| Naiive Bayes Classification | 8.3 | Data mining  by Jaiwei Han |
|---|---|---|

Text book:  Machine Learning by  TOM M MITCHELL Pg No 158

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.– Parametric estimation
- Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

# Naive Bayes algorithm

Assumes that the occurrence of a certain feature is independent of the occurrence of other features.
Depends on bayes Theorem

$$\underbrace{P(c \mid x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x \mid c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- *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*.

- Players will play if weather is sunny.

P(Yes | Sunny) = P( Sunny | Yes) * P(Yes) / P (Sunny)

P (Sunny |Yes) = 2/9 =
P(Sunny) = 5/14 = 0.36,
P( Yes)= 9/14 = 0.64

P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60

| Day | *Outlook* | *Temperature* | *Humidity* | *Wind* | *PlayTennis* |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

Buy computer if he/she ⟹ student)

(i) $P(Y / \text{student}) = \dfrac{P(\text{student} / \text{yes}) \cdot P(\text{yes})}{P(\text{student})} = 0.85$

⇓ p.p   Likely

$= Y$

$P(\text{yes}) = 9/14$ =

$P(\text{student}) = 7/14.$ =

$P(\text{student} | \text{yes}) = 6/9$ =

(ii) $P(N / \text{Senior}=\text{age}) = \dfrac{P(\text{Sen} / N) \cdot P(N)}{P(\text{age}=\text{Sen})}$

$= \dfrac{2/5 \times 5/14}{5/14} = 0.40$

$< 0.5$

$\boxed{\text{NO}}$

| RID | age | income | student | credit-rating | class:buy_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

# Naïve Baysian Estimation – Classification Problem Example 1

- classify is whether the customer X (35 years, 40 K income ) will buy a computer or not
- **X (age=youth, income=medium, student=yes, credit=fair)**
- Maximize P(X|Ci)P(Ci), for i=1,2

classify is whether the customer X (35 years, 40 K income ) will buy a computer or not

**X (age=youth, income=medium, student=yes, credit=fair)**

Maximize P(X|Ci)P(Ci), for i=1,2

$P \cdot P = \frac{(\ )(\ )}{(\ )}$

| RID | age | income | student | credit-rating | class:buy_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Given **X (age=youth, income=medium, student=yes, credit=fair)**

Maximize **$P(X|C_i)P(C_i)$**, for i=1,2

**First step**: Compute $P(C_i)$. The prior probability of each class can be computed based on the training tuples:

P(buys_computer=yes)=9/14=0.643

P(buys_computer=no)=5/14=0.357

**Second step**: compute $P(X|C_i)$ using the following c

P(age=youth | buys_computer=yes)=0.222

P(age=youth | buys_computer=no)=3/5=0.666

P(income=medium | buys_computer=yes)=0.444

P(income=medium | buys_computer=no)=2/5=0.400

P(student=yes | buys_computer=yes)=6/9=0.667

P(tudent=yes | buys_computer=no)=1/5=0.200

P(credit_rating=fair | buys_computer=yes)=6/9=0.667

P(credit_rating=fair | buys_computer=no)=2/5=0.400

| RID | age | income | student | credit-rating | class:buy_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

**X (age=youth, income=medium, student=yes, credit=fair)**

$$P(X \mid buys\_computer=yes) = P(age=youth \mid buys\_computer=yes) \times P(income=medium \mid buys\_computer=yes) \times P(student=yes \mid buys\_computer=yes) \times P(credit\_rating=fair \mid buys\_computer=yes) = 0.044$$

$$P(X \mid buys\_computer=no) = P(age=youth \mid buys\_computer=no) \times P(income=medium \mid buys\_computer=no) \times P(student=yes \mid buys\_computer=no) \times P(credit\_rating=fair \mid buys\_computer=no) = 0.019$$

**Third step**: compute $P(X \mid C_i)P(C_i)$ for each class

$P(X \mid buys\_computer=yes)P(buys\_computer=yes) = 0.044 \times 0.643 = \mathbf{0.028}$

$P(X \mid buys\_computer=no)P(buys\_computer=no) = 0.019 \times 0.357 = \mathbf{0.007}$

**The naïve Bayesian Classifier predicts buys_computer=yes for tuple X**

$\Rightarrow$ **X (age=youth, income=medium, student=yes, credit=fair)**

| | Y | N |
|---|---|---|
| Age (Y) | 2 | 3 |
| Income (M) | 4 | 2 |
| St(Y) | 6 | 1 |
| CR(fair) | 6 | 2 |

Liklihood table

$Y = 9$

$N = 5$

$P(Y) = 9/14$

$P(N) = 5/14$

| RID | age | income | student | credit-rating | class:buy_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

Liklihood Table.

| | Y | N |
|---|---|---|
| age (Y) | 2/9 | 3/5 |
| I (M) | 4/9 | 2/5 |
| S (Y) | 6/9 | 1/5 |
| CR (Y) | 6/9 | 2/5 |
| $P(X\|C_i) =$ | $\frac{2}{9} \cdot \ldots$ | |
| | $P(X\|Y)$ | $P(X\|N)$ |

$$\rightarrow P(X/Y) \cdot P(Y) = \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14} = \cancel{0.044} \; 0.028$$

$$\rightarrow P(X|N) \; P(N) = \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = \cancel{0.007} \; 0.007$$

$$P(X|Y) \cdot P(Y) > P(X|N) \cdot P(N)$$

$\therefore X \Rightarrow$ Buys computer

$X = P(\text{Age} = \text{Youth}, \text{Income} = \text{low}, \text{Student} = \text{NO}, \text{Credit rating} = \text{Excellent}.$

| RID | age | income | student | credit-rating | class:buy_computer |
|---|---|---|---|---|---|
| 1 | youth | high | no | fair | no |
| 2 | youth | high | no | excellent | no |
| 3 | middle-aged | high | no | fair | yes |
| 4 | senior | medium | no | fair | yes |
| 5 | senior | low | yes | fair | yes |
| 6 | senior | low | yes | excellent | no |
| 7 | middle-aged | low | yes | excellent | yes |
| 8 | youth | medium | no | fair | no |
| 9 | youth | low | yes | fair | yes |
| 10 | senior | medium | yes | fair | yes |
| 11 | youth | medium | yes | excellent | yes |
| 12 | middle-aged | medium | no | excellent | yes |
| 13 | middle-aged | high | yes | fair | yes |
| 14 | senior | medium | no | excellent | no |

| | Y | N |
|---|---|---|
| A(Y) | 2 | 3 |
| I(L) | 3 | 1 |
| S(N) | 3 | 4 |
| CR(E) | 3 | 3 |

$P(Y) = 9/14$

$P(N) = 5/14$

Liklihood tables

| | Y(9) | N(5) |
|---|---|---|
| A(Y) | 2/9 | 3/5 |
| I(L) | 3/9 | 1/5 |
| S(N) | 3/9 | 4/5 |
| CR(E) | 3/9 | 3/5 |
| $P(X \mid C_i)$ | $2/9 \times 5/4$ ... $9/4$ | |

(i) $P(X/Y) \cdot P(Y) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0.00529$

(ii) $P(X/N) \cdot P(N) = \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0.021$

(ii) > (i)

$X \Rightarrow$ 'NO'

# Example2

| Day | *Outlook* | *Temperature* | *Humidity* | *Wind* | *PlayTennis* |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

3/5

- Task is to predict the target value (yes or no) of the target concept PlayTennis for the following instance
- (Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

**(Text book: Machine Learning by TOM M MITCHELL Pg No 178)**

- P(P1ayTennis = yes) = 9/14 = .64
- P(P1ayTennis = no) = 5/14 = .36
- P(Wind = stronglPlayTennis = yes) = 319 = .33
- P(Wind = strongl PlayTennis = no) = 315 = .60

$$P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

The naive Bayes classifier assigns the target value PlayTennis = no

- P(c/y)=0.5
- P(c/n)=0
- P(g/y)=0.25
- P(g/n)=0.67
- P(t/y)=0.25
- P(t/n)=0.67
- P(hr/y)=0.5
- P(hr/n)=0.33
- p(s=y)=0.57
- P(s=n)=3/7=0.42

2

$X = (\text{Rain}, \text{Cool}, \text{Norl}, \text{weak}) = \text{'Y'}$

| | Y (9) | N (5) |
|---|---|---|
| R | 3 | 2 |
| C | 3 | 1 |
| N | 6 | 1 |
| W | 6 | 2 |

$P(X/Y)\,P(Y) = \frac{3}{9} \times \frac{3}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14} = 0.0317$

$P(X/N) \cdot P(N) = \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{5}{14} = 0.0023$

| Day | *Outlook* | *Temperature* | *Humidity* | *Wind* | *PlayTennis* |
|---|---|---|---|---|---|
| D1 | Sunny | Hot | High | Weak | No |
| D2 | Sunny | Hot | High | Strong | No |
| D3 | Overcast | Hot | High | Weak | Yes |
| D4 | Rain | Mild | High | Weak | Yes |
| D5 | Rain | Cool | Normal | Weak | Yes |
| D6 | Rain | Cool | Normal | Strong | No |
| D7 | Overcast | Cool | Normal | Strong | Yes |
| D8 | Sunny | Mild | High | Weak | No |
| D9 | Sunny | Cool | Normal | Weak | Yes |
| D10 | Rain | Mild | Normal | Weak | Yes |
| D11 | Sunny | Mild | Normal | Strong | Yes |
| D12 | Overcast | Mild | High | Strong | Yes |
| D13 | Overcast | Hot | Normal | Weak | Yes |
| D14 | Rain | Mild | High | Strong | No |

$9 \rightarrow Y$

$5 \rightarrow N.$

# Calculate likelihood

| Students | CGPA | Written test marks | Technical Round | HR Rating | Selected |
|---|---|---|---|---|---|
| N1 | 9.2 | 16 | Weak | High | No |
| N2 | 8.7 | 19 | Strong | High | No |
| N3 | 7.9 | 17 | Weak | High | Yes |
| N4 | 5.9 | 10 | Weak | High | Yes |
| N5 | 6.2 | 6 | Weak | Normal | Yes |
| N6 | 6.4 | 7 | Strong | Normal | No |
| N7 | 8.1 | 2 | Strong | Normal | Yes |

- P(CGPA=7.5|Selected=Y)

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{0.982\sqrt{2\pi}} e^{-\frac{(7.5-7.025)^2}{2(0.9812)^2}}$$

$$=$$

- P(CGPA=7.5|Selected=N)

$$= \frac{1}{1.219\sqrt{2\pi}} \cdot e^{-\frac{(7.5-8.1)^2}{2(1.219)^2}} =$$

Which is higher $\Rightarrow$

| | CGPA(Y) | CGPA(N) |
|---|---|---|
| | 7.9, 5.9, 6.2, 8.1 | 9.2, 8.7, 6.4 |
| Mean | 7.025 | 8.1 |
| SD($\sigma$) | 0.9832 | 1.219 |

$X =$

# P( CGPA=7.5, WT=12,TR=Strong, HR=High)

| Students | CGPA | Written test marks | Technical Round | HR Rating | Selected |
|---|---|---|---|---|---|
| N1 | 9.2 | 16 | Weak | High | No |
| N2 | 8.7 | 19 | Strong | High | No |
| N3 | 7.9 | 17 | Weak | High | Yes |
| N4 | 5.9 | 10 | Weak | High | Yes |
| N5 | 6.2 | 6 | Weak | Normal | Yes |
| N6 | 6.4 | 7 | Strong | Normal | No |
| N7 | 8.1 | 2 | Strong | Normal | Yes |

| | Y (4) | N (3) |
|---|---|---|
| TR(S) | 1/4 | 2/3 |
| HR(H) | 2/4 | 2/3 |
| GCPA(7.5) | $x_1$ | $x_2$ |
| WT (12) | $x_3$ | $x_4$ |

$P(Y) = 4/7$

$P(N) = 3/7$

| | CGPA (Y) | CGPA(N) | WT(Y) | WT(N) |
|---|---|---|---|---|
| | 7.9, 5.9, 6.2, 8.1 | 9.2, 8.7, 6.4 | 17, 10, 6, 2 | 16, 19, 7 |
| M | 7.025 | 8.1 | | |
| SD (σ) | 0.98 | 1.21 | | |

$$x_1 = P(CGPA=7.5 \mid Y) = \frac{1}{0.98\sqrt{2\pi}} e^{-\frac{(7.5-7.025)^2}{2(0.98)^2}}$$

$$x_2 = P(CGPA=7.5 \mid N) = \frac{1}{1.21\sqrt{2\pi}} e^{-\frac{(7.5-8.1)^2}{2(1.21)^2}}$$

$$x_3 = P(WT=12 \mid Y) =$$

$$x_4 = P(WT=12 / N) =$$

$$P(X \mid Y) \cdot P(Y) = \frac{1}{4} \times \frac{2}{4} \times x_1 \times x_3 \times \frac{4}{7}$$

$$=$$

$$P(X \mid N) \cdot P(N) = \frac{2}{3} \times \frac{2}{3} \times x_2 \times x_4 \times \frac{3}{7}$$

$$=$$

# Example3

## Problem formulation:

**Classification of high risk and low risk customers in a bank:**

**Parameters to decide risk** : Salary, Saving, state of economy in full detail and about the customer, his or her intention, moral codes

IF income>θ 1 AND savings>θ 2 THEN low-risk ELSE high-risk

Let the credibility of a customer is denoted by a Bernoulli random va

where Y=1 indicates a high-risk customer

Y =0 indicates a low-risk customer

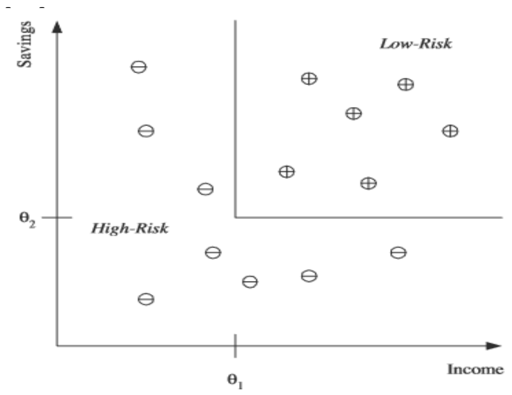
X are the customer attributes , X={ X1,X2,…..}

P(Y |X), where Classification as learning an association from X to Y

For a given X = x, if we have P(Y =1|X =x)=0.8

Then customer has an 80 percent probability of being high-risk

or equivalently a 20 percent probability of being low-risk.



Thus if we know $P(Y | X1,X2)$, when a new application arrives with $X1 = x1$ and $X2 = x2$,

we can choose $Y = 1$ if $P(Y = 1 | x1,x2) > 0.5$

$Y = 0$ otherwise or equivalently

choose $Y = 1$ if $P(Y = 1 | x1,x2) > P(v=0 | x1,x2)$

$Y = 0$ otherwise

The problem then is to be able to calculate $P(Y | x)$. UsingBayes' rule

$P(Y | x) = P(Y) p(x | v) / p(x)$ Bayes' rule

**Prior probability :** P(Y =1) is called the prior probability that Y takes the value 1

The probability that a customer is highrisk, regardless of the x value.

It is the knowledge about the value of Y before looking at the observables

**Class likelihood** : p(x| Y) is the conditional probability class likelihood that an event belonging to Y has the associated observation value x.

In this case, p(x1,x2| Y = 1) is the probability that a high-risk customer has his or her X1 = x1 and X2 = x2. It is what the data tells us regarding the class.

**Evidence :** p(x) is the marginal probability that an observation x is evidence seen, regardless of whether it is a positive or negative example.

$$p(x)=p(x|\ Y=1)P(Y=1)+p(x|\ Y=0)P(Y=0)$$

Combining the prior and what the data tells us using Bayes' rule, calculate the posterior probability of the concept, P(Y |x),

**posterior= prior×likelihood/ evidenc**

$$P(Y|x)= p(x|\ Y)P(Y) /\ p(x)$$

Objective function is Maximise **posterior**

Text book: Machine Learning by TOM M MITCHELL Pg No 158

| | |
|---|---|
| 0.98 | |
| | 0.97 |

test with two possible outcomes: $\oplus$ (positive) and $\ominus$ (negative). We have prior knowledge that over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \qquad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \qquad P(\ominus|\neg cancer) = .97$$

Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not? The maximum a posteriori hypothesis can be found using Equation (6.2):

$$\rightarrow \quad P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$

Thus, $h_{MAP} = \neg cancer$. The exact posterior probabilities can also be determined by normalizing the above quantities so that they sum to 1 (e.g., $P(cancer|\oplus) = \frac{.0078}{.0078+.0298} = .21$). This step is warranted because Bayes theorem states that the posterior probabilities are just the above quantities divided by the probability of the data, $P(\oplus)$. Although $P(\oplus)$ was not provided directly as part of the problem statement, we can calculate it in this fashion because we know that $P(cancer|\oplus)$ and $P(\neg cancer|\oplus)$ must sum to 1 (i.e., either the patient has cancer or they do

# Naïve Bayes

- **1. Missing Data**
- Naive Bayes can handle missing data.
- Attributes are handled separately by the algorithm at both model construction time and prediction time.
- As such, if a data instance has a missing value for an attribute, it can be ignored while preparing the model, and ignored when a probability is calculated for a class value.
- **2. Use Log Probabilities**
- Probabilities are often small numbers. To calculate joint probabilities, you need to multiply probabilities together. When you multiply one small number by another small number, you get a very small number.
- It is possible to get into difficulty with the precision of your floating point values, such as under-runs. To avoid this problem, work in the log probability space (take the logarithm of your probabilities).
- This works because to make a prediction in Naive Bayes we need to know which class has the larger probability (rank) rather than what the specific probability was.

https://machinelearningmastery.com/better-naive-bayes/

- **3. Use Other Distributions**
- **4. Use Probabilities For Feature Selection**
- **5. Segment The Data**
- **6. Re-compute Probabilities**
- **7. Use as a Generative Model**
- **8. Remove Redundant Features**

- **Gaussian:** It is used in classification and it assumes that features follow a normal distribution.
- **Multinomial**: It is used for discrete counts. For example, let's say, we have a text classification problem. Here we can consider Bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number $x_i$ is observed over the n trials".
- **Bernoulli**: The binomial model is useful if your feature vectors are binary (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.