L2 Naïve Bayes Classifier

Greg Ridgeway

2025-01-04

Table of contents

1 Introduction

Consider the aim of wanting to compute the probability that an individual is likely to reoffend (perhaps violently) during some period (e.g., while awaiting trial, in consideration for parole, or when setting probation terms). In statistical notation, this is $P(Y = 1|\mathbf{x})$, where Y is the indicator for recidivism, and $\mathbf{x}$ is the vector of features associated with the subject. Bayes' theorem states that:

$$P(Y = 1|\mathbf{x}) = \frac{P(\mathbf{x}|Y = 1)P(Y = 1)}{P(\mathbf{x})}$$

It is convenient to rewrite the naïve Bayes classifier as the odds that $Y = 1$:

$$\frac{P(Y = 1|\mathbf{x})}{P(Y = 0|\mathbf{x})} = \frac{\frac{P(\mathbf{x}|Y=1)P(Y=1)}{P(\mathbf{x})}}{\frac{P(\mathbf{x}|Y=0)P(Y=0)}{P(\mathbf{x})}} = \frac{P(\mathbf{x}|Y = 1)P(Y = 1)}{P(\mathbf{x}|Y = 0)P(Y = 0)}$$

This says that we need to know the rate at which recidivism occurs in the population, $P(Y = 1)$, and how often it does not occur, $P(Y = 0)$. We also need the probability that a recidivist has the set of features $\mathbf{x}$, and the probability that a non-recidivist has features $\mathbf{x}$.

The difficulty in implementation occurs when the dimension of $\mathbf{x}$ is large. In that case, the naïve Bayes classifier has seen widespread use. It forms the basis of the system described in Spiegelhalter and Knill-Jones (1984). The naïve Bayes assumption is that

$$P(\mathbf{x}|Y = y) = P(x_1|Y = y) \cdots P(x_d|Y = y)$$

In other words, the components of the feature vector $\mathbf{x}$ are independent given y. For example, the assumption says that given that a person recidivates, knowing that they were being held on a violent charge gives you no additional information about their employment. Although this assumption does not always hold, the naïve Bayes model has shown itself to be consistently robust to violations in the conditional independence assumption.

There are several benefits to using such a model.

- Estimating the components of the model requires a single scan of the dataset
- Prediction for a new subject is linear in the dimension of the feature vector
- The model inferences are transparent through the use of evidence balance sheets. These devices, explained later, itemize the observed features that have values that favor a particular decision and in a separate column itemize the observed features that are against the decision
- Both estimation and prediction can handle missing data without special modifications

2 Prediction

Prediction for a new subject is efficient. For numerical reasons as well as interpretation, we often compute the prediction rule on the log-odds scale. On the log-odds scale the prediction

rule is often called the "weight of evidence" (WOE). The following derivation shows that evidence for Y accumulates additively on the log-odds scale.

$$\begin{aligned}
\text{WoE} &= \log \frac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} \\
&= \log \frac{P(Y=1)P(\mathbf{x}|Y=1)}{P(Y=0)P(\mathbf{x}|Y=0)} \\
&= \log \frac{P(Y=1)}{P(Y=0)} + \log \frac{P(\mathbf{x}|Y=1)}{P(\mathbf{x}|Y=0)} \\
&= \log \frac{P(Y=1)}{P(Y=0)} + \log \frac{P(x_1|Y=1)}{P(x_1|Y=0)} + ... + \log \frac{P(x_d|Y=1)}{P(x_d|Y=0)} \\
&= w_0 + w_1(x_1) + ... + w_d(x_d)
\end{aligned}$$

The w_j are the weights of evidence described by Good (1965). Madigan, Mosurski, and Almond (1996) and Becker, Kohavi, and Sommerfield (1997) further discuss and develop the explanatory strengths of weights of evidence. The prediction rule derivation here shows that the total weight of evidence is a sum of the weights of evidence of each component. On the log-odds scale, a positive total weight of evidence equates to $P(Y=1|\mathbf{x}) > \frac{1}{2}$ and a negative total weight of evidence equates to $P(Y=1|\mathbf{x}) < \frac{1}{2}$. Computing the prediction requires a sum of $d+1$ weights each of which can be stored in a lookup table for constant time access. The next section discusses how to estimate the necessary weights of evidence to utilize this method.

3 Estimation

If we have data then estimation requires a single scan of the dataset. We need to estimate the prior rate of Y, $P(Y=y)$, and the conditional probabilities of each of the features, $P(x_j|Y=y)$. The usual estimate of $P(Y=y)$ is simply the fraction of observations in the dataset for which Y takes on the value y, the maximum likelihood estimator for p.

We can estimate the remaining terms as

$$\hat{P}(x_j = x|Y=y) = \frac{\sum(x_{ij}=x)(y_i=y)}{\sum(y_i=y)} \tag{1}$$

When the dataset is small or there are some values of x that rarely occur, analysts frequently use the Laplace-corrected frequency.

$$\hat{P}(x_j = x|Y=y) = \frac{1+\sum(x_{ij}=x)(y_i=y)}{m_j+\sum(y_i=y)}$$

where m_j is the number of possible values that x_j can have. For example, if x_j is a 0/1 variable then $m_j = 2$.

This estimation step **is** machine learning. As new observations accumulate, we can update the probabilities in Equation 1, which directly feed into the weights of evidence.

The naïve Bayes classifier is particularly easy to estimate and update. It is certainly the simplest machine learning approach that we will encounter. It is particularly useful to start our exploration of machine learning with the naïve Bayes classifier because it sets up the issues that we will regularly encounter.

Characteristic	Naïve Bayes
What is the structure of the machine learning method?**	Additive on the log odds scale
What is the objective?**	Produce good probabilities of class labels
How does it learn from data?**	Simple calculation of probabilities like Equation 1
How computationally difficult is it to learn from data?**	Easy, involving a single scan of the dataset
Is the method interpretable?**	Yes. Simple addition of weights of evidence
Can it handle different types of data sources?	Limited to categorical data. Continuous features need to be discretized. Easily handles missing values
Can it uncover the "true" relationship?	No. It can only get to a close linear approximation on the log odds scale

4 Example: NIJ Recidivism Challenge

Table 2 shows the weights of evidence. Here the weights of evidence are multiplied by 100 to make them easier to read.

Table 2: NIJ Challenge weights of evidence

Feature	Value	WoE
Prior	NA	31
Gender	F	-49
	M	7
Age at Release	18-22	63
	23-27	38
	28-32	17
	33-37	-1
	38-42	-16
	43-47	-30

Feature	Value	WoE
	48 or older	-67
Education Level	At least some college	-51
	High School Diploma	12
	Less than HS diploma	11
Prior Conviction Episodes (Viol)	false	-7
	true	16
Prison Offense	NA	-1
	Drug	-15
	Other	15
	Property	26
	Violent/Non-Sex	-15
	Violent/Sex	-106
Prison Years	1-2 years	11
	Greater than 2 to 3 years	-9
	Less than 1 year	27
	More than 3 years	-47

A positive $w_j(x_j)$ implies that the state of x_j is evidence in favor of $Y = 1$ and a negative $w_j(x_j)$ is evidence in favor of $Y = 0$ (assuming equal costs for misclassification). After obtaining estimates of the probabilities we can construct an evidence balance sheet for a newly observed subject as described in Spiegelhalter and Knill-Jones (1984). From the features of the new subject we can assemble those pieces of features with weights of evidence that favor recidivism and those features associated with no recidivism. Since the weights are additive we can simply sum the weight totals for a full accounting of the evidence bearing on the particular subject. Table 3 shows the weights of evidence, each multiplied by 100 for readability, for a specific case.

Table 3: Evidence balance sheet

Feature	WoE	Feature	WoE
Prior	31		
Offense = Property	26	Education = At least some college	-51
Years = 1-2 years	11	Gender = F	-49
		Age at Release = 38-42	-16
		Prior Conviction Viol = false	-7
Total positive weight	68	Total negative weight	-123
		Total weight of evidence	-55
		Probability =	0.37

The conversion from total weight of evidence to probability is

$$p = \frac{1}{1 + \exp(-\text{WoE})}$$

or by the conversion table shown in Table 4.

Table 4: Table: Conversion from probability to total weight of evidence

Probability	Total Weight of Evidence
10%	-220
20%	-139
30%	-85
40%	-41
50%	0
60%	41
70%	85
80%	139
90%	220

5 Missing data

Missing data is common. Even though we may be interested in 20 different pieces of evidence, for a particular subject we may have information on only three of the features. The naïve Bayes classifier can still handle such a scenario without modification. For features that are frequently missing, we may allow that categorical feature to have a missing level and compute $w_j(NA) = \frac{P(x_j=NA|Y=1)}{P(x_j=NA|Y=0)}$. Otherwise, the naïve Bayes assumption allows us to trivially skip unobserved features. Let's say we have x_1, x_2, x_3, but for a particular case x_3 is missing. We can simply predict using

$$\frac{P(Y=1|x_1,x_2)}{P(Y=0|x_1,x_2)} = \frac{P(Y=1)}{P(Y=0)}\frac{P(x_1|Y=1)}{P(x_1|Y=0)}\frac{P(x_2|Y=1)}{P(x_2|Y=0)}$$

The naïve Bayes classifier is unconcerned that x_3 is unavailable.

6 Evaluating performance

Typically there is no single metric that summarizes the performance of a classifier. This section will review several of the most common ways to describe a classifier's performance.

Of fundamental importance is evaluating the classifier on data that was not used in training the dataset. We will always evaluate "out-of-sample performance," performance on data held

back from the model fitting process, sometimes called a "validation dataset." Particularly for more complex machine learning methods, they can become "overfit" to a training dataset to the point that they do not predict well on a validation dataset.

6.1 Misclassification rate and misclassification cost

The most straightforward performance measure is the misclassification rate, the fraction of cases for which the predicted value does not equal the true value.

Misclassification rate

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2}$$

Baked into this calculation is some decision on where to set the threshold for predicting $\hat{y} = 1$. If we decided that $\hat{y} = I(\hat{p} > \frac{1}{2})$, equivalent to believing that false positives and false negatives had equal costs, then we could compute the misclassification rate for the Georgia parolee data as

```
datRecid |>
  group_by(Training_Sample) |>
  summarize(misclass=mean((Recidivism_Within_3years=="false" & p>0.5) |
                          (Recidivism_Within_3years=="true" & p<0.5)))
```

This breaks down the misclassification rate separately for training data and validation data. Note that the classification error on the training data is slightly lower than than on the validation data (but really not by much in this example).

Sample	Misclassification Rate
Training	0.3641003
Validation	0.3692840

We may also compare the **false positive** and **false negative** rates. The false positive rate is the fraction among those who really are 0s, but we in error predict them to be 1s. That is, by mistake we labeled them as a 1. False negatives are those cases we mistakenly label as a 0. The false negative rate, therefore, is the fraction of cases that are truly 1s that we predict erroneously to be 0s.

False positive rate

$$\frac{\sum_{i=1}^{n} I(y_i = 0 \cap \hat{y}_i = 1)}{\sum_{i=1}^{n} I(y_i = 0)} \tag{3}$$

This is also known as a "Type I error"
"Specificity" is $1 -$ false positive rate

False negative rate

$$\frac{\sum_{i=1}^{n} I(y_i = 1 \cap \hat{y}_i = 0)}{\sum_{i=1}^{n} I(y_i = 1)} \tag{4}$$

This is also known as a "Type II error"
"Sensitivity" or "recall" is $1 -$ false negative rate

6.2 Receiver Operating Characteristic (ROC)

It is easy to make either the false positive or false negative rates 0. We can just predict everyone to be a 0 to get eliminate all of our false positive errors. Or we can predict everyone to be 1s and eliminate all of our false negative errors. Clearly, there is a trade-off in these two kinds of errors. Reducing one invariably results in increasing the other. The Receiver Operating Characteristic, or ROC, curve shows this tradeoff.

To construct the ROC curve, we vary the probability threshold used to classify a case as a 1. For numerous values of the threshold, we compute the false positive rate and the true positive rate (1-false negative rate). Along the x-axis we plot the false positive rate and along the y-axis we plot the false negative rate. Figure 1 shows the result. The red dot in Figure 1 corresponds to the decision $\hat{y} = I(p > 0.5)$, the equal misclassification cost decision rule.

Different machine learning methods can produce different ROC curves. Ideally we would like it to be pushed well up into the top left corner, low false positives with high true positives.

At Area Under the ROC Curve (AUC) is a common summary measure for overall performance, rather than judging the classifier's performance at only one threshold the way misclassification rate does. AUC turns out to be equal to the probability that the classifier ranks a random selected case with $y_i = 1$ to have higher probability than a random selected $y_i = 0$ case.

In R, the `pROC` package calculates AUC and displays ROC curves.

```
library(pROC)
nbROC <- roc((Recidivism_Within_3years=="true")~p, data=datRecid)
nbROC$auc
```

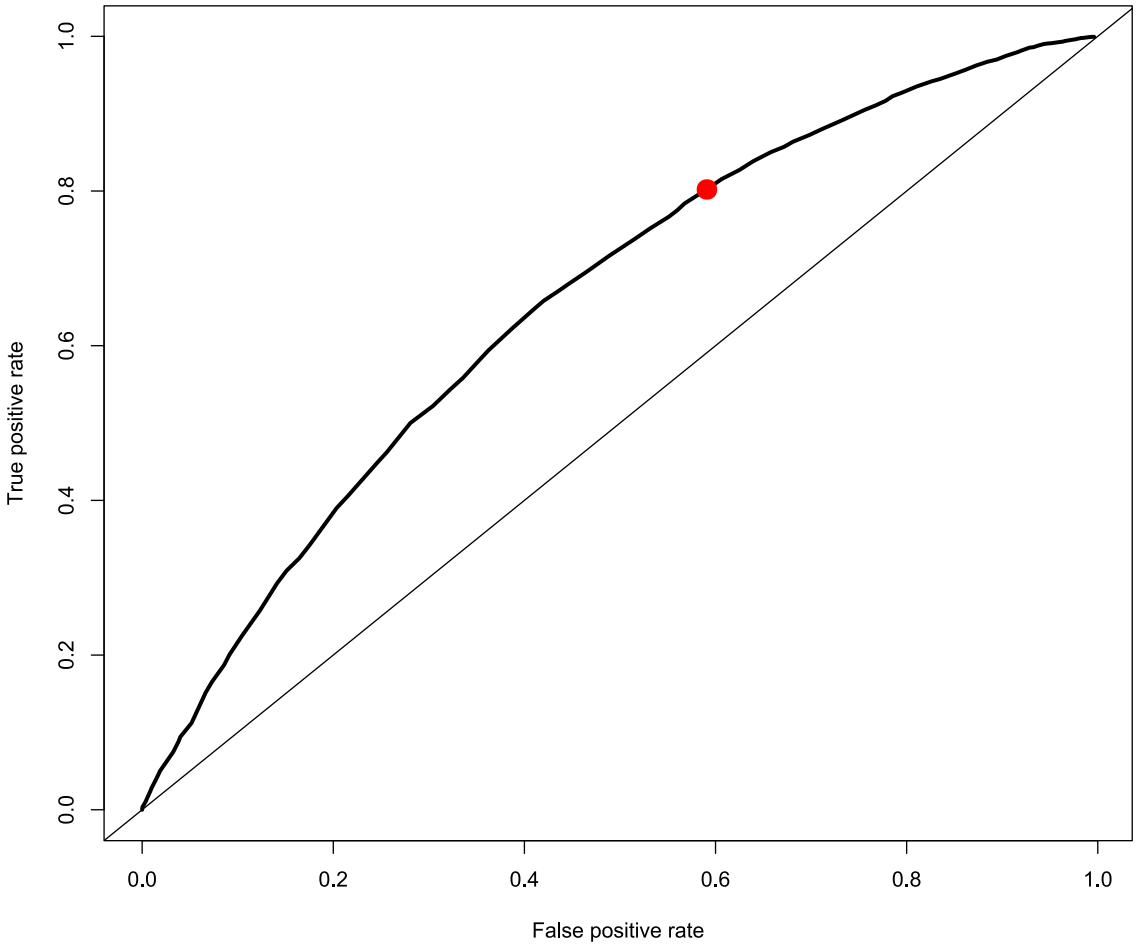

Figure 1: ROC curve for a naïve Bayes classifier on Georgia parolee data

```
# note x-axis is specificity, 1-FPR, and labeled from 1 down to 0
plot(nbROC)
```

6.3 Calibration

If we consider all of the parolees that we predicted to have a 70% chance of reoffending, then if our probabilities are meaningful 70% of those parolees should reoffend and 30% should not. Calibration gets at this concept. Are our predicted probabilities meaningful as probabilities? We will explore this characteristic of our naive Bayes classifier graphically.

To place the red dots in Figure 2 I created bins for the predicted probabilities, $(0.1, 0.2], \ldots, (0.8, 0.9]$. For each parolee with predicted probability of reoffending in $(0.1, 0.2]$ I computed the fraction that actually reoffended. In reality 20% of the parolees with predicted probabilities in this range reoffended. So the calibration of the predicted probabilities in this range is a little off. The predicted probabilities are a little too low. I repeated this process for each of the other bins. The eight red dots in Figure 2 show the actual rate of reoffending within each bin.

The blue curve in Figure 2 is a smooth version of this using natural splines, which we will cover later. If the probabilities from the naïve Bayes classifier were perfectly calibrated, then they would fall along the black diagonal line. It is not perfectly calibrated, but also the predicted probabilities are off by at most 0.05.

It is possible to calibrate the probabilities by inverting the blue curve. That is, if you want to know about parolees with a 30% chance of reoffending, then look up 0.30 on the vertical axis and find the associated predicted probability along the x-axis. This recalibrates the probabilities so that they match with the observed reoffense rates.

It is trivial to obtain a perfectly calibrated prediction. In the dataset, 57.8% of the training sample parolees reoffended within 3 years. So, predict everyone to reoffend with probability 0.578, a perfectly calibrated probability. Clearly, calibration as a performance measure on its own is not useful as such a predictive model has no ability to separate parolees who have higher or lower risk. Like all of the other measures described here, improving performance in one aspect sometimes decreases performance in another aspect.

7 Summary

For more information on the topic, the article by Spiegelhalter and Knill-Jones (1984) contains a lengthy description of decision systems in use in the early 1980s. In addition their section 4 offers an extensive discussion on the use of the naïve Bayes model in decision systems with various special cases. In particular they discuss the case of branching questions, those features that would be further inspected if some other feature turned up positive (e.g. If the subject

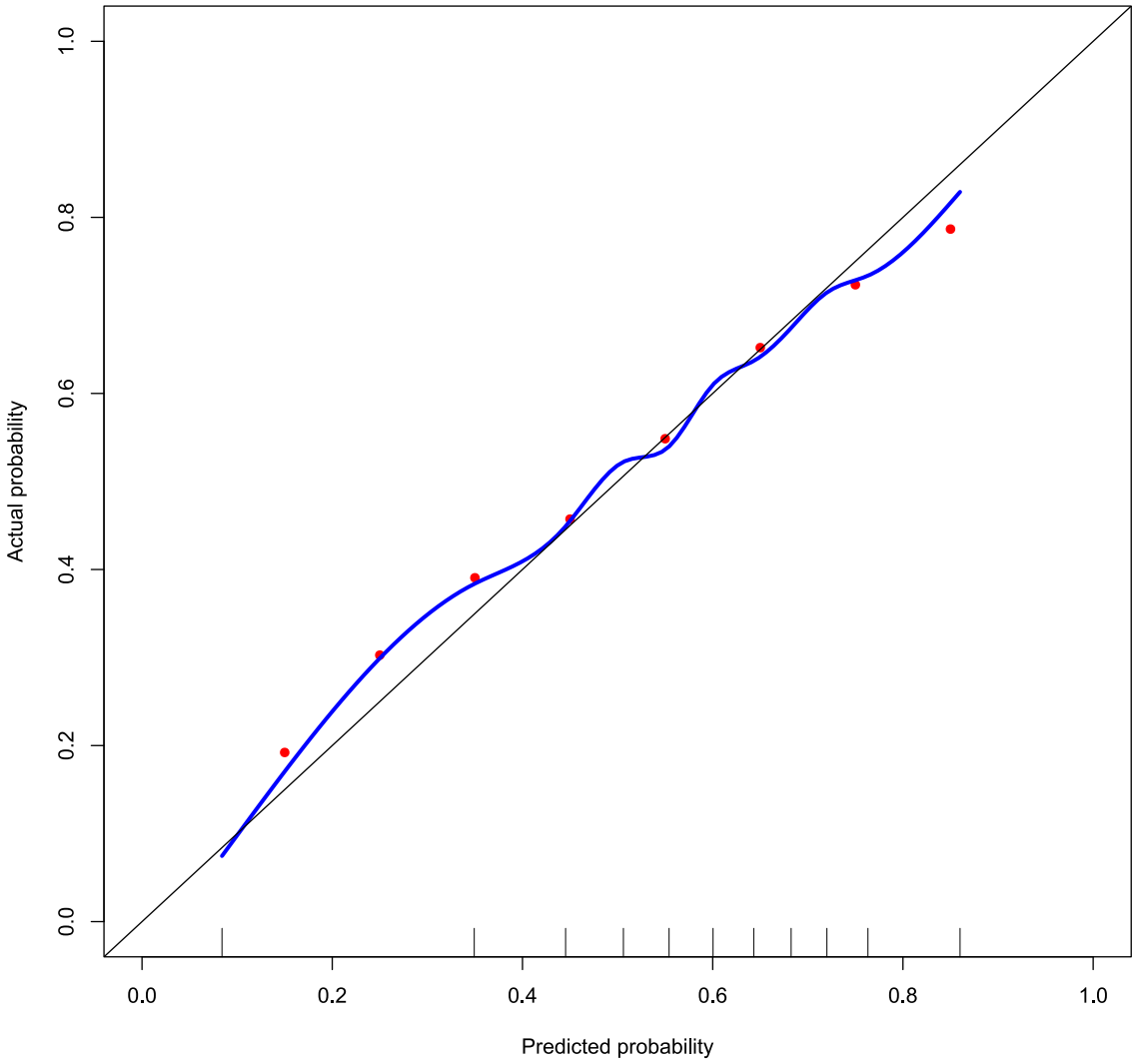

Figure 2: Calibration curve for a naïve Bayes classifier on Georgia parolee data

indicated that they had pain, where is the location of the pain?). In addition they discuss enhanced estimates of the weights of evidence that offer improved predictive performance. Although their work is many decades old, the naïve Bayes classifier is still used as a competitive classifier due to its robustness and simplicity. See Domingos and Pazzani (1997) for example.

References

Becker, B., R. Kohavi, and D. Sommerfield. 1997. "Visualizing the Simple Bayesian Classifier." In *KDD 1997 Workshop on Issues in the Integration of Data Mining and Data Visualization*.

Domingos, P., and M. Pazzani. 1997. "On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss." *Machine Learning* 29: 103–30.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press.

Madigan, D., K. Mosurski, and R. G. Almond. 1996. "Explanation in Belief Networks." *Journal of Computational and Graphical Statistics* 6: 160–81.

Spiegelhalter, D. J., and R. P. Knill-Jones. 1984. "Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, with an Application in Gastroenterology (with Discussion)." *Journal of the Royal Statistical Society (Series A)* 147: 35–77.