Lecture Notes

Supervised Classification – Naïve Bayes

In this module you understood another supervised learning classification algorithm called Naïve Bayes. Naïve Bayes is a probabilistic classifier which returns the probability of a test point belonging to a class rather than the label of the test point.

Bayes Theorem and Its Building Blocks

So here you learnt about probability, conditional probability, joint probability and how these two combines to make Bayes Theorem using a two way contingency table.

The **probability** of an event is the measure of the chance that the event will occur as a result of an experiment. The **probability** of an event A is the number of ways event A can occur divided by the total number of possible outcomes.

Probability is the chance of occurrence of an event, it can be defined as a **ratio of the number of desired outcomes to the number of total possible outcomes**. It is denoted as p(E), indicating the probability of an event E. In the numerator we have the number of favourable outcomes for this event E and in the denominator, we have the total number of outcomes corresponding to our data.

In the coin toss for the cricket match, we have only two possible outcomes namely Heads and Tails. So the denominator is right away fixed, i.e. the number of possible outcomes is 2. When your team captain called for Heads, the number of favourable outcomes become 1. So the probability of a Head occurring in this setup is the number of favourable outcomes which is 1 divided by the number of total possible outcomes which is 2 which gives us 1 over 2 or 0.50. So the probability is 50%.

Conditional probability:

$$\boldsymbol{P(A|B) = P(A \cap B)/P(B)}$$

P (A|B) is the probability of event A occurring, given that event B occurs. Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

Joint probability:
$\boldsymbol{P(A \cap B) = P(A|B) * P(B)}$

P (A and B). The probability of event A **and** event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written p(A ∩ B). Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

Bayes Theorem : Given Probability of an event B occurring given event A has already occurred and individual Probabilities of A and B we can find the reverse conditional probability P(A|B) by using what is called Bayes Theorem which is shown below.

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

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

Bayes Theorem

Naïve Bayes For Categorical Data

In this session, we looked at how Naïve Bayes can be used on categorical data to classify new test data points and developed an intuition on how naïve Bayes would work for continuous data. For categorical data, the probability can be calculated by simple counting.

Bayes Theorem is defined as P(Ci/X) = P(X/Ci) *P(Ci)/P(X) where Ci denotes the classes and X denotes the features of the data point..

- The effect of the denominator P(x) is not incorporated while calculating probabilities as it is a scaling factor.
- Naïve Bayes follows an **assumption** that the variables are conditionally independent given the class i.e. P(X=convex, smooth/C=edible) can be written as P(X=smooth/C=edible)*P(X=convex/C=edible). The terms P(X=smooth/C=edible) and P(X=convex/C=edible) can simply be calculated by counting the data points. Therefore, the name 'naïve'.
- P(Ci) is known as the prior probability. It is the probability of an event occurring before the collection of new data. Prior plays an important role while classifying, when using Naïve Bayes, as it highly influences the class of the new test point.

- P(X/Ci) represents the likelihood function. It tells the likelihood of a data point occurring in a class. The conditional independence assumption is leveraged while computing the likelihood probability.
- The effect of the denominator P(x) is not incorporated while calculating probabilities as it is the same for both the classes and hence can be ignored without affecting the final outcome.

- P(Ci/X) is called the posterior probability, which is finally compared for the classes and the test point is assigned the class whose Posterior probability is greater.
- The prior probability incorporates our '**prior beliefs**' before we collect specific information.
- The likelihood function updates our prior beliefs with the new information.
- The posterior probability is the final outcome which combines **prior and beliefs information i.e. the likelihood function.**

Classification is done in Naïve Bayes using the MAP or Maximum aposteriori rule . This rule suggests that a point will be classified to a category whose posterior is greater than the posterior of the other class.

Naïve Bayes For Text Classification

In this session you understood how the Multinomial Naïve Bayes Classifier works in the backend with help of a hand worked out example. Also we saw the Python implementation of the worked out example and also built a SMS spam ham classifier.

Let us quickly recall the pre processing steps given the test documents. See the below picture for test documents.

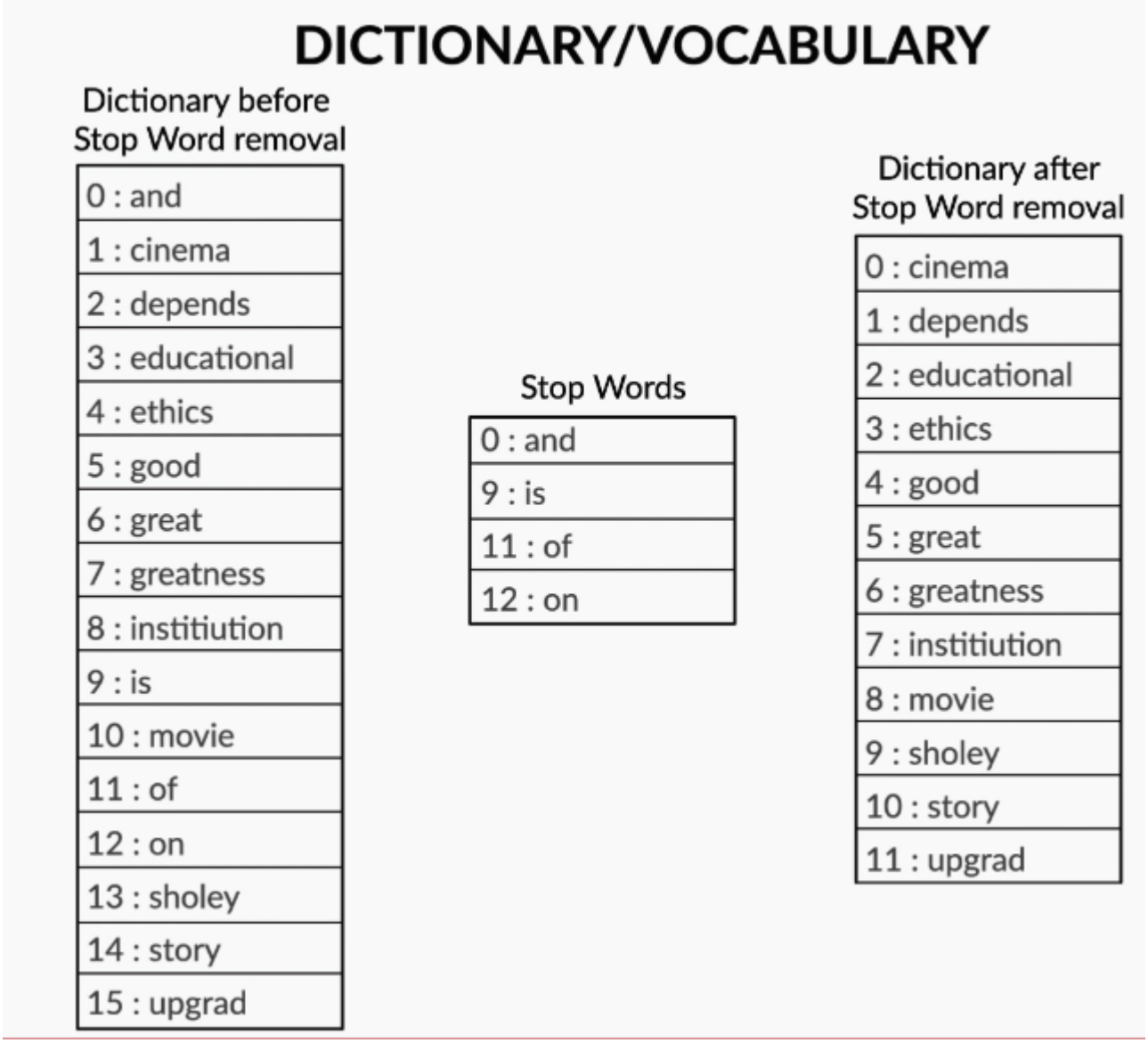
Test Dataset

	Document	Class
0	UpGrad is a great educational institution.	education
1	Educational greatness depends on ethics	education
2	A story of great ethics and educational greatness	education
3	Sholey is a great cinema	cinema
4	good movie depends on good story	cinema

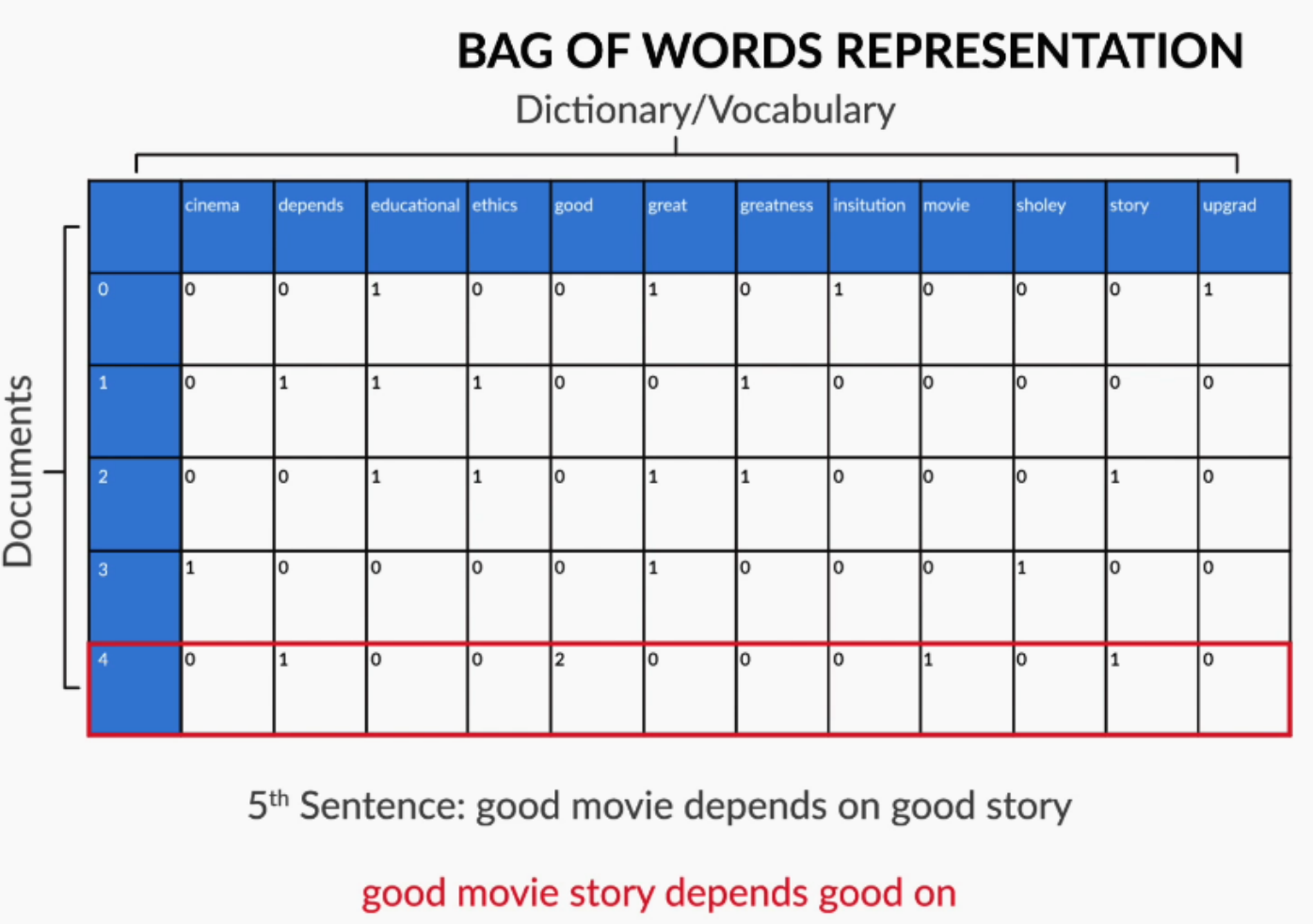
After this we built a dictionary/vocabulary denoted by |V| from the above documents after removing the **STOP WORDS.**

Stop Words are basically the words which are inconsequential in terms of providing any discriminatory information helpful in classification process.

You can see the dictionary below with and without stop words.



Now that we have the dictionary ready , what we do next is to convert it into a Bag of Word Representation shown below



	cinema	depends	educational	ethics	good	great	greatness	insitution	movie	sholey	story	upgrad
0	0	0	1	0	0	1	0	1	0	0	0	1
1	0	1	1	1	0	0	1	0	0	0	0	0
2	0	0	1	1	0	1	1	0	0	0	1	0
3	1	0	0	0	0	1	0	0	0	1	0	0
4	0	1	0	0	2	0	0	0	1	0	1	0

Bag of Word Representation basically breaks the sentences available into words and makes order of the irrelevant as if the words were put in a Bag and shuffled. All that we are concerned about is the no. of occurrences of the words present (Multinomial way). If you notice that both the sentences shown in the figure above have same feature representation.

Post this you saw the same representation in Array format for separate classes with total no. of occurrences of each words in the respective classes and the total no. of occurrences of all words in the respective classes.

$$[\text{cinema}, \ldots\ldots\ldots., \text{upgrad}]$$

$$D^{\text{education}} = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \end{bmatrix} \quad 13$$

$$0,1,3,2,0,2,2,1,0,0,1,1$$

$$D^{\text{cinema}} = \begin{bmatrix} 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix} \quad 8$$

$$1,1,0,0,2,1,0,0,1,1,1,0$$

So now we can calculate the posterior probability to identify that which class the new document belongs to.

Prior

$$P(\text{education}) = 3/5$$

$$P(\text{cinema}) = 2/5$$

$$\left.\begin{matrix} P(\text{education}|w_1,w_2,\ldots\ldots,w_n) \\ P(\text{cinema}|w_1,w_2,\ldots\ldots,w_n) \end{matrix}\right] \text{Posterior}$$

$$P(\text{cinema}|w_1,w_2,\ldots\ldots,w_n) = \frac{P(w_1,w_2,\ldots\ldots,w_n|\text{class})\ P(\text{class})}{P(w_1,w_2,\ldots\ldots,w_n)}$$

$$= P(w_1|c)\ P(w_2|c).\ldots\ldots P(w_n|c) \times P(c)$$

Note that due to the assumption of Naïve Bayes we could separate the likelihoods like above. Also, that we have ignored denominator as it is the same for both the classes and hence can be ignored without affecting the final outcome.

Let us now look into the tabular representation of all the numbers which helped in eventually classifying the documents.

Dictionary/Vocabulary	$n_{education}^{(w)}$	p(w|c = education)	$n_{cinema}^{(w)}$	p(w|c = cinema)
w_1 = cinema	0	0	1	1/8
w_2 = depends	1	1/13	1	1/8
w_3 = educational	3	3/13	0	0
w_4 = ethics	2	2/13	0	0
w_5 = good	0	0	2	2/8
w_6 = great	2	2/13	1	1/8
w_7 = greatness	2	2/13	0	0
w_8 = institiution	1	1/13	0	0
w_9 = movie	0	0	1	1/8
w_{10} = sholey	0	0	1	1/8
w_{11} = story	1	1/13	1	1/8
w_{12} = upgrad	1	1/13	0	0

$$P(education|w_1,w_2,.....,w_n)$$

$$P(cinema|w_1,w_2,.....,w_n) \propto P(w_1,w_2,.....,w_n|c)\,P(cinema)$$

$$\propto P(w_1|c) \times P(w_2|c)..... \times P(cinema)$$

$$P(w_3|cinema) = 0$$

$$P(w_3|education) = 3/13$$

- 2nd and 4th column contains the total number of occurrences of individual words of dictionary in their respective classes.
- 3rd and 5th columns contain the probability of the word being present in the individual classes

Using the above table we classified couple of documents to education class based on the maximum aposteriori rule which states that a point will be classified to a category whose posterior is greater than the posterior of the other class.

Laplace Smoothing:

This helps in overcoming the zero sum probability issue which hinders in classification due to inconclusive aposteriori values.

You basically add +1 to the numerator and +|V| to the denominator. And by doing this you ensure that some probability is reserved for unseen words in a class but which present in our vocabulary.

Let us see how the tabular representation looks after Laplace smoothing

	$n_{education}^{(w)+1}$	p(w|c = education)	$n_{cinema}^{(w)+1}$	p(w|c = cinema)
w_1 = cinema	0+1=1	1/(12+13)=1/25	1+1=2	2/(12+8)=2/20
w_2 = depends	1+1=2	2/(12+13)=2/25	1+1=2	2/(12+8)=2/20
w_3 = educational	3+1=4	4/(12+13)=4/25	0+1=1	1/(12+8)=1/20
w_4 = ethics	2+1=3	3/(12+13)=3/25	0+1=1	1/(12+8)=1/20
w_5 = good	0+1=1	1/(12+13)=1/25	2+1=3	3/(12+8)=3/20
w_6 = great	2+1=3	3/(12+13)=3/25	1+1=2	2/(12+8)=2/20
w_7 = greatness	2+1=3	3/(12+13)=3/25	0+1=1	1/(12+8)=1/20
w_8 = institiution	1+1=2	2/(12+13)=2/25	0+1=1	1/(12+8)=1/20
w_9 = movie	0+1=1	1/12+13)=1/25	1+1=2	2/(12+8)=2/20
w_{10} = sholey	0+1=1	1/(12+13)=1/25	1+1=2	2/(12+8)=2/20
w_{11} = story	1+1=2	2/(12+13)=2/25	1+1=2	2/(12+8)=2/20
w_{12} = upgrad	1+1=2	2/(12+13)=2/25	0+1=1	1/(12+8)=1/20
	13+12 = 25		8+12 = 20	

Some takeaways from the hand worked out examples studied in video :

- Ignore the words which are not present in our vocabulary but present in the test document
- If the word appears n no. of times in the test document the probability is raise to n times in the posteriori calculation
- Add a +1 to the numerator and +|V|to the denominator of the individual likelihoods of the words to avoid zero sum probability

Introduction to Bernoulli Theorem

- Recall the difference between Multinomial and Bernoulli way of building feature vector. Unlike Multinomial way which is concerned about the no. of occurrences of the word in the class , in Bernoulli we are just concerned about whether the word is present or not.

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

$$\uparrow$$

$$0,1,0,0,1,0,0,0,1,0,1,0$$

----------THE END----------