In Bernoulli's Naive Bayes classification initial steps are the same. We first need to construct our dictionary. If the training documents remain same dictionary will be same. Let us reproduce our training documents and the dictionary constructed using those documents.

D_1 = Upgrad is a great educational institution

D_2= Educational greatness depends on ethics

D_3= A story of great ethics and educational greatness

D_4=Sholay is a great cinema

D_5=good movie depends on good story

These documents will give following dictionary vector as we have already seen in the case of Multinomial Naive Bayes.

0	1	2	3	4	5	6	7	8	9	10	11
cinema	depends	educational	ethics	good	great	greatness	institution	movie	sholay	story	upgrad

Now we need to vectorize our training documents according to this dictionary. Vectorization process for Bernoulli's Naive Bayes is different from Multinomial Naive Bayes. Let us first see the vectorized feature vector for Multinomial Naive Bayes.

$$D = \begin{pmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \\ 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{pmatrix}$$

Notice a '2' at the bottom row of the matrix. Last row is the feature vector for our fifth document. As the fifth document (D_5=***good*** movie depends on ***good*** story) contains word "***good***" twice , it has appeared as 2 in 5th column of the last row. In Multinomial Naive Bayes, feature vectors capture the frequency of the words in the document.

Bernoulli's feature vectors are a bit different. It generates a Boolean indicator about each term of the vocabulary and equals to 1 if the term belongs to the examining document and 0 if it does not. Let us see the feature vector of our document for better understanding.

$$D = \begin{pmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \\ 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,1,0,0,0,1,0,1,0 \end{pmatrix}$$

See the difference. We have only 0's and 1's. The '2' of the last row has disappeared. We are only interested whether a word appears in a document or not.

Thus each document is represented as a 12-dimensional binary vector. Let us separate them acco rding to their class.

$$D^{education} = \begin{pmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \end{pmatrix}$$

$$D^{cinema} = \begin{pmatrix} 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{pmatrix}$$

Now how do we calculate the probability of a word occurring in a class. For example what is the probability of the word "ethics" appearing in the class "education"? Means we are formally trying to find out***P(w=ethics | C=education).*** How do we calculate this? Not as we did while calculating the same expression in Multinomial Naive Bayes. It is calculated using following formula

$$\textbf{\textit{P (w=ethics | C=education)}} = \frac{n_{education}(w_{ethics})}{N_{education}}$$

. $n_{education}(w_{ethics})$ means total no of word "***ethics***" in all the documents of class "***education***" and $N_{education}$ means total no of documents in class " ***education"***. $n_{education}(w_{ethics})$ our example is 2 and $N_{education}$ is 3. So, the probability of the word "ethics" appearing in the class "education" is $\frac{n_{education}(w_{ethics})}{N_{education}} = \frac{2}{3}$.

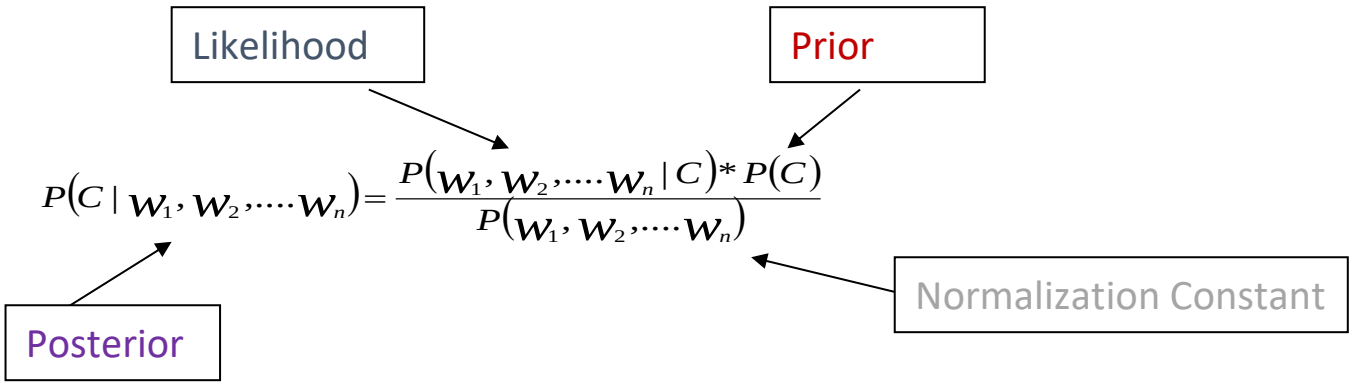
Now let us make a table showing probability of all the words in all the classes of documents using our feature vector of documents as we did for Multinomial Naive Bayes.

	$n_{education}(w)$	$p(w \mid c = education)$	$n_{cinema}(w)$	$p(w \mid c = cinema)$
w_1= cinema	0	0	1	1/2
w_2= depends	1	1/3	1	1/2
w_3= educational	3	3/3 =1	0	0
w_4= ethics	2	2/3	0	0
w_5= good	0	0	2	2/2=1
w_6= great	2	2/3	1	1/2
w_7= greatness	2	2/3	0	0
w_8= institution	1	1/3	0	0
w_9= movie	0	0	1	1/2
w_{10}=sholey	0	0	1	1/2
w_{11}= story	1	1/3	1	1/2
w_{12}=upgrad	1	1/3	0	0

Suppose we want to classify following document into "education" or "cinema".

“very good educational institution”

Let us first recall our Bayes’s formula written to suit hour this discussion.



Denominator of the right-hand side expression is generally ignored as that will be same for all cases and hence doesn't affect the final outcome. Let us calculate the "Prior". "Prior" is our prior knowledge of probability of a document of belonging to a certain class. For this discussion we will restrict to two classes "education" and "cinema". So what is the probability of a document belonging to "education "class?

$$P(education) = \frac{\#\ of\ document\ belonging\ to\ the\ class\ "education"}{\#\ of\ total\ documents}$$

$$= \frac{3}{5} = 0.6$$

Similarly

$$P(cinema) = \frac{\#\ of\ document\ belonging\ to\ the\ class\ "cinema"}{\#\ of\ total\ documents} = \frac{2}{5} = 0.4$$

These figures are our prior beliefs about any documents belonging to a certain class. Given any new document we believe that its probability of belonging it to education class is 60%.

Now let us focus on the likelihood term. Let P(w |C) be the probability of word w occurring in a document of class C; the probability of w not occurring in a document of this class is given by (1−P(w |C)). If we make the naive Bayes assumption, that the probability of each word occurring in the document is independent of the occurrences of the other words, then we can write the document likelihood P(D | C) in terms of the individual word likelihoods P(w |C):

$$P(w_1, w_2, \ldots w_n | C) = P(d|C) = \prod_{i=1}^{|V|} [\ d_i P(w_i|C) + (1-d_i)(1-P(w_i|C))]$$

What is d in the above expression ?

$\boldsymbol{d = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12})}$ is a feature vector of any document.

$\boldsymbol{d_i}$could be either "0" or "1". When $\boldsymbol{d_i}$**=1** then$(1 - \boldsymbol{d_i})$**=0** and When $\boldsymbol{d_i}$**=0**then$(1 - \boldsymbol{d_i})$**=1.** That means any one of the terms $d_i P(w_i|C)$ or $(1-d_i)(1-P(w_i|C)$ in the likelihood expression will be non-zero.This product goes over all words in the vocabulary. If word w_i is present, then $\boldsymbol{d_i}$**=1** and the required probability is $P(w_i|C)$; if word w_i is not present, then $\boldsymbol{d_i}$**=0**and the required probability is $1-P(w_i|C)$.

So how should we proceed to calculate likelihood of our test document "***very good educational institution"***. Its going to be different than how we calculated it for Multinomial Naïve Bayes. That might have been obvious from the likelihood equation for Bernoulli's Naïve Bayes. We didn't

bother about feature vector of test document then. But now we need to transform our test document to a feature vector as we did for training documents. So feature vector for our test document will be

d = (0,0,1,0,1,0,0,1,0,0,0,0) . It was to know what our feature vector would be. There are 4 words.

First, we should ignore those words in the document which are not part of the dictionary. In this case the word "very" is not part of the dictionary so we should ignore it.We only need to consider remaining words which are "good" , "educational" and "institution" . Feature vector has three "1" corresponding to these words at $d_3, d_5,$ and $d_8,$ .All other values of feature vector are zero as these words don't appear in our test document.

d = (0,0,1,0,1,0,0,1,0,0,0,0)

P(education|d) $\propto$ ***P(education)*** $\prod_{i=1}^{12}[\ d_i P(w_i|C=education)+(1-d_i)(1-P(w_i|C=education))]$

$$\propto \frac{3}{5}\times 1\times\frac{2}{3}\times 1\times\frac{1}{3}\times 0\times\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}\times 1\times 1\times\frac{2}{3}\times\frac{2}{3}$$

Let us see reasoning for few terms to understand the above expression.

First term -> $\frac{3}{5}$ This is prior for education and equal to ***P(education)***

Second term -> 1 First term $d_1,$ of the feature vector is 0 so we use $(1-d_i)(1-P(w_i|C=education))$. By putting $d_1,$**=0** and $P(w_1|C=education)$ =0 we get 1

Third term-> $\frac{2}{3}$ Second term $d_2,$ of the feature vector is 0 again so we again use $(1-d_i)(1-P(w_i|C=education))$. By putting d_2**=0** and $P(w_2|C=education)$ =$\frac{1}{3}$ form the probability table we get $\frac{2}{3}$**.**

Fourth term -> 1 Third term $d_3,$ of the feature vector is 1 a so we use $d_i P(w_i|C=education)$. By putting d_3**=1** and $P(w_3|C=education)$ =1 form the probability table we get **1.**

Fifth term -> $\frac{1}{3}$ Fourth term d_4 of the feature vector is 0 so we again use $(1-d_i)(1-P(w_i|C=education))$. By putting d_4=0 and $P(w_4|C=education)$ =$\frac{2}{3}$ form the probability table we get $\frac{1}{3}$.

Sixth term -> 0 . This is special and needs our extra attention. This will make the whole expression zero. Fifth term d_5 of the feature vector is 1 a so we use $d_i P(w_i|C=education)$. By putting d_5=1 and $P(w_5|C=education)$ =0 from the probability table we get 0.

There is no point for any further calculations. In Multinomial Naive Bayes we have already seen how to handle this. We use Laplace Smoothing. Formula of Laplace Smoothing for Bernoulli's Naive Bayes is a bit different and as follows.

$$P(w_t \mid C)=\frac{n_C(w_t)+1}{N_C+2}$$

$n_C(w_t)$ is the number of documents in class ***C*** in which word w_t is present and

N_C is total no of documents in class ***C*** . Derivation of this formula is beyond the scope of this discussion. Let us use this formula to recalculate our probability table for dictionary words for different classes.

	$n_{education}(w)$	$p(w\mid c=education)$	$n_{cinema}(w)$	$p(w\mid c=cinema)$
w_1= cinema	0+1=1	1/(2+3)=1/5	1+1=2	2/(2+2)=1/2
w_2= depends	1+1=2	2/(2+3)=2/5	1+1=2	2/(2+2)=1/2
w_3= educational	3+1=4	4/(2+3)=4/5	0+1=1	1/(2+2)=1/4
w_4= ethics	2+1=3	3/(2+3)=3/5	0+1=1	1/(2+2)=1/4
w_5= good	0+1=1	1/(2+3)=1/5	2+1=3	3/(2+2)=3/4
w_6= great	2+1=3	3/(2+3)=3/5	1+1=2	2/(2+2)=1/4
w_7= greatness	2+1=3	3/(2+3)=3/5	0+1=1	1/(2+2)=1/4

w_8= institution	1+1=2	2/(2+3)=2/5	0+1=1	1/(2+2)=1/4
w_9= movie	0+1=1	1/(2+3)=1/5	1+1=2	2/(2+2)=1/2
w_{10}=sholey	0+1=1	1/(2+3)=1/5	1+1=2	2/(2+2)=1/2
w_{11}= story	1+1=2	2/(2+3)=2/5	1+1=2	2/(2+2)=1/2
w_{12}=upgrad	1+1=2	2/(2+3)=2/5	0+1=1	1/(2+2)=1/4

Now let us use this table to calculate the class of the test document which feature vector is given by ***d = (0,0,1,0,1,0,0,1,0,0,0,0)***

We will use same formula and steps.

$$P(education|d) \propto P(education)\prod_{i=1}^{12}[\ d_i P(w_i|C=education)+(1-d_i)(1-P(w_i|C=education))]$$

$$\propto \frac{3}{5}\times\frac{4}{5}\times\frac{3}{5}\times\frac{4}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{2}{5}\times\frac{2}{5}\times\frac{2}{5}\times\frac{4}{5}\times\frac{4}{5}\times\frac{3}{5}\times\frac{3}{5}$$

$$= 0.0002717908992$$

$$P(cinema|d) \propto P(cinema)\prod_{i=1}^{12}[\ d_i P(w_i|C=cinema)+(1-d_i)(1-P(w_i|C=cinema))]$$

$$\propto \frac{2}{5}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{1}{4}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{3}{4}$$

$$= 0.000185394287109375$$

As ***P(education|d) > P(cinema|d)*** we can conclude that the document belongs to "***education***" class.

As you notice that the model of this variation is significantly different from Multinomial not only because it does not take into consideration the number of occurrences of each word, but also because it takes into account the non-occurring terms within the document. While in Multinomial model the non-occurring terms are completely ignored, in Bernoulli model they are factored when computing the conditional probabilities and thus the absence of terms is taken into account.