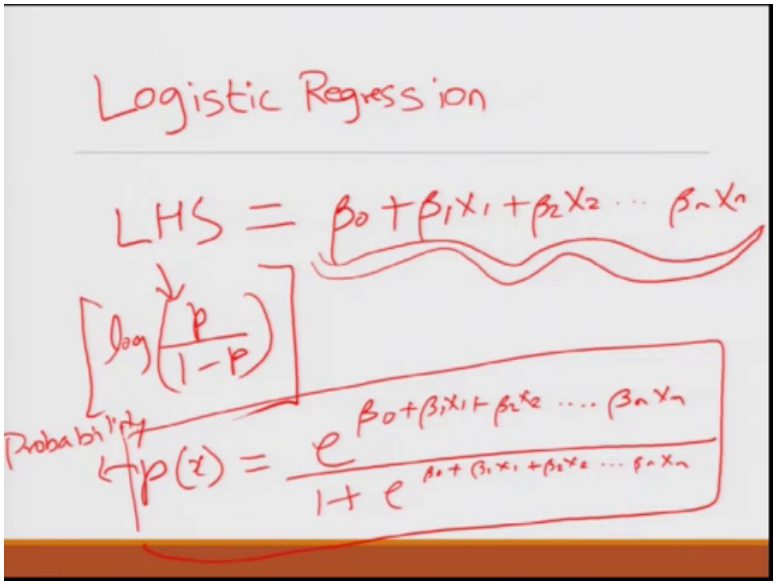
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

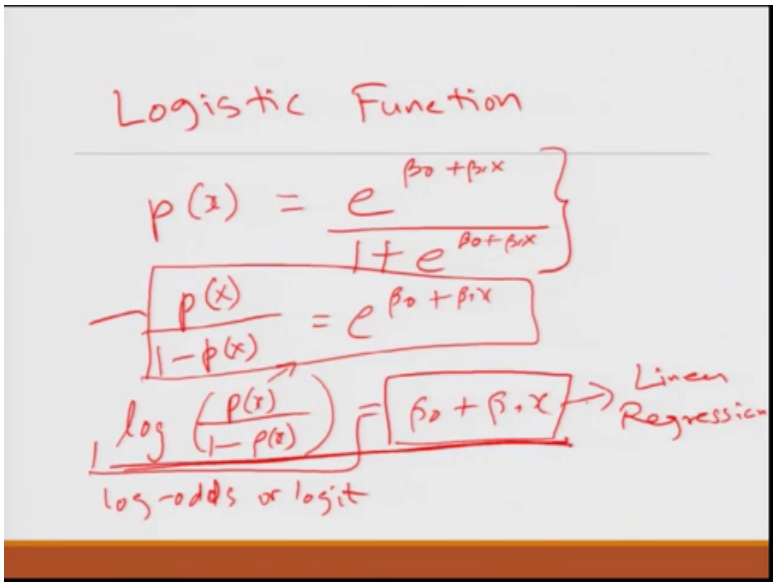
Prof. Deepu Philip
Department of Industrial and Management Engineering
Indian Institute of Technology Kanpur
Dr. Amandeep Singh Oberoi
National Institute of Technology Jalandhar
Mr. Sanjeev Newar
C E O, Galton Analytics New Delhi

Lecture – 28
Machine Learning – (Part 5)

Hello welcome you all to another session on machine learning in this practitioners course on analytics, we were talking about logistic regression last time as 1 of the most popular classification rather machine learning methods being used in the industry, if you just let me may be rewind a bit and take you through what we studied.

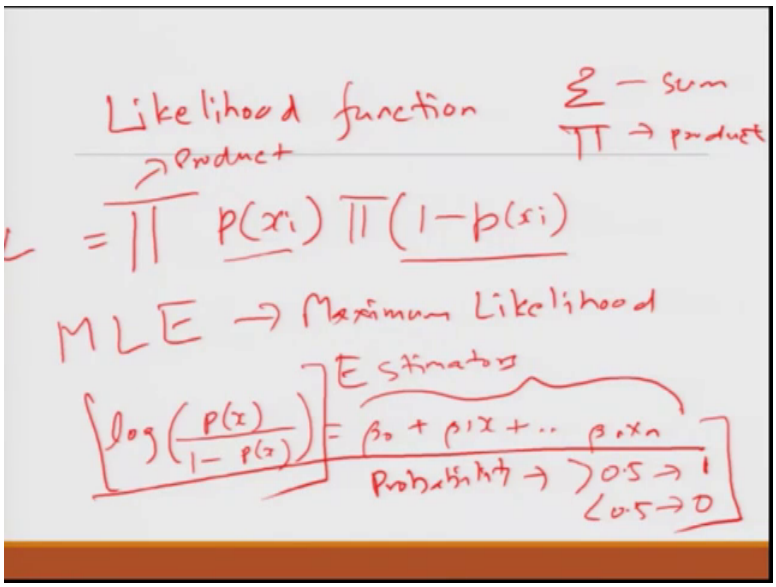


(Refer Slide Time: 00:24)



We discussed that it is basically a logistic function is a function of this particular form and so what happens is when you take a log of this after some simple algebraic transformation, this form that you get this form is if you look at the right hand side it is just like any normal linear regression form.

(Refer Slide Time: 01:12)



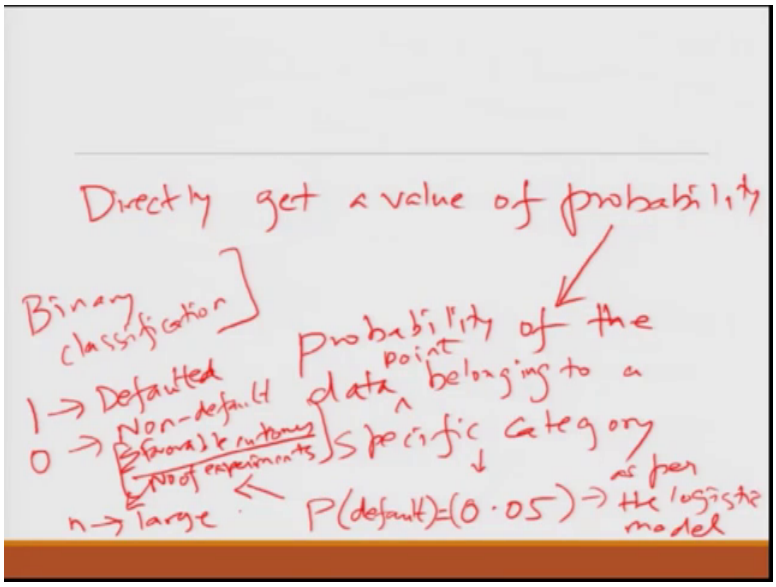
So, if there are multiple parameters this particular equation the right hand side which is this part this is almost likely in a regression. So, what we do is we do this calculation of say the left hand side be LHS. So, LHS is equal to RHS which was nothing but beta not

plus beta 1 X1 plus beta 2 X2 and so on. Here just to remind you again we were trying to optimize or maximize the likelihood which was this product of probability of an X I event happening and not happening.

So, we were ma trying to maximize the likelihood that was the certain difference compared to normal linear regression, the output that we would get if you look at the left hand side this was log of p by 1 minus p p is a function of x am just for sake of simplicity and clarity am removing that function of X, because p is the function of X. So, we get this form of an of a term on the left hand side after we do this complex this seemingly simple, but your mathematically slightly complex it is a 1 line code for if you are using r or python. So, we get this left hand value and then what we do is we transform this in to a probability function.

So, if you recall let me just get back again if you recall we started with this particular term. So, from there if you again move we get a value of p it is a function of x is equal to e to the power whatever this beta not plus beta 1 X1 plus beta 2 X2 up to beta n xn divided by 1 plus e to the power same term beta not plus beta 1 X1 plus beta 2 X2 up to beta n Xn. So, this p is my probability so what happens is from a practi practitioners stand point unit to understand that when you use logistic regression and you do the necessary transformation which the software code will do for you, you directly get a value of probability.

(Refer Slide Time: 04:14)



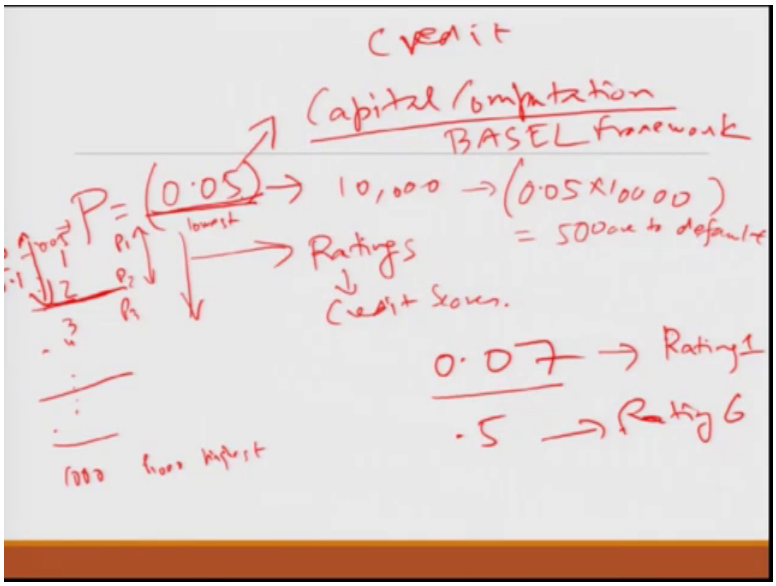
Now, what is this probability of the data or data point being or let me may be write in more express it in belonging to a specific category.

So, if you recall in logistic regression it was binary classification, binary classification. So, for example, if you are modeling default then 1 was the case that the person defaulted you can take otherwise also depending upon the way you formulate the problem, but say 1 was default then 0 was non default. So, the value of probability you get suppose you get a value of 0.05. So, this means that probability of default is equal to 0.05 as per the logistic model, so as per the model the probability of default is 0.05.

So, now this is where a lot of people also trip and the trip is that what exactly probability means, if you go back to the fundamentals probability is the frequency is the ratio is the denominator is the total number of times you repeat an experiment and the number of favourable outcomes that you get so when and when that happens when you do a large number of experiments.

So, if you go back to your class this probability is some of favorable outcomes by number of experiments, when number of experiments if this be n when n is large; which means if you get from logistic regression p value of say 0.05, it means that if you were to say you were modeling default in case of a bank.

(Refer Slide Time: 07:13)



So, it means that if you were give loan to a person with this exactly this characteristic and if you are able to choose say 10000 people then of this 10000 people 0.05 into 10000 people are likely to default. So, whatever be that number so 0.05 you have 1010. So, may be around are to default this does not mean a person because a person can either default or non default.

So, what happens is now there are various ways to interpret is 1 is that you can directly use this value for example, in banking you need this value for calculating your capital computation. This is an important very important calculation that each bank has to do which is nothing but or the credit capital computation, which is nothing but the amount that a bank should keep as buffer to account for the fact that some people will default. So, it is a complex calculation let us not get into that you can read more about Basel or Basel I mean it is pronounced in different ways Basel frame work for more of that, but the point is that you need to calculate this numbers. So, you can use directly this number often what they do is they calculate this p for a large number of customers.

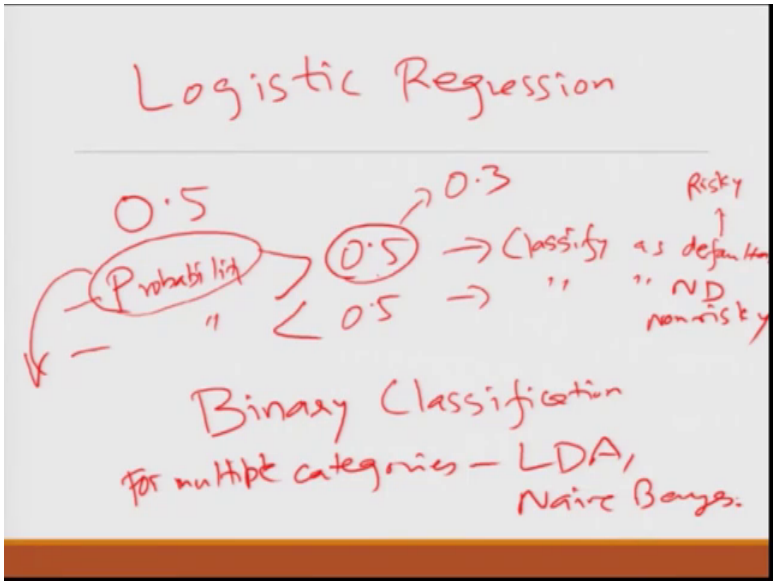
So, if you have say thousand customers you do it for 1000 customers and then you break your for example, this was for customer 1 customer 2 customer 3 up to customer 1000 you had 1 value of p 1 you another value of this was p 2 p 3 up to p1000, what you do is you often sort it in increasing order. So, this is lowest this is highest and then you create splits, now these splits are what become almost ratings or credits course.

So, what you do is you take an average of these things and allocate to all of these peoples. So, that is how to you come with different credits course, you know that you say that a bank follows a 5 points credits scales. So, 5 are nothing but these divisions that you do and then for each of those credit scales the bank calculates the average default rate and then uses that in calculation in practical stand point given that say suppose you got a customer whose score was 0.07, then what you do is you see whether the average this value.

So, so these are the cut off points, so if so for example this range was between 0 to 0.1. So, average was say 0.05 you got 0.07 you say that point 1 is the cut off. So, probably he belongs to class 1 or rating 1 if it was 0.5 he may be belonging to sixth class. So, that may be a rating class 6, so depends upon how you cut.

So, this is how decisions are typically taken in financial institutions the seed is this calculation, this calculation which comes from this logistic regression model. So, you calculate this these X1 X2 are the different parameters of the borrower like credit history like this demographic information like his salary and other stuff things that you get from know your customers information and then you use this model you come with a value and you do the prediction, often it also happens that you many banks they there is a different way of looking at it they put a cutoff of say 0.5, 0.5 so and then they say that if the probability value that you got if the probability is greater than 0.5, then classify as defaulter if probability is less than 0.5 classify as non defaulter.

(Refer Slide Time: 12:15)

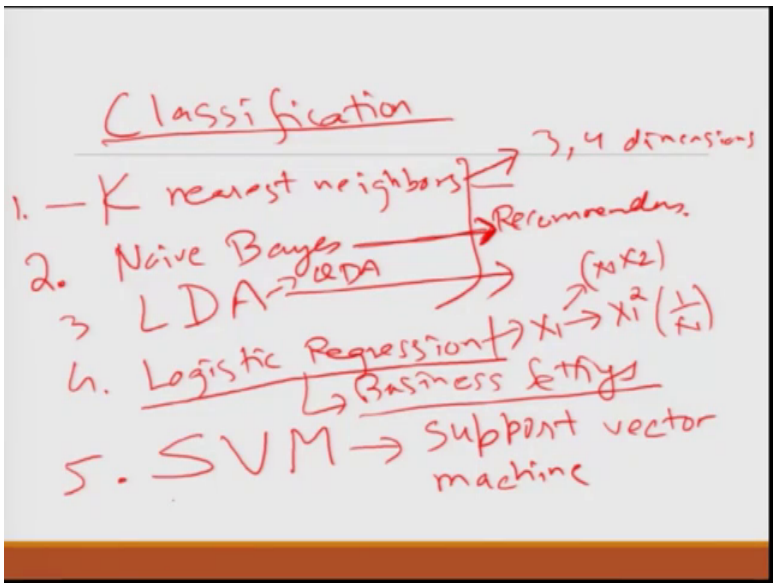


So, some banks uses multiple points scale you may in the beginning also have a simple scale like this and if you want to be conservative then you can actually play around with this number and you can say that let me put cut off it 0.3. So, anybody above 0.3 he becomes risky for me I may not call him defaulter you may call him risky and non risky. So, depending upon your risk tolerance you can play around this number you can make multiple instead of using just 1.0 classification, defaulter non defaulter risky non risky you can have a 10.0 scale, you can also directly use this number you can directly use the probability figure and take your decisions on the portfolio.

So, this is how logistic regression is used, logistic regression is typically used when it is binary classification; you have logistic models for multiple classes, but they are not very

popularly used for multiple categories LDA linear discriminant analysis or naïve bayes that we talked about these are more popularly used or another way that people do is they use logistic regression successively. So, they classify into 2 first and then within those 2 they again do a further classification. So, you have different approaches depending upon the kind of data you have ok.

(Refer Slide Time: 14:17)

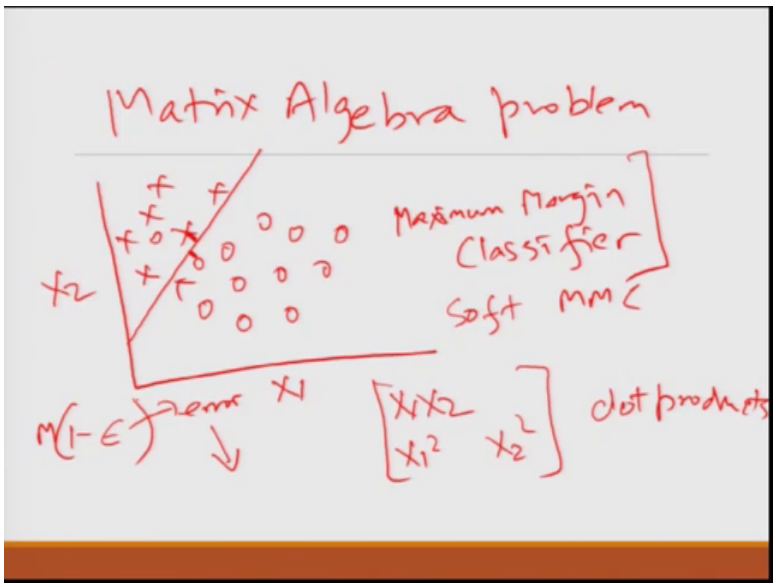


So, so far we discussed in classification we discussed k nearest neighbors, were we say that we just do a poling of who are your closest neighbors, it may be 5 neighbors 10 neighbors 20 neighbors typically 5 to 10 is the ideal range and using (Refer Time: 14:39) to a poling and then use a majority vote and whosever are the more number of neighbors for you I will classify you according to that. So, this was 1 we discussed second we discussed naïve bayes method, we also briefly discussed that the same extents to linear discriminant analysis based methods and we discussed logistic regression.

Now, logistic regression is typically used a lot in business settings, these are used more in recommender systems recommenders; it is used in business because just like linear regression the model is simple it is intuitive you directly get the parameters the list of parameters and relative weight age of parameters. So, from a management stand point it is easy to have a business sense of what we are trying to do and manipulate things accordingly, most of the cases it works beautifully especially on human data on data like credit default and stuff it seems to be very effectively used.

This one more method which was very popular and it is still popular it is computationally slightly expensive it is called support vector machines. In fact, before support vector machines was that the method which actually started this rage about machine learning this. So, before deep learning took over support vector machines were supposed to the most cutting edge ways.

(Refer Slide Time: 16:39)



Now, support vector machine I will not get to this is purely ah, I would say a matrix algebra problem unlike other methods which had their origins from statistics this had origin purely from machine learning and it was actually a Russian scientist who created it and then America actually got him out of that was the time of US s are and he got out because, of this innovation that he did and who are a time because it was computationally much smarter given the kind of computational power that machines had that time.

It was very very appealing and it actually created a lot of I would say revolution there are lot of improvements that happened, after that I will give you a very high level concept of what this is about.

So, what it says is that if say X1 X2 be 2 parameters and let me you know let there be 2 different kinds of variables or 2 different kinds of classes out there. So, you have to classify something into 2 parts, so 1 am representing by cross the other I represent by

circle. So, what this method tries to do is that it first of all tries to create something called a support vector.

So, what it is support vector is that it is a plane in this case in 2 dimensional situation a plane is nothing but a line if it is 3 dimensional then it is a plane a normal plane that we see in n dimensional it will in n dimensional hyper plane. So, it tries to create a line may this point here a line that divides the 2 classes, such that this distance the distance between the perpendicular distance between the 2 nearest neighbors to that line is maximum, so it is called maximum margin classifier.

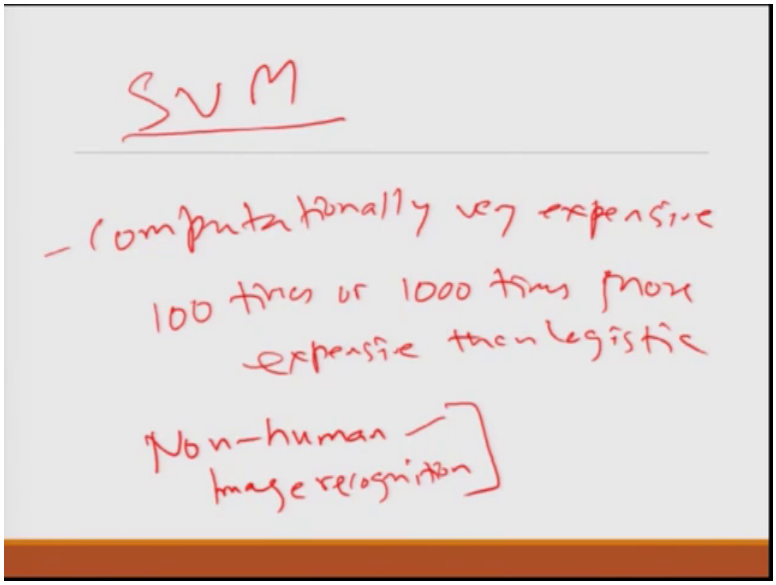
Now this works very beautifully if it can be linearly separable, then all you do is you create this kind of line now there can be multiple lines actually which classifies. So, they try to find the line which maximizes the perpendicular distance. So, that is the whole idea I will not get in to the maths of it what they do is because, sometimes this they may not be so beautifully separable there may be some point here some point here. So, what then they do is they still try to create a line and allow for errors, so they assume that there will be an error.

So, instead of trying to maximize the distance they try to dis maximize the distance of 1 minus epsilon is a term of error. So, they allow for errors depending on the kind of error that you are accommodative about they will have a different line altogether. So, so creation of this line in an n dimensional space it becomes matrix algebra problem, they also try to then create extend the same problem. So, in this case there are the 2 variables you know are they are only 2 variables X1 X2, but then what about correlations between different the 2 different variables what about.

So, so instead of just using 2 variables as such they also try to they create a hyper plane. So, if example if there were only X1 and X2 there would be another dimension called X1 X2 there would be dimension of X1 square there would be a dimension of X2 square. So, in this way depending upon to the level of polynomials that, you want to go you try to create all these combinations through dot products I know am getting slightly technical here, but unfortunately this is a very mathy subject. So, they try to get in to the different components of dot product and make that massive hyper plane and then they try to create this maximum margin classifier they account for this error, so it is called a soft maximum margin classifier. So, because soft means that they allow for some kind of an error.

So, the intuitions should be important but ya having said that this because, it becomes huge matrix computational exercise.

(Refer Slide Time: 21:33)



It is support vector machines are computationally very expensive and more the liberty that you start giving in terms of the number of polynomial levels you want to reach it becomes really expensive. So, typically the computationally it may be 100 times or even 1000 times more expensive than logistic or even more depends on the number of variables in parameters that you have.

So, typically that is why it is used is very knish cases it is used in lot of I would say non human cases, it is used in image recognition for example, but for most of the human cases logistic regression I would say still is much more popular much more easy to do. But for if your if your dataset small you can try svm it will still take some time, but in r or in python it is just 1 line code. So, that is all about I think the major classification algorithms that are there then there are lot of modification and complex systems based out of their, what you need to understand is more than the maths of it. You need to understand what are the various usages of them where which should be applied.

So, for example k nearest neighbor you should be an you know because, all of them for you from a practitioners stand point they will be just 1 or 2 lines of code. What is important is that you put the data very carefully in to that, what is important for you to is that you know which are right problems for each of them. So, for example k nearest

neighbor will be good if you have 3 or 4 dimensions not more than that, the moment you have more number of dimensions the neighbors as I said becomes.

So, far apart it does not make much of sense similarly in naïve bayes naïve bayes is good for recommendation system, it assumes independence of different parameters which is rarely the case in real life business scenarios. Hence I would say naïve bayes again beyond online methods beyond say image detection beyond say cancer pattern recognition stuff you should use very sparingly LDA, LDA is very powerful. But LDA requires is lot of computation because, again it assumes linear the moment from LDA you move towards something more complex like a QDA it becomes computationally expensive.

So, at that time as we discussed in the beginning in the genesis that we need to understand that accuracy is not everything, our goal should not just to make the most accurate model most complex model because complexity has a prize; often logistic regression still kind of provides the right balance in most cases the world is not linear, but then as we discussed in linear regression that you can always adjust for non-linearitys also through variable transformation.

So, instead of X1 you can transform it to something like X1 square. So, this parameter you can transform or you can take 1 by X1 inverse of it. So, you can take you can make more complex this thing, if you have interaction variables you can create a new variable called X1 X2, it also gives you lot of flexibility in terms of working with dummy variables creating additional categorical variables. So, logistic regressions is still I would say from a practitioners stand point, the 1 that you would in business settings at least in enterprise settings you would be using more because you can control it, with other algorithms another problem is that the moment data goes out of hand and the model stops looking you do not even know, whether it is a temporary problem whether there is some systematic change that has happened because models are largely black box.

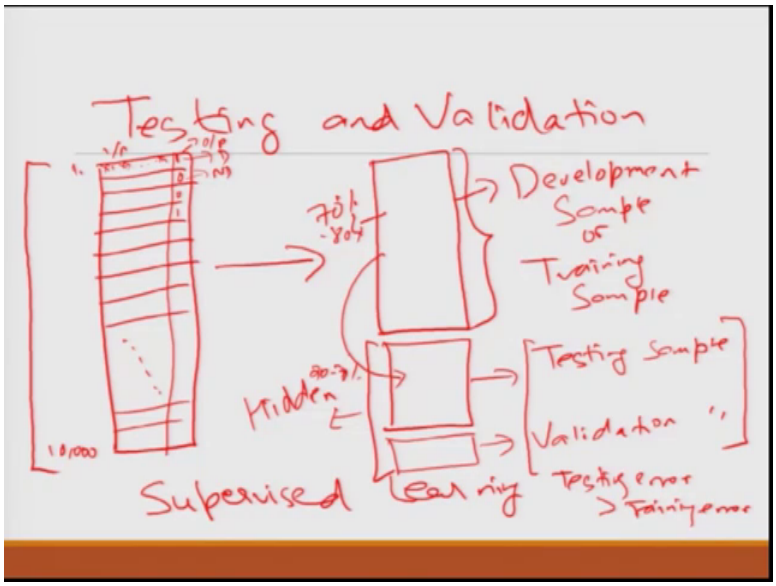
So, these are the main classification algorithms what is most important for you is to have this intuition with you because, tomorrow a lot of new algorithms and lot of new variations of the same algorithms may come.

(Refer Slide Time: 26:22)

You may also go for something which are called meta models, you instead of using 1 model you actually use a combination of different models for more robust results. So, all these would require you to have understanding of knowing you know what the limitations are it is like think of you being you know being the manager of football team, you may not yourself play football you may not be the best goal keeper or the best player in the center forward or attacker, but you need to know which are your players what strength each player has and then work accordingly.

So, just think of yourself that way from a practitioners stand point you need to understand and of course, you need to have a feel of what playing is all about.

(Refer Slide Time: 27:01)



So, next before we move ahead an important part which you need to know in your process of model building is testing and validation. Now this is again often ignored and hence it is important that we briefly touch upon this aspect, now testing and validating means that now let us let this be your data set. So, these are different records may be have you have 1 to say 10000 records.

So, each record so I am I am right now talking about supervised learning, where this is your output these are your input variables, so these may be X1 X2 to Xn. So, multiple parameters and this is the output, so let us take an example of say modeling default something that we just discussed. So, these X1 X2 X3 Xn may be different parameters about the borrowers his account history his demographic information his salary and stuff and this is 1 historical record you have something like a 1001, 1 means the fellow defaulted 0 means it did not default. So, you get this kind of data and suppose you have 10000 records from history you want a build a model.

So, that when a new customer comes were able to classify him properly or take other kinds of decision, so 1 of the typical mistakes that they do and often despite a lot of warning, I have seen this thing happening is that what they do is they choose this and entire model and then they build the model because, the model learning is a simple 1 line code or 3 line code or whatever transformations you do, but what they do is they choose all the data points they build the model and then the model gives fantastic results. So, the

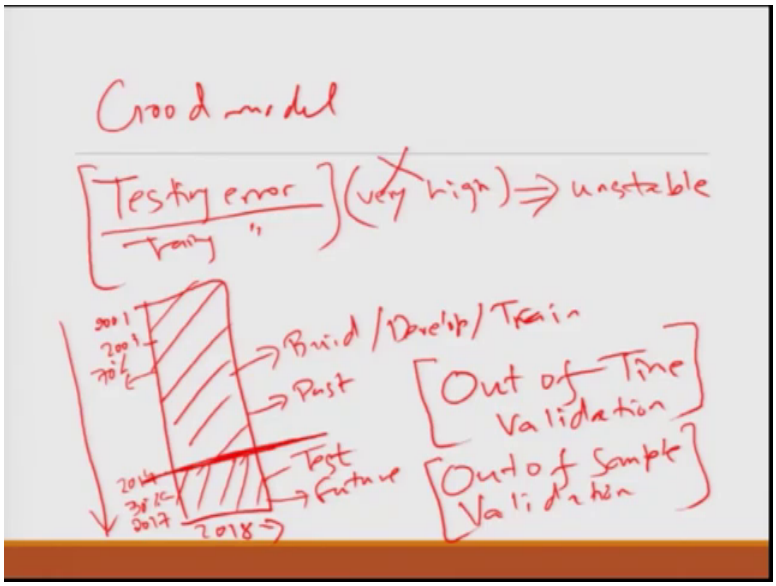
problem with this is that once if you build the model and all the datasets, there is no way for you to know whether the model will run on future data set or not, so tomorrow what comes because your model already running on this. So, often what they do is they run the model they calculate the miss classifications and we talked about you know we calculating the error rate and so the error rate is always good because, the model is optimized on this on the data on which you are testing the error rate also.

So, in a more un biased setting what you need to do is you need to split this into 2 different parts, so this is called your development sample or training sample this is called your testing sample, sometimes you also have another sample and call it validation sample.

But often you merge them together so when you call testing and validation sample it largely is assumed unless you are working with very huge number of data points clean data points, it largely used assumed that you are meaning the same thing. So, the goal is now that you build your development sample means your model is built on only this data. So, this data is hidden from you this is hidden. So, even before building the model you take this data points out, you do not even know about you simply build the model on this and then you run the model on this data.

Now on this data the model gives an output you compare with the actual output and then try to see what error rates look like. So, training your testing error in all probability will be above training error. So, when you build a more complex model, we talked about bias variance and model fitting curve fitting and stuff you build a very complex model with flexible model it may give you very low training error very low, but the moment you do a testing on your data you will find that you know suddenly there is huge divergence.

(Refer Slide Time: 31:56)



So, if you have to make a good model your testing error to training error this should not be very high, 1 is impossible obviously testing error will be high, but ya it should not be like it becomes double or triple then; that means, if it is very high it should not be because very high implies unstable model. If it is reverse that means you have done some mistake in the model because, you cannot predict the future to be much even better than the present.

So, you have to see that it is within range, typically what they do is this is normally 70 percent of the data or 80 percent of the data this is around 20 to 30 percent of the data. So, this is how you split now in case of your default modeling your data you put the data in a sequential order, say you are you got the data for a customers between 2001 to 2017. So, what you do is you arrange them sequentially probably your say 2014 was the cut off points.

So, this is where you have around 70 percent of data this was the 30 percent of data, you cut off sequentially build the model on this build and then test build means develop or trade. So, these are just synonym words, now it is important that instead of randomly picking any 70 percent 30 percent you pick it sequentially because, the concept is that your model is should be designed to predict future.

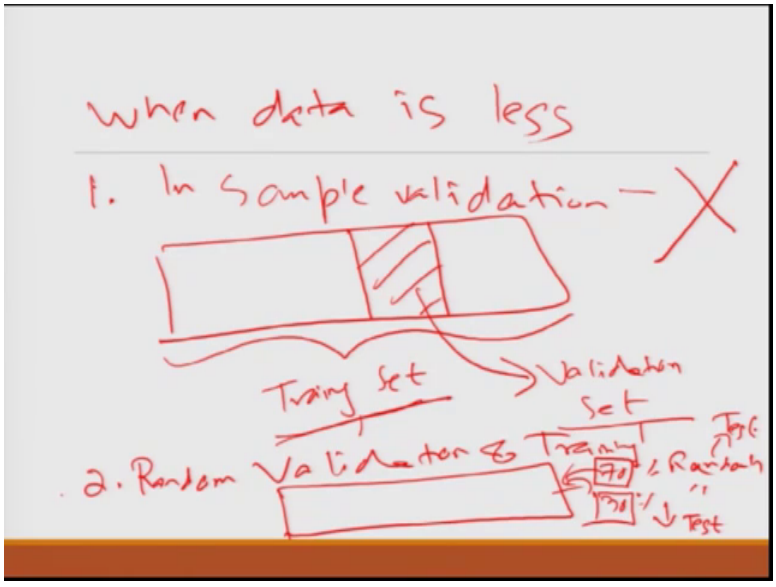
So, based on past data this is past data this is past from a modeling perspective this is past data and this is future. So, you want to see whether based on my past data can I

predict the future because, if you do vice versa for example, you had taken data between say here the cutoff was 2003 you took a data between 2000 to 2017 to build the model and tested on old data that even if the predictions were good you are in no position to say that into 2018 my predictions will be equally good.

So, though a model build on more recent data will be more accurate because, you get the more recent data the more the relevant data the earlier the more recent data is the more relevant data. So, ya it will be a better model from purely from a modeling perspective, but from a business implementation perspective and you do not know about it future robustness as I said robustness of the model stability of the model becomes important.

So, that is why you build the model on historical data and test on the future data. So, this is called out of time validation it is also because these 2 samples are mutually exclusive it is also called out of sample validation. So, out of sample validation is something that should do with all models and if it is data with that importance of time it should be both out of time and out of sample validation, now different kinds of model actually because there are different ways of running the model.

(Refer Slide Time: 36:06)



There is another way in which when the data points are cars when data is less, than people try to figure out a d all this methods actually compromise with the quality of model and there are different methods that people use. So, 1 is in sample validation and this is very popular among many consulting forms I would strongly recommend that this

is actually to me to some extent, actually fraudulent validation instead of in sample validation in sample validation means that if this be the training set on which you build model.

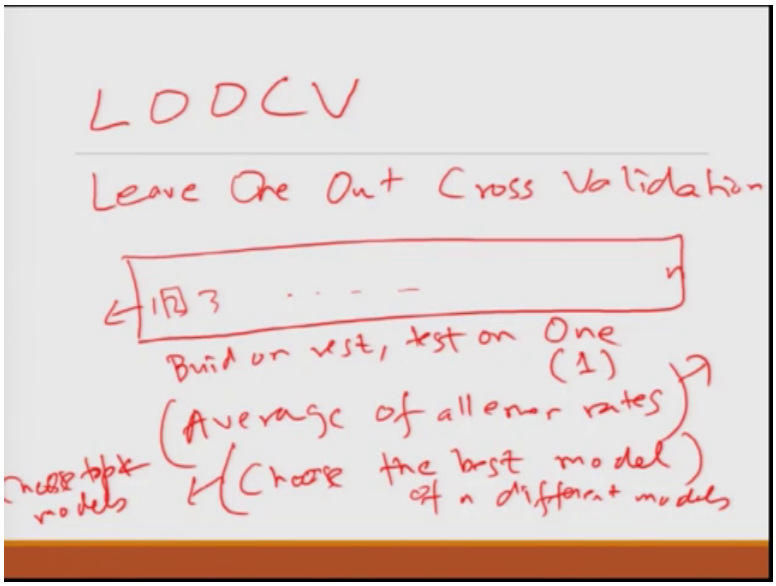
So, this is my training set you take a sample and you validate on this, this is my validation set or testing set. So, because this is the already part of the training set actually does not make sense, I mean you do not even need to do this you can just actually go with this result when why even take this case.

So, it may impress your clients and you know may be if the fellow is if your client is not has not undertaken, this course on practitioners views on the data analytics that we are undergoing he may actually fall for this and say that well the model is robust because, it is part of the same data training set the errors that you get here the errors that you get here will also be almost of the same type. So, it may all look good but actually this is not right. So, in sample something you should not do then there is also something called random validation and training.

So, in this what they do is if this be the dataset the randomly picked 70 percent 70 percent randomly picked and then on 30 percent on this 70 percent you build, the model 30 percent again randomly you test the model, but because it is again randomly picked though it is very popular in many stats books also you will find this thing.

Actually given that the data is randomly picked technically this data set and this data set do not actually differ too much in their characters and hence the results though may not be as there may be some difference, but it is not rigorously right way. But then often in many cases when it is not a very time bound data or something ah, for example for image recognition for non human data again this can work, but for human data and stuff again this is slightly mislead and you need to take it with pinch of salt. If you again you know because, the number of data points asked sometimes cars there are other ways that people try to do the validation the 1 another method is called L O O C V validation.

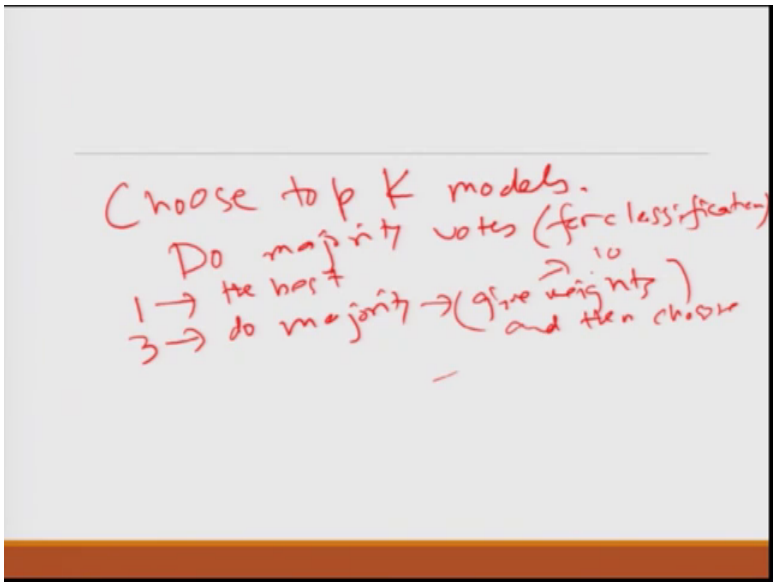
(Refer Slide Time: 39:05)



Which is leave out leave 1 out cross validation, in this case what they do is if this be the data set say 1 2 3 this is the nth data point. So, what happens is first of all you leave this out, build the model on this test on this then you leave this out and include this. So, build on rest test on 1. So, each time you take out 1 sample point and build the model on the rest of the points and then test on this and then in this way what you do then is the average find out the average of all the error rates. So, you find an average of all the error rates and so that becomes your actually the reasonable error for the models that you build. But now you have since you have build if there are n points you build actually n different models technically.

So, what you do is then you choose the best model because you have n different models, so you choose the best models the model which gave the model which gave the best result, but you say that the error that I get is this error. So, choose the best model of n different models you cannot average it because, the model is not a linear thing each model comes out of different complex mathematics. So, you have to choose just 1 of them, sometimes what they also do is they choose not the best, but they choose top k models k may be 1 2 3 and they get the prediction of all.

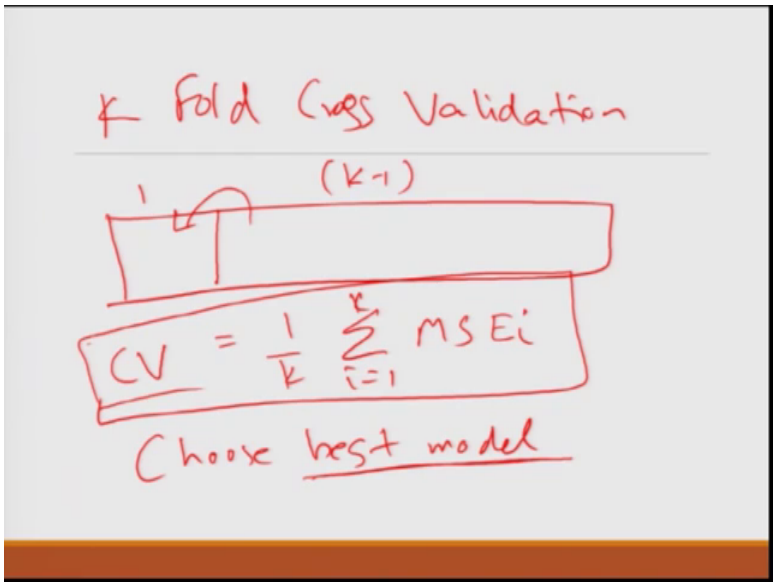
(Refer Slide Time: 41:27)



So, another way is that you choose top k models do majority vote for classification methods. So, this k may be 1 model so if it is 1 well you choose the best the best model, you normally do not go for 2 models or because then there may be a tie in case of 2 models in actually does not make sense you go by 3 models and then you do majority. Sometimes you give weights you can play around with weights and then chose not for 3, but for say you choose 10 models give weight and then you come with something.

So, you can actually this is where you know people I would say start getting crazy about it because, the model themselves are not very accurate, but ye 2 elegant simple things is that you choose 3 or 4 or 5 or normally yeah, but I think 1 and 3 they suffice. So, what you are doing here is that this is called leave 1 out cross validation.

(Refer Slide Time: 42:56)



There is another method called k fold cross validation, in k fold cross validation was do is instead of breaking into 1 you know taking just 1 out because the 1 out the downs are held, with leave 1 out is that all the models because only 1 data point has been left out all models are actually almost the same. So, one of the ways is that you then divide into k different groups, so this is be the k so this is one group and then these are k minus 1 groups k minus 1 and test on this then again randomly choose k minus 1 and test on this and this will this way this way you know you choose 2 possible combinations, then your estimate is 1 by k of. So, you basically an average of all of the case and then you again then choose best model, but your validation your error you go by this and the model may be the best, but you do not choose the error rate of best model you choose assume that the error rate is going to be this.

So, these are different ways which you should be aware of when doing classification, well I think with this we conclude our discussion on classification based algorithms. I think with that you would be fairly equipped to be state of the art with machine learning of course, the path is long and then there lot that you can learn yourself, but you have you will have the frame work that you can confidently get into the industry and solve more serious kinds of problems.

Thank you very much.