersed binomial model, a "natural" continuous mixture of binomials with the integrated binomial likelihood – i. e. the Beta density (6), whose normalizing term involves the Beta function – as the mixing distribution.

$$\mathrm{Beta}(\alpha,\beta)(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \qquad (6)$$

[1] Many of the data sets used by McCallum and Nigam (1998) are available at `http://www.cs.cmu.edu/~TextLearning/datasets.html`.

The resulting mixture model (7) is known as the Pólya–Eggenberger distribution (Johnson and Kotz, 1969) or as the beta-binomial distribution. It has been used for a comparatively small range of NLP applications (Lowe, 1999) and certainly deserves more widespread attention.

$$\begin{aligned}
&\mathrm{BetaBin}(\alpha,\beta)(x \mid n) \\
&\quad = \int_0^1 \mathrm{Binom}(p)(x \mid n)\ \mathrm{Beta}(\alpha,\beta)(p)\ dp \\
&\quad = \binom{n}{x} \frac{\mathrm{B}(x+\alpha, n-x+\beta)}{\mathrm{B}(\alpha,\beta)} \qquad (7)
\end{aligned}$$

As was the case with the negative binomial (which is to the Poisson as the beta-binomial is to the binomial), it is convenient to reparameterize the distribution. We choose a slightly different parameterization than Lowe (1999); we follow Ennis and Bi (1998) and use the identities

$$\begin{aligned}
p &= \alpha/(\alpha+\beta), \\
\gamma &= 1/(\alpha+\beta+1).
\end{aligned}$$

To avoid confusion, we will refer to the distribution parameterized in terms of p and γ as BB:

$$\mathrm{BB}(p,\gamma) = \mathrm{BetaBin}\left(p\frac{1-\gamma}{\gamma},\ (1-p)\frac{1-\gamma}{\gamma}\right) \qquad (8)$$

After reparameterization the expectation and variance are

$$\begin{aligned}
\mathrm{E}[x; \mathrm{BB}(p,\gamma)(x \mid n)] &= n\,p, \\
\mathrm{Var}[x; \mathrm{BB}(p,\gamma)(x \mid n)] &= n\,p\,(1-p)\ (1+(n-1)\,\gamma).
\end{aligned}$$

Comparing this with the expectation and variance of the standard binomial model, it is obvious that the beta-binomial has greater variance when $\gamma > 0$, and for $\gamma = 0$ the beta-binomial distribution coincides with a binomial distribution.

Using the method of moments for estimation is particularly straightforward under this parameterization (Ennis and Bi, 1998). Suppose one sample consists of observing x successes in n trials (x occurrences of the target word in a document of length n), where the number of trials may vary across samples. Now we want to estimate parameters based on a sequence of s samples $\langle x_1, n_1\rangle, \ldots, \langle x_s, n_s\rangle$. We equate sample moments with distribution moments

$$\begin{aligned}
\sum_i n_i\,\hat{p} &= \sum_i x_i, \\
\sum_i n_i\,\hat{p}\,(1-\hat{p})\ (1+(n_i-1)\,\hat{\gamma}) &= \sum_i (x_i - n_i\,\hat{p})^2,
\end{aligned}$$

and solve for the unknown parameters:

$$\hat{p} = \frac{\sum_i x_i}{\sum_i n_i}, \qquad (9)$$

$$\hat{\gamma} = \frac{\sum_i (x_i - n_i\,\hat{p})^2 / (\hat{p}\,(1-\hat{p})) - \sum_i n_i}{\sum_i n_i^2 - \sum_i n_i}. \qquad (10)$$

In our experience, the resulting estimates are sufficiently close to the maximum likelihood estimates, while method-of-moment estimation is much faster than maximum likelihood estimation, which requires gradient-based numerical optimization[2] in this case. Since we estimate parameters for up to 400,000 models (for 20,000 words and 20 classes), we prefer the faster procedure. Note that the maximum likelihood estimates may be suboptimal (Lowe, 1999), but full-fledged Bayesian methods (Lee and Lio, 1997) would require even more computational resources.

The **fourth** and final model is a zero-inflated binomial distribution, which is derived straightforwardly via equation (3):

$$\begin{aligned}
&\mathrm{ZIBinom}(z,p)(x \mid n) \\
&\quad = z\,(x \equiv 0) + (1-z)\,\mathrm{Binom}(p)(x \mid n) \\
&\quad = \begin{cases} z + (1-z)(1-p)^n & \text{if } x = 0 \\ (1-z)\binom{n}{x} p^x\,(1-p)^{n-x} & \text{if } x > 0 \end{cases} \qquad (11)
\end{aligned}$$

Since the one parameter p of a single binomial model can be estimated directly using equation (9), maximum likelihood estimation for the zero-inflated binomial model is straightforward via the EM algorithm for finite mixture models. Figure 5 shows pseudo-code for a single EM update.

Accuracy results of Naive Bayes document classification using each of the four word frequency models are shown in Table 2. One can observe that the differences between the binomial models are small,

[2]Not that there is anything wrong with that. In fact, we calculated the MLE estimates for the negative binomial models using a multidimensional quasi-Newton algorithm.

```
Z ← 0; X ← 0; N ← 0
{E step}
for i ← 1 to s do
  if x_i = 0 then
    ẑ_i ← z/(z + (1 − p)^{n_i})
    Z ← Z + ẑ_i
    X ← X + (1 − ẑ_i) x_i
    N ← X + (1 − ẑ_i) n_i
  else {x_i ≠ 0, ẑ_i = 0}
    X ← X + x_i
    N ← N + n_i
  end if
end for
{M step}
z ← Z/s
p ← X/N
```

Figure 5: Maximum likelihood estimation of ZI-Binom parameters z and p: Pseudo-code for a single EM iteration that updates the two parameters.

	Bernoulli	Binom	ZIBinom	BetaBin
20	30.94	28.19	29.48	29.93
50	45.28	44.04	44.85	45.15
100	53.36	52.57	53.84	54.16
200	59.72	60.15	60.47	61.16
500	66.58	68.30	67.95	68.58
1,000	69.31	72.24	72.46	73.20
2,000	71.45	75.92	76.35	77.03
5,000	73.80	80.64	80.51	80.19
10,000	74.18	82.61	82.58	82.58
20,000	74.05	83.70	83.06	83.06

Table 2: Accuracy of the four models on the Newsgroup data set for different vocabulary sizes.

	Binom ZIBinom	Binom BetaBin	ZIBinom BetaBin
20	✗	✗	
50	✗	✗	
100	✗	✗	
200	✗		
500			
1,000	✗		
2,000	✗		
5,000			
10,000			
20,000		✗	

Table 3: Pairwise McNemar test results. A ✗ indicates a significant difference of the classification results when comparing a pair of of models.

but even small effects can be significant on a test set of about 4,000 messages. More importantly, note that the beta-binomial and zero-inflated binomial models outperform both the simple binomial and the Bernoulli, except on unrealistically small vocabularies (intuitively, 20 words are hardly adequate for discriminating between 20 newsgroups, and those words would have to be selected much more carefully). In light of this we can revise McCallum and Nigam's McCallum and Nigam (1998) recommendation to use the Bernoulli distribution for small vocabularies. Instead we recommend that neither the Bernoulli nor the binomial distributions should be used, since in all reasonable cases they are outperformed by the more robust variants of the binomial distribution. (The case of a 20,000 word vocabulary is quickly declared unreasonable, since most of the words occur precisely once in the training data, and so any parameter estimate is bound to be unreliable.)

We want to know whether the differences between the three binomial models could be dismissed as a chance occurrence. The McNemar test (Dietterich, 1998) provides appropriate answers, which are summarized in Table 3. As we can see, the classification results under the zero-inflated binomial and beta-binomial models are never significantly different, in most cases not even approaching significance at the 5% level. A classifier based on the beta-binomial model is significantly different from one based on the binomial model; the difference for a vocabulary of 20,000 words is marginally significant (the χ^2 value of 3.8658 barely exceeds the critical value of 3.8416 required for significance at the 5% level). Classification based on the zero-inflated binomial distribution differs most from using a standard binomial model. We conclude that the zero-inflated binomial distribution captures the relevant extra-binomial variation just as well as the overdispersed beta-binomial distribution, since their classification results are never significantly different.

The differences between the four models can be seen more visually clearly on the WebKB data set

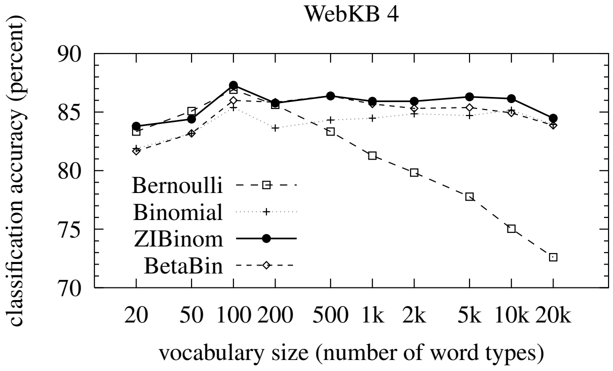

Figure 6: Accuracy of the four models on the WebKB data set as a function of vocabulary size.

(McCallum and Nigam, 1998, Figure 4). Evaluation results for Naive Bayes text classification using the four models are displayed in Figure 6. The zero-inflated binomial model provides the overall highest classification accuracy, and clearly dominates the beta-binomial model. Either one should be preferred over the simple binomial model. The early peak and rapid decline of the Bernoulli model had already been observed by McCallum and Nigam (1998).

We recommend that the zero-inflated binomial distribution should always be tried first, unless there is substantial empirical or prior evidence against it: the zero-inflated binomial model is computationally attractive (maximum likelihood estimation using EM is straightforward and numerically stable, most gradient-based methods are not), and its z parameter is independently meaningful, as it can be interpreted as the degree to which a given word is "on topic" for a given class of documents.

4 Conclusion

We have presented theoretical and empirical evidence for zero-inflation among linguistic count data. Zero-inflated models can account for increased variation at least as well as overdispersed models on standard document classification tasks. Given the computational advantages of simple zero-inflated models, they can and should be used in place of standard models. For document classification, an event model based on a zero-inflated binomial distribution outperforms conventional Bernoulli and binomial models.

Acknowledgements

Thanks to Chris Brew and three anonymous reviewers for valuable feedback. Cue the usual disclaimers.

References

Kenneth W. Church. 2000. Empirical estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2. In *18th International Conference on Computational Linguistics*, pages 180–186. ACL Anthology C00-1027.

Kenneth W. Church and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1:163–190.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1924.

Pedro Domingos and Michael J. Pazzani. 1996. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In *13th International Conference on Machine Learning*, pages 105–112.

Daniel M. Ennis and Jian Bi. 1998. The beta-binomial model: Accounting for inter-trial variation in replicated difference and preference tests. *Journal of Sensory Studies*, 13:389–412.

Ashutosh Garg and Dan Roth. 2001. Understanding probabilistic classifiers. In *12th European Conference on Machine Learning*, pages 179–191.

Norman L. Johnson and Samuel Kotz. 1969. *Discrete Distributions*, volume 1. Wiley, New York, NY, first edition.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *12th International Conference on Machine Learning*, pages 331–339.

Jack C. Lee and Y. L. Lio. 1997. A note on Bayesian estimation and prediction for the beta-binomial model. *Journal of Statistical Computation and Simulation*, 63:73–91.

David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *10th European Conference on Machine Learning*, pages 4–15.

Stephen A. Lowe. 1999. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval. In *6th European Conference on Speech Communication and Technology*, pages 2443–2446.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for naive Bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, pages 41–48.

Andrew Kachites McCallum. 1996. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. `http://www.cs.cmu.edu/~mccallum/bow/`.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers*. Springer, New York, NY, second edition.

Yudi Pawitan. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York, NY.