Perplexed Bayes Classifier

Cohan Sujay Carlos
Aiaioo Labs
Bangalore
India
cohan@aiaioo.com

Abstract

Naive Bayes classifiers estimate posterior probabilities poorly (Zhang, 2004).

In this paper, we propose a modification to the Naive Bayes classification algorithm which improves the classifier's posterior probability estimates without affecting its performance.

Since the modification involves the use of the reciprocal of the *perplexity* of the class-conditional feature probabilities, we call the resulting classifier the *Perplexed Bayes* classifier.

We demonstrate that the modification results in better calibrated posterior probabilities on a gender categorization task.

1 Introduction

Probabilistic classifiers work by selecting the most probable *class* given the *features* of the data point being classified, as shown in Equation 1.

$$\arg\max_{c} P(C|F) \tag{1}$$

Bayesian classifiers transform $P(F|C)$ into $P(C|F)$ as shown in Equation 2.

$$P(C|F) = \frac{P(F|C) \times P(C)}{P(F)} \tag{2}$$

Naive Bayes classifiers additionally assume that the features f_1, f_2, f_3, etc. are all independent of one another, conditional on the class C, yielding the following equation.

$$P(F|C) = \prod_i P(f_i|C) \tag{3}$$

Equation 3 can be substituted into Equation 2 to obtain Equation 4.

$$P(C|F) = \frac{\left(\prod_i P(f_i|C)\right) \times P(C)}{P(F)} \tag{4}$$

The posterior probability estimates obtained using Equation 4 tend to be extreme as observed in Eyheramendy et al (2003).

Improving the posterior probability estimates of Naive Bayes classifiers might make them more useful for NLP (Nguyen and O'Connor, 1999).

In this paper, we present the *Perplexed Bayes* classification algorithm that produces better calibrated posterior probabilities than the Naive Bayes algorithm and operates with the same accuracy.

2 Related Work

The Naive Bayes classification algorithm is still commonly used as a baseline algorithm for many classification tasks (Rennie et al, 2003), and is reputed to perform surprisingly well (McCallum and Nigam, 1998; Rennie et al, 2003; Zhang, 2004) though the posterior probabilities might be estimated poorly (Eyheramendy et al, 2003; Rennie et al, 2003; Zhang, 2004).

Attempts to improve the Naive Bayes classifier have relied on augmentations to relax independence assumptions (Peng and Schuurmans, 2003; Peng et al, 2004), transformations to correct systemic errors (Rennie et al, 2003), the weighting of counts or probabilities (Zaidi et al, 2013; Frank et al, 2003; Webb and Pazzani, 2004) and the subsetting of features (Langley and Sage, 1994).

It has been proposed that one might use corrective sigmoid functions (Platt, 1999; Bennett, 2000; Niculescu-Mizil and Caruana, 2005; Caruana and Niculescu-Mizil, 2006)), isotonic regression (Zadrozny and Elkan, 2002), asymmetric distributions (Bennett, 2003) and binning (Zadrozny and Elkan, 2001; Bella et al, 2009) to obtain calibrated posterior probabilities (Rüping, 2006) from SVMs, decision trees and Naive Bayes classifiers.

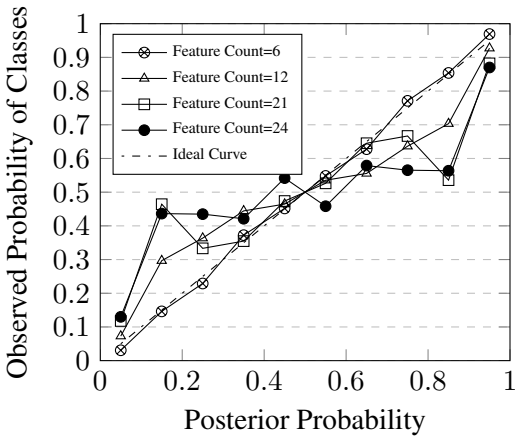

Figure 1: Reliability diagram for a Naive Bayes (NB) classifier.

Our approach is closer to that of Zaidi et al (2013) who used weighted class-conditional feature probabilities. One of the equations that Zaidi suggests could be used (but does not go on to explore) is identical to Equation 8 in this paper.

None of the previous studies has, to our knowledge, explored in detail, attempted to generalize, or developed a theoretical foundation for the approach that we describe in this paper.

3 Naive Bayes

In this section, we show that the posterior probabilities of the Naive Bayes classification algorithm are not well calibrated.

A Naive Bayes classifier's posterior probabilities were measured on a classification task (identifying the gender of names using the dataset described in Section 5) and a reliability diagram (Bröcker and Smith, 2007) plotted for different numbers of features used as shown in Figure 1.

A perfectly calibrated classifier's reliability diagram would show a straight line (like the ideal curve of Figure 1). As can be seen, the Naive Bayes classifier does not produce well-calibrated posterior probabilities, except for the feature count of 6. The calibration is seen to deteriorate as the number of features increases.

In the next section, we propose a modification to the Naive Bayes algorithm to attempt to solve the problem of poor posterior probability estimation.

4 Perplexed Bayes

The perplexity $PP(p_1, p_2, \ldots p_n)$ of a set of probabilities $\{p_1, p_2, \ldots, p_n\}$ is computed as shown in Equation 5.

$$PP = \frac{1}{(p_1 \times p_2 \times \ldots \times p_n)^{\frac{1}{n}}} \tag{5}$$

So, the *reciprocal of the perplexity* of the probabilities is merely their *geometric mean* as shown in Equation 6.

$$PP^{-1} = (p_1 \times p_2 \times \ldots \times p_n)^{\frac{1}{n}} \tag{6}$$

In the Perplexed Bayes classifier, we combine the class conditional feature probabilities using the geometric mean, as shown in Equation 7.

$$P(F|C) = \left(\prod_{1 \le i \le n} P(f_i|C) \right)^{\frac{1}{n}} \tag{7}$$

So, the posterior probability equation can be written as shown in Equation 8, where n is the number of features, and N is the normalizer.

$$P(C|F) = \frac{\prod_i P(f_i|C)^{\frac{1}{n}} \times P(C)}{N} \tag{8}$$

We call a classifier that uses the posterior probability equation in Equation 8 the fully *Perplexed Bayes* classifier.

4.1 Assumption

We can show that Equation 8 can be derived from Equation 2 if we assume that *the class C is independent of all features but one, and that none of the features is special* as shown in Equation 9, where $1 \le i \le n$.

$$P(C|f_1, f_2, \ldots f_n) = P(C|f_i) \tag{9}$$

We can write Equation 9 in n different ways, as follows, because no feature is special.

$$\begin{aligned} P(C|f_1, f_2, \ldots f_n) &= P(C|f_1) \\ &= P(C|f_2) \\ &\vdots \\ &= P(C|f_n) \end{aligned} \tag{10}$$

We show below that the assumption embodied in Equation 9 is sufficient for the derivation of Equation 8 (but we have not shown that it is also necessary).

4.2 Derivation

Multiplying together all the terms on both sides of Equation 10 we get Equation 11.

$$P(C|f_1, f_2, \ldots f_n)^n = \\ = \prod_{1 \leq i \leq n} P(C|f_i) \tag{11}$$

Inverting the terms on the right-hand side of Equation 11 using the Bayesian inversion equation (2), we get Equation 12.

$$P(C|F)^n = \left(\prod_{1 \leq i \leq n} \frac{P(f_i|C) \times P(C)}{P(f_i)} \right) \tag{12}$$

Since $P(C)$ and $P(F)$ are independent of i, we can write Equation 12 as Equation 13.

$$P(C|F)^n = \left(\prod_{1 \leq i \leq n} P(f_i|C) \right) \\ \times \frac{P(C)^n}{\prod_{1 \leq i \leq n} P(f_i)} \tag{13}$$

$$P(C|F) = \left(\prod_{1 \leq i \leq n} P(f_i|C) \right)^{\frac{1}{n}} \times \frac{P(C)}{N} \tag{14}$$

Finally, taking the nth root on both sides, we get Equation 14 (where N is the normalizer) and this is substantially the same as Equation 8.

So we have shown that the assumption that *the class C is independent of all features but one, and that none of the features is special* (written as Equation 10) can give us Equation 8.

It is interesting to note that Equation 15, representing the posterior probabilities of a classifier that uses the *arithmetic mean* instead of the *geometric mean*, can be derived by a similar sequence of steps from Equation 10 as well.

$$P(C|F) = \left(\sum_{1 \leq i \leq n} P(f_i|C) \right) \times \frac{P(C)}{n \times N} \tag{15}$$

4.3 Interpretation

It can be shown that the independence of classes and features $P(C|F) = P(C)$ is a direct result of Equation 10 as follows.

$$P(c_i) = \sum_F P(c_i, F) = \sum_F P(c_i|F)P(F) \tag{16}$$

But, since $P(c_i|F)$ is a constant m_i by reason of Equation 10, we get:

$$P(c_i) = m_i \times \sum_F P(F) \tag{17}$$

But, $\sum_F P(F) = 1$.

So, $P(c_i) = m_i = P(c_i|F)$ for all i.

So, it has been shown that Equation 10 implies that $P(C|F) = P(C)$ and therefore the features are independent of the classes.

Moreover, it can be seen that the constraints in Equation 10 are only constraints on the classes.

It follows that the features are not constrained in any way by Equation 10 and do not have to be class-conditionally independent of one other.

4.4 Generalization

It appears possible to model assumptions that fall between those of the Naive Bayes classifier and the fully Perplexed Bayes algorithm described above through the use of an *attenuation coefficient* k in the geometric mean as shown in Equation 18.

$$PP^{-\mathrm{k}} = (p_1 \times p_2 \times \ldots \times p_n)^{\frac{\mathbf{k}}{n}} \tag{18}$$

By plugging Equation 18 into Equation 2, we get the following posterior probablity equation.

$$P(C|F) = \left(\prod_{1 \leq i \leq n} P(f_i|C) \right)^{\frac{\mathbf{k}}{n}} \times \frac{P(C)}{N'''} \tag{19}$$

The attenuation coefficient k ranges from 1 to n, where lower values correspond to more perplexity.

It may be noted that if we set $k = n$, Equation 19 becomes Equation 4 used in the Naive Bayes classifier.

On the other hand, if we set $k = 1$, Equation 18 becomes the same as Equation 8 used in the fully Perplexed Bayes classifier.

4.5 Approximation

It is possible to obtain the same accuracy as a Naive Bayes classifier and yet retain the excellent posterior probability characteristics of the Perplexed Bayes classifier using the approximation shown in Equation 20.

$$P(C|F) = \frac{\left((\prod_i P(f_i|C)) \times P(C)\right)^{\frac{\mathbf{k}}{n+1}}}{N''} \quad (20)$$

It can be seen from Equation 20 that the numerator is the $k/(n+1)$th root of the numerator of the posterior probability equation of the Naive Bayes classifier as shown in Equation 4.

So, the posterior probability approximation of Equation 20 provably makes classification decisions about data points in exactly the same way as Equation 4 because if a positive real number a/N' is greater than b/N', then a^k/N'' must also be greater than b^k/N'' where k, N' and N'' are constants.

5 Experimental Results

For our experiments, we used a collection of 7944 gender-labelled names with 2943 marked male and 5001 marked female.

The data set was randomized and then split into a training set consisting of the first 6354 names and a test set consisting of the remaining 1590 names[1].

In all experiments, the approximation in Equation 20 was used. Unless otherwise stated, for all experiments where the attenuation coefficient was automatically computed, it was chosen (through a binary search) so as to minimize the standard deviation of the normalized histogram of posterior probabilities on the training data.

5.1 Distribution Experiments

The curves of the standard deviation of normalized histogram counts of posterior probabilities plotted against feature counts in Figure 2 show that posterior probabilities are more evenly distributed in Perplexed Bayes classifiers than in Naive Bayes classifiers for higher feature counts.

5.2 Accuracy Experiments

It is to be expected that as the Perplexed Bayes classifier's confidence in its results increases, so would its accuracy. So, the accuracy of the classifier for different ranges of posterior probabilities was computed and is presented in Table 1. It can be seen from Table 1 that with higher thresholds, it is possible to obtain higher accuracies in the Perplexed Bayes classifier.

[1]The randomized collection of names used may be downloaded from http://www.aiaioo.com/downloads/namesfile.txt

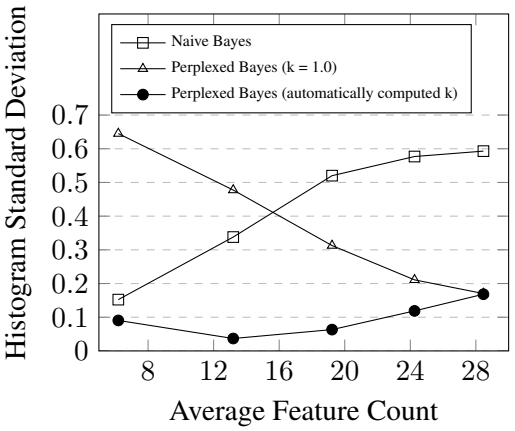

Figure 2: The standard deviation of the normalized histogram counts of the posterior probabilities plotted against the average number of features.

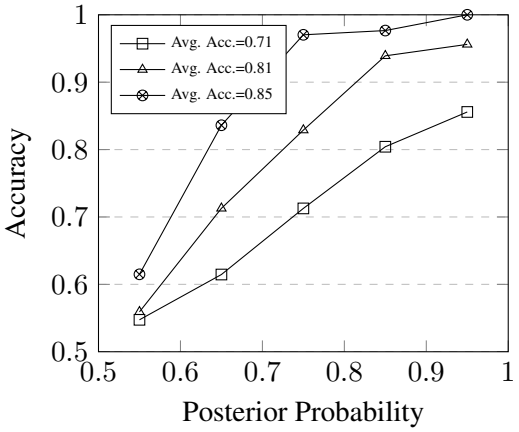

Figure 3: The accuracy of the Perplexed Bayes classifier against its posterior probability.

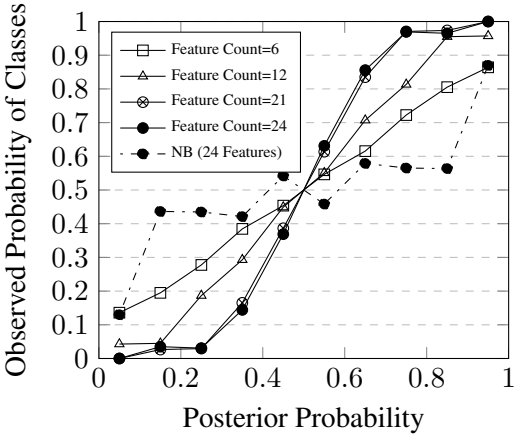

Figure 4: Reliability diagram for a Perplexed Bayes classifier with the *attenuation coefficient* optimized for flatness of the posterior probability histogram, alongside a Naive Bayes (NB) curve.

$P(C\|F)$	Points	PB Acc	Points	NB Acc
0.5-0.6	387	0.6149	26	0.5000
0.6-0.7	421	0.8361	22	0.3636
0.7-0.8	439	0.9703	26	0.5000
0.8-0.9	300	0.9766	42	0.5238
0.9-1.0	43	1.0000	1474	0.8792

Table 1: Perplexed and Naive Bayes classifier accuracies for different confidence intervals (average of 24.4 features, and overall accuracy of 0.85).

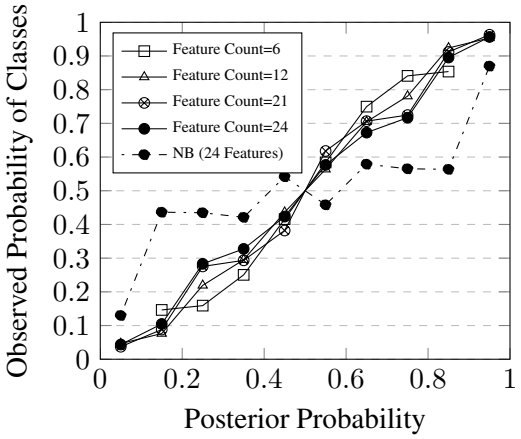

Figure 5: Reliability diagram for a Perplexed Bayes classifier with the *attenuation coefficient* optimized for good calibration on a validation set, and a Naive Bayes (NB) curve for comparison.

In contrast, measurements for the Naive Bayes classifier, also shown in Table 1, indicate that 92.7% of the data points are classified with a confidence of above 0.9, and that the remaining data points are assigned to classes almost randomly, so the accuracy is not very sensitive to threshold changes between 0.5 and 0.9. Figure 3 shows an increase in the accuracy of classification with an increase in the posterior probability.

The reliability diagram for the Perplexed Bayes classifier with the Y-axis values representing the probability of a data point's real class equalling the class for which the classifier's posterior probability is plotted on the X-axis, is shown in Figure 4.

The reliability diagram in Figure 5 was obtained similarly using a Perplexed Bayes classifier where the attenuation coefficient was estimated so as to mimimize the Root Mean Square Error (RMSE) of the observed posterior probabilities from ideal values over a held-out validation set of data points.

The RMSEs of the observed posterior probabilities of Figure 5 were 0.069, 0.043, 0.049 and 0.064 for 6, 12, 21 and 24 features respectively, whereas the RMSEs for a Naive Bayes classifier were 0.016, 0.093, 0.173 and 0.164 (at accuracies of 71.5%, 81.5%, 85.7% and 84.5%), establishing that on this data set *the Perplexed Bayes classifier produced better calibrated posterior probabilities for higher feature counts than a Naive Bayes classifier of the same accuracy*.

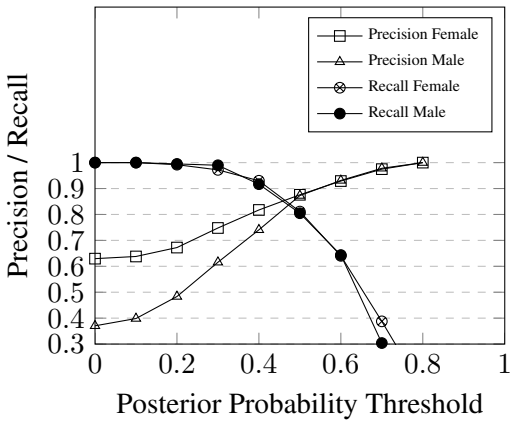

Figure 6: The precision and recall of the Perplexed Bayes classifier against decision thresholds.

5.3 Precision & Recall Experiments

The precision and recall of the Perplexed Bayes classifier plotted against confidence thresholds for the selection of one class over the others are as shown in Figure 6.

6 Conclusions

We have shown that it is possible to build a classifier (the Perplexed Bayes classifier) that makes classification decisions that are identical to those of a Naive Bayes classifier without assuming that the features used are class-conditionally independent, by combining the class-conditional feature probabilities into posterior probabilities using their geometric mean unlike the Naive Bayes classifier that takes their product, and that such a classifier incorporating an attenuation coefficient can produce better calibrated posterior probabilities on the given data set than a Naive Bayes classifier for higher feature counts.

7 Future Work

We should like to see if the mathematics used in the Perplexed Bayes classifier could be used to make improvements to Hidden Markov Models and in Probabilistic Graphical Models.

Acknowledgments

The author is grateful to Srivatsan Laxman for the assistance he willingly offered with matters related to probability theory and mathematics, to Sumukh Ghodke for his feedback, and to the reviewers for their helpful and very useful comments.

References

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning (ICML '05)*,625–632.

Andrew McCallum and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*,752:41–48.

Antonio Bella and Cèsar Ferri and José Hernández-Orallo and Marïa José Ramírez-Quintana. 2009. Similarity-binning Averaging: A Generalisation of Binning Calibration. In *Proceedings of the 10th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'09)*,341–349. Springer-Verlag.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classiers. *ICML*, 609–616.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classier scores into accurate multiclass probability estimates. *KDD*, 694–699.

Eibe Frank and Mark Hall and Bernhard Pfahringer. 2003. Locally Weighted Naive Bayes. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 249–256. Morgan Kaufmann Publishers Inc.

Fuchun Peng and Dale Schuurmans. 2003. Combining Naive Bayes and N-Gram Language Models for Text Classification. In *Proceedings ofThe 25th European Conference On lnformmion Retrieval Research (ECIR03)*.

Fuchun Peng and Dale Schuurmans and Shaojun Wang. 2004. Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, 7(3-4):317–345. Springer.

Geoffrey I. Webb and Michael J. Pazzani. 1998. Adjusted probability naive Bayesian induction. In *Proceedings of the Eleventh Australian Joint Conference on Artificial Intelligence*, 285–295. Springer-Verlag.

Harry Zhang. 2004. The Optimality of Naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*. AAAI Press.

Jason D. M. Rennie and Lawrence Shih and Jaime Teevan and David R. Karger. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning*, 20:616–623.

Jochen Bröcker and Leonard A. Smith. 2007. Increasing the Reliability of Reliability Diagrams. In *Weather and Forecasting*, 22:651661.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classiers*, 61–74. MIT Press.

Khanh Nguyen, Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. *Proceedings of EMNLP 2015*.

Nayyar A. Zaidi and Jesús Cerquides and Mark J. Carman and Geoffrey I. Webb. 2013. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. *Journal of Machine Learning Research*, 14(1):1947–1988. JMLR.org.

Pat Langley and Stephanie Sage. 1994. Induction of Selective Bayesian Classifiers. *Conference on Uncertainty in Artificial Intelligence*, 399–406. Morgan Kaufmann.

Paul N. Bennett. 2000. Assessing the calibration of naive Bayes posterior estimates. *Technical Report*. Carnegie Mellon University.

Paul N. Bennett. 2003. Using Asymmetric Distributions to Improve Text Classifier Probability Estimates. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*, 111–118. ACM.

Rich Caruana and Alexandru Niculescu-Mizil. 2006. An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd international conference on Machine learning (ICML '06)*,161–168. ACM.

Stefan Rüping. 2006. Robust Probabilistic Calibration. In *Proceedings of the 17th European Conference on Machine Learning*, 743–750. Springer.

Susana Eyheramendy and David D. Lewis and David Madigan. 2003. On the Naive Bayes Model for Text Categorization. In *9th International Workshop on Artificial Intelligence and Statistics*.