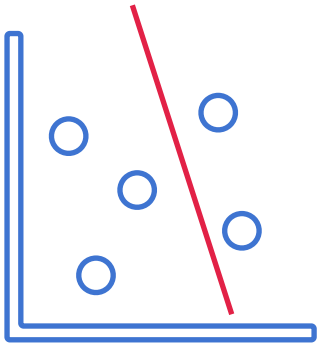
Naive Bayes Classifier

B. Tech. Sem VI Computer Engineering

Brijesh Bhatt
Computer Engineering Department
30 December 2022

Prelude



Classification is a task of identifying correct category label for the given input observation. In binary classification there are only two category labels to choose from. For example, assigning a given email to the "spam" or "non-spam" class is an example of binary classification problem. On the other hand, if there are more than two category labels or class labels to choose from, such classification is known as multinomial classification. For example assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.).

Usually, the classification algorithms assign exactly one category label to an example. Such classifiers are known as Hard Classifier. In real world scenario often a thing can belong to more than one category with varying degree of belongingness. In such situations we need soft classifiers which can produce more than one category labels for the given input observation. In particular, a soft classification rule generally estimates the class conditional probabilities explicitly and then makes the class prediction based on the largest estimated probability. In contrast, hard classification bypasses the requirement of class probability estimation and directly estimates the classification boundary.

A probabilistic classifier is a classifier that is able to predict, given an observation of an input, a probability distribution over a set of classes, rather than only outputting the most likely class that the observation should belong to. Naive Bayes classifiers are a family of simple probabilistic classifiers. They are among the simplest Bayesian network models.

1. Introduction

"All that glitters is not gold"

In a classification task we identify the correct class for the given object. Humans, as the intelligent being, do such classification all the time with a remarkable ease. For example, we divide house hold items into various categories like, kitchenware, drawing-room furniture, etc. While eating food we immediately identify the test, e.g., sweet, sour, etc. We look at the metal object and identify which metal its is made up of. Here, the characteristics or information that we use to classify objects are called features or attributes; and the category that we assign to the object is called class label. For example, "Glitter" and "colour" can be two indicative features to identify a gold ornament. So, we can say that, "Gold" metal causes the ornaments to "glitter". In other words, "Gold" is the cause which creates the effect of "Glittering" and "Golden" colour. [another example of cause and effect is, "Corona is the cause which creates effects of headache and fever"].

In probabilistic setting we can calculate the probability of "Glittering" given that the ornament is made up of "Gold" as P(Glittering / Gold) and similarly we can define P(Golden Colour / Gold). Note that, here "Gold" is the cause which creates effects of "Glittering" and "Golden".

When we look at an ornament and try to figure out wether it is made up of gold or not, we are doing a classification tasks. Here, we look at the features like "Glittering' and "Golden" and try to predict if the ornament is made up of "Gold"? So here, we are looking at the effect and try to find out the cause. The Probabilistic question that we are asking is, P(Gold / Glittering & Golden). To answer this "flipped question", we can make use of the prior knowledge of gold and the knowledge of how gold makes an ornament glitter.

The probabilistic model to solve the problem would be something like this,

$$P(Gold/Glitter, Golden) = \frac{P(Gold, Glitter, Golden)}{P(Glitter, Golden)}$$

This can be solved using Bayes Rule. The following section provides a brief overview of probability concepts required to further discussion.

2. Probability Basics

Marginal Probability: The probability of an event irrespective of the outcomes of other random variables, e.g. P(A)

A	1	0	1	1	0	0	1	0	1
B	1	1	0	0	1	1	0	0	1

For the table shown above, assuming that A is an independent random variable, the marginal probability of A, P(A), can be calculated as follows,

$$P(A = 1) = \frac{No\ of\ Examples\ in\ which\ A = 1}{Total\ Examples}$$

$$\therefore P(A = 1) = \frac{5}{9}$$

Similarly, marginal probability P(B =1) = 5/9

Conditional Probability: Conditional Probability is defined as the probability of one (or more) event given the occurrence of another event, e. g. P(A given B) or P(A / B).

$$P(A = 1/B = 1) = \frac{No\ of\ times\ A = 1\ and\ B = 1}{No\ of\ times\ B = 1}$$

$$\therefore P(A = 1/B = 1) = \frac{2}{5}$$

Joint Probability: Joint probability is calculated when we want to study two or more simultaneous events. For example, “What is the probability of both A & B are 1?”, P(A=1, B=1).

If A & B are independent random variable, the joint probability P(A=1, B=1) can be calculated as,

$$P(A = 1,B = 1) = P(A = 1) \times P(B = 1)$$

If A & B are dependent random variables, then the joint probability P(A=1, B=1) can be calculated as,

$$P(A=1,B=1) = P(A=1/B=1) \times P(B=1)$$

OR

$$P(A=1,B=1) = P(B=1/A=1) \times P(A=1)$$

For the example data shown in the table, $P(A=1/B=1) = \frac{2}{5} \times \frac{5}{9}$

3. Naive Bayes Classifier

Naive Bayes classifier is a probabilistic supervised learning algorithm which calculates the posterior probability $P(y/x_1, x_2, \ldots, x_n)$, where y is the class label to be predicted and $x_1, x_2, \ldots, x_n$ are the observed feature variables. In other words, Naive Bayes classifier decides outcome of Y based on the observed values of features $X_1, X_2, \ldots, X_n$. $P(y/x_1, x_2, \ldots, x_n)$ is calculated by applying Bayes' theorem and "Naive" assumption of conditional independence between every pair of features given the value of class label, hence the classifier is named Naive Bayes classifier.

Using the definition of conditional probability and joint probability discussed in previous section, Bayes theorem can be stated as follows,

The probability of outcome Y given the observation X is defined as, $P(Y/X) = \frac{P(X,Y)}{P(X)}$. As we know, the joint probability of X & Y is defined as, $P(X,Y) = P(X/Y) \times P(X)$. Hence,

$$P(Y/X) = \frac{P(X/Y) \times P(Y)}{P(X)}$$

The above equation is known as Bayes Theorem or Bayes Rule. ***Here,*** $P(X/Y)$ ***is called Likelihood, and*** $P(Y)$ ***and*** $P(X)$ ***are called class prior and predictor prior, respectively.***

For n feature variables, X1… Xn the Bayes rule can be rewritten as,

$$P(y/x_1, x_2 \ldots . x_n) = \frac{P(x_1, x_2 \ldots x_n/y)P(y)}{P(x_1, x_2 \ldots . x_n)}$$

Probability $P(x_1, x_2, \ldots, x_n/y)$ can be calculated using the chain rule as follows,

$$P(x_1, x_2, \ldots, x_n/y) = P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \times P(y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

To expand and solve the above equation becomes quite a cumbersome task. The calculations can greatly be simplified if the features variables are independent of each other. So we make a Naive assumption that features variables are independent and identically distributed (IID Assumption). With i.i.d. assumption the $P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y)$,

Hence,

$$P(x_1, x_2, \ldots, x_n/y) = P(x_1/y) \times P(x_2/y) \ldots P(x_n/y) \times P(y)$$

$$\therefore P(y/x_1, x_2 \ldots . x_n) = \frac{P(x_1/y)P(x_2/y) \ldots P(x_n/y)P(y)}{P(x_1, x_2 \ldots . x_n)}$$

$$\text{Here, } P(x_1, x_2 \ldots . x_n) = \sum_{\forall y} P(x_1, x_2 \ldots . x_n, y)$$

$$\text{And } P(x_1, x_2 \ldots . x_n, y) = P(x_1, x_2 \ldots . x_n/y)P(y)$$

4. Training a Naive Bayes classifier

As mentioned in the previous section the posterior probability of the class label can be predicted using the likelihood and prior probabilities using the following formula.

$$\therefore P(y/x_1, x_2 \ldots . x_n) = \frac{P(x_1/y)P(x_2/y) \ldots P(x_n/y)P(y)}{P(x_1, x_2 \ldots . x_n)}$$

Likelihood P(x/y) and class label p(y) are known as the model parameters which we aim to learn through training. For example, consider the following data set of 8 example, which contains two features X1 & X2 and the class label Y.

R. V.	1	2	3	4	5	6	7	8
X_1	1	1	1	1	0	0	0	0
X_2	1	0	1	1	1	0	1	0
Y	1	1	1	0	1	1	0	0

From the data we can calculate the conditional probability table as shown below,

Conditional Probability Tables
Class Prior : P(y)
Likelihood : P(x1/y), P(x2/y)

	Y=0	Y=1
P(y)	3/8	5/8

$P(x_1/y)$	Y=0	Y=1
X_1=0	2/3	2/5
X_1=1	1/3	3/5

$P(x_2/y)$	Y=0	Y=1
X_2=0	1/3	2/5
X_2=1	2/3	3/5

This conditional probability table is learned through the training and can be used to answer any inference query to perform the classification.

4. Inference

Once the Naive Bayes Classifier is trained and Conditional Probability Tables are learned, the classifier can be used to make predictions. The classifier can predict the class label for the given features/attribute value. The prediction is done based on statistical inference, where the posterior probability of class label is calculated using the prior probability and likelihood.

The following example shows the inference on the Naive Bayes classifier discussed in the previous section.

Inference : What is the probability of Y=1, given that X1=1 & X2=1?

$$P(Y = 1/X_1 = 1, X_2 = 1)?$$

$$P(Y = 1/X_1 = 1, X_2 = 1) = \frac{P(X_1 = 1, X_2 = 1, Y = 1)}{P(X_1 = 1, X_2 = 1)}$$

$$= \frac{P(X_1 = 1, X_2 = 1/Y = 1)P(Y = 1)}{P(X_1 = 1, X_2 = 1)}$$

$$= \frac{P(X_1 = 1/Y = 1)P(X_2 = 1/Y = 1)P(Y = 1)}{P(X_1 = 1, X_2 = 1)}$$

$$A = P(X_1 = 1/Y = 1)P(X_2 = 1/Y = 1)P(Y = 1)$$
$$B = P(X_1 = 1, X_2 = 1)$$

Computation of Numerator :

$$A = P(X_1 = 1/Y = 1)P(X_2 = 1/Y = 1)P(Y = 1)$$
$$= (3/5)(3/5)(5/8)$$
$$= (9/40)$$

Computation of Denominator :

$$B = P(X_1 = 1, X_2 = 1)$$

$$= P(X_1 = 1, X_2 = 1, Y = 1) + P(X_1, X_2 = 1, Y = 0)$$

$$= P(X_1 = 1/Y = 1)P(X_2 = 1/Y = 1)P(Y = 1)$$
$$+P(X_1 = 1/Y = 0)P(X_2 = 1/Y = 0)P(Y = 0)$$

$$= (3/5)(3/5)(5/8) + (1/3)(2/3)(3/8)$$
$$= 9/40 + 1/12$$
$$= 0.31$$

$$P(Y = 1/X_1 = 1, X_2 = 1) = \frac{9/40}{9/40 + 1/12}$$

$$= 0.725$$

So we can say, “Y=1, with a probability of 0.725”. When we mention probability with the outcome, it is known as **soft classification**. In **Hard classification**, we say Y = 1, if the probability is greater than 0.5 and Y = 0 otherwise.

5. Exercise

(A) State true or false and justify your answer, “Predictor prior probability does not impact the inference in the Naive Bayes Classifier”

(B) What will be the Conditional Probability Tables learned by the Naive Bayes Classifier for the data set shown below?

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

(C) For the Naive Bayes Classifier trained in the question B, answer the following inference questions,

- (A) Would you Play Golf if the outlook is Rainy, Temp is Cool, Humidity is High and Windy is True?
- (B)Would you Play Golf is the outlook is Overcast, Temp Cool, Humidity is Normal and Windy is False?

(D) Implement the Naive Bayes Classifier in Python using SciKit Library. Compare “Categorical”, “Multinomial” and “Gaussian” Classifiers.