

Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 40
Regression – IV

In the process of obtaining the best regression model either a line or some other non-linear function what we did was to define some kind of goodness of fit in the case we considered it was the squared error.

(Refer Slide Time: 00:26)

Overfitting

- Squared Error score (as an example: we could use other scores)
$$S(\theta) = \sum_i [y^{(i)} - f(x^{(i)}; \theta)]^2$$

where $S(\theta)$ is defined on the training data D
Handwritten note: Minimize Sum squared error for training set
- We are really interested in finding the $f(x; \theta)$ that best predicts y on **future data**, i.e., minimizing
$$E[S] = E [y - f(x; \theta)]^2$$
 (where the expectation is over future data)
- Empirical learning
 - Minimize $S(\theta)$ on the training data D_{train}
 - If D_{train} is large and model is simple we are assuming that the best f on training data is also the best predictor f on future test data D_{test}

Handwritten note: Minimize Sum squared error for training set

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

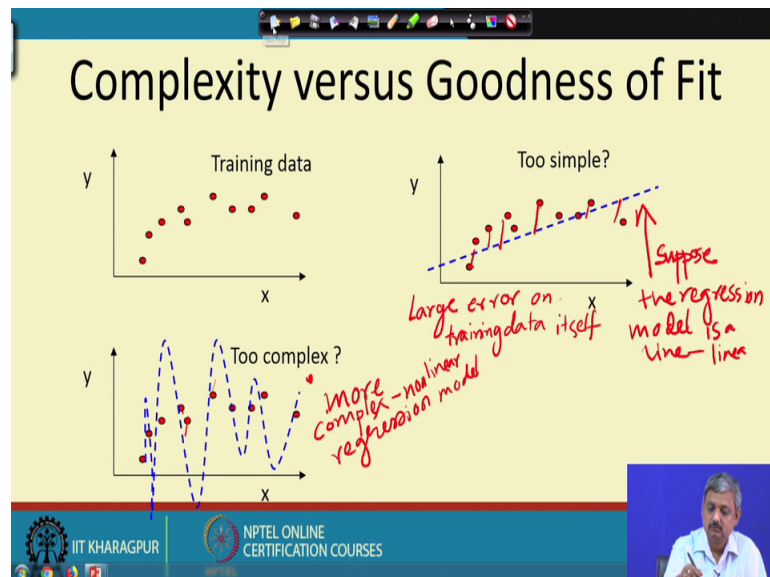
That means, you take the actual value of y , using the model regression function f you get an some parameter of the function f you get a predicted value. Take the difference, square it, add up over all the trading data. So, what we do is to for each training data find the error. Take the square add up over the importantly training set. So, this summation is minimized and then that will give you some value of the theta parameters the coefficients of the regression.

So, this process is actually sometimes known as empirical learning, empirical means something which is based on data on experiments and the known experiments previous. But what we are actually interested in is not to minimize the error on this training data, but to give the best prediction on an unknown future data ok. So, that is what because we know how stock prices depend on say things like national GDP and other things, today's

stock price, we are not interested today's stock price we are actually find a model that fits the stock price over the past one year very good very nicely that is fine. But what we actually want is to predict tomorrows stock price very well ok. So, that is the goal for the future data you want to do well.

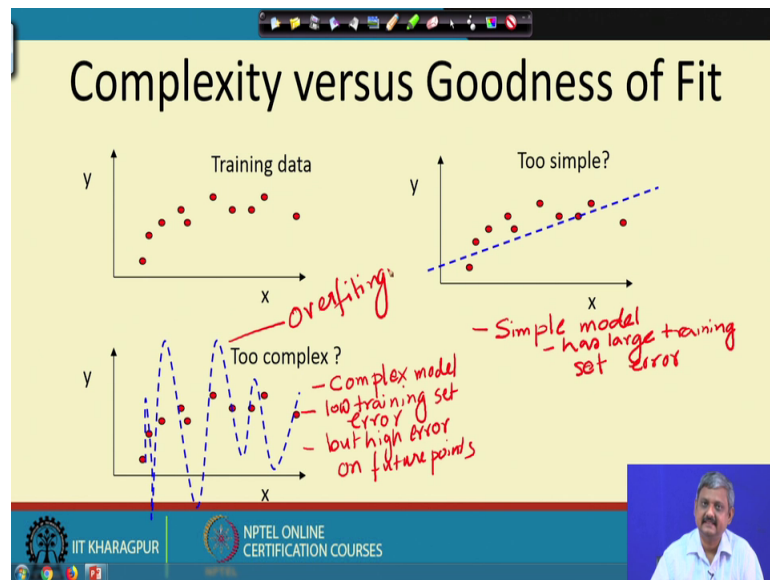
The usual assumption is the way data past data behaved future data will also behave similarly ok. So, if we have a model which minimizes error on past data it will also minimize the error on future data ok. So, that is the assumption, but that is not always true and we have the problem of overfitting in certain cases.

(Refer Slide Time: 03:52)



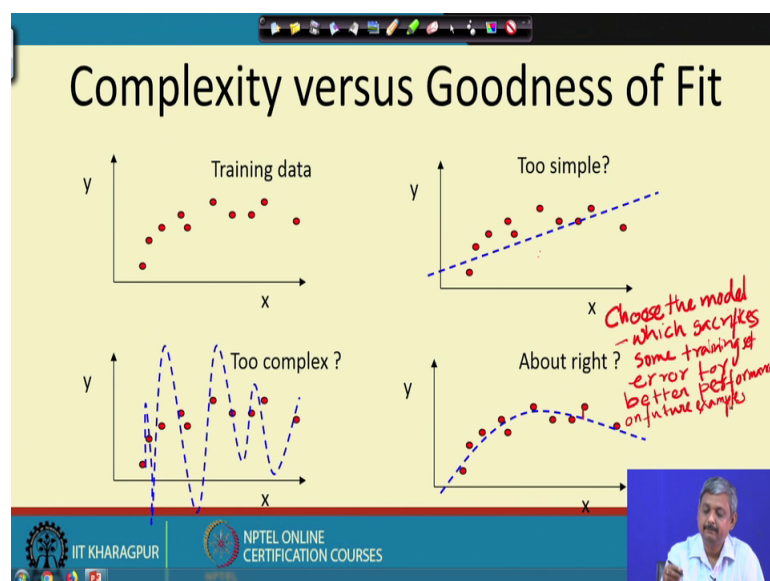
For example, I have a training data like this is a line linear then we will have something like this. But the problem is and this may have a high error on the training data itself ok, forget about future. So, to minimize the training data what we can have is to have a more complex model than non-linear regression model ok, it will fit the training data very well almost together, but it may fail on a future data. So, this may be a complex model; future points ok. So, this is a problem overfitting.

(Refer Slide Time: 06:30)



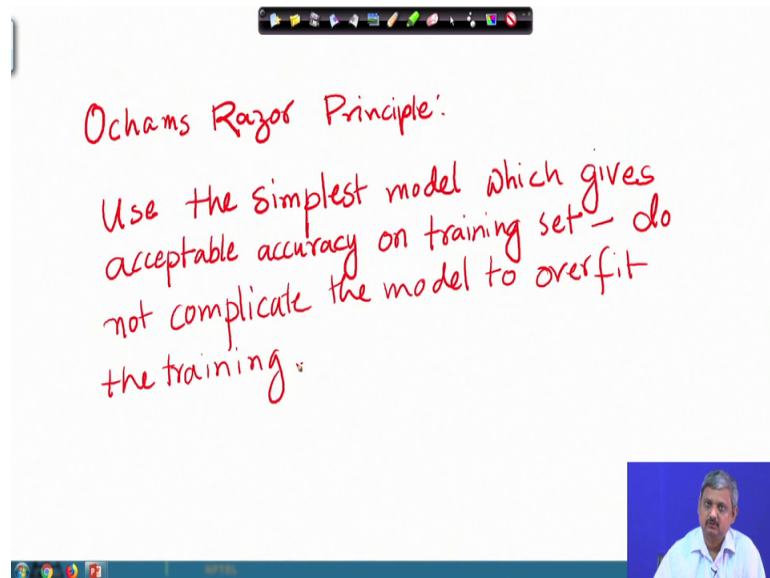
So, overfitting means we have matched the training data to well we have got a very low training set error, but we are failing to perform on a new data it is like rote learning. So, on the exercises of your book you have learned very well you have made no error, but you cannot generalize you cannot apply this in future. So, you get a very poor performance when you perform on a new example. So, that is the problem of overfitting. So, that is a common problem in regression also that what is the regression model, should it be too simple or should it be too complex.

(Refer Slide Time: 08:30)



So, sacrifices some training set error for a better purpose. So, you see there is some error not as high error as this, but not as low error as this ok. So, the principle is actually like this, the principle is called.

(Refer Slide Time: 09:44)

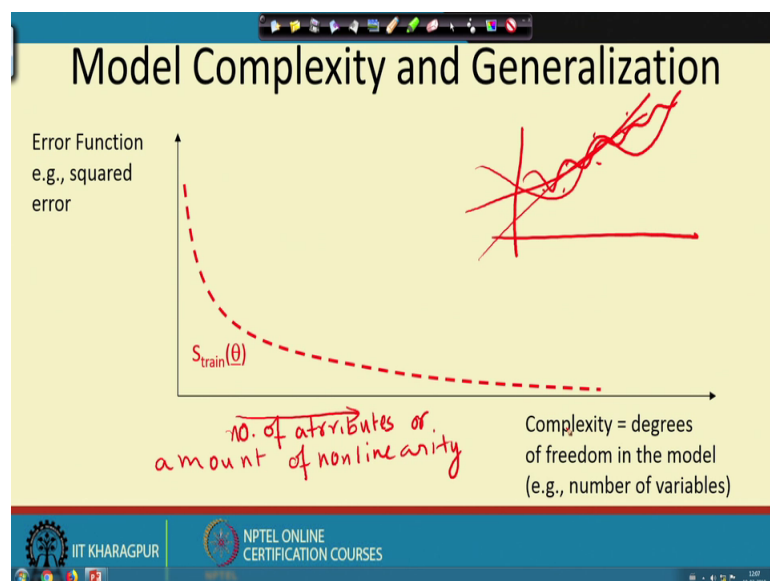


Ockham's Razor Principle:
Use the simplest model which gives acceptable accuracy on training set - do not complicate the model to overfit the training.

A small video inset in the bottom right corner shows a man with grey hair and a white shirt speaking.

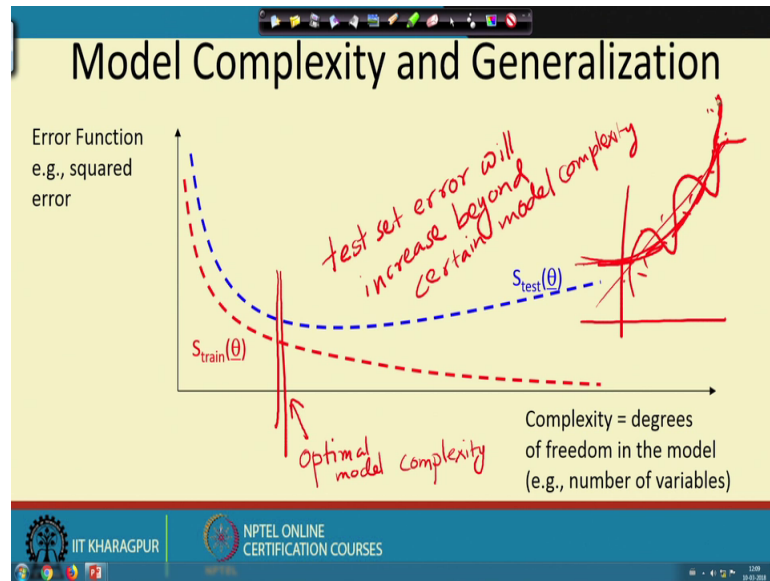
So, Ockham was a barber in England in 1300 C, gave a principle. Use the simplest model, this is the general principle to be followed in any predictive data binding task ok. It is a very important principle as a practitioner you should always use this.

(Refer Slide Time: 11:23)



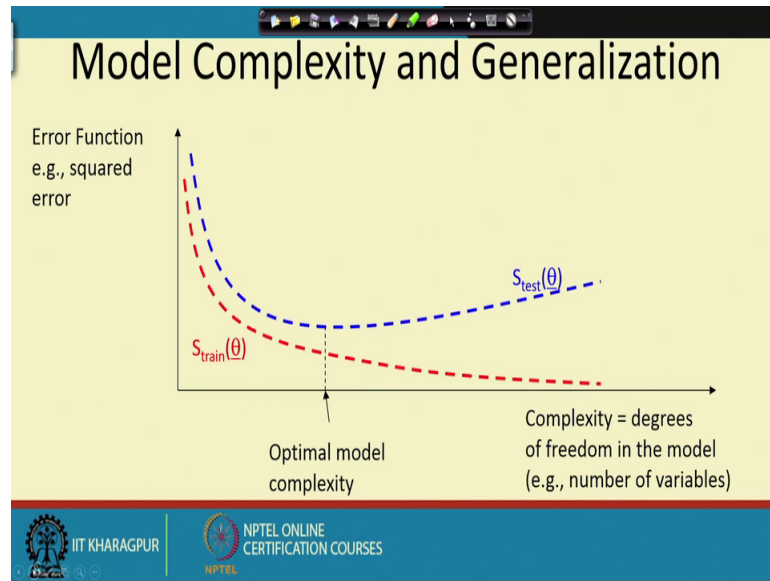
So, it is generally the curve will look like this if you increase the model complexity which may be the number of attributes. So, model complexity is, so it can be, so how does model complexity this is the simplest model, this is little more complex, this is little more complex, this is little more complex and so on. So, this is this and y axis is the training error.

(Refer Slide Time: 12:35)



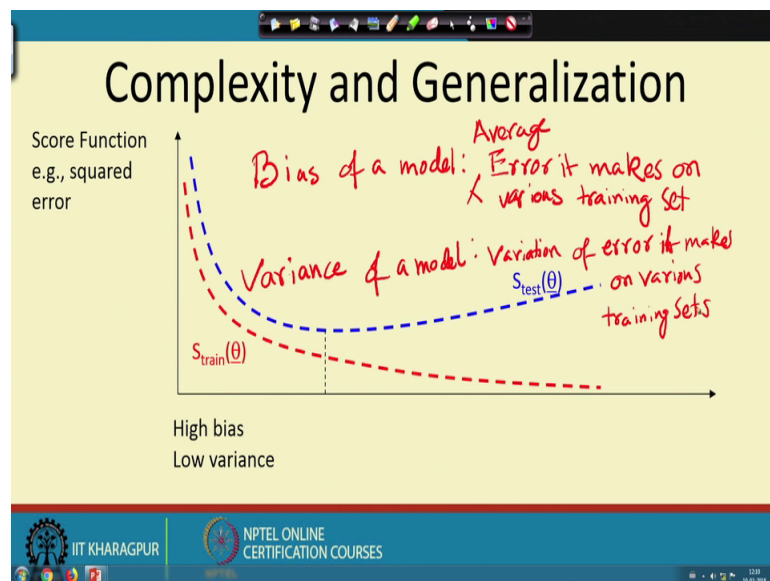
As you increase model complexity you can better and better fit the training set, but it will test set error will increase. So, this is the, this is not optimal, this is optimal and again this is not optimal. So, this is the this is the optimal point, not too simple, not too complex.

(Refer Slide Time: 13:54)



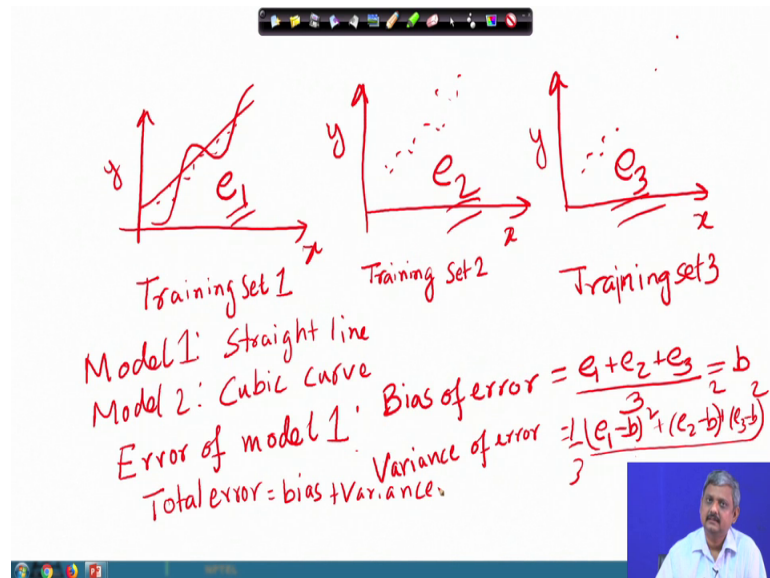
There is another term people use training sets.

(Refer Slide Time: 14:02)



Let me explain what it is.

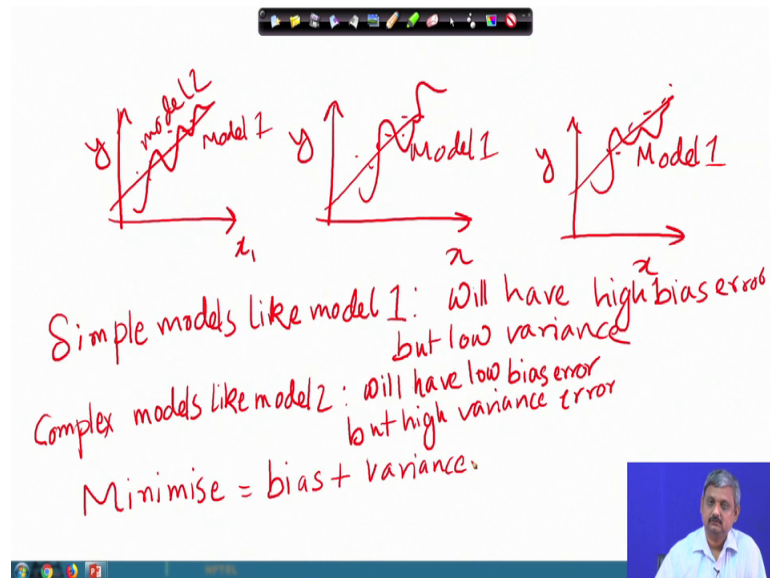
(Refer Slide Time: 15:30)



3 ok, straight line, model two is a cubic curve ok. So, so the model 1, will have some sum squared error on this training said let me call it e_1 . Model 2 will have some error on another training set sorry model one itself will have error on the training set and it will have a third error on the third training set. So, either of model 1 if I want to calculate which error will you quote e_1 or e_2 or e_3 which error you will quote. So, what you quote is 3 components two components of the error. Error is this is the bias of the error and error is if I call it biases b the variation from the mean value of the error variation one-third, ok.

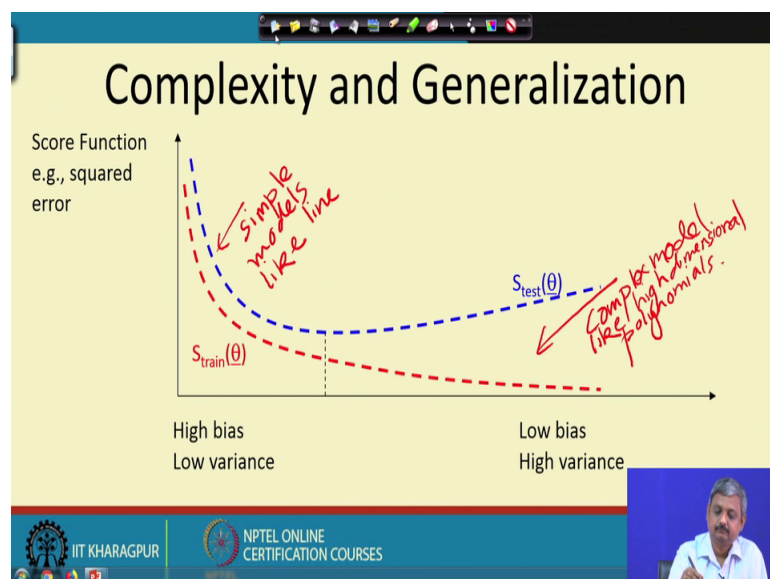
So, total error is bias plus variance because see it see error on single training set does not matter on different-different training sets for what error are you getting ok. So, your error of any model is the bias average error our training sets plus the variation of error about the training sets.

(Refer Slide Time: 19:06)



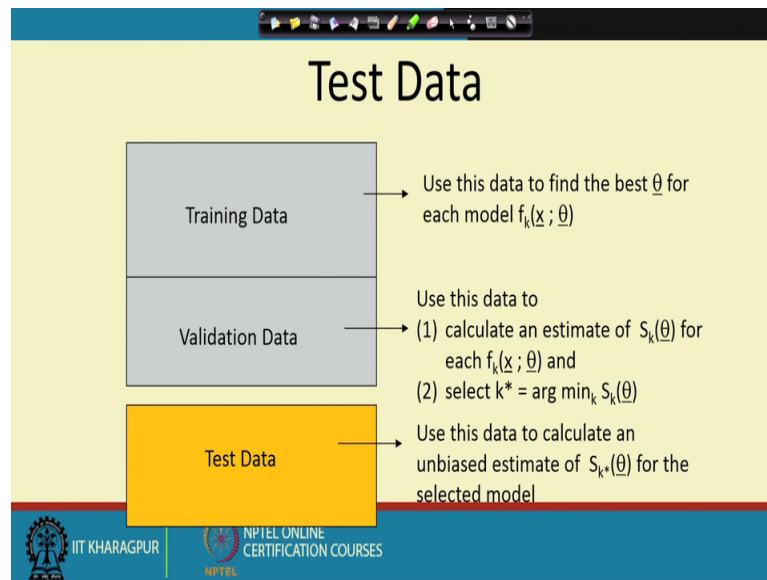
Now, simple models like model 1 will have high bias error, but low variance ok, whereas model 2 they will fit at an each of the training data they will fit well, but there will be a large variation ok. So, there is a trade off if you have simple model you have high bias low variance if you have complex model you have low bias high variance. So, minimize select the model which has lowest bias plus variance ok. So, these are the two extremes ok.

(Refer Slide Time: 21:40)



So, this has a simple model, and this is, yes ok, ok. So, that that is the principle.

(Refer Slide Time: 22:40)



So, usually what we do is that you use not only training and test data you do your validation data also. You find that training data to find the best model parameters use the validation data to have the best model among all the models and finally, use the test data to calculate the error. So, it is like when you prepare for an exam you do the exercises of a book, how well you are prepared you give a mock test that is the validation data, and your actual performance is the marks on the actual test which is the test data ok.

(Refer Slide Time: 23:30)

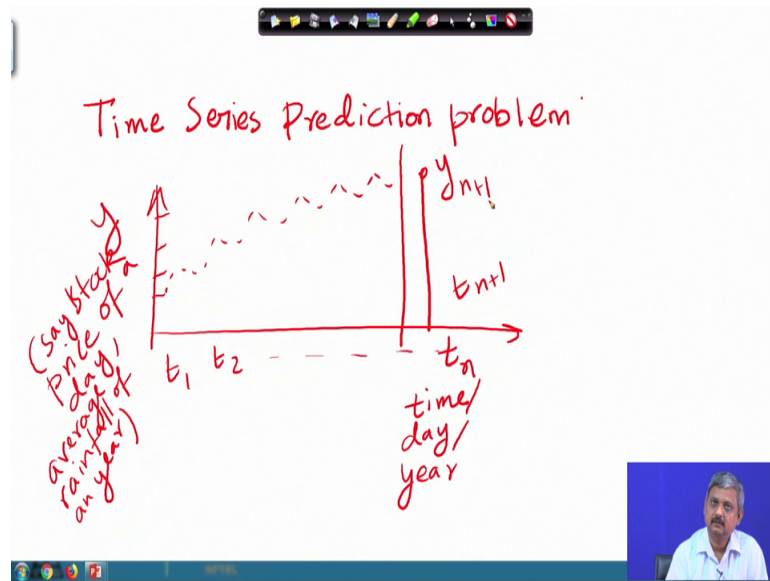
Time-series prediction as regression

- Measurements over time x_1, \dots, x_t
- We want to predict x_{t+1} given x_1, \dots, x_t
- Autoregressive model
$$x_{t+1} = f(x_1, \dots, x_t; \theta) = \sum \alpha_k x_{t-k}$$
 - Number of coefficients K = memory of the model
 - Can take advantage of regression techniques in general to solve this problem (e.g., linear in parameters, score function = squared error, etc)
- Generalizations
 - Vector x
 - Non-linear function instead of linear
 - Add in terms for time-trend (linear, seasonal), for “jumps”, etc

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

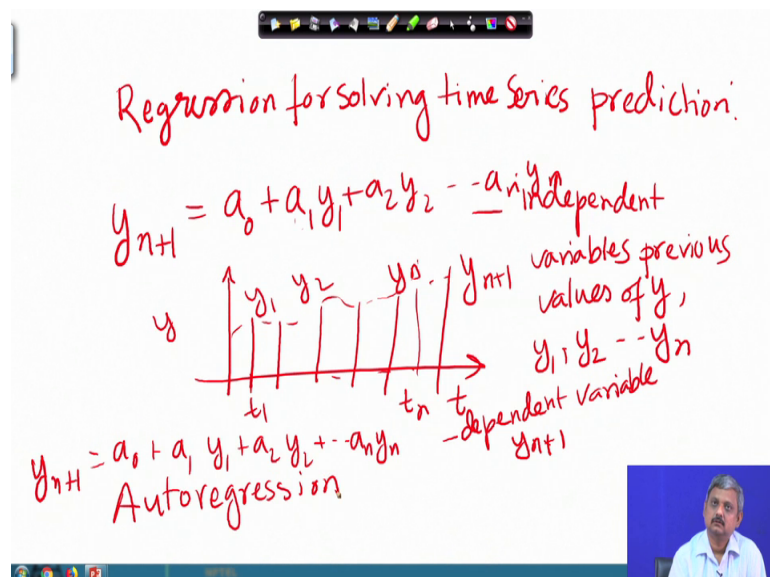
So, now some more applications of time of this regression.

(Refer Slide Time: 23:49)



So, y axis is some quantity which varies with time and x axis at the time instant. If we know the value of y up to nth time you have to predict the value of y at n plus 1th time that is the problem. So, how do you solve it in time series in using regression?

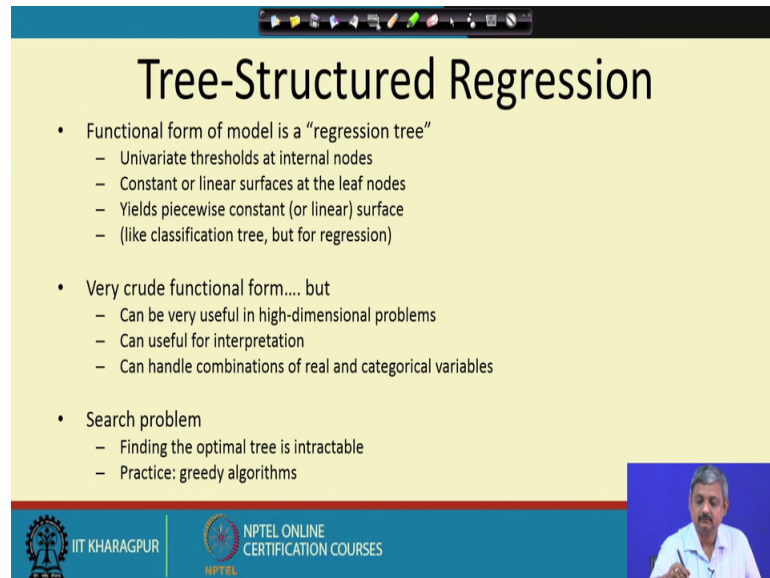
(Refer Slide Time: 25:30)



You make a model like this, you have to predict y at n plus 1. You have to predict and your independent variables are now this values of y previous values of y dependent variable is value you want to predict, next time instant. So, let me write down properly y n. So, independent variables are previous value dependent variable is next value. So, this

type of regression is called auto regression, regression on itself so that you can use for time series.

(Refer Slide Time: 28:10)



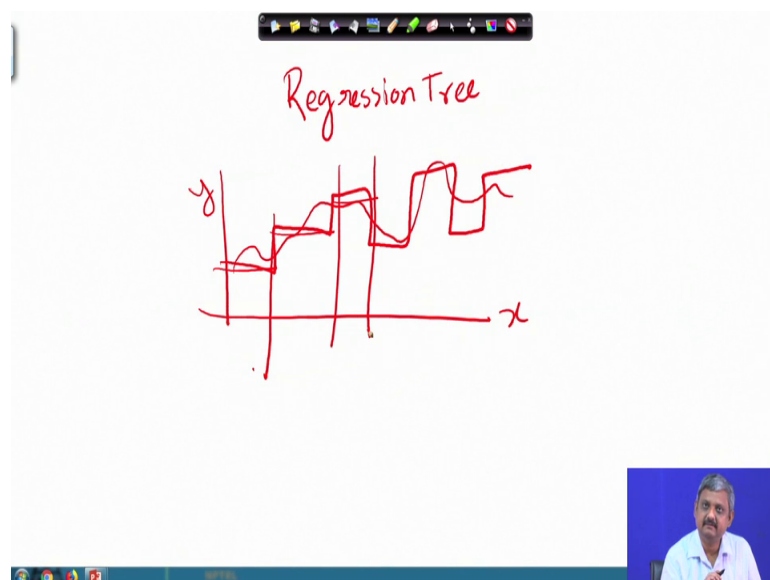
Tree-Structured Regression

- Functional form of model is a “regression tree”
 - Univariate thresholds at internal nodes
 - Constant or linear surfaces at the leaf nodes
 - Yields piecewise constant (or linear) surface
 - (like classification tree, but for regression)
- Very crude functional form... but
 - Can be very useful in high-dimensional problems
 - Can be useful for interpretation
 - Can handle combinations of real and categorical variables
- Search problem
 - Finding the optimal tree is intractable
 - Practice: greedy algorithms

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Another is a regression tree, where you say approximate any function by a regions and constant values on that region.

(Refer Slide Time: 28:15)

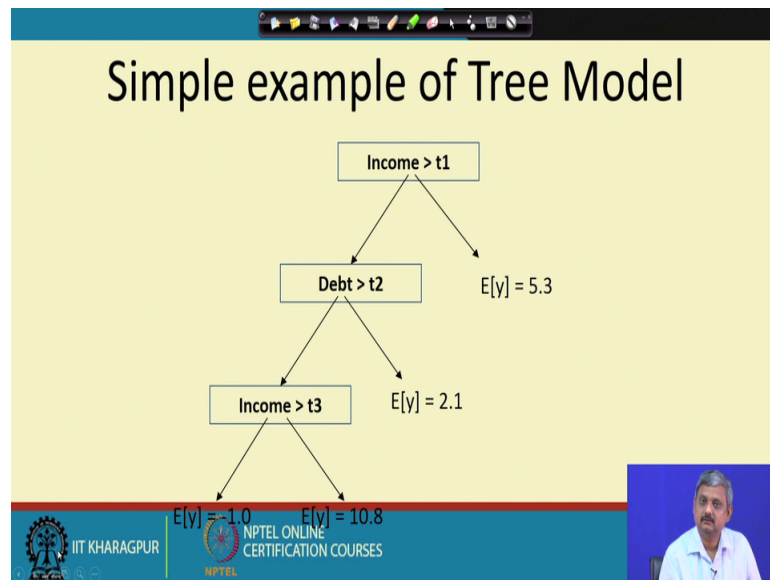


Regression Tree

A hand-drawn graph on a whiteboard showing a piecewise constant function. The vertical axis is labeled 'y' and the horizontal axis is labeled 'x'. The function is drawn as a red line that is constant in each of several intervals along the x-axis, with vertical lines indicating the boundaries between these intervals. The overall shape of the function is irregular, resembling a step function that approximates a more complex underlying function.

Regions you find by branches of the decently and taking at a constant value like this ok.

(Refer Slide Time: 28:48)



(Refer Slide Time: 28:52)

Greedy Search for Learning Regression Trees

- Binary_node_splitting, real-valued variables
 - For each variable x_j
 - For each possible threshold t_{jk} , compute
$$MSE(y; t_{jk}) = P(x \leq t_{jk})MSE(y|x \leq t_{jk}) + P(x > t_{jk})MSE(y|x > t_{jk})$$

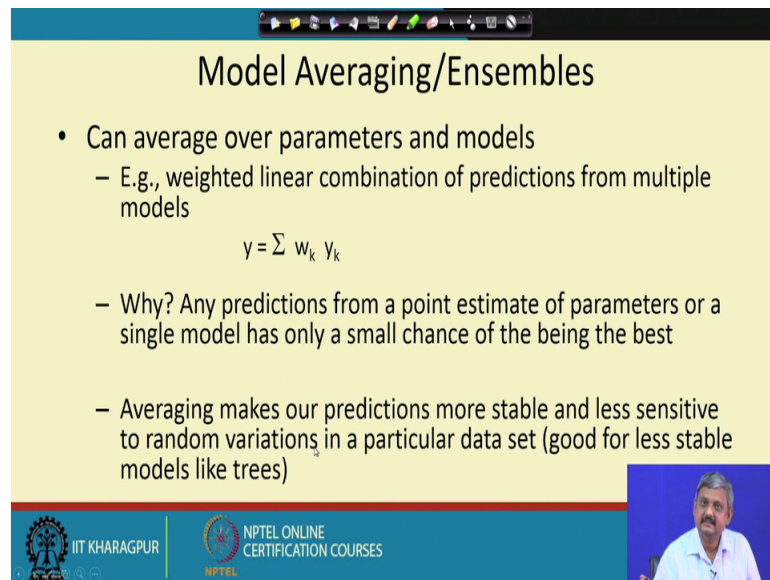
MSE in left branch MSE in right branch

 - Select t_{jk} with the lowest MSE for that variable
 - Select variable x_j and t_{jk} with the lowest MSE
 - Split the training data into the 2 branches
 - For each branch
 - If leaf-node: prediction at this leaf node = mean value of y data points
 - If not: call binary_node_splitting recursively
- Time complexity?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, then there are different ways of finding.

(Refer Slide Time: 28:56)



Model Averaging/Ensembles

- Can average over parameters and models
 - E.g., weighted linear combination of predictions from multiple models

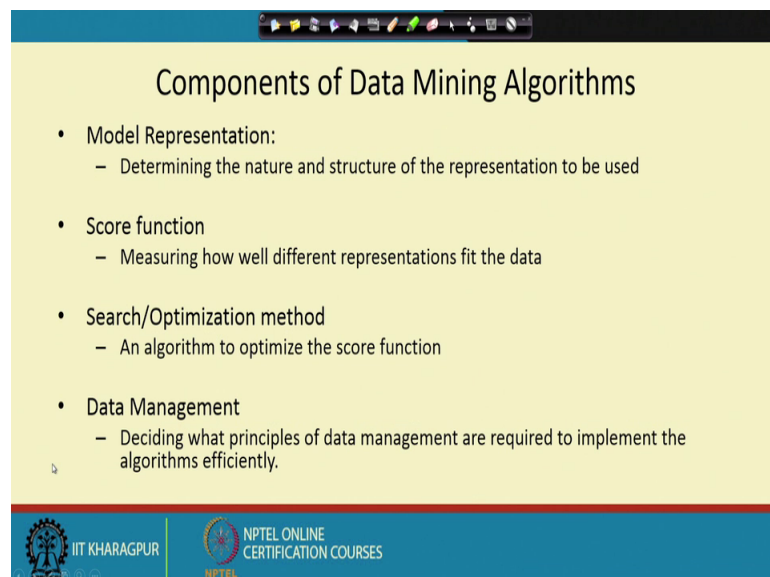
$$y = \sum w_k y_k$$

- Why? Any predictions from a point estimate of parameters or a single model has only a small chance of the being the best
- Averaging makes our predictions more stable and less sensitive to random variations in a particular data set (good for less stable models like trees)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And another way is you use several regression lines and then add them up called ensemble regression ok.

(Refer Slide Time: 29:02)



Components of Data Mining Algorithms

- Model Representation:
 - Determining the nature and structure of the representation to be used
- Score function
 - Measuring how well different representations fit the data
- Search/Optimization method
 - An algorithm to optimize the score function
- Data Management
 - Deciding what principles of data management are required to implement the algorithms efficiently.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

These are the different techniques and you have to go through these steps if you want to do data mining algorithm in the following as we have discussed.

So, thank you. In the next lecture we will discuss some techniques of dimensionality reduction.

Thank you.