Natural Language Processing

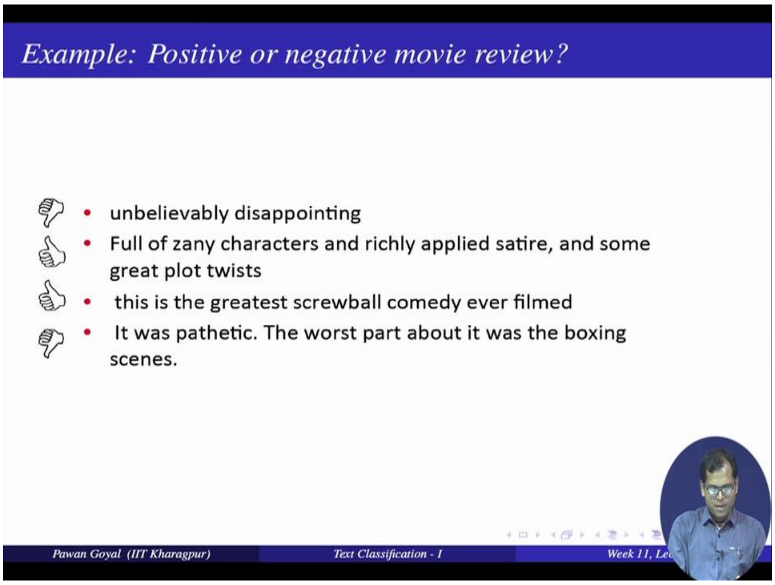
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 54
Text Classification – I

Welcome back to the fourth lecture of this week, so we have been talking about various applications and we finished our discussions on text summarization. So, in the last two lectures of this week we will focus on one very important application in NLP that is called text classification. So, we will discuss what are the different applications were this whole approach to text classification can be used and we will talk about one simple baseline that is how do you use a Naive Bayes classifier for text classification and we will also spend some time in discussing; how do you evaluate your models. So, once you have built your system and how do you evaluate your system.

So, starting with what do we mean by text classification, what are some of the examples you can you can think of. So, text classification as the name would say, so you are given a piece of text; it can be a sentence, it can be a document, it can be a paragraph or any other unit of text and even to classify it into certain categories and the categories can depend on what is your application and that is where this becomes a very very generic problem because many of the problems in the NLP, you can treat them as text classification problems. We will see some examples.

(Refer Slide Time: 01:30)



Example: Positive or negative movie review?

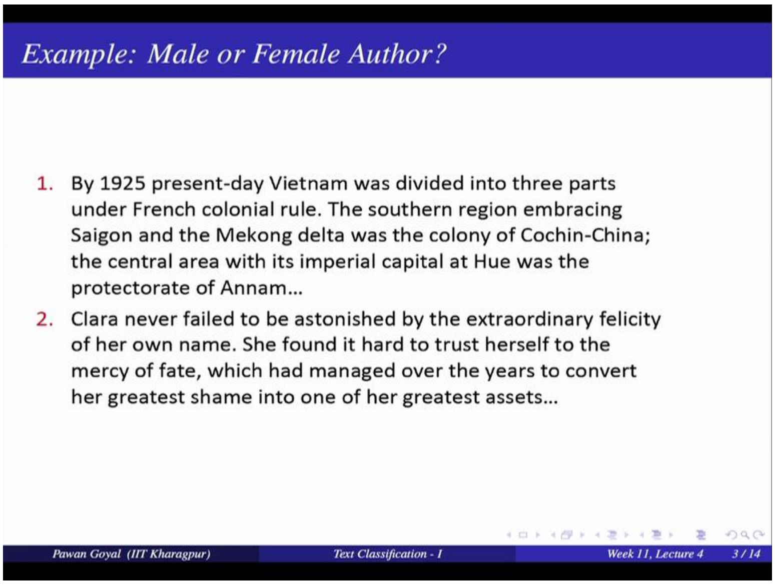
- 👎 unbelievably disappointing
- 👍 Full of zany characters and richly applied satire, and some great plot twists
- 👍 this is the greatest screwball comedy ever filmed
- 👎 It was pathetic. The worst part about it was the boxing scenes.

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Le

So suppose the problem is you are giving a movie review and you have to find out is it a positive review or a negative review. So, here are some examples, so this is the field of sentiment analysis. So, you have different reviews here like unbelievably disappointing and this you would say immediately; this is a negative feeling, then full of zany characters and richly applied satire and some great plot twists and we will say this is a positive review, this is a greatest screwball comedy ever filmed again positive review it was pathetic and so on this becomes a negative review.

So, now the problem here is, suppose you are given a lot of movie reviews or say product reviews or hotel reviews and we have to find out each and individual sentence is it talking about some positive sentiments for this product or negative sentiments for the product, how do I do that automatically. So, this becomes a classification problem, so given a text I want to classify it into one of these classes; positive, negative or neutral.

(Refer Slide Time: 02:32)



Example: Male or Female Author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

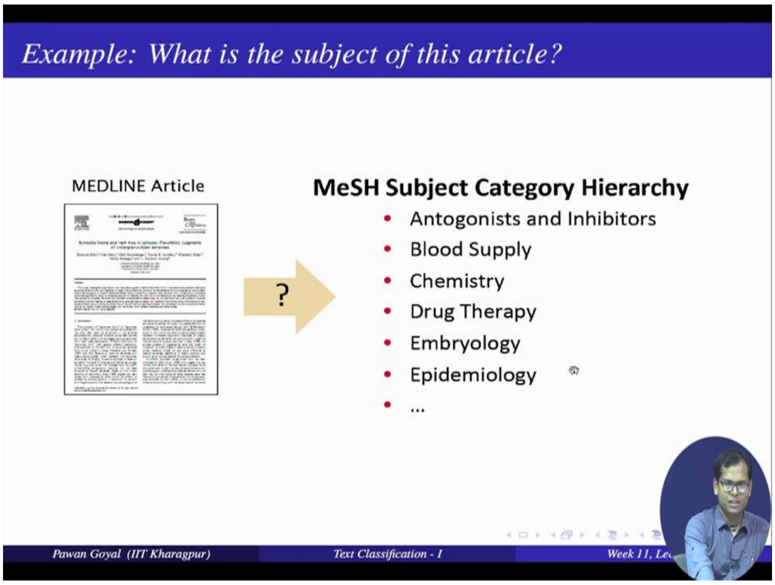
Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 3 / 14

Let us take another example, so here this is about authorship attribution. So given a piece of text; find out certain demographics of the author. So, what is the gender of the author is it a female author or the male author, what is the age of the author, is he in the early 20’s or 30’s or is he a very old author and there can be many other things like what is the country that author belongs to and then you can also go to very very fine label to saying who is the author for this particular paragraph or text. Now why do you want to treat as a classification problem? So you will say that authors for a particular demography will have certain traits. So, they will have a particular style of writing, so can I try to seek given a text what style of writing it matches and match to the corresponding class.

So, here for example there are two pieces of text and the problem is whether this is a male author or a female author and then you can use certain heuristics and some sort of studies that people have done that when they are male authors, they will write lot of about lot of different facts and then female authors; they will be some certain opinions; opinions will be in majority. So, when you see these two piece of text and you will say the first one is talking about lot of different facts about the year, information and the place information and the second text is about certain opinions; certain subjective information and then this might be used to say which one is the male author, which one is the female author.

But in general when you are given a lot of data about writings from male and female author, you can try to learn a model that can classify a piece of text; is it from a male author or a female author.

(Refer Slide Time: 04:21)



Example: What is the subject of this article?

MEDLINE Article

?

MeSH Subject Category Hierarchy

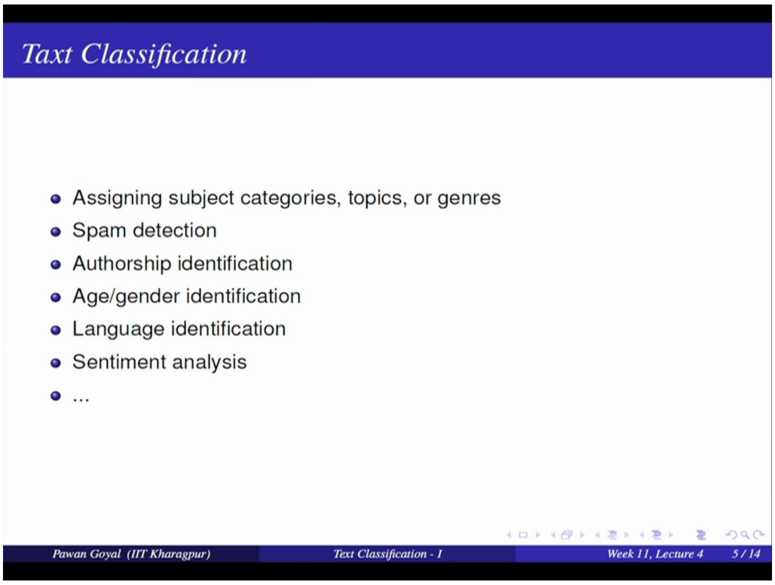
- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Pawan Goyal (IIT Kharagpur) *Text Classification - I* *Week 11, Le*

Similarly so here you have a research article like Medline article and it can be from any research domain and what is the problem? For that, suppose for your digital library; you have a set of categories or broad topics and given a new article, you want to assign it to one of these topics. So, that is putting this in somewhere in your hierarchy of categories, so here for example, the categories can be antagonists and inhibitors; blood supply, chemistry, drug therapy, embryology, epidemiology. So, all these are different categories that you have in your mesh hierarchy; given a scientific article you want to put it in one of these categories. So, this is again a classification problem all these categories are your classes and given a piece of text, you want to find out what is the category it matches to.

So, immediately start seeing that this text classification is a very very wide problem and different, different applications you can always convert them to some sort of text classification problem.

(Refer Slide Time: 05:25)



Taxt Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language identification
- Sentiment analysis
- ...

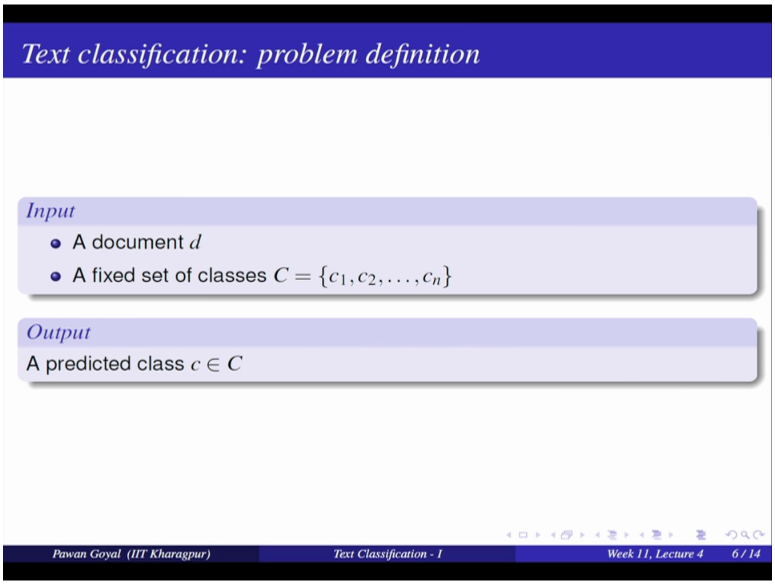
Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 5 / 14

So, like assigning subject categories, topics or genres this is one category of text classification; then it can be spam detection, you are given an email or it can be now tweets, movie reviews or product reviews, is it spam or not. So, it is classification problem is it spam or not spam then authorship identification, so we talked about this; who is the author, find out the demographies of the author and so on, age and gender identification, language identification so that is you are given the piece of text, find out what is the language it belongs to.

So, here what is important is that when you are, so for example, let us talk about English versus Hindi. So, whether language is English or Hindi, so what do you think is it a simple problem or difficult problem? So, suppose you are writing Hindi in Devanagari then this basis on the script, you can find out whether its English or Hindi, but when you type Hindi in your say in your comments, Facebook posts or twitter. So, do you write in Devanagari or do you mainly use the roman script, you mainly use the roman script and that is we use the transliteration for writing in Hindi.

So that means, when have the English text and the Hindi text they have the same script. So, script is same; that means, you have to go to the actual word level to find out is it coming from Hindi or English and that is where the problem starts, the same words can be written both in English and Hindi, so same way For example, take pure in English and pure in Hindi; they will have roughly the same transliteration P u r e. So, given this word, it will be difficult to find out if it is a English or Hindi word, so like that there will be different problems and given a piece of text; what is the language there you need to build up some different language specific models. Then there is a field of sentiment analysis, that is what we talked about all the people give an entire week for talking about opinions and sentiment analysis that will be the last week of this course.

(Refer Slide Time: 07:46)



Text classification: problem definition

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$

Output

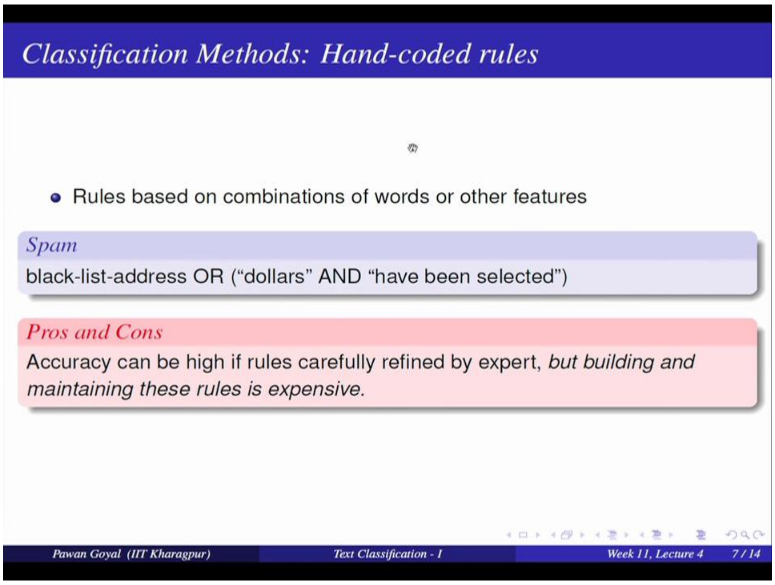
A predicted class $c \in C$

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 6 / 14

So, let us formally define this problem of text classification, so what do you have? You are given a document, by document I mean a piece of text that is the unit that you want to classify. So, in general it can be a sentence paragraph or whatever, so you have a document d and you have fixed set of classes C; C equal to C n. So, this document d you want to classify it into one of these n classes of course, there can be some variations where you may want to assign it to many of the classes or so we will talk about it, but let us take the simplistic assumption that each document belongs to one and only one class. So, given a document I want to assign it to one of these classes, this is a text classification problem.

Now, so my output will be one of the classes, one of the class from all these, from this set of classes.

(Refer Slide Time: 08:45)



Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

black-list-address OR (“dollars” AND “have been selected”)

Pros and Cons

Accuracy can be high if rules carefully refined by expert, *but building and maintaining these rules is expensive.*

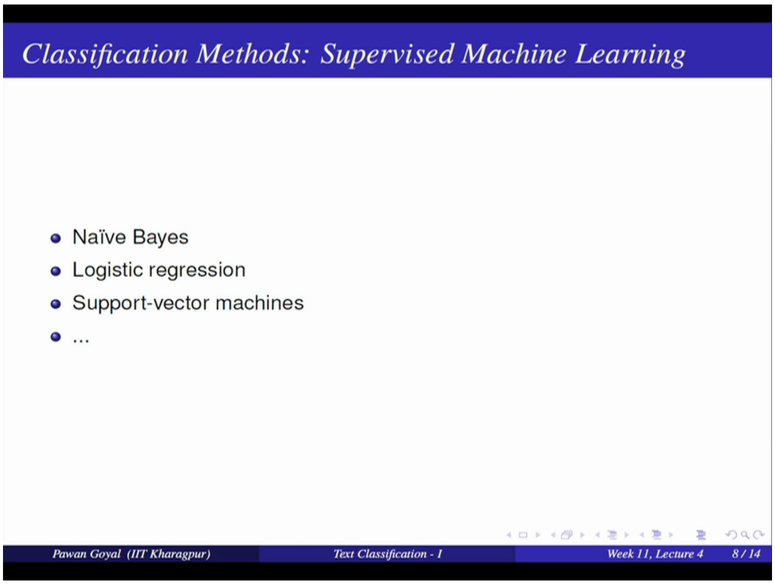
Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 7 / 14

So, what can be the simplistic method that I can use, so like many other application that we have seen. So, we can also use some hand-coded rules, so let us say you have the problem of spam detection, you have an incoming email and you want to classify into spam or not a spam. So, what will be the simplistic model, so simplistic model could be, try to see some spam emails and make some hand-coded rules. So, what is the common spam that you remember? So it is like you have been selected for some: 100 million dollars or 750000 dollars and so on, this is a very common form of a spamming email. So you can have some hand-coded rules like; if the email is coming from some blacklist addresses. So, suppose you have a list of blacklist either URLs and email ids. So, if it is coming from there or if it contains something like dollars and have been selected, so then it is a spam.

So, this is one form of hand-coded rule that you can build by seeing what are the different spams; that email that I see regularly and like that you can try to find out many other spams and build these various rules; that is one approach.

So, like the other application that we have seen, so what are the pros and cons for using hand-coded rules? So, pros are that if you are an expert and you can write very good rules, then in general they will be very very precise, they will be quite accurate, so you will get a high precision. So, whenever you have something like dollars and have been selected you will probably be; it will probably be a spam email, but other problem is that you do not know how many such rules you have to build. So, if you do not have sufficient rules, you will not have a very high recall, so you will not be able to find out all the spams. So, we will talk about these evaluation measures precision spams also in detail during this topic. So, maintaining is also quite expensive.

(Refer Slide Time: 11:01)



Classification Methods: Supervised Machine Learning

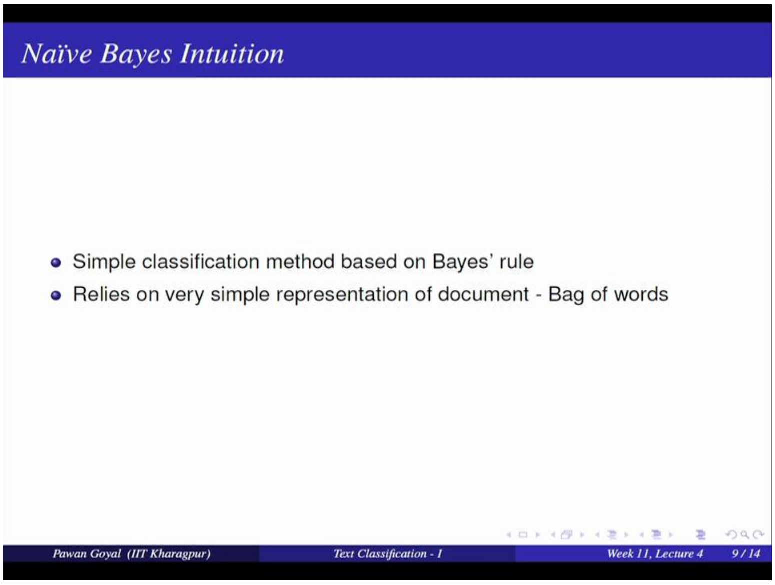
- Naïve Bayes
- Logistic regression
- Support-vector machines
- ...

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 8 / 14

So if you do not want to use hand-coded rules; what is the other option? So, you will use some supervised classification method and so some very popular methods are like Naive Bayes classification, logistic regression that is similar to what we talked in terms of maximum entropy model. So, what is the probability of a class given an input and a history x, so you will say this is linear function over the various features; sigma, lambda f i and then you have an exponent term; exponent of summation lambda i f i.

So, this is a Maxent model or logistic regression model, but you can also use other models like Naive Bayes models and support vector machines, SVM and all these require thinking of various features that would be helpful for this particular task. So in these two lectures, we will focus on one model that is Naive Bayes model and that is one of the very powerful baselines for text classification. So, this is the default model that you would use for any text classification task, so we will see what is the Naive Bayes model and we will also do a working example for this.

(Refer Slide Time: 12:15)



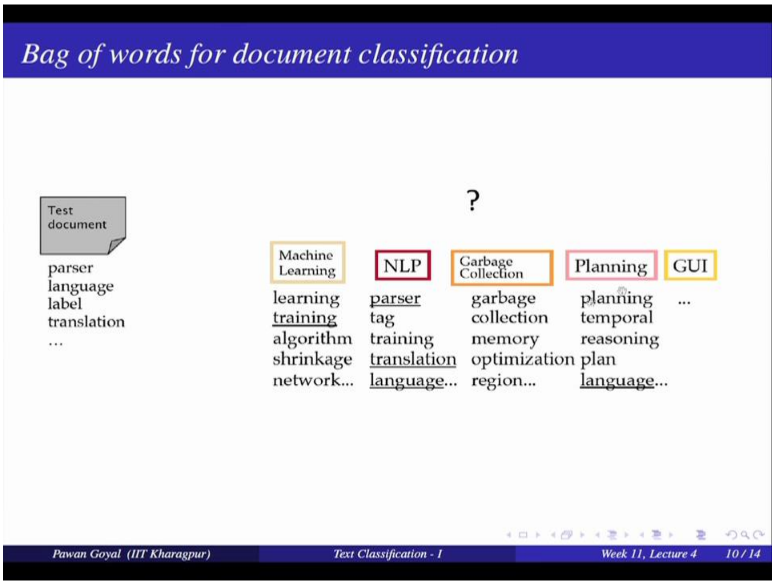
Naïve Bayes Intuition

- Simple classification method based on Bayes' rule
- Relies on very simple representation of document - Bag of words

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 9 / 14

So, what is the Naive Bayes model? So that is; it uses Bayes rule to do classification of text into certain classes and this relies on a very simple representation of documents that is bag of words and by now you understand what is bag of words model that is; you are given a piece of text, now in bag of words model the order in which the words occur in the document do not matter, what matters is that what all words are there. So, it is like a set of words; that is the bag of words assumption, now what are the other assumption that Naive Bayes model uses; let us see that model in detail.

(Refer Slide Time: 12:57)



Bag of words for document classification

Test document

parser
language
label
translation
...

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	...
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network...	language...	region...	language...	

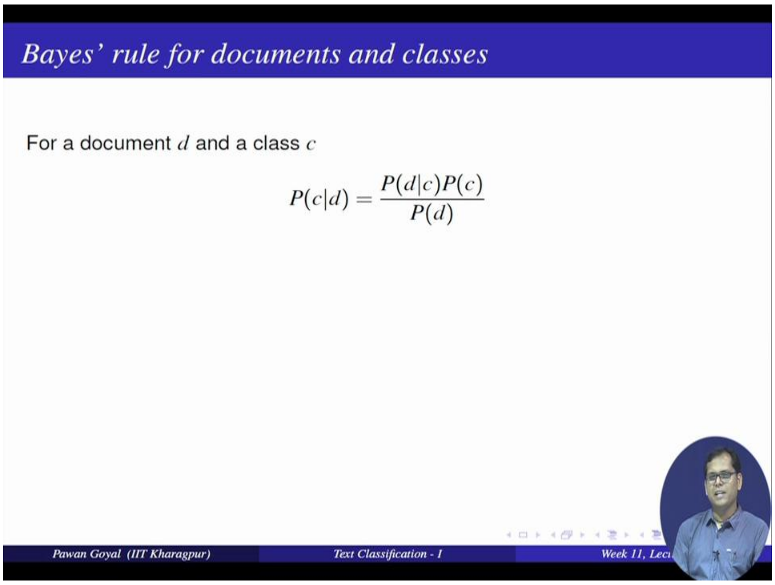
Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 10 / 14

So, let us quickly see what do I mean by bag of words model for document classification. So assume that you have certain documents with you and you know also the category, so suppose you are talking about certain research domains and you want to put all your documents or scientific articles into one of these domains. For example, here you have an article from machine learning, another from NLP then garbage collection, planning GUI and so on.

Now these documents will contain certain words, so what is bag of words model? So, you assume that document is nothing, but this bag of words. So, here all these words that occur in the document; corresponds to this document learning, training, algorithm, shrinkage, network etcetera; they all this is all the document and then this is assigned a category of machine language. Similarly here parser, tag, training, translation, language these are all the words that occur in the document and document is assigned the category of NLP and so on for all these documents.

So, now what is your task? So you know there will be some data that is already labeled with these classes; at test time you will get a document but you do not know what is the class for this document. So, for example, at test time you get a document with these words parser, language, label, translation and you do not know what is the class of this document; among all these possible categories. So here based on what are the words that are occurring in this document, you want to assign it to one of these classes and there you are using the bag of words assumption that I know what are all the words that are there and I do not care about the order information.

(Refer Slide Time: 14:48)



Bayes' rule for documents and classes

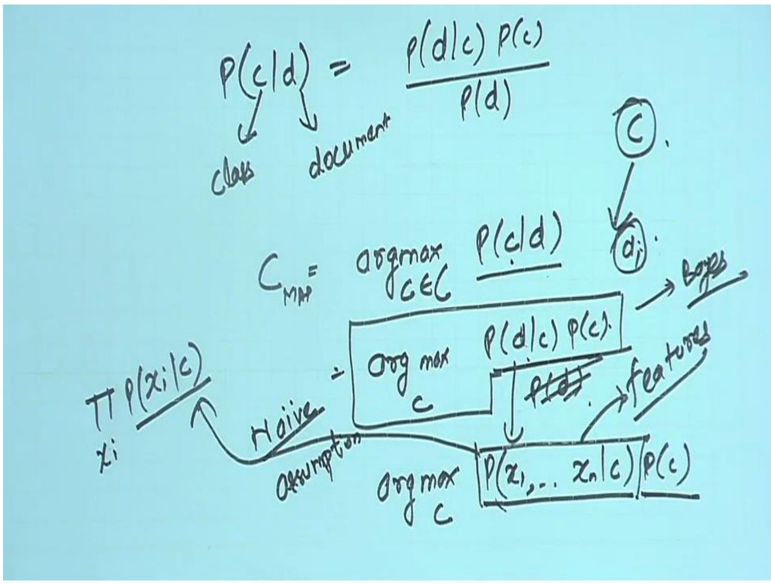
For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lec

Now once I have the scenario, how do I use the Naive Bayes model for classification? So, all of you know the Bayes rule by now, so what is the Bayes rule.

(Refer Slide Time: 15:00)


P(c|d) = p(d|c) P(c) / P(d)
class
document
c
d_i
C_MAP = argmax_{c∈C} P(c|d)
= arg max_c P(d|c) P(c) / P(d)
Bayes
features
arg max_c P(x_1,... x_n|c) P(c)
Naive
assumption
∏_{x_i} P(x_i|c)

So, it would say that probability of a class given a document would be; so how do you write it; P; c given d probability of the class given a document is nothing, but probability d given c P c given P d. So, now why are we using the Bayes rule here, c is a particular class and d is a document. Now I want to find out the probability for all the different classes given this test document and I want to take the one that is having the maximum probability. So, my problem is take the class that is giving the maximum probability argmax; c in the set such that probability c given d is the highest and you can also give the maximum of posterior probability; this is the class that you get by Bayes rule.

Now so the question is how do you compute this probability; probability of a class given d for all the different classes. Now what is the Naive Bayes model, so if you remember we talked about two different families of model; one was generative models, another was discriminative models, so Naive Bayes is a kind of generative model. So, that is it will say that you have the class first and then you generate different documents for that class. So, what would happen in the model? So you have the class C and then you are generating the document d from this class.

So, that is suppose I want to generate a review; it says that first you think that whether you want to generate a positive or negative review. So, class will come first and then for that class you will generate the document, so that is how the probabilities will flow. So, you can find the probability with document given the class from the generative model. So, that is why you cannot compute this directly, so you have to compute probability d given C. So, that is where the Bayes rule comes into picture, so I want to convert that into probability d given c and how do I do that using the Bayes rule. So, this becomes argmax over all classes; probability d given c, probability c given probability d.

And now because P d is common for all the classes; yes you know the given document, so this you can remove. So, this is the probability that I have to estimate argmax over c probability d given c, probability of c; what is probability of c? That is the prior probability of the class, how much this class is prevalent in this collection, this is if I do not know about this document, what can I say about which class is more probable than other this is the prior probability and then this tells you given this data; when you have seen the document then what is the probability of document, getting generated from this class. Together they give you the posterior probability of class given the document. So, P c you can easily compute; if you are given a training data, you can find out how often this class c occurs among all the classes.

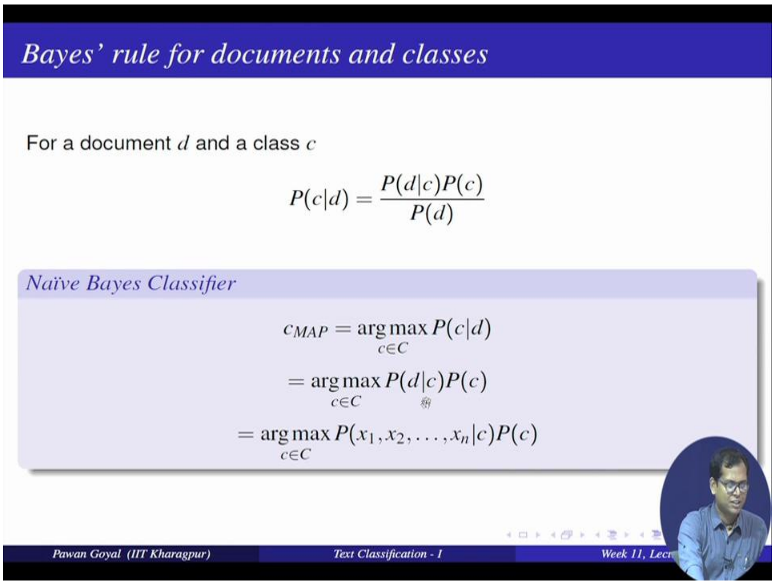
Now, how do you compute probability d given c? Now for that you have to see what is your document, your document is nothing but bag of words. So, suppose the document contains some words x 1 to x n. So, then you can convert d to x 1 to x n and you can write like; this is same as argmax over c probability x 1 to x n; given c probability c and where x 1 to x n are various. So, in general you can call these as various features, if you are only using your words then features will be only words, if you are using some other information like it may be the time of the document or the length of the document and many other things then this is also possible here.

So, this is that you get by this simply using the Bayes rule. So, now you have this, so again this one is easy, but this might be difficult to compute; what is the probability of getting all these features together given this class and that is where you make an Naive Bayes assumption. So, this is the naive assumption that is why this is called the Naive Bayes model. So, what is the naive assumption here? So that is the probability of all these features given the class can be written as; so this whole thing can be written at multiplication of probability x i given c; for all x i.

So, that is each feature is conditionally independent of each other given the class. So, this becomes; this simplifies this model to a very large extent. So, now you have to worry about the joint probability of the all the features, you can talk about the individual probability of different features and then you multiply. So, this is from the naive assumptions and that is where you are using the Bayes rule and together that is why this is called the Naive Bayes model.

So, you know what is the naive assumption here and you are using the Bayes rule for computing this probability.

(Refer Slide Time: 20:43)



Bayes' rule for documents and classes

For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes Classifier

$$c_{MAP} = \arg\max_{c \in C} P(c|d)$$

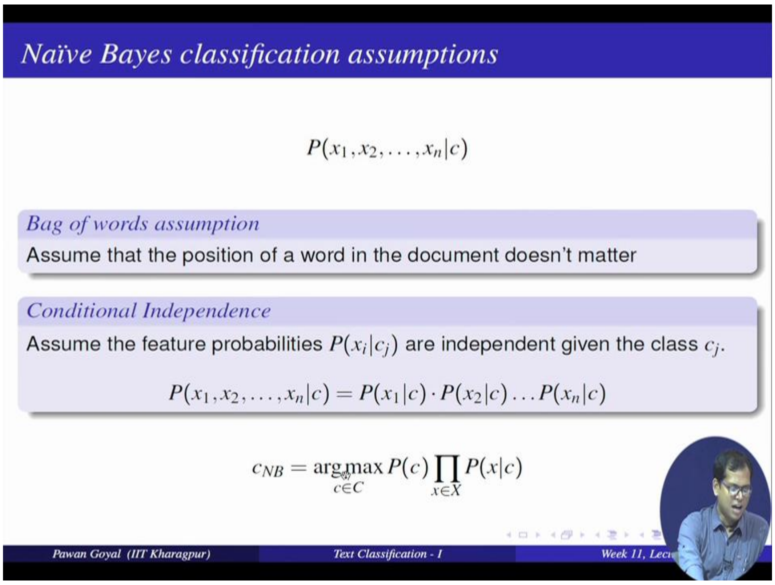
$$= \arg\max_{c \in C} P(d|c)P(c)$$

$$= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c)$$

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture

So, now if we go back, so we want to compute the find out the class that gives the highest probability; probability c given d and then we write using Bayes rule in this form and then we say document d is nothing, but all the features that are there in the document, so this is collection of the features x 1 to x n given the class c.

(Refer Slide Time: 21:02)



Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n|c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c_j.

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \ldots P(x_n|c)$$

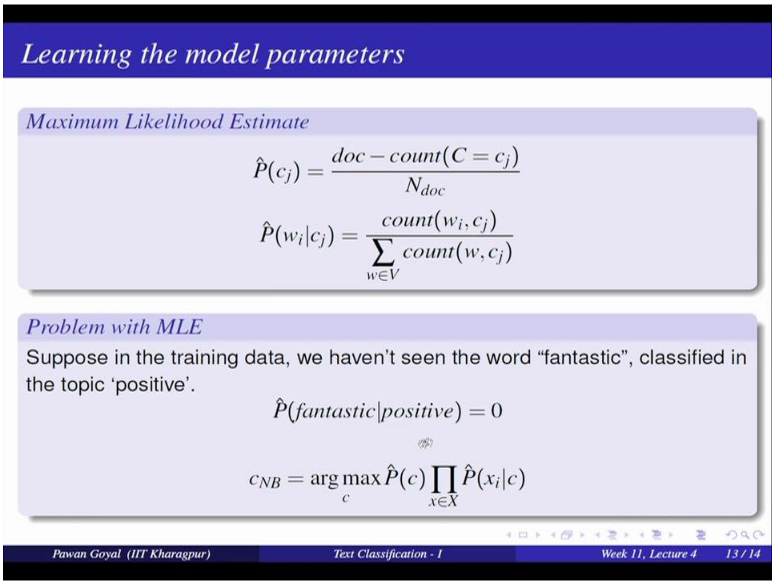
$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in X} P(x|c)$$

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture

Then how do you compute probability x 1 to x n given the class c? So, there you are making two assumptions one is bag of words that is the position of a word in a document does not matter. So, x 1 to x n whatever order they occur is immaterial, so I am only talking about a set of words here and then what is the other conditional independence, the naive assumption that the feature probabilities x i given c j are conditionally independent of each other; given the class c j. So, I can write probability x 1 to x n given c as probability x 1 given c; times probability x 2 given c up to probability x n given c.

So, now I have taken all these assumptions and now I have a very simplified model that is I will take the class that gives the maximum argmax over P c times P x given c and x here are all my different words. Now the next question is how do I compute these probabilities from my training data, so how would you compute probability c. So, P c would be; what is the probability of this class in the training data.

(Refer Slide Time: 22:20)



Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc-count(C=c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i,c_j)}{\sum_{w\in V} count(w,c_j)}$$

Problem with MLE

Suppose in the training data, we haven't seen the word “fantastic”, classified in the topic ‘positive’.

$$\hat{P}(fantastic|positive) = 0$$

$$c_{NB} = \arg\max_{c} \hat{P}(c)\prod_{x\in X}\hat{P}(x_i|c)$$

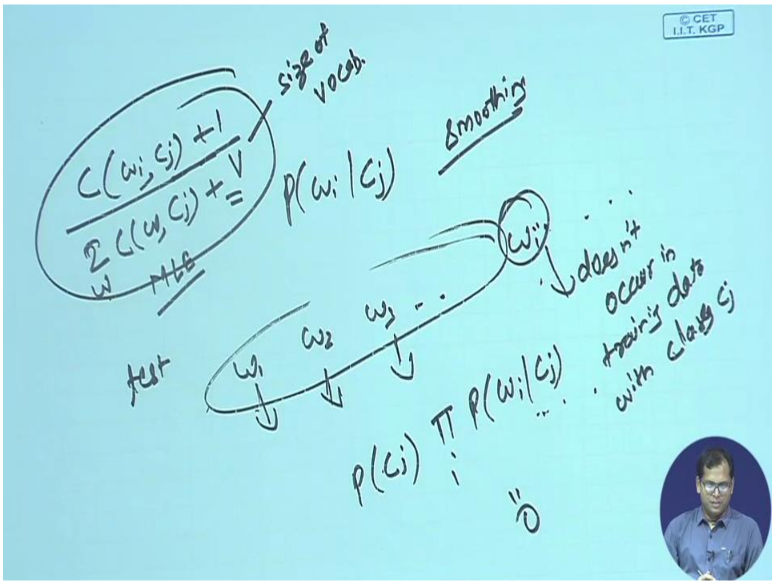
Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 13 / 14

So, that is in my training data how many documents have been labeled with this class. So, this is doc count c is equal to c j i; how many documents in my training data labeled with this class divided by how many documents do I have in my training data. So, this is what is probability of this class in my training corpus. Then how do you compute probability x i given c j; probability of a word given with the class and that will again be very easy similar to what we did in the case of language modeling.

So, number of times this word occurs in the class divide by number of all the words that occur in that class; that is number of times this word is occurring w i occurring in class c j divided by count of all the words occurring in class c j, this can also be taken as the whole corpus count of class c j; how many unique not unique, how many different total tokens are there in this class c j. So, this will give me both the probabilities and so this I can estimate from my training data. So, I have these probabilities, I can multiply these and get the best class for a given test document.

So, what can be one problem with this approach only taking these probabilities as such? So, assume that in training data all the classes occur at least once, so the prior probability of a class will not be 0.

(Refer Slide Time: 23:58)


© CET
I.I.T. KGP
size of vocab.
Smoothing
$\frac{C(w_i, c_j)+1}{\sum_w C(w, c_j) + V}$
MLE
$P(w_i|c_j)$
test
w_1 w_2 w_3 . .
w_i . . .
doesn't occur in training data with class c_j
$P(c_j) \prod_i P(w_i|c_j)$
0

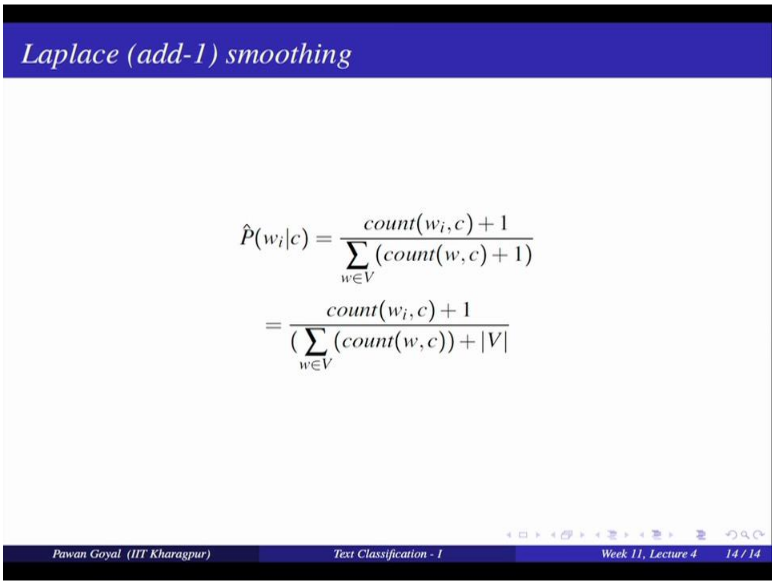
But what about this particular term probability w i given c j. Suppose you are given a test document and test document I mean there are different words w 1, w 2, w 3 and so on. There is a word w i that does not occur in training data with class c j, but all the words they occur some number of times with class c j. So what will happen with my estimate of the posterior estimate? So this will become probability of c j times, probability w i; given c j for all i.

Now while all the other w i are having a positive probability, this one is giving me 0 probability. So, I multiply this everything becomes 0; irrespective of even if the other words were very highly suggesting that class c j. So, only one word if it does not occur in that class, it will make this whole probability go to 0 so that means, we need to do something to avoid this scenario and what should be that we will be doing and so that will again remind you are what we did in the case of language model to avoid the 0’s.

So, we do something like a smoothing, so we use the smoothing to get these counts, to get these probabilities. So, how do I use smoothing? So till now I am computing probability w i am given c j as, right now this probability is computed as count of word i in class c j divided by summation over count of any word w in class c j; for all words, that is the current probability that is the MLE; maximum likelihood estimate. Now how do you use smoothing? Suppose I use the simplest smoothing that is add one smoothing. So, I will now write it as plus 1 in the denominator I have to add 1 to all the w words, so this will be plus V; where V is my size of vocabulary and this will be now add-1 smoothing and that is what I will be using. So, once I do that this will have some small probability 1 by V and this whole thing will not go to 0.

So that is what is the problem with using maximum likelihood estimate, so suppose in your training data, we have not seen the word fantastic in class positive. So, you know this probability will be 0; probability of fantastic given positive will be 0; while other words in the review or in the sentence might indicate positive.

(Refer Slide Time: 27:04)



Laplace (add-1) smoothing

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)}$$
$$= \frac{count(w_i,c)+1}{(\sum_{w\in V}(count(w,c))+|V|}$$

Pawan Goyal (IIT Kharagpur) Text Classification - I Week 11, Lecture 4 14 / 14

So, this whole thing will go to 0, so instead we use a; add-1 smoothing. So, count plus 1 divided by summation over count w c plus 1; that will give me a plus 1 in the numerator and plus V in the denominator and that is my add-1 smoothing.

So, we talked about the Naive Bayes model and how do we compute various probabilities and we saw one problem with the probabilities; that some probabilities might go to 0 and that will take the whole thing to go to 0. That is why we can do n simple add-1 smoothing, so now in the next lecture, we will take an example and see how do we actually compute all these different probabilities from a simple training data and we find the probability for the test sentence or document.

Thank you.