Perplexed Bayes Classifier - Mathematics Companion

Cohan Sujay Carlos
Aiaioo Labs
Bangalore
India
cohan@aiaioo.com
Aiaioo Labs Technical Report - ALTR2

Abstract

We show that it is possible to derive a classifier (the *Perplexed Bayes classifier*) that makes classification decisions that are identical to those of the Naive Bayes classifier but with better-calibrated posterior probability estimates. We also show that:

- The classifier's posterior probability equation is possible to derive starting from a certain set of assumptions.
- The aforesaid assumptions imply the absence of naive class-conditional feature independence assumptions.
- The classifier, with the introduction of an approximation to assign the prior probability the same weightage it is assigned in the Naive Bayes classifier, provably makes the same decisions as a Naive Bayes.
- This approximated classifier has a reliability curve that is closer to the ideal than the Naive Bayes'.

Thus we show that there is good reason to believe that the naive independence assumption (that the features are all class-conditionally independent of one another) used in the Naive Bayes classifier is not essential to its performance, whereas discarding it could result in better posterior probability estimates.

1 Naive Bayes

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.

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

The Naive Bayes classifier also uses the assumption 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 (i.e., estimates of $P(C|F)$) obtained using Equation 4 tend to be extreme, as observed in Eyheramendy et al (2003).

$P(C\|F)$	**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).

We can also observe the tendency to return extreme posteriors in the last two columns of Table 1 which shows the number of data points that were classified with posterior probabilities (confidences) in various ranges between 0.5 and 1.0 by the Naive Bayes classifier (on the name gender classification task (Carlos, 2015)).

It can be seen that most of the data points were classified with a confidence of greater than 0.9.

What we'd like to have is a classifier whose confidences are distributed evenly between 0.5 and 1.0 as shown in the first two columns of Table 1. We were able to create such a classifier by using the Perplexed Bayes posterior probability equation described in the next section.

However, the intuition behind the Perplexed Bayes equation was obtained through the simulation described below.

1.1 Naive Bayes Simulations

We attempted to understand through a simulation why the posterior probabilities returned by a Naive Bayes classifier tend to be extreme.

For the simulation of the Naive Bayes classifier, we use a table of all possible combinations of classes and features and assigned the joint probabilities by picking a random value between the number obtained by multiplying the probabilities of all the features (given a class) together for each row, and 80% of that number, and then normalizing the values in all the rows so that they add up to 1 as shown in Table 2.

Class	f_1	f_2	$P(f_1, f_2, C)$	$P(C\|f_1, f_2)$
c_1	0	0	0.4565	0.8120
c_1	1	0	0.0040	0.0180
c_1	0	1	0.0264	0.2106
c_1	1	1	0.0003	0.0028
c_2	0	0	0.1057	0.1880
c_2	1	0	0.2159	0.9820
c_2	0	1	0.0991	0.7894
c_2	1	1	0.0922	0.9972

Table 2: Simulation Probabilities Table.

This method yielded a standard deviation of the posterior probabilities $P(C|F)$ of 0.4345. This turns out to be higher than the standard deviation of the class conditional feature probabilities used to generate the joint probability table, which is 0.2880 and higher than the marginal probabilities of the individual features given the class $P(f_i|C)$ computed from it, which is 0.3648.

This suggests that the procedure used to generate the joint probability table (the combination by multiplication of marginal probabilities into joint probabilities) results in posterior probabilities that tend towards 0 or 1.

It is seen from Figure 1 that as the number of features increases, so does the standard deviation of the simulated posterior probabilities.

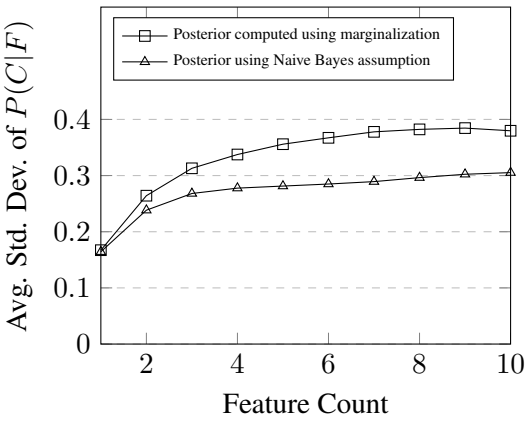

Figure 1: The standard deviation of posterior probabilities simulated from class conditional feature probabilities with average divergence of approximately 0.2.

It can be seen that the experimental results in Figure 2 closely resemble the simulation results in Figure 1. In both, as the number of features increases, the posterior probabilities tend to get increasingly extreme.

In the simulation results in Figure 1, the posterior probabilities appear extreme even when computed through marginalization and not just when computed using the naive bayes assumption.

This perhaps shows that the cause of the extremeness in predictions is the multiplication operator used to combine class-conditional feature probabilities, because the joint probability distribution was generated by choosing random values between the number obtained by multiplying the probabilities of all the features (given a class) together for each row, and 80% of that number, and then normalizing the values in all the rows so that they add up to 1.

So, the multiplicative combination of probabilities was used to generate the joint probability distribution in Table 2. As we see in Figure 1, that alone sufficed to make the posterior probabilities (computed merely through marginalization) as extreme as those in the experiment whose measured posterior probabilities are as shown in Figure 2.

This suggests that the log ratio of the multiplicative combinations of two sets of random probabilities (of the same cardinality) tends to the extremes of the space of real numbers as the cardinality of those sets tends to infinity.

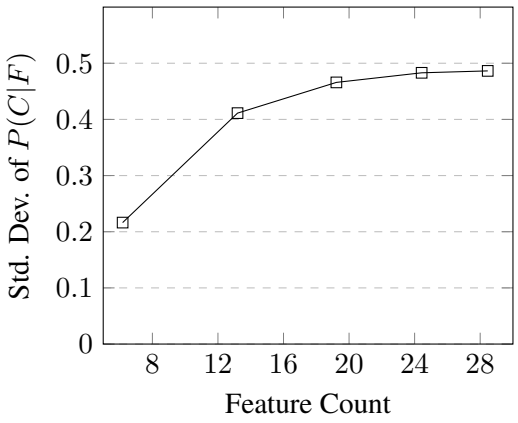

Figure 2: The standard deviation of the posterior probabilities of a Naive Bayes classifier plotted against the number of features (on the gender classification task (Carlos, 2015)).

2 Perplexed Bayes

As we saw in the preceding section, the reason for the extreme posterior probability estimates of the Naive Bayes classifier seems to be the multiplicative combination of the class-conditional feature probabilities.

In this section, we propose a posterior probability equation that does not suffer from the problem described above. The equation uses the *geometric mean* of the class-conditional feature probabilities instead of their product. We show using a simulation that the resulting posterior probabilities tend towards the center (0.5) as the number of features increases.

We later show experimentally that the posterior probabilities can be spread out evenly for the range of feature counts encountered in the gender classification task (Carlos, 2015) through the use of an *attenuation coefficient* as shown in Equation 37.

First, we shall explain how the geometric mean is related to the perplexity operator (which gives the Perplexed Bayes classifier its name).

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 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.

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

As a result, the posterior probability equation of the Perplexed Bayes classifier becomes the one shown in Equation 8, where n is the number of features, and N is the normalizer.

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

Equation 8 was, to our knowledge, first reported in the work of Zaidi et al (2013).

We will attempt **three proofs** in the rest of this report and argue that they strongly suggest that the 'naive' independence assumptions in the Naive Bayes classifier do not contribute to its accuracy.

The three proofs are as follows:

- **Proof 1: That the Perplexed Bayes posterior probability equation can be derived from the assumption that *the class C is independent of all features but one, and none of the features is special*.**
- **Proof 2: That the above assumption can be shown to not place any constraints on features; that joint probability distributions that comply with the conditions of class-conditional feature independence and joint probability distributions that don't, are both compatible with the above assumption.**
- **Proof 3: That the classification decisions of the Perplexed Bayes classifier (with an approximation to reduce the weightage given to the prior) are identical to the classification decisions of the Naive Bayes classifier.**

These three proofs fall short of allowing us to claim with certainty that the accuracy of the Naive Bayes is not in any way connected to its naive class-conditional feature independence assumption, because we have only been able to show that the Perplexed Bayes posterior probability equation can be derived from the assumption and not the other way around, and because the decisions of

the Naive and Perplexed Bayes classifiers are the same only if the assumption changing the prior's weightage as shown in Equation 40 is made.

3 Proof 1:

We attempt to prove below that Equation 8 can be derived from the assumption that *the class C is independent of all features but one, and none of the features is special*.

The assumption can be encoded as shown in Equation 9 (where $1 \leq i \leq n$).

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

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

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

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

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

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 $P(C)$ is independent of i, we can write Equation 12 as Equation 13.

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

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

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.

Q.E.D: 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* leads to Equation 8.

Below, we describe an alternate formulation of Equation 8.

3.1 Alternate Formulation

It is interesting to note that Equation 19, representing the posterior probabilities of a classifier that uses the *arithmetic mean* instead of the *geometric mean*, can be derived from Equation 10 as well.

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

Adding together all the terms on both sides of Equation 10 we get Equation 16.

$$\begin{aligned} P(C|f_1, f_2, \ldots f_n) \times n = \\ = \sum_{1 \leq i \leq n} P(C|f_i) \end{aligned} \tag{16}$$

Inverting the terms on the right-hand side of Equation 16 using the Bayesian inversion equation (2), we get Equation 17.

$$P(C|F) \times n = \left(\sum_{1 \leq i \leq n} \frac{P(f_i|C) \times P(C)}{P(f_i)} \right) \tag{17}$$

Since $P(C)$ is independent of i, we can write Equation 17 as Equation 18.

$$\begin{aligned} P(C|F) \times n = P(C) \times \left(\sum_{1 \leq i \leq n} P(f_i|C) \right) \\ \times \frac{1}{\sum_{1 \leq i \leq n} P(f_i)} \end{aligned} \tag{18}$$

Finally, dividing by n on both sides, we get:

$$P(C|F) = \left(\sum_{1 \leq i \leq n} P(f_i|C) \right) \times \frac{1}{n} \times \frac{P(C)}{N} \tag{19}$$

Now, we have proved that a certain assumption leads to the Perplexed Bayes posterior probability equation.

However, we have not shown that the assumption itself can be realized in the form of a probability distribution, that probability distributions that satisfy that assumption exist.

Below, we show that probability distributions that satisfy the assumption that *the class C is independent of all features but one, and that none of the features is special* exist and that the constraints do not require that the features be class-conditionally independent.

3.2 Illustration

Here, we provide an example of a distribution that satisfies Equation 10.

Assume the set of all possible categories to be $C = \{c_1, c_2\}$ and the set of features to be $F = \{f_1, f_2\}$ where each of the features f_i is boolean and can take the values $\{true, false\}$.

Equation 10 yields the following equations:

$$\begin{aligned} &P(C = c_1 | f_1 = true, f_2 = true) \\ &\quad = P(C = c_1 | f_1 = true) \\ &\quad = P(C = c_1 | f_2 = true) \end{aligned} \tag{20}$$

$$\begin{aligned} &P(C = c_1 | f_1 = false, f_2 = true) \\ &\quad = P(C = c_1 | f_1 = false) \\ &\quad = P(C = c_1 | f_2 = true) \end{aligned} \tag{21}$$

$$\begin{aligned} &P(C = c_1 | f_1 = false, f_2 = false) \\ &\quad = P(C = c_1 | f_1 = false) \\ &\quad = P(C = c_1 | f_2 = false) \end{aligned} \tag{22}$$

$$\begin{aligned} &P(C = c_1 | f_1 = true, f_2 = false) \\ &\quad = P(C = c_1 | f_1 = true) \\ &\quad = P(C = c_1 | f_2 = false) \end{aligned} \tag{23}$$

$$\begin{aligned} &P(C = c_2 | f_1 = true, f_2 = true) \\ &\quad = P(C = c_2 | f_1 = true) \\ &\quad = P(C = c_2 | f_2 = true) \end{aligned} \tag{24}$$

$$\begin{aligned} &P(C = c_2 | f_1 = false, f_2 = true) \\ &\quad = P(C = c_2 | f_1 = false) \\ &\quad = P(C = c_2 | f_2 = true) \end{aligned} \tag{25}$$

$$\begin{aligned} &P(C = c_2 | f_1 = false, f_2 = false) \\ &\quad = P(C = c_2 | f_1 = false) \\ &\quad = P(C = c_2 | f_2 = false) \end{aligned} \tag{26}$$

$$\begin{aligned} &P(C = c_2 | f_1 = true, f_2 = false) \\ &\quad = P(C = c_2 | f_1 = true) \\ &\quad = P(C = c_2 | f_2 = false) \end{aligned} \tag{27}$$

By equating the right-hand side, you get two sets of equations as shown in Equation 28 and Equation 29.

$$\begin{aligned} &P(C = c_1 | f_1 = true) = \\ &\quad P(C = c_1 | f_1 = false) \\ &\quad = P(C = c_1 | f_2 = true) \\ &\quad = P(C = c_1 | f_2 = false) \end{aligned} \tag{28}$$

$$\begin{aligned} &P(C = c_2 | f_1 = true) = \\ &\quad P(C = c_2 | f_1 = false) \\ &\quad = P(C = c_2 | f_2 = true) \\ &\quad = P(C = c_2 | f_2 = false) \end{aligned} \tag{29}$$

Equation 28 and Equation 29 are satisfied by any values of $P(C = c_1 | f_1 = true)$ and $P(C = c_2 | f_1 = true)$ that are positive and sum to 1.0.

Thus we have shown that distributions that satisfy Equation 10 do exist.

4 Proof 2:

We attempt to show below that the assumption that *the class C is independent of all features but one, and none of the features is special* is compatible with class-conditional feature probability distributions that do not reflect naive independence assumptions.

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.

$$P(c_i) = \sum_F P(c_i, F) = \sum_F P(c_i|F)P(F) \tag{30}$$

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

$$P(c_i) = m_i \times \sum_F P(F) \tag{31}$$

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.

Q.E.D: Thus we have shown that the assumption that *the class C is independent of all features but one, and none of the features is special* is compatible with class-conditional feature probability distributions that do not reflect naive independence assumptions.

We can also demonstrate the independence of classes from features by using the bayesian inverse of the above equations as follows:

The above equations also show that if Equation 28 and Equation 29 are true, then Equation 32 and Equation 33 (below) are also true.

$$\begin{aligned} P(C = c_1|f_1 = true) = \\ P(C = c_1|f_1 = false) \\ = P(C = c_1|f_2 = true) \\ = P(C = c_1|f_2 = false) \\ = P(C = c_1) \end{aligned} \tag{32}$$

$$\begin{aligned} P(C = c_2|f_1 = true) = \\ P(C = c_2|f_1 = false) \\ = P(C = c_2|f_2 = true) \\ = P(C = c_2|f_2 = false) \\ = P(C = c_2) \end{aligned} \tag{33}$$

Moreover, by using Bayesian inversion as shown in Equation 2, we can obtain the following equations from Equation 28 and Equation 29.

$$\begin{aligned} P(C = c_1|f_1 = true) = \\ P(C = c_1) = \\ \frac{P(f_1 = true|C = c_1) \times P(C = c_1)}{P(f_1 = true)} = \\ \frac{P(f_1 = false|C = c_1) \times P(C = c_1)}{P(f_1 = false)} \\ = \frac{P(f_2 = true|C = c_1) \times P(C = c_1)}{P(f_2 = true)} \\ = \frac{P(f_2 = false|C = c_1) \times P(C = c_1)}{P(f_2 = false)} \end{aligned} \tag{34}$$

$$\begin{aligned} P(C = c_2|f_1 = true) = \\ P(C = c_2) = \\ \frac{P(f_1 = true|C = c_2) \times P(C = c_2)}{P(f_1 = true)} = \\ \frac{P(f_1 = false|C = c_2) \times P(C = c_2)}{P(f_1 = false)} \\ = \frac{P(f_2 = true|C = c_2) \times P(C = c_2)}{P(f_2 = true)} \\ = \frac{P(f_2 = false|C = c_2) \times P(C = c_2)}{P(f_2 = false)} \end{aligned} \tag{35}$$

Cancelling $P(C = c_1)$ and $P(C = c_2)$ everywhere, we obtain Equation 36.

$$\begin{aligned} P(f_1|C = c_1) = P(f_1) \\ P(f_2|C = c_1) = P(f_2) \\ P(f_1|C = c_2) = P(f_1) \\ P(f_2|C = c_2) = P(f_2) \\ \ldots P(F|C) = P(F) \end{aligned} \tag{36}$$

The above equations also show that the classes are independent of the features.

Thus we see from the above that the Perplexed Bayesian assumption, though it is an independence assumption, does not assume the class-conditional independence of the features used.

In other words, the features do not have to be class-conditionally independent of one another to satisfy Equation 9.

We illustrate below through examples that class-conditional feature probability distributions that conform to naive independence constraints and class-conditional feature probability distributions that don't both satisfy Equation 9.

4.1 Examples

We shall illustrate the lack of constraints on features with a few examples.

Take two fair coins tossed simultaneously. The probability of either of the coins turning up heads $P(H)$ is 0.5.

If the two coins were independent of each other, the probability of both coins turning up heads $P(H, H)$ would be as depicted in Table 3 (bearing in mind that $P(f_1, f_2|c_1) = P(f_1, f_2)$ by Equation 36).

On the other hand, if the coins were to be welded side by side, so that when one fell heads,

Coin 1 (f_1)	Coin 2 (f_2)	$P(f_1, f_2 \mid c_1)$
H	H	0.25
H	T	0.25
T	T	0.25
T	H	0.25

Table 3: Joint Probabilities for Independent Coins.

the other would as well, the joint probability distribution would be as shown in Table 4 (again bearing in mind that $P(f_1, f_2|c_1) = P(f_1, f_2)$ by Equation 36).

Coin 1 (f_1)	Coin 2 (f_2)	$P(f_1, f_2 \mid c_1)$
H	H	0.5
H	T	0.0
T	T	0.5
T	H	0.0

Table 4: Joint Probabilities for Welded Coins.

Now if we were to use a conditional probability distribution table that satisfied Equation 9, like Table 5, we would find that the probability distribution in Table 3 and the one in Table 4 both yield joint probability distributions that meet all the constraints of Equation 9.

$\mathbf{f}_1$	$\mathbf{f}_2$	$P(c_1 \mid f_1, f_2)$	$P(c_2 \mid f_1, f_2)$
H	H	0.3	0.7
H	T	0.3	0.7
T	T	0.3	0.7
T	H	0.3	0.7

Table 5: Perplexed Bayes Conditional Probabilities.

The joint probability distribution obtained by combining the distribution in Table 3 with Table 5 is Table 6.

The joint probability distribution obtained by combining the distribution in Table 4 with Table 5 is Table 7.

The joint probability distributions in Table 6 and Table 7 both satisfy Equation 9.

For example from Table 7 (the joint probability table computed from features welded together), it can be easily seen that $P(C = c_1) = 0.3$ and that $P(C = c_1|f \in \{H,T\}) = 0.3$ and that $P(C = c_1|f_1 \in \{H,T\}, f_2 \in \{H,T\}) = 0.3$.

The same values of $P(C = c_1) = 0.3$, $P(C = c_1|f \in \{H,T\}) = 0.3$ and $P(C = c_1|f_1 \in \{H,T\}, f_2 \in \{H,T\}) = 0.3$ are obtained from Table 6 where the features are class-conditionally independent of each other.

$\mathbf{f}_1$	$\mathbf{f}_2$	$P(c_1, f_1, f_2)$	$P(c_2, f_1, f_2)$
H	H	0.3/4	0.7/4
H	T	0.3/4	0.7/4
T	T	0.3/4	0.7/4
T	H	0.3/4	0.7/4

Table 6: Joint Probabilities for Independent Features.

$\mathbf{f}_1$	$\mathbf{f}_2$	$P(c_1, f_1, f_2)$	$P(c_2, f_1, f_2)$
H	H	0.3/2	0.7/2
H	T	0	0
T	T	0.3/2	0.7/2
T	H	0	0

Table 7: Joint Probabilities for Welded Features.

So, the Perplexed Bayes assumption, unlike the Naive Bayes assumption, does not forbid complete dependence between the features used in classification.

We show below using simulations that the posterior probabilities of the fully Perplexed Bayes classifier described above tend to the middle (to 0.5) as the number of features increases.

4.2 Perplexed Bayes Simulations

By simulating the posterior probabilities produced by Equation 14, it can be seen that the posterior probabilities are less extreme as compared to the posterior probabilities produced by a Naive Bayes classifier for higher feature counts, as shown in Figure 3, and that they decrease as the number of features increases.

4.3 Generalization

In order to distribute the probabilities more evenly, so that they tend neither to the extremes nor to the middle, we attempted to find a way to mitigate the degree of averaging of the class-conditional feature probabilities.

It appeared possible to find that middle ground between 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 37.

$$PP^{-\mathbf{k}} = (p_1 \times p_2 \times \ldots \times p_n)^{\frac{\mathbf{k}}{n}} \qquad (37)$$

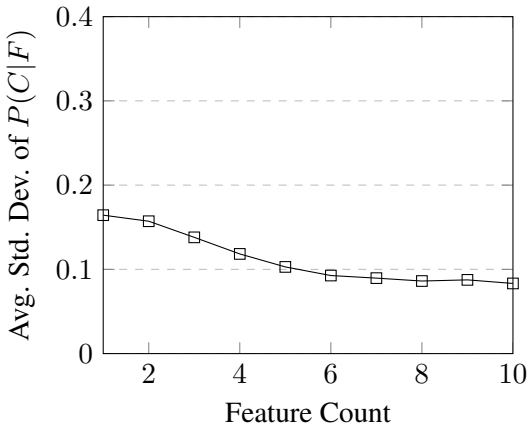

Figure 3: The standard deviation of the posterior probabilities produced by Equation 14 plotted against the number of features simulated.

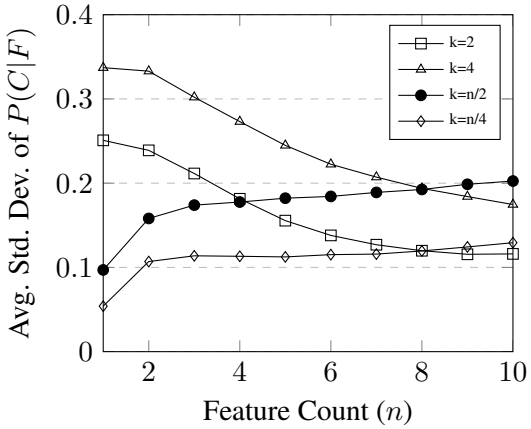

Figure 4: The standard deviation of the posterior probabilities produced by Equation 38 plotted against the number of features simulated, for different values of the attenuation coefficient k.

By substituting Equation 37 in Equation 2, the posterior probablity equation for the generalized case of the Perplexed Bayes classifier can be written as shown in Equation 38.

$$P(C|F) = \left(\prod_{1 \leq i \leq n} P(f_i|C) \right)^{\frac{\mathbf{k}}{n}} \times \frac{P(C)}{P(F)} \quad (38)$$

Again, a simulation showed that different attenuation coefficients might be needed to hold the standard deviation of the posterior probability at a desired value.

4.4 Generalized Form Simulations

Plots of the standard deviation of the posterior probabilities of the generalized Perplexed Bayesian classifier computed using Equation 38 for different values of the attenuation coefficient k are shown in Figure 4.

The plots show that the magnitude of the standard deviation can be increased or decreased, for a given set of features, by picking a suitable value of k.

In each of the above simulations, each point in the graph was computed by averaging the results of 1000 experiments.

5 Proof 3:

In this section, we attempt to prove that the classification decisions of the Perplexed Bayes classifier (with an approximation to reduce the weightage given to the prior) are identical to the classification decisions of the Naive Bayes classifier.

A comparison of Equation 4 with Equation 8 shows that the class conditional feature probabilities have a far greater say in the outcome of classification in the Naive Bayes classifier than in the fully Perplexed Bayes classifier.

In the former, as the number of features increases, the prior probability distribution has an increasingly negligible impact on the posterior probability distribution since the prior probability is being combined with an increasing number of probabilities each representing a feature probability. So, if there are ten features being used in classification, the features have ten times the influence on the posterior probability distribution than the prior.

In the case of the Perplexed Bayes classifier, however, the likelihood is represented by a geometric average of the class conditional feature probabilities. So, the features and the prior have an equal say in the computation of the posterior probability distribution.

Simulations of classification accuracy using generated joint probability distribtions show that giving the prior equal weightage is not conducive to classification accuracy. The accuracy ratios in Figure 5 indicate that the fully Perplexed Bayes classifier has significantly lower accuracy than the Naive Bayes classifier.

So two approximations, both of which can reduce the role of the prior in classiciation, are presented below. The first approximation involves flattening the prior so that it doesn't favour one category over another by replacing $P(C)$ by a con-

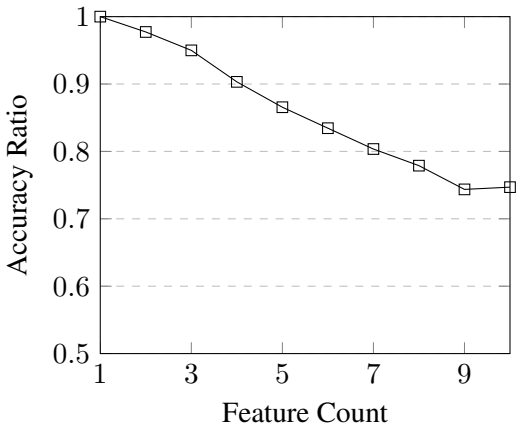

Figure 5: Perplexed Bayes / Naive Bayes accuracy ratio in simulations against feature counts (each accuracy ratio computed by averaging the results of one thousand simulations).

stant in Equation 38. The result is Equation 39 where N' is the normalizing factor.

$$P(C|F) = \left(\prod_{1 \leq i \leq n} P(f_i|C) \right)^{\frac{\mathbf{k}}{n}} \times \frac{1}{N'} \quad (39)$$

This is equivalent to giving the prior no weightage at all.

However, it is seen, in simulations, that a Perplexed Bayes classifier with the prior removed from consideration as shown in Equation 39 is as accurate as the Naive Bayes classifier (there is no significant improvement or deterioration in accuracy averaged over 1000 simulations).

However, it is possible to obtain the same accuracy as a Naive Bayes classifier and yet retain the posterior probability characteristics of the Perplexed Bayes classifier using the approximation shown in Equation 40.

$$P(C|F) = \frac{\left(\left(\prod_i P(f_i|C) \right) \times P(C) \right)^{\frac{\mathbf{k}}{n+1}}}{N''} \quad (40)$$

It can be seen from Equation 40 that the posterior probability of a Perplexed Bayes classifier computed in this way is nothing but the $k/(n+1)$th root of the posterior probability of a Naive Bayes classifier. So, their accuracies must be the same, because **if a positive real number a is greater than b, then a^k/N must also be greater than b^k/N where k and N are constants**.

Q.E.D.: Thus we have shown that the classification decisions of the Perplexed Bayes classifier with an approximation to reduce the weightage given to the prior are identical to the classification decisions of the Naive Bayes classifier.

6 Conclusions

We have shown that it is possible to build a classifier that we call the Perplexed Bayes classifier, that can (with an approximation) make classification decisions that are identical to those of a Naive Bayes classifier.

We have shown that if a certain assumption that the Perplexed Bayes classifier's posterior probability equation can be derived from holds, then the Perplexed Bayes classifier can be shown to not assume that the features used are class-conditionally independent.

We have also shown experimentally in other work (Carlos, 2015) that a Perplexed Bayes 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.

All the above suggests that naive independence assumptions do not necessarily contribute to the accuracy of the Naive Bayes classifier, and possibly have a deleterious effect on posterior probability estimates instead.

7 Future Work

Since the Perplexed Bayes correction rectifies a shortcoming in the multiplicative combination of class-conditional feature probabilities, it is believed that it might be possible to apply with the attendant benefits of better posterior probability estimates, to any system where the product operator is used to combine probabilities, including hierarchical Bayesian classifiers, Probabilistic Graphical Models and Hidden Markov Models with multiple emissions. Experimental studies of the performance of such systems before and after the application of the Perplexed Bayes correction would be very interesting.

It does not follow from the three proofs that the Perplexed Bayes classifier does not make independence assumptions, because the proofs derive the posterior probability equation from the assumptions that imply the absence of the naive assumption, but not the other way round.

Further work to establish or reject the hypothesis that the Perplexed Bayes classifier does not make the same naive independence assumptions as the Naive Bayes classifier would be essential to understanding why the Perplexed Bayes classifier produces better posterior probabilities on the data set that we ran the experiments on. Experiments to test the calibration of the Perplexed Bayes classifier on data sets with very large numbers of features are also needed.

References

Cohan Sujay Carlos. 2015. Perplexed Bayes Classifier. *12th International Conference on Natural Language Processing (ICON-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.

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.