Naïve Bayes Classifier

Dr. JASMEET SINGH
ASSISTANT PROFESSOR, CSED
TIET, PATIALA

Naïve Bayes Classifier- Introduction

- Naïve Bayes classifier is a probabilistic classifier that uses Bayes theorem and Naïve assumption to classify test examples using the training examples.
- According to Bayes Theorem,

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

where P(A|B) is called *posterior probability* of A given B; P(A) is the *prior probability* of A; P(B|A)is the *likelihood* of B given A; and P(B) is the *evidence* of B.

- For machine learning tasks; A is the target variable (y_i) and B is the input test case ($X = x_1 x_2 x_3 x_4 \dots\dots\dots\dots\dots\dots x_k$)
  - Therefore we find, $P(y_i|X) = \frac{P(y_i)\,P(X|y_i)}{P(X)} \qquad for\ all\ \ y_i \in Y$

Naïve Bayes Classifier- Introduction (Contd….)

- Since P(X) is constant w.r.t different values of y_i. Hence it can be ignored.
- Therefore, $P(y_i|X) \propto P(y_i)\, P(X|y_i)$
- According to **Naïve assumption**, the probability of each feature in the input is conditionally independent of each other.

Therefore,

$$P(y_i|X) \propto P(y_i) \prod_{j=1}^{k} P(x_j\,|y_i)$$

The final predicted label (y*) for a given input X is thus computed as:

$$y^* = \arg\max_{y} \; P(y_i) \prod_{j=1}^{n} P(x_j \mid y_i)$$

Training Phase of Naïve Baye Classifier

- In the training phase of Naïve Bayes Classifier, we compute prior probability and likelihood probabilities from the training data.

- **Computing Class Prior Probabilities**
  - Class prior probability of each unique value y_i of the output variable Y is computed as:

$$P(y_i) = \frac{number\ of\ training\ examples\ labeled\ as\ y_i}{total\ training\ examples} = \frac{n_{y_i}}{N}$$

Training Phase of Naïve Baye Classifier (Contd…)

- **Computing Likelihoods**
  - The likelihood of each unique value of each feature given each class label is computed as follows:

$$P(x_j = c|y_i) = \frac{number\ of\ training\ examples\ for\ which\ feature\ x_j\ has\ value\ c\ and\ labeled\ as\ yi}{total\ number\ of\ training\ examples\ labeled\ as\ yi} = \frac{n_{x_j=c,y_i}}{n_{y_i}}$$

For all, $x_j \in X\ (feature\ set), c\ \in unique\ values\ of\ x_j$, and $y_i \in unique\ vales\ of\ Y$

Testing Phase of Naïve Bayes Classifier

- In the test phase, for each test example X_test=$x_1x_2x_3\ldots\ldots x_k$, probability of each class label given the test example is computed as:

$$P(y_i|X_{test}) \propto P(y_i)\prod_{j=1}^{k} P(x_j|y_i)$$

- The final predicted label (y*) for a given input X is thus computed as:

$$y^* = \arg\max_{y} P(y_i)\prod_{j=1}^{n} P(x_j \mid y_i)$$

Numerical Example-I

Consider the following training set, that classify the output variable play golf as Yes or No depending upon weather conditions such as Outlook, Temperature, Humidity, and Wind Status.

Using Naïve Bayes Classifier, classify that whether on a Rainy, Cool, High Humidity, and Windy day we can play golf or not.

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

Example 1-Solution

- Training Phase:

Computing Class Prior Probability

- For each unique value of output variable Play Golf i.e., Yes or No, the prior probability is computed as follows:

$$P(play\ golf = yes) = \frac{number\ of\ training\ examples\ labelled\ yes}{total\ training\ examples} = \frac{9}{14}$$

$$P(play\ golf = no) = \frac{number\ of\ training\ examples\ labelled\ no}{total\ training\ examples} = \frac{5}{14}$$

Example 1-Solution (Contd….)

- **Computing Likelihoods**
  - The likelihood of each unique value of each feature given each class label is computed.
  - For instance, for the feature outlook, unique values are Rainy, Overlook, and Sunny.
  - Therefore, P(Outlook=Rainy|Yes), P(Outlook=Rainy|No), P(Outlook=Overlook|Yes), P(Outlook=Overlook|No), P(Outlook=Sunny|Yes), P(Outlook=Sunny|No) are computed. The same is repeated for all features as shown in figure).

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Example 1-Solution (Contd….)

Test Example: Outlook=Rainy, **Temperature**=Cool, **Humidity**=High, and **Windy** =True

$P(play_{golf} = yes|$Outlook=Rainy, Temperature=Cool, Humidity=High, and Windy =True)

$= P(yes) \times P($Outlook=Rainy|yes$) \times P($Temperature=Cool|yes$) \times P($Humidity=High|yes$) \times P($Windy=True|yes$)$

$$= \frac{9}{14} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = \frac{729}{91854} = \mathbf{0.007936}$$

$P(play_{golf} = no|$Outlook=Rainy, Temperature=Cool, Humidity=High, and Windy =True)

$= P(no) \times P($Outlook=Rainy|no$) \times P($Temperature=Cool|no$) \times P($Humidity=High|no$) \times P($Windy=True|no$)$

$$= \frac{5}{14} \times \frac{2}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = \frac{120}{8750} = \mathbf{0.013714}$$

Therefore, the given test example should be labeled as Play Golf = No

Advantages of Naïve Bayes Classifier

1. It is simple and easy to understand.
2. No hyper-parameter tuning is required.
3. It is scalable i.e. if a new instance is added it is easy to adjust class prior and likelihood probabilities.
4. It can be used for real time classifications.
5. It is very suitable for multi-class classification (as we need not to apply techniques like one vs. rest to fit multiple binary classifiers).

Naïve Bayes Classifier- Problems

- **Problem I: Zero Frequency Problem**
  - If an individual feature value for a particular class label is missing, then the frequency-based probability estimate will be zero. And we will get a zero when all the probabilities are multiplied. This problem is called zero frequency problem.
  - For example, in the figure (shown in slide 9), the *P(outlook=Overcast | no) =0* because there is no training example which has *overcast outlook* for label *no*.
  - To handle this zero frequency problem, we apply smoothing technique.

Naïve Bayes Classifier- Problems (Contd..)

- **Problem I: Zero Frequency Problem (Solution)**
  - Smoothing is a technique that handles the problem of zero probability in Naïve Bayes.
  - In smoothing, while computing likelihood of any feature given label, we add a parameter α in numerator and α X number of features in denominator i.e.,

$$P(x_j = c|y_i) = \frac{n_{x_j=c,y_i} + \alpha}{n_{y_i} + \alpha \times k}$$

Where k is the number of features. α is added so that probability is never 0 and $\alpha \times k$ is added in denominator so that probability is never greater than 1.

  - When $\alpha = 1$, it is called Laplace Smoothing (correction) and if $\alpha < 1$, it is called Lidstone Smoothing.
  - α should not be taken greater than 1 because it will give higher probability mass to zero frequency counts.

Naïve Bayes Classifier- Problems (Contd..)

- **Problem II: Independence Assumption**
  - Naïve Bayes Classifier, is based on the Naïve assumption, that the features are independent of each other.
  - But in real case scenarios, input features are not always independent.
  - For instance, if we have to label a person as *adult* or *child* on the basis of height and weight of person, then features height and weight are not independent of each other.
  - In order to handle this problem, we must apply dimensionality reduction if the features are corelated.
  - Due to the Naïve assumption, this classifier is most suitable for Text Classification as the words are features in text and these words can be considered independent for classification.

Naïve Bayes Classifier- Problems (Contd..)

- **Problem III: Numerical Underflow**
  - We know, likelihood probability is computed as:

$$P(y_i|X) \propto P(y_i) \prod_{j=1}^{k} P(x_j|y_i)$$

If k (number of features are large). Then the product is very small and approximate to zero. This problem is called numerical underflow.

In order to solve this problem, we compute log likelihoods (so as to convert product to sum).

$$\log\big(P(y_i|X)\big) \propto \log\big(P(y_i)\big) + \sum_{j=1}^{k} \log(P(x_j|y_i))$$

Numerical Example II: Text Classification

Consider the training dataset (shown in figure), where in each training document is labeled as Sports or Politics category.

Using Naïve Bayes Classifier, classify the document '*A very close game* '

(Pre-processing: Convert lower case, no stopword removal, no stemming).

Text	Category
"A great game"	Sports
"The election was over"	Politics
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Politics

Example 2: Solution

Training Phase

Computing Prior Probability

$$P(Sports) = \frac{number\ of\ documents\ labeled\ as\ 'Sports'}{total\ number\ of\ docuemnts} = \frac{3}{5}$$

$$P(Politics) = \frac{number\ of\ documents\ labeled\ as\ 'Politics'}{total\ number\ of\ docuemnts} = \frac{2}{5}$$

Computing Likelihood Probabilities:

Number of features (words) = 14 = |V|

{a, great, game, the, election, was, over, very, clean, match, but, forgettable, it, close}

Example 2: Solution (Contd…)

Probability of each feature (word) given Sports Label

Document→ Feature↓	D1	D3	D4	Frequency	Probability(word/Sports)
a	1	0	1	2	2+1/11+14=0.12
great	1	0	0	1	1+1/11+14=0.08
game	1	0	1	2	2+1/11+14=0.12
the	0	0	0	0	0+1/11+14=0.04
election	0	0	0	0	0+1/11+14=0.04
was	0	0	0	0	0+1/11+14=0.04
over	0	0	0	0	0+1/11+14=0.04
very	0	1	0	1	1+1/11+14=0.08
clean	0	1	1	2	2+1/11+14=0.12
match	0	1	1	2	2+1/11+14=0.12
but	0	0	1	1	1+1/11+14=0.08
forgettable	0	0	1	1	1+1/11+14=0.08
it	0	0	0	0	0+1/11+14=0.04
close	0	0	0	0	0+1/11+14=0.04
Total				11	

Example 2: Solution (Contd…)

Probability of each feature (word) given Politics Label

Document→ Feature↓	D2	D5	Frequency	Probability(word/Politics)
a	0	1	1	1+1/9+14=0.09
great	0	0	0	0+1/9+14=0.04
game	0	0	0	0+1/9+14=0.04
the	1	0	1	1+1/9+14=0.09
election	1	1	2	2+1/9+14=0.13
was	1	1	2	2+1/9+14=0.13
over	1	0	1	1+1/9+14=0.09
very	0	0	0	0+1/9+14=0.04
clean	0	0	0	0+1/9+14=0.04
match	0	0	0	0+1/9+14=0.04
but	0	0	0	0+1/9+14=0.04
forgettable	0	0	0	0+1/9+14=0.04
it	0	1	1	1+1/9+14=0.09
close	0	1	1	1+1/9+14=0.09
Total			9	

Example 2: Solution (Contd…)

Testing Phase

Test Sentence= X = a very close game

$P(Sports|X) = P(S) \times P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports)$

$= 0.6 \times 0.12 \times 0.08 \times 0.04 \times 0.12$

$= 2.76 \times 10^{-5}$

$P(P|X) = P(Politcs) \times P(a|Politcs) \times P(very|Politics) \times P(close|Politics) \times P(game|Politics)$

$= 0.4 \times 0.09 \times 0.04 \times 0.09 \times 0.04$

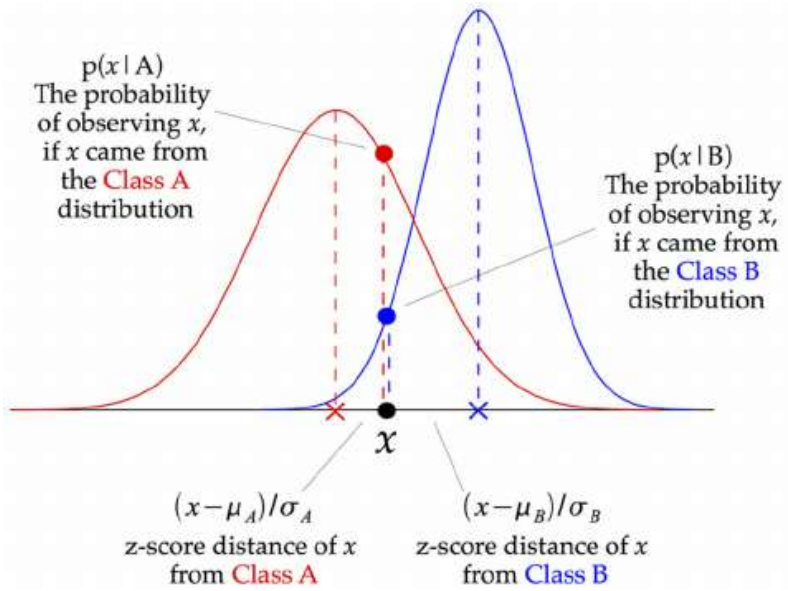
$= 5.18 \times 10^{-6}$

Since $P(Sports|X) > P(Politics|X)$

Therefore, the test sentence can be labeled as Sports

Gaussian Naïve Bayes Classifier

- In case the feature variable are continuous, then it is not possible to compute likelihood of each feature value given label for continuous range.
- The version of Naïve Bayes algorithm that deal with continuous feature values is called Gaussian Naïve Bayes Classifier.
- In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a **Gaussian distribution.**
- So, we make use of z-score for observing each feature value given a label.



Gaussian Naïve Bayes Classifier (contd…)

- In particular, Probability of observing any feature value c for feature x_j given a class label y_i is computed as:

$$P(x_j = c|y_i) = \frac{1}{\sqrt{2\pi\sigma_{x_j,y_i}^2}} e^{\frac{-1}{2}\left(\frac{c-\mu_{x_j,y_i}}{\sigma_{x_j,y_i}}\right)^2}$$

Where μ_{x_j,y_i} denote mean of x_j feature values labeled as y_i and σ_{x_j,y_i} is standard deviation of x_j feature values labeled as y_i.

Numerical Example - 3

Consider the following training set, that classify the output variable play golf as Yes or No depending upon weather conditions such as Temperature, Humidity (same as example 1 but the features are continuous instead of categorical).

Using Gaussian Naïve Bayes Classifier, classify that whether on a day when temperature is 66 and humidity is 90 is we can play golf or not.

Temperature	Humidity	Play
85	85	no
80	90	no
83	86	yes
70	96	yes
68	80	yes
65	70	no
64	65	yes
72	95	no
69	70	yes
75	80	yes
75	70	yes
72	90	yes
81	75	yes
71	91	no

Example 3: Solution

- Training Phase:

Computing Class Prior Probability

- For each unique value of output variable Play Golf i.e., Yes or No, the prior probability is computed as follows:

$$P(play\ golf = yes) = \frac{number\ of\ training\ examples\ labelled\ yes}{total\ training\ examples} = \frac{9}{14}$$

$$P(play\ golf = no) = \frac{number\ of\ training\ examples\ labelled\ no}{total\ training\ examples} = \frac{5}{14}$$

Example 3: Solution (Contd…)

	Yes	*No*
Temp:	64, 68, 69, 70, 72, …	65, 71, 72, 80, 85, …
	$\mu = 73$, $\sigma = 6.2$	$\mu = 75$, $\sigma = 7.9$

	Yes	*No*
Humidity:	66, 70, 70, 75, 80, …	70, 85, 90, 91, 95, …
	$\mu = 79$, $\sigma = 10.2$	$\mu = 86$, $\sigma = 9.7$

Example 3: Solution (Contd…)

Testing Phase:
Test Example: T=66,H=90

$$P(T=66|yes)=\frac{1}{\sqrt{2\pi}\times 6.2}e^{\frac{-1}{2}\left(\frac{66-73}{6.2}\right)^2}=0.034$$

$$P(T=66|no)=\frac{1}{\sqrt{2\pi}\times 7.9}e^{\frac{-1}{2}\left(\frac{66-75}{7.9}\right)^2}=0.0279$$

$$P(H=90|yes)=\frac{1}{\sqrt{2\pi}\times 10.2}e^{\frac{-1}{2}\left(\frac{90-7}{10.2}\right)^2}=0.0221$$

$$P(H=90|yes)=\frac{1}{\sqrt{2\pi}\times 9.7}e^{\frac{-1}{2}\left(\frac{90-8}{9.7}\right)^2}=0.0381$$

$$P(yes|T=66,H=90)=P(yes)P(T=66|yes)P(H=90|yes)=\frac{9}{14}\times 0.034\times 0.0221=0.00048$$

$$P(no|T=66,H=90)=P(no)P(T=66|no)P(H=90|no)=\frac{5}{14}\times 0.0279\times 0.0381=0.00037$$

Therefore, on given temperature and humidity, we can play golf