Week 3 : Session 3

5. Naive Bayes

As the name implies, this is a Bayesian Technique, which is characterized by the Prior and Posterior Probability. (i) Before the observation and After the observation. (ii) Apriori and Aposteriori.

Naive Bayes is best suited for ML applications, wherein the feature vectors $\bar{x}$ are DISCRETE.

Comparison Table

	Response Variable (y)	Feature vector ($\bar{x}$)
Linear Regression	Continuous	Continuous
Logistic Regression	Discrete	Continuous
Naive Bayes	Discrete	Discrete

Examples:

① ML-based e-mail Spam filter

(i) $y = 1 \rightarrow$ SPAM

$y = 0 \rightarrow$ GENUINE.

② Like / Dislike an item based on other preference.

(i) Say a particular person likes Harry Potter movie and the person also likes Avenger movie, then what is the Probability that the person would like Top Gun?

Let us now consider Naive Bayes in the context of e-mail classification. Other examples can also be deduced easily though.

Consider a feature vector $\bar{x}$ of size N, where N is the number of words in the English Dictionary. Say, N ~ 100,000 words.

$$\bar{x} = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \end{bmatrix}_{100{,}000 \times 1}$$

Each entry corresponds to a word.

The labels $y = 1, 0$ indicate SPAM, GENUINE emails, respectively.

The entry $x_j = 1$, if the email contains the j^{th} word of the dictionary.

Else $x_j = 0$ (i.e. the email doesn't contain the j^{th} word)

Each j corresponds to a word in Dictionary.

For instance, the email contains the statement "I am able to access the office space", the corresponding feature vector would be as follows.

$$\bar{x} = \begin{bmatrix} \vdots \\ 1 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

	index
able	4562
above	4563
abroad	4564
access	4565

$x_{4562} = 1$

$x_{4563} = 0$

$x_{4564} = 0$

$x_{4565} = 1$

N → No. of Features

The word 'Naive' meaning Simple, NOT complicated, Derogatory.

Naive Bayes assumption:

The different words are **Conditionally independent** given the label y. This is the simplistic assumption.

Without this simplistic assumption, life becomes very complicated. Infact, we may think its derogatory, but without this assumption, things become very complicated. We'll see why this assumption is important.

We assume that the different words are conditionally independent which is reasonable true in practice. and not completely true.

One can always achieve a better result with more complicated algorithm. (ie) In Machine Learning, it is always a trade-off between performance vs complexity. But if the complexity is high, it prohibits the implementation in practical systems. So, one has to also pay attention to that.

The **Naive Bayes assumption** is given by:

$$p(\bar{x} = \bar{v} \mid y = u), \text{ where } v_i \in \{0,1\} \text{ and } u \in \{0,1\}$$

($\bar{x}$: Feature vector, y: Response)

$$= p\left(x_1 = v_1, x_2 = v_2, \ldots, x_N = v_N \mid y = u\right)$$

$$= p\left(x_1 = v_1 \mid y = u\right) \times \ldots\ldots \times p\left(x_N = v_N \mid y = 0\right) \quad \Big| \because \text{conditionally independent}$$

$$= \prod_{j=1}^{N} p\left(x_j = v_j \mid y = u\right) \quad \Big| \text{conditionally independent, Given the RESPONSE}$$

The conditional independence may not be completely True but greatly simplifies our life. Because, otherwise, one has to compute all the Joint probabilities. For instance, if there are 100 words in an email, computing the 100th order Joint probability in an email is numerically impossible, coz we've to take all possible combinations of the dictionary words taken 100 at a time. (i.e) ${}^{n}C_{100}$ which is very very hard.

Thus, without this Naive Bayes assumption, the problems like e-mail classification problem becomes very very hard / Not possible at all, as they are computationally complex. Because the feature vectors are discrete, each element can take several possibilities, which means the total number of possibilities becomes humongous which cannot be stored / processed by any realistic / existing system.

Therefore, the probabilities of occurences of these words (i.e) x_i's are conditionally independent given if it is a genuine / spam email. This is the Naive Bayes assumption.

The quantities $p(x_j = v_j | y = c)$ are the Prior probabilities. How to calculate these?

$p(x_{4562} = 1 | y = 1)$ → What is the Probability the word "able" occurs in a SPAM email.

$p(x_{4562} = 1 | y = 0)$ → What is the Probability the word "able" occurs in a GENUINE email.

$$P(x_{4562} = 0 \mid y = 1)$$
$$= 1 - P(x_{4562} = 1 \mid y = 1)$$

→ What is the probability the word "able" does not occur in a SPAM email.

$$P(x_{4562} = 0 \mid y = 0)$$
$$= 1 - P(x_{4562} = 1 \mid y = 0)$$

→ What is the probability the word "able" does not occur in a GENUINE email.

So, we are interested in calculating all the probabilities ranges from $j = 1, 2, \ldots, N$. (ii) $P(x_j = v_j \mid y = u) \;\forall\; j = 1, \ldots, N$

- Essentially what it means is, for every word in the dictionary, we have to calculate two probabilities. (i) (ii) (i) what is the probability that each word occurs in a GENUINE email and (ii) what is the probability that each word occurs in a SPAM email.

Probabilities that j^{th} word occurs in GENUINE / SPAM emails

Consider the availability of M training pairs $(\bar{x}(i), y(i))$, where $i = 1, 2, \ldots, M$.

M training points = M total emails (GENUINE + SPAM)

Now, the various prior probabilities can be calculated as follows.

(i) Prior probability that j^{th} word occurs in SPAM email

$$P(x_j = 1 | y = 1) = \frac{\text{No. of SPAM emails with } j^{th} \text{ word}}{\text{Total No. of SPAM emails}}$$

Note:
Indicator function
$1(x) = 1$; if x is True
$= 0$; else.

$$= \frac{\sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 1\right)}{\sum_{i=1}^{M} 1\left(y(i) = 1\right)}$$

(ii) Prior probability that j^{th} word occurs in GENUINE email

$$P(x_j = 1 | y = 0) = \frac{\text{No. of GENUINE emails with } j^{th} \text{ word}}{\text{Total No. of GENUINE emails.}}$$

Prior Probabilities of [redacted] occurence

$$= \frac{\sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 0\right)}{\sum_{i=1}^{M} 1\left(y(i) = 0\right)}$$

(iii) Prior probability of getting a SPAM email

$$P(y = 1) = \frac{\text{No. of SPAM emails}}{\text{Total No. of emails}}$$

$$= \frac{\sum_{i=1}^{M} 1\left(y(i) = 1\right)}{M}$$

(iv) Prior probability that j^{th} word doesnot occur in SPAM email

$$P(x_j = 0 | y = 1) = 1 - \left(\text{Prior Probability that } j^{th} \text{ word occurs in SPAM email}\right)$$

Prior Probabilities of NON occurence

$$= 1 - P(x_j = 1 | y = 1)$$

(v) Prior probability that j^{th} word doesnot occur in GENUINE email

$$P(x_j = 0 | y = 0) = 1 - \left(\text{Prior probability that } j^{th} \text{ word occurs in GENUINE email}\right)$$

$$= 1 - P(x_j = 1 | y = 0)$$

(vi) Prior probability of getting GENUINE email

$$P(y = 0) = 1 - \text{Prior Probability of getting SPAM email}$$

$$= 1 - P(y = 1).$$

The Posterior probabilities are calculated as follows.

(i) Posterior Probability that SPAM occurs, given the email $\bar{x}$

$$P(y=1|\bar{x}=\bar{v}) = \frac{P(\bar{x}=\bar{v}|y=1)\cdot P(y=1)}{P(\bar{x}=\bar{v})}$$

Follow from Bayes Rule

(ii) Posterior Probability that GENUINE occurs, given the email $\bar{x}$

$$P(y=0|\bar{x}=\bar{v}) = \frac{P(\bar{x}=\bar{v}|y=0)\cdot P(y=0)}{P(\bar{x}=\bar{v})}$$

$x_j = 1$, if j^{th} word occurs in email

$x_j = 0$, otherwise

Now, E-mail is classified as SPAM if

$$P(y=1|\bar{x}=\bar{v}) > P(y=0|\bar{x}=\bar{v})$$

$$\Rightarrow \frac{P(\bar{x}=\bar{v}|y=1)\cdot P(y=1)}{P(\bar{x}=\bar{v})} > \frac{P(\bar{x}=\bar{v}|y=0)\cdot P(y=0)}{P(\bar{x}=\bar{v})}$$

$$\Rightarrow P(\bar{x}=\bar{v}|y=1)\cdot P(y=1) > P(\bar{x}=\bar{v}|y=0)\cdot P(y=0)$$

CONDITION FOR SPAM

$$P(\bar{x}=\bar{v}|y=1)\cdot P(y=1) \leq P(\bar{x}=\bar{v}|y=0)\cdot P(y=0)$$

CONDITION FOR GENUINE

$$\Rightarrow \left(\prod_{j=1}^{N} P(x_j=v_j|y=1)\right) P(y=1) > \left(\prod_{j=1}^{N} P(x_j=v_j|y=0)\right) P(y=0)$$

To compute the product might be very cumbersome. So, easier way to process this is to take Log, so that the

product becomes sum, which is known as Log Aposteriori probability.

$$\left(\sum_{j=1}^{N} \ln p(x_j = v_j | y=1)\right) + \ln p(y=1) \quad >$$

$$\left(\sum_{j=1}^{N} \ln p(x_j = v_j | y=0)\right) + \ln p(y=0)$$

Week 4 : Session 2

Laplace Smoothing

Naive Bayes has a problem.

Let's say, a new word "IITK" appears in your email, which is not present in any training e-mails (GENUINE OR SPAM)

Let's say index of "IITK" in the dictionary is j. say $j \approx 25,634$.

The Prior Probabilities are calculated as follows.

(i) Prior Probability that the word "IITK" occurs in SPAM email.

$$p(x_j = 1 | y=1) = \frac{\text{No. of SPAM emails containing the word "IITK"}}{\text{Total No. of SPAM emails}}$$

$$= 0 \quad (\because \text{Typically No spam email contains IITK})$$

(ii) Prior Probability that the word "IITK" occurs in GENUINE email

$$p(x_j = 1 | y=0) = \frac{\text{No. of GENUINE emails containing the word "IITK"}}{\text{Total No. of GENUINE emails}}$$

$$= 0 \quad (\because \text{None of the GENUINE email contains IITK})$$

Mathematically, this can be represented as

$$P(x_j = 1 \mid y = 1) = \frac{\sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 1\right)}{\sum_{i=1}^{M} 1\left(y(i) = 1\right)} = 0$$

$$P(x_j = 1 \mid y = 0) = \frac{\sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 0\right)}{\sum_{i=1}^{M} 1\left(y(i) = 0\right)} = 0$$

Note that both the Prior Probabilities are 0, which implies both the Posterior Probabilities are 0. (ii)

$$\left(\prod_{j=1}^{N} P(x_j = v_j \mid y = 1)\right) P(y = 1) > \left(\prod_{j=1}^{N} P(x_j = v_j \mid y = 0)\right) P(y = 0)$$

$$\Rightarrow \quad 0 > 0$$

Therefore, we cannot classify.. And this happens very frequently coz not all the words are there in the training set, there might be so many words that are absent in the training set.

Prior Probabilities = 0 implies Posterior Probabilities = 0. This causes problems in classification. Therefore, we use the following prior probabilities instead. (ii)

$$P(x_j = 1 \mid y = 1) = \frac{(1) + \sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 1\right)}{(2) + \sum_{i=1}^{M} 1\left(y(i) = 1\right)}$$

$$P(x_j = 1 \mid y = 0) = \frac{(1) + \sum_{i=1}^{M} 1\left(x_j(i) = 1, y(i) = 0\right)}{(2) + \sum_{i=1}^{M} 1\left(y(i) = 0\right)}$$

(1) → Because of addition of 1, Prior probabilities will never be 0.

(2) → Since the Total Probability is supposed to be 1.

(a) $P(x_j = 1 | y = 1) + P(x_j = 0 | y = 1) = \frac{1}{2} + \frac{1}{2} = 1$

when the word "IITK" is not present in any training emails (SPAM or GENUINE).

This is basically termed as "LAPLACE SMOOTHING". (very important for practical implementation)

Laplace Smoothing avoids Prior probabilities = 0.

This is equivalent to adding the following emails to Training Set.

(i) Add a SPAM email which contains all the words in Dictionary

(ii) Add a SPAM email which contains None of the words in Dictionary

(iii) Add a GENUINE email which contains all the words in Dictionary

(iv) Add a GENUINE email which contains None of the words in Dictionary

Week 4 : Session 3

6. Naive Bayes Examples

Recall, Naive Bayes is a ML technique, which is used to predict DISCRETE responses, based on DISCRETE feature vector.

Let us consider a training set of people's opinion regarding Movies.

x_1 → Top Gun
x_2 → Avengers
y → AVATAR

Like = 1
Don't Like = 0
"y"

Assume we have the preferences for two movies x_1 (Top Gun) and x_2 (Avengers). What is the Probability that the person will Like / Dislike the movie y (AVATAR)?

(ii) Say, what is the Probability that a person Likes x_1 (Top Gun) and Dislikes x_2 (Avengers), who likes the movie y (AVATAR)?

The estimates of the prior probabilities are as follows.

	$x_1 = 0$	$x_1 = 1$	$P(x_1 = 0 \mid y)$	$P(x_1 = 1 \mid y)$
$y = 0$	No. of people who dislike x_1 and y $\} = 3$	No. of people who like x_1 and dislike y $\} = 10$	$P(x_1 = 0 \mid y = 0)$ No. of people who dislike x_1 and y / Total No. of people who dislikes y $= \frac{3}{13}$	$P(x_1 = 1 \mid y = 0)$ No. of people who like x_1 and dislike y / Total No. of people who dislikes y $= \frac{10}{13}$
$y = 1$	No. of people who dislike x_1 and like y $\} = 4$	No. of people who like x_1 and y $\} = 13$	$P(x_1 = 0 \mid y = 1)$ No. of people who dislike x_1 and like y / Total No. of people who like y $= \frac{4}{17}$	$P(x_1 = 1 \mid y = 1)$ No. of people who like x_1 and y / Total No. of people who like y $= \frac{13}{17}$

	$x_2 = 0$	$x_2 = 1$	$P(x_2 = 0 \mid y)$	$P(x_2 = 1 \mid y)$
$y = 0$	No. of people who dislike x_2 and y $\} = 5$	No. of people who like x_2 and dislike y $\} = 8$	$P(x_2 = 0 \mid y = 0)$ No. of people who dislike x_2 and y / Total No. of people who dislikes y $= \frac{5}{13}$	$P(x_2 = 1 \mid y = 0)$ No. of people who like x_2 and dislike y / Total No. of people who dislikes y $= \frac{8}{13}$
$y = 1$	No. of people who dislike x_2 and like y $\} = 7$	No. of people who like x_2 and y $\} = 10$	$P(x_2 = 0 \mid y = 1)$ No. of people who dislike x_2 and like y / Total No. of people who like y $= \frac{7}{17}$	$P(x_2 = 1 \mid y = 1)$ No. of people who like x_2 and dislike y / Total No. of people who like y $= \frac{10}{17}$

Total Number of people = 30

		P(y)
$y=0$	No. of people who dislike y } = 13	$P(y=0)$ = No. of people who dislike y / Total No. of people = $\frac{13}{30}$
$y=1$	No. of people who like y } = 17	$P(y=1)$ = No. of people who like y / Total No. of people = $\frac{17}{30}$

Now, what is the probability that a new person, who likes the movie x_1 ($x_1=1$) and dislikes the movie x_2 ($x_2=0$), will like the movie y?

(i) $P(y=1 \mid x_1=1, x_2=0)$ → Aposteriori Probability of person likes y, given the person likes x_1 and dislikes x_2.

$= P(y=u \mid \bar{x}=\bar{v})$

$= \frac{P(x_1=1, x_2=0 \mid y=1) \cdot P(y=1)}{P(x_1=1, x_2=0)}$ → Using Bayes rule

$= \frac{P(x_1=1 \mid y=1) \cdot P(x_2=0 \mid y=1) \cdot P(y=1)}{P(x_1=1, x_2=0)}$ → Simplified using Naive Bayes Assumption, since x_i's are conditionally independent given y.

Similarly, what is the probability that a new person, who likes the movie x_1 ($x_1=1$) and dislikes the movie x_2 ($x_2=0$), will dislike the movie y?

(ii) $P(y=0 \mid x_1=1, x_2=0)$ → Aposteriori probability of person dislikes y, given the person likes x_1 and dislikes x_2

$= P(y=u \mid \bar{x}=\bar{v})$

$$= \frac{P(x_1=1, x_2=0 \mid y=0) \cdot P(y=0)}{P(x_1=1, x_2=0)}$$ → Using Bayes Rule

$$= \frac{P(x_1=1 \mid y=0) \cdot P(x_2=0 \mid y=0) \cdot P(y=0)}{P(x_1=1, x_2=0)}$$ → Simplified using Naive Bayes Assumption

Now, the new person likes the movie y if

$$P(y=1 \mid x_1=1, x_2=0) > P(y=0 \mid x_1=1, x_2=0)$$

$$\Rightarrow \frac{P(x_1=1 \mid y=1) \cdot P(x_2=0 \mid y=1) \cdot P(y=1)}{P(x_1=1, x_2=0)} >$$

$$\frac{P(x_1=1 \mid y=0) \cdot P(x_2=0 \mid y=0) \cdot P(y=0)}{P(x_1=1, x_2=0)}$$

$$\Rightarrow P(x_1=1 \mid y=1) \cdot P(x_2=0 \mid y=1) \cdot P(y=1) > P(x_1=1 \mid y=0) \cdot P(x_2=0 \mid y=0) \cdot P(y=0)$$

Q_1 → Aposterior for $y=1$

Q_0 ← Aposterior for $y=0$

$$\Rightarrow \frac{13}{17} \times \frac{7}{17} \times \frac{17}{30} > \frac{10}{13} \times \frac{5}{13} \times \frac{13}{30}$$

$$\Rightarrow 0.178 > 0.128$$

Since the Aposteriori $Q_1 > Q_0$ (i) Q_1 is higher, the new person is "Likely" to like the movie y.

Laplace Smoothing Example

We do Laplace Smoothing in Prior probabilities, which avoids Prior probabilities being zero. This is because, if the Prior probabilities are zero, then the posterior probability would also be zero. And this causes problem in classification.

In Laplace Smoothing, we add 1 in the Numerator and 2 in the denominator.

The estimates of the prior probabilities are as follows.

	$x_1=0$	$x_1=1$	$P(x_1=0\|y)$	$P(x_1=1\|y)$
$y=0$	3	10	$\frac{3+1}{13+2}=\frac{4}{15}$	$\frac{10+1}{13+2}=\frac{11}{15}$
$y=1$	4	13	$\frac{4+1}{17+2}=\frac{5}{19}$	$\frac{13+1}{17+2}=\frac{14}{19}$

	$x_2=0$	$x_2=1$	$P(x_2=0\|y)$	$P(x_2=1\|y)$
$y=0$	5	8	$\frac{5+1}{13+2}=\frac{6}{15}$	$\frac{8+1}{13+2}=\frac{9}{15}$
$y=1$	7	10	$\frac{7+1}{17+2}=\frac{8}{19}$	$\frac{10+1}{17+2}=\frac{11}{19}$

Note: 2 is added in the denominator in order to maintain total probability = 1.

Say, $\frac{8}{19}+\frac{11}{19}=\frac{19}{19}=1$

Now, what is the probability that a new person, who likes the movie x_1 ($x_1=1$) and dislikes the movie x_2 ($x_2=1$), will like the movie y?

We can readily compute this using the Naive Bayes formula! (Refer previous example)

For $y=1$, $Q_1 = P(x_1=1|y=1) \cdot P(x_2=0|y=1) \cdot P(y=1)$

$$= \frac{14}{19} \times \frac{8}{19} \times \frac{17}{30}$$

$$= 0.1758$$

For $y=0$, $Q_0 = P(x_1=1|y=0) \cdot P(x_2=0|y=0) \cdot P(y=0)$

$$= \frac{11}{15} \times \frac{6}{15} \times \frac{13}{30}$$

$$= 0.1271$$

Since, the Aposteriori $Q_1 > Q_0$ (i.e) Q_1 is higher, the new person is "likely" to like the movie y.

Here, we've done computation for one particular movie. For an actual recommender system, we can calculate the Aposteriori for a whole lot of different movies, and we can recommend those movies with higher Aposteriori probability. This essentially becomes a Recommender System.