Assessing the Calibration of Naive Bayes' Posterior Estimates

Paul N. Bennett
September 12, 2000
CMU-CS-00-155

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

In this paper, we give evidence that the posterior distribution of Naive Bayes goes to zero or one exponentially with document length. While exponential change may be expected as new bits of information are added, adding new words does not always correspond to new information. Essentially as a result of its independence assumption, the estimates grow too quickly. We investigate one parametric family that attempts to downweight the growth rate. The parameters of this family are estimated using a maximum likelihood scheme, and the results are evaluated.

Email: pbennett+@cs.cmu.edu

Keywords: Naive Bayes, calibration, well-calibrated, reliability, posterior, text classification, Reuters

1 Introduction

The Naive Bayes classifier has shown itself to be a top competitor in some text domains. In other cases, it is used because its simplicity and its computational efficiency make it an attractive choice. However, researchers have often found when used for learning the actual posterior distribution, performance has been poor or average. Many people have specifically noted the fact that Naive Bayes tends to generate a bimodal posterior with scores clustered either arbitrarily close to 0 or arbitrarily close to 1. Of course, these estimates would be fine if the Naive Bayes classifier were always correct, but as it is not, it tends to produce uncalibrated probability estimates.

We give a theoretical justification of why this bimodal distribution is extremely likely to occur when using the Naive Bayes classifier, and given this justification, show why this suggests learning a posterior distribution based on the log odds produced by Naive Bayes may be one way to compensate for the overly confident predictions of the classifier.

Finally, we evaluate a sigmoid family that seems to be a likely candidate for obtaining better posterior estimates. The evaluation of this family indicates that while it often reduces the mean squared error of the estimates, it often does not lead to improved classification performance. Further investigation shows that this family may not be appropriate for Naive Bayes and indicates directions for future research to pursue.

2 Related Work

DeGroot and Fienberg (1983) discuss comparing the probability estimates of two classifiers in terms of calibration and refinement. A classifier is said to be *well-calibrated* if as the number of predictions goes to infinity the predicted probability goes to the empirical probability. For example, if we look at all of the times a weatherperson predicts a 40% chance of rain and the relative frequency of rain for those days is 40%, then the prediction is calibrated. If all probability predictions are calibrated, then the weatherperson is well-calibrated.

It is often easiest to depict the calibration of a classifier in a reliability diagram [Murphy and Winkler, 1977]. A reliability diagram plots the predicted probabilities versus the empirical probabilities. Thus, if all points fall on the $x = y$ line then the classifier is well-calibrated.

Refinement is essentially a measure of how close the probability estimates are to zero or one. Within well-calibrated classifiers, it is preferable to have a more refined classifier. However, it must be emphasized that the primary concern is calibration then refinement.

Platt (1999) discusses the use of a sigmoid family to fit the posterior distribution given the scores output by a SVM classifier. A maximum likelihood approach is used to determine the values of the parameter. It is essentially this technique that we follow here.

3 Data and Basic Tools

The data set for evaluating these ideas will be the Reuters 21578 data set [Lewis, 1997]. The train/test split used (which we refer to as ModApte-90) was obtained by refining the ModApte split as described in [Yang and Liu, 1999]. This data set has a standard train/test split of 7769/3019 documents. There are 90 classes; each of which has at least one document in the training and testing split. Each document can have multiple classes. The average is 1.234 classes/doc. Throughout the remainder of this paper we will focus on the class *Earn* (37.03% training data) and *Corn* (2.33% training data). We choose these two classes so we can form an idea of the behavior for very little training data and with a large amount of data. There are classes rarer than *Corn*, but it is difficult to examine trends with less data. It should be noted that the conclusions presented are consistent with results over all classes.

The Naive Bayes results presented throughout result from applying the unigram model of Naive Bayes [McCallum and Nigam, 1998] as implemented in the Rainbow classification system [McCallum, 1996]. Stemming and stop word removal were performed, and all words occurring less than 3 times in the training data were removed. No other feature processing or optimizations were done. All other parameters used (i.e. smoothing) were the defaults of the Rainbow system. The classification was done as a collection of binary classifiers (i.e. a two-class classifier is built for each of the 90 classes).

4 The Behavior of the Naive Bayes Posterior

4.1 Characterization of Behavior

First, we note that the Naive Bayes unigram Model is:

$$P(c_i \mid d) = \frac{P(c_i)\prod_{t=1}^{|d|} P(w_t \mid c_i)}{\eta(d)}, \tag{1}$$

where c_i is a class, d is a document, w_t is the t^{th} word of a document, and $\eta(d)$ is a normalization term that depends only on the document d.

Histograms for the posterior estimates for classes *Earn* and *Corn* are shown in figure 1. It should be noted

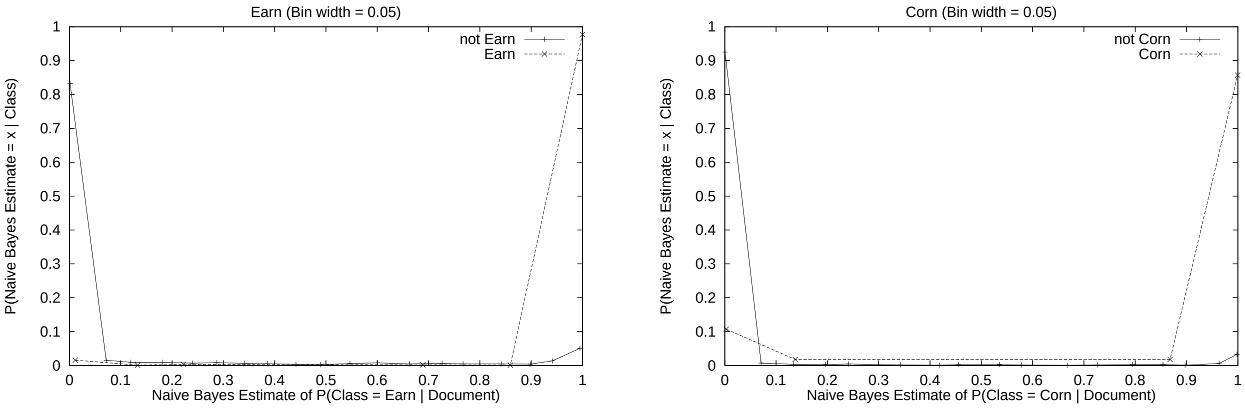

Figure 1: Conditional Distribution of Naive Bayes Posterior Estimates for *Earn* and *Corn* Binary Classifiers

that the classifier almost always outputs a class posterior of 1 or 0. Most of the examples in both cases are being classified correctly, but those that are being classified incorrectly have estimates that are completely wrong. We can show that this behavior can be expected in general as the length of the input feature vector grows. For text this means that we can expect to see this behavior as document length increases. Since longer feature vectors roughly correspond to higher dimensional spaces, we expect to see this behavior commonly in high-dimensional spaces and less frequently in low-dimensional spaces.

Note that we need only concentrate on the likelihood ratios

$$R_d(c_i, c_j) \doteq \frac{P(c_i \mid d)}{P(c_j \mid d)} \tag{2}$$

Since, if for some class δ, we have $R_d(c_i, c_\delta)$ (for all i), then we use $P(c_i \mid d) = P(c_\delta \mid d)R_d(c_i, c_\delta)$ to find

$P(c_i \mid d)$. To obtain $P(c_\delta \mid d)$, we solve the equation

$$P(c_\delta \mid d) \sum_i R_d(c_i, c_\delta) = 1. \tag{3}$$

This is convenient since the likelihood ratio allows us to ignore the normalization term. For the two classifier case, this means we are concentrating on, $P(c_\delta \mid d) + P(c_\delta \mid d) R_d(c_{\neg\delta}, c_\delta) = 1$. Specifically, how do the likelihood ratios change with documents?

We can define this problem formally by asking what happens to the posteriors when we add one single word to document d to form d'.

Problem

Let $d \doteq w_1 \ldots w_{n-1}$ and $d' \doteq w_1 \ldots w_n = dw_n$. How do we quantify the following relationship?

$$R_d(c_\alpha, c_\beta) \quad ? \quad R_{d'}(c_\alpha, c_\beta) \tag{4}$$

We will show with a straightforward derivation that the relationship is:

$$\frac{P(w_n \mid c_\alpha)}{P(w_n \mid c_\beta)} R_d(c_\alpha, c_\beta) = R_{d'}(c_\alpha, c_\beta) \tag{5}$$

For any document, d_i, we have:

$$\begin{aligned} R_{d_i}(c_\alpha, c_\beta) &= \frac{P(c_\alpha \mid d_i)}{P(c_\beta \mid d_i)} \text{, by definition.} \\ &\text{Substituting by defn. and cancelling the term } \eta(d_i) \\ &= \frac{P(c_\alpha) \prod_{t=1}^{|d|} P(w_t \mid c_\alpha)}{P(c_\beta) \prod_{t=1}^{|d|} P(w_t \mid c_\beta)}. \end{aligned} \tag{6}$$

So,

$$\begin{aligned} R_{d'}(c_\alpha, c_\beta) &= \frac{P(c_\alpha) \prod_{t=1}^{|d'|} P(w_t \mid c_\alpha)}{P(c_\beta) \prod_{t=1}^{|d'|} P(w_t \mid c_\beta)} \text{ by eq. 6} \\ &\text{Since } d' = dw_n, \\ &= \frac{P(w_n \mid c_\alpha) P(c_\alpha) \prod_{t=1}^{|d|} P(w_t \mid c_\alpha)}{P(w_n \mid c_\beta) P(c_\beta) \prod_{t=1}^{|d|} P(w_t \mid c_\beta)} \\ &\text{Substituting by eq. 6 again,} \\ &= \frac{P(w_n \mid c_\alpha)}{P(w_n \mid c_\beta)} R_d(c_\alpha, c_\beta). \end{aligned} \tag{7}$$

This yields the relationship we desired to show in eq. 5

We can consider the expected value of $\frac{P(w_i \mid c_\alpha)}{P(w_i \mid c_\beta)}$ over all words. Then as we add words, the ratio will grow exponentially on average with the average conditional word ratio. When viewed in terms of equation 3 , we can see that this will quickly push $P(c_\delta \mid d)$ to zero or one as the document length increases. It should be noted that feature selection methods that work well with Naive Bayes tend to maximize these conditional word ratios [Mladenic, 1998], and thus, they make the growth occur even more quickly.

Essentially this may be viewed as a direct byproduct of the independence assumption. We expect to see the posteriors quickly going to zero or one as new information is added (i.e. new evidence is additive logarithmically), but adding a word does not correspond to as much new evidence as the Naive Bayes method believes since the word occurrences are not truly independent.

4.2 Possible Corrective Methods

The above evidence and derivation indicates that we may want to consider some method that is essentially like slowing the growth rate according to the actual information being learned incrementally. Given this view point it is a natural development to examine the distribution of the log odds, $\log \frac{P(c|d)}{P(\neg c|d)}$, predicted by the Naive Bayes classifier to gain a sense of the growth rate of information for any given class.

Figure 2 shows the histogram based approximation of the class conditional log odds distribution for the Naive Bayes classifier for classes *Earn* and *Corn* obtained by hold-one-out testing. We could consider a

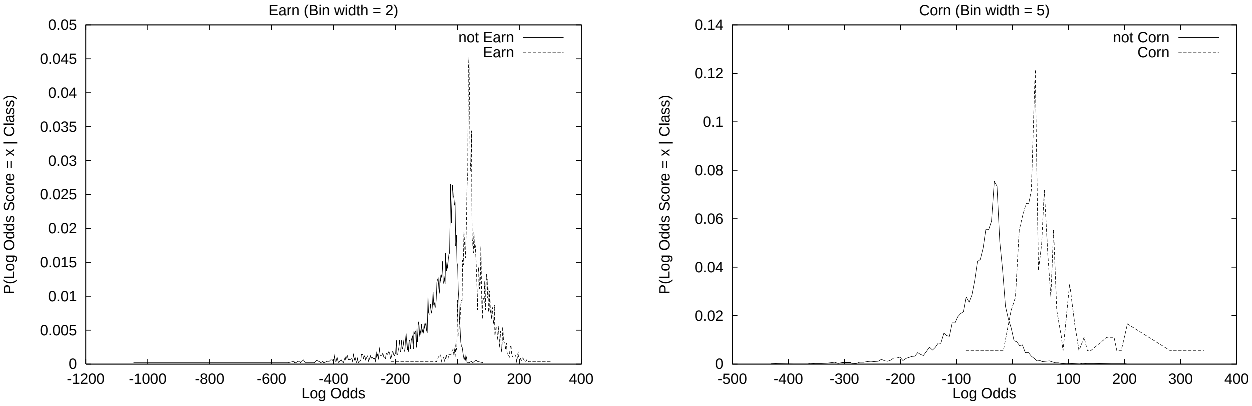

Figure 2: Conditional Distribution of Log Odds Scores for *Earn* and *Corn* Binary Classifiers

method of directly modeling these conditional score distributions and then use Bayes rule to invert the distributions [Hastie and Tibshirani, 1996]. Another alternative is to directly model the posterior distribution, $P(c \mid \text{Log Odds Score} = x)$. Learning the posterior distribution based simply on the log odds can be viewed as using the log odds as a sufficient statistic for Naive Bayes.

Since classifiers are trained to linearly separate the data, there is strong reason to suspect beforehand that the posterior distribution will be sigmoidal (assuming the classifier is performing reasonably well) [Platt, 1999]. Therefore, we may want to try a parametric sigmoid fit to validation data. In addition, if we assume that the conditional score distributions are Gaussian with a tied variance (although Platt, 1999, makes claims to the contrary – see alternative interpretation below), then Bayes rule applied to two such distributions gives the form:

$$P(c \mid \text{Log Odds Score} = x) = \frac{1}{1 + \exp\{Ax + B\}}, \tag{8}$$

where $A = \frac{\mu_{\neg c} - \mu_c}{\sigma^2}$, $B = -\frac{\mu_{\neg c}^2 - \mu_c^2}{2\sigma^2} + \log \frac{1-P(c)}{P(c)}$, and c indicates the "positive" class. Clearly, the distributions in figure 2 are not Gaussian, but it might be a reasonable first case approximation. In addition, the same sigmoid family results from applying Bayes rule to two exponentials with a different interpretation of the parameters [Platt, 1999]. Future methods may consider different assumptions; for example, the curves on the outside slopes are clearly very near exponential, but the interior part of the curves show a much steeper curve than standard exponentials. Finally, using the model given in equation 8 is equivalent to assuming that the output of the classifier is proportional (effect of parameter A) to the log of a biased (effect of parameter B) likelihood ratio. In other words, we can view it as a corrective slope to the growth rate.

The above reasons give good theoretical reasons for investigating a sigmoidal fit, but we can also justify these empirically by inverting the class conditional distributions given in figure 2. The results of this are

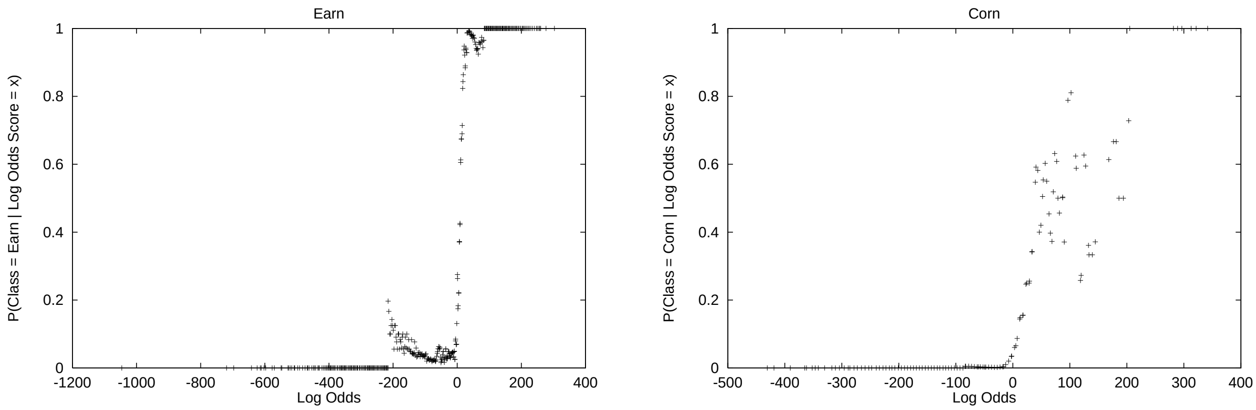

Figure 3: Posterior Distribution given Log Odds Scores for *Earn* and *Corn* Binary Classifiers

given in figure 3. These posterior distributions do seem primarily sigmoidal with several exceptional notable features. The plot for *Earn* shows a clear non-monotonic behavior around $x = -200$. Given the amount of data for *Earn* it seems unreasonable to assume this is simply noise. Therefore, any sigmoid fit will pay a penalty between balancing between this aberration and the primary upward curve of the sigmoid around $x = 0$. For *Corn*, there appears to be a decent amount of deviation from the sigmoid in the transitional area, however, as the class is rare, assuming this is noise is more reasonable here. The curve for *Corn* also seems to indicate an asymmetric sigmoid might be a better choice. If we were to actually choose an asymmetric sigmoid, this would likely show better performance for class *Earn* as well since the lower aberration would not as strongly effect the fit of the upper portion of the curve. This is a promising direction for future research. Finally, it should be noted that we could use a nonparametric method such as we did for creating these graphs. However, such methods are not easily applicable to rare classes, and we would like to apply these methods for rare classes as well.

5 Experiments

5.1 Methodology

Using the method described in Platt (1999), we use a maximum likelihood method to estimate the parameters of a sigmoid of the form given in equation 8. The data used for the fit is obtained from doing hold-one-out testing for each binary classifier. This raises another reason why it may be advantageous to use Naive Bayes. Since Naive Bayes is so efficient, we can perform hold-one-out testing with little training time penalty. This is not the case for many other methods including SVMs [Platt, 1999]. This means we are more likely to get a better approximation of the final classifier's behavior. Finally, the only notable bias that arises in this approach is that hold-one-out testing for a class with only one positive example will underestimate the log odds for positive examples.

In order to evaluate our method, we examine the mean squared error of the estimates as well as classifier performance via the F1 scores [Yang, 1997].

5.2 Results

The learned parameters for classes *Earn* and *Corn* fit the validation data they were trained on as shown in figure 4. The fit of the sigmoid to class *Earn* is subject to the flaws discussed above. However, it should be

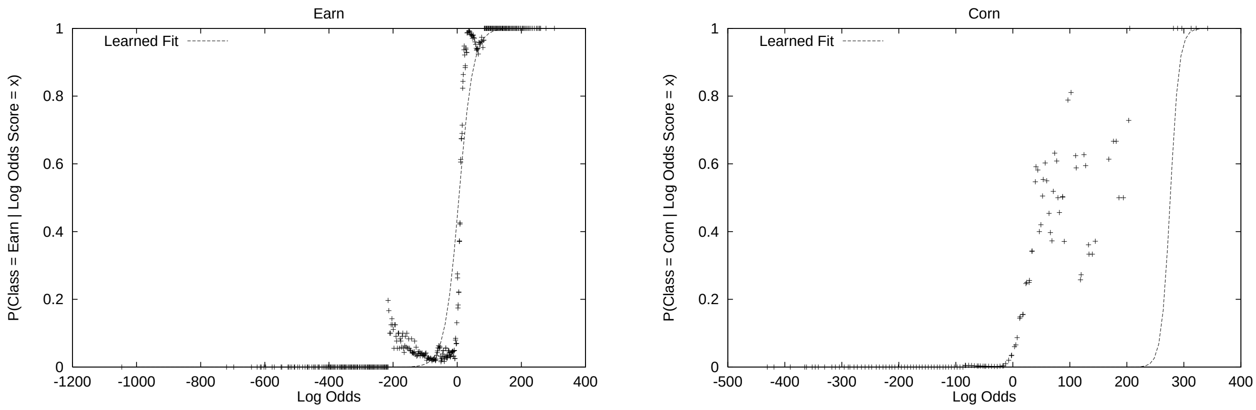

Figure 4: Data Distribution with Fitted Sigmoid

noted that it perfectly separates the data with respect to the threshold for the 0-1 Loss function (i.e. 0.5). The sigmoid for class *Corn* seems suspicious at first, but after inspection it understandably (since *Corn* is rare) shifts the curve far enough to the right that the only points with more than a non-zero probability are those completely correct. This is an indication that this method may not be appropriate for rare classes.

The calibration for classes *Earn* and *Corn* for the two methods can be seen in the reliability diagrams given in figures 5 and 6, respectively. The labels on the points are the number of examples the point

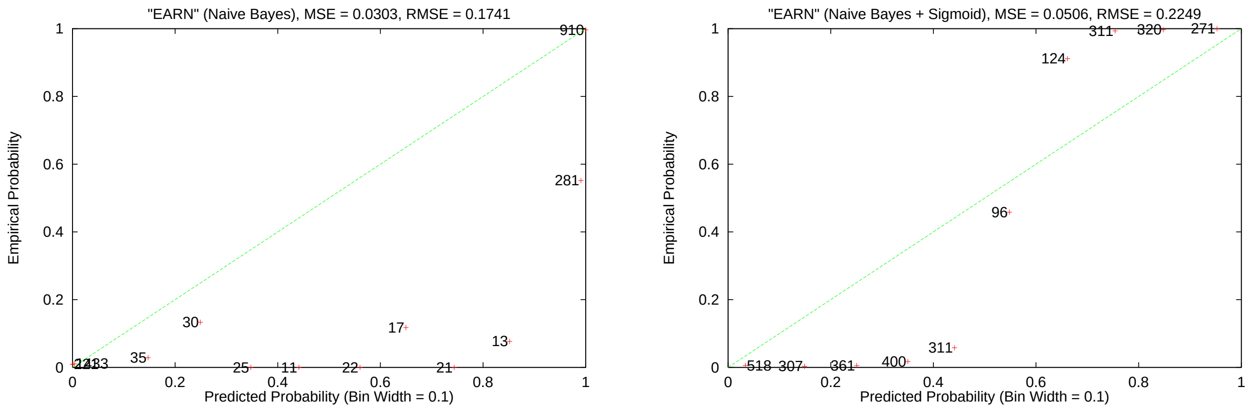

Figure 5: Reliability Diagrams for Class "Earn"

represents. The mean squared error (MSE) and root mean squared error (RMSE) are given in the title. The reliability diagram for *Earn* is consistent with the sigmoidal fit. Initially the sigmoid overestimates the true posterior, then gets it approximately correct in the middle region, and in the upper region underestimates it. The reliability diagram for *Corn* clearly shows that the sigmoid method improves the MSE because of the rarity of class *Corn*, i.e. it never predicts a non-zero probability except when it's absolutely certain.

For 83 of the 90 classes, the sigmoid method reduced the MSE of the estimates. Five of these classes were *extremely* rare classes ($< 0.3\%$ of the training data). The two notable exceptions were classes *Earn* and *Acq* which are the two most common classes. The first of these can be explained by the aberrational behavior discussed above. However, since it was able to find an optimal threshold, the F1 score increased

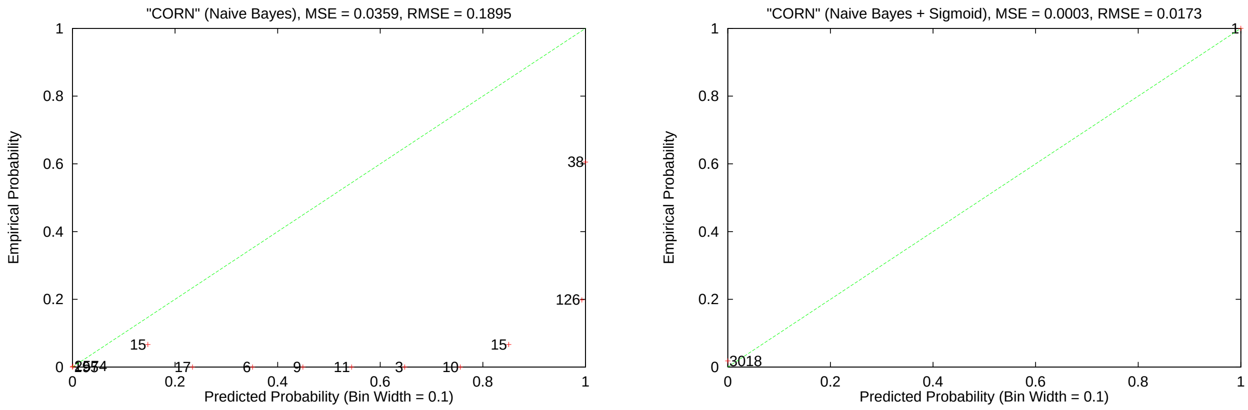

Figure 6: Reliability Diagrams for Class "Corn"

significantly (0.906 vs. 0.956). For *Acq*, the posterior distribution shows aberrational behavior as well, but on the positive curve in the sigmoid (i.e. there is a significant dip in the posterior). Therefore, the method was unable to find either a good fit or an optimal threshold. As a result, F1 decreased significantly (0.923 vs. 0.405). This at least calls into question using a symmetric sigmoid and possibly fitting the posterior directly with a parametric family.

On average, the sigmoid method significantly depressed the F1 scores (both micro and macro). This resulted primarily from the method's inability to handle rare classes very well. Thus the number of correct positives were severely decreased, and the number of correct positives is the most important component of F1.

5.3 Discussion

Overall, the sigmoid method here would indicate that the family of equation 8 cannot capture the behavior of Naive Bayes. As discussed, an asymmetric sigmoid seems to be a more appropriate choice.

In general, this method does not seem to work well for rare classes. This is especially true if the data are extremely non-separable based on the log odds score.

In addition, the aberrations (from a sigmoid) in the two most common classes indicate that directly fitting the posterior may not be possible at all. More investigation is needed on this issue, but feature selection methods may lead to a more consistent sigmoid form as they increase linear separation (since they are increasing performance).

The behavior of the conditional log odds score distributions seemed stable enough that modeling them and then using Bayes rule to invert the distributions seems more promising.

6 Future Work

Clearly, future work will want to explore use of other sigmoid families in approximating the posterior distribution given the Naive Bayes log odds approximation. Specifically asymmetric families seem promising as discussed above.

In addition using parametric families to fit the conditional score distributions $P(\text{Log Odds Score} = x|\text{class})$ and Bayes Rule to "invert" them seems to be another promising avenue.

Evaluation over other data sets is also necessary to determine the extent to which Naive Bayes' behavior is consistent across domains. Finally, investigating the effects of feature selection on the shape of the posterior curve given the log odds score would be interesting as feature selection may give a more sigmoidal curve (i.e. since it increases performance, it's better at separating the data).

7 Conclusion

In the above, we have explained the conditions and theoretical reasons that give rise to the nearly binary posterior estimates of Naive Bayes. Using this explanation, we justified modeling the posterior given the log odds scores of Naive Bayes.

Finally, we considered and evaluated an initially promising candidate sigmoid family for this approach. After evaluation, the results indicate other methods that may overcome the Naive Bayes nearly binary estimates without being subject to the shortcomings of the particular sigmoid family investigated here.

References

[Hastie and Tibshirani, 1996] Hastie, T. and Tibshirani, R. (1996). Classification by pairwise coupling. Technical report, Stanford University and University of Toronto. http://www-stat.stanford.edu/~trevor/Papers/2class.ps.

[Lewis, 1997] Lewis, D. D. (1997). Reuters-21578, Distribution 1.0. http://www.research.att.com/ lewis.

[McCallum and Nigam, 1998] McCallum, A. and Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. In *Working Notes of AAAI 1998, Workshop on Learning for Text Categorization.*

[McCallum, 1996] McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow.

[Mladenic, 1998] Mladenic, D. (1998). *Machine Learning on Non-homogeneous, Distributed Text Data.* PhD thesis, University of Ljubljana, Department of Computer and Information Science.

[Murphy and Winkler, 1977] Murphy, A. H. and Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26:41–47.

[Platt, 1999] Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. J., Bartlett, P., Scholkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers.* MIT Press.

[Yang, 1997] Yang, Y. (1997). An Evaluation of Statistical Approaches to Text Categorization. Technical Report CMU-CS-97-127, Carnegie Mellon University.

[Yang and Liu, 1999] Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *ACM SIGIR Conference on Research and Development in Information Retrieval.* http://www.cs.cmu.edu/ yiming/publications.html.