



Naïve Bayes Classifier

A probabilistic Approach

Naw varsha Pipada

Asst. Prof., dept. of CSE

Engineering College Bikaner



Introduction

- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes model is easy to build and particularly useful for very large data sets.
- Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.
- Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems.



Introduction

- The technique is easiest to understand when described using binary or categorical input values.
- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is a **probabilistic classifier**, which means it predicts on the basis of the probability of an object.

Conditional Probability

- For two events A and B,

$$P(B|A) = P(A \text{ and } B) / P(A)$$

- Which is also equal to

$$P(A|B) = P(A \text{ and } B) / P(B)$$

- From both the equations above,

$$P(A|B) = P(B|A) * P(B) / P(A)$$

Why is it called Naïve Bayes

- **Naïve:** It is called Naïve because it assumes that **all features are conditionally independent of one another**, that is, the probability of occurrence of a certain feature given the target feature is independent of the probability of occurrence of other features given the target feature.
 - Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes:** It is called Bayes because it depends on the principle of Bayes' Theorem.

Conditional Independence: Example

- Let the two events be the probabilities of persons A and B getting home in time for dinner, and the third event is the fact that a snow storm hit the city.
- While both A and B have a lower probability of getting home in time for dinner, the lower probabilities will still be independent of each other. That is, the knowledge that A is late does not tell you whether B will be late. (They may be living in different neighborhoods, traveling different distances, and using different modes of transportation.)
 - Conditional Probability of A getting home given there will be snow storm can be written as
$$P(A=\text{getting_home}|\text{Weather}=\text{snow_storm})$$
 - given there will be snow storm can be written as
$$P(B=\text{getting_home}|\text{Weather}=\text{snow_storm})$$

Conditional Independence: Example

- If the two above two events are independent, then theorem of conditional independence says,

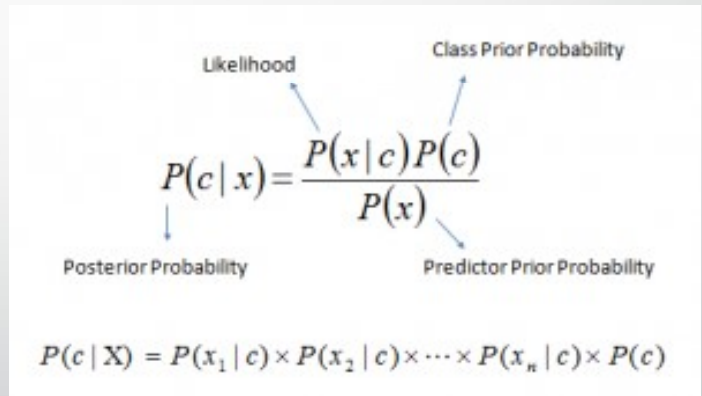
$$\begin{aligned} P(A=\text{getting_home and } B=\text{getting_home} | \text{Weather}=\text{snow_storm}) = \\ P(A=\text{getting_home} | \text{Weather}=\text{snow_storm}) * \\ P(B=\text{getting_home} | \text{Weather}=\text{snow_storm}) \end{aligned}$$

- Similarly , if x_1, x_2, \dots, x_n are the input features and Y is the target feature and the input features are conditionally independent, then

$$\begin{aligned} P(x_1 \text{ and } x_2 \text{ and } x_3 \dots \text{and } x_n | Y) = P(x_1, x_2, x_3, \dots, x_n | Y) = \\ P(x_1 | Y) * P(x_2 | Y) * P(x_3 | Y) \dots P(x_n | Y) \end{aligned}$$

Bayes Theorem

- Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:
- Such that
- $P(c|x)$ is the posterior probability of class (c , target) given predictor (x , attributes).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.



The diagram shows the Bayes Theorem equation $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with four labels and arrows: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$

How probability used for classification

- After calculating the posterior probability for a number of different classes, you can select the class with the highest probability.
- This is the maximum probable class and may formally be called the maximum a posteriori (MAP) hypothesis.
- This can be written as:

$$\text{MAP}(h) = \max(P(c|x))$$

or

$$\text{MAP}(h) = \max((P(x|c) * P(c)) / P(x))$$

How probability used for classification

- The $P(x)$ is a normalizing term which allows us to calculate the probability.
- We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.

$$\text{MAP}(h) = \max((P(x|c) * P(c)))$$

- Also, if we have an even number of instances in each class in our training data, then the probability of each class (e.g. $P(c)$) will be equal.
- Again, this would be a constant term in our equation and we could drop it so that we end up with:

$$\text{MAP}(c) = \max(P(x|c))$$

How it works

- Step 1: Convert the data set into a frequency table
- Step 2: Create Likelihood table by finding the probabilities
- Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Example

- Below is the frequency table for a given dataset,

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- How much is the probability, that the test fruit is banana? $500/1000=0.5$
- How much is the probability, that the test fruit is long? $500/1000=0.5$
- How much is the probability, that the test fruit is long given that it is Banana? $400/500=0.8$
- How much is the probability, that the test fruit is Banana given that it is long, sweet and yellow? $P(Y|X)=?$

Example

- The possible input features for given example are Shape(x_1), Taste(x_2) and Color(x_3)
- And Target Feature is Type(Y)

Example: 1. Compute the 'Prior' probabilities

- $P(Y=\text{Banana}) = 500 / 1000 = 0.50$
- $P(Y=\text{Orange}) = 300 / 1000 = 0.30$
- $P(Y=\text{Other}) = 200 / 1000 = 0.20$

Example: 2. Compute the probability of predictors

- $P(x_1=\text{Long}) = 500 / 1000 = 0.50$
- $P(x_2=\text{Sweet}) = 650 / 1000 = 0.65$
- $P(x_3=\text{Yellow}) = 800 / 1000 = 0.80$

Example: 3. Compute the probability of likelihood

- **Probability of Likelihood for Banana**

- $P(x_1=\text{Long} \mid Y=\text{Banana}) = 400 / 500 = 0.80$
- $P(x_2=\text{Sweet} \mid Y=\text{Banana}) = 350 / 500 = 0.70$
- $P(x_3=\text{Yellow} \mid Y=\text{Banana}) = 450 / 500 = 0.90$

- **Probability of Likelihood for Orange**

- $P(x_1=\text{Long} \mid Y=\text{Orange}) = 0 / 300 = 0$
- No need to calculate others

- **Probability of Likelihood for Other**

- $P(x_1=\text{Long} \mid Y=\text{Other}) = 100 / 200 = 0.5$
- $P(x_2=\text{Sweet} \mid Y=\text{Other}) = 150 / 200 = 0.75$
- $P(x_3=\text{Yellow} \mid Y=\text{Other}) = 50 / 200 = 0.25$

Example: 4. Substitute the values in Bayes Theorem

- Banana

$$P\left(\frac{\text{Banana}}{\text{Long, Sweet, Yellow}}\right) = \frac{P\left(\frac{\text{Long}}{\text{Banana}}\right) \times P\left(\frac{\text{Sweet}}{\text{Banana}}\right) \times P\left(\frac{\text{Yellow}}{\text{Banana}}\right) \times P(\text{Banana})}{P(\text{Long}) P(\text{Sweet}) P(\text{Yellow})}$$

$$P\left(\frac{\text{Banana}}{\text{Long, Sweet, Yellow}}\right) = \frac{(0.8) \times (0.7) \times (0.9) \times (0.5)}{0.25 \times 0.33 \times 0.41}$$

$$P\left(\frac{\text{Banana}}{\text{Long, Sweet, Yellow}}\right) = 0.252$$

$$P\left(\frac{\text{Other}}{\text{Long, Sweet, Yellow}}\right) = \frac{P\left(\frac{\text{Long}}{\text{Other}}\right) \times P\left(\frac{\text{Sweet}}{\text{Other}}\right) \times P\left(\frac{\text{Yellow}}{\text{Other}}\right) \times P(\text{Other})}{P(\text{Long}) P(\text{Sweet}) P(\text{Yellow})}$$

$$P\left(\frac{\text{Other}}{\text{Long, Sweet, Yellow}}\right) = \frac{(0.5) \times (0.75) \times (0.25) \times (0.2)}{0.25 \times 0.33 \times 0.41}$$

$$P\left(\frac{\text{Other}}{\text{Long, Sweet, Yellow}}\right) = 0.01875$$

- Orange

$$P\left(\frac{\text{Orange}}{\text{Long, Sweet, Yellow}}\right) = 0$$

In this case, based on the higher score (0.252 for banana) we can assume this Long, Sweet and Yellow fruit is in fact, a Banana

Types of Naïve Bayes Model

- **Gaussian**

- The Gaussian model assumes that features follow a normal distribution.
- This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

- **Multinomial**

- The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed.
- It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.
- The classifier uses the frequency of words for the predictors.

- **Bernoulli**

- The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables.
- Such as if a particular word is present or not in a document.
- This model is also famous for document classification tasks.

Pros

- It is easy and fast to predict class of test data set. It also perform well in multi class prediction
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Cons

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.
 - To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent
- Require to **remove correlated features** because they are voted twice in the model and it can lead to **over inflating importance**.

Example of Smoothing

- Let us consider that test case has Medium long as one of the value for Shape attribute. Then
- $P(\text{Shape}=\text{Medium_Long}|\text{Y}=\text{Banana}) = P(\text{Shape}=\text{Medium_Long and Y=Banana})/P(\text{Y}=\text{Banana}) = \text{count}(\text{Shape}=\text{Medium_Long in class Y=Banana})/\text{count}(\text{Y}=\text{Banana})$
- But, since medium long value was not in training data set, so $P(\text{Shape}=\text{Medium_Long and Y=Banana})$ will be 0.
- To consider this,
- $P(\text{Shape}=\text{Medium_Long}|\text{Y}=\text{Banana})$ given by

$$\frac{\text{count}(\text{Shape}=\text{Medium_Long in class Y=Banana})+\alpha}{\text{count}(\text{Y=Banana})+\text{number_of_features}*\alpha}$$

- Where α is the smoothing parameter, which is generally set as 1 to remove this zero probability problem

Thank You!!!!

