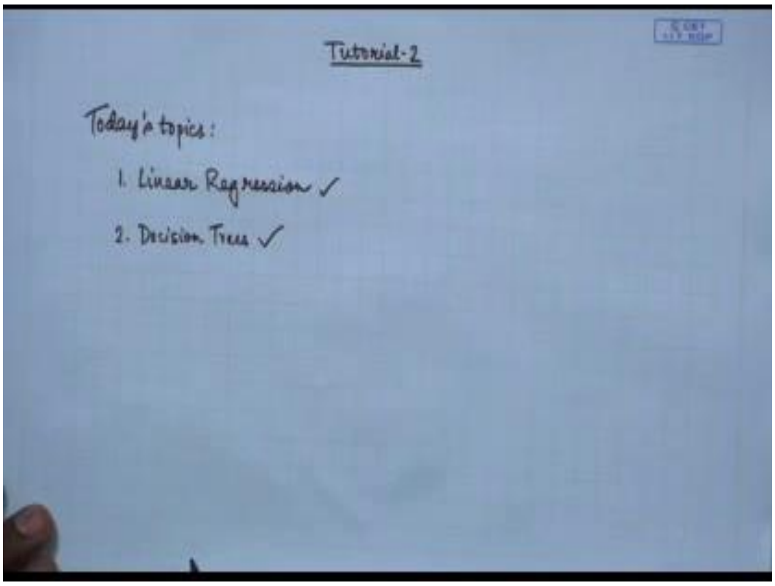
Introduction to Machine Learning
Prof. Mr. Anirban Santra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 11
Tutorial II

Hello everyone. Welcome to the second tutorial class of this course. I am Anirban and I am doing my PhD in machine learning. I am TA for this course. In this class, we are going to summarize all the topics that have been covered in the second week of this course and then we will study how to solve problems, and those kinds of problems, which you can expect in the assignments and the exam.

(Refer Slide Time: 00:43)



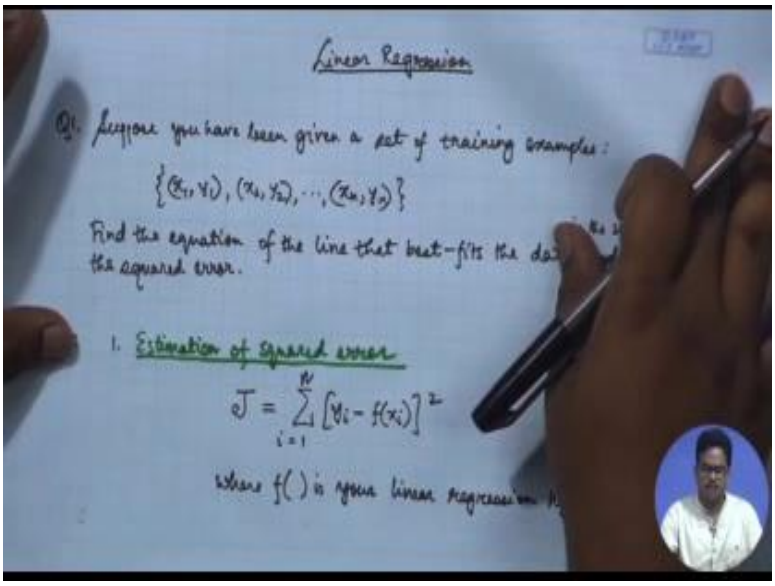
Tutorial-2

Today's topics:

1. Linear Regression ✓
2. Decision Trees ✓

So, the topics that we are going to cover today are linear regression and decision trees. So, let us take up the first topic first – linear regression.

(Refer Slide Time: 00:52)



Linear Regression

Q1. Suppose you have been given a set of training examples:

$\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Find the equation of the line that best-fits the da
the squared error.

1. Estimation of squared error

$$J = \sum_{i=1}^{N} [y_i - f(x_i)]^2$$

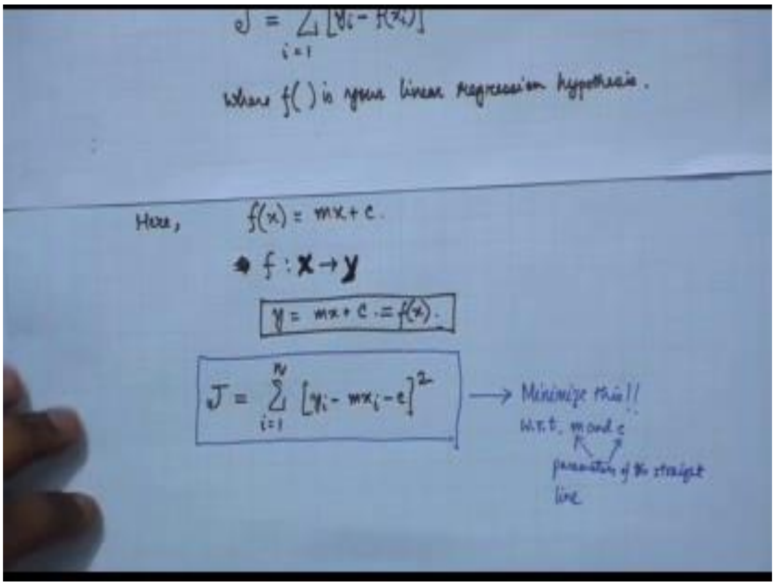
where f( ) is your linear regression h

As we all know, linear regression is all about fitting a straight line to the data. And, in n dimensions, straight line is generalized to a hyper plane. So, we have a set of training examples. Our motivation is to fit one straight line that best fits the data, which minimizes the least square error in approximation. So, you will face problems in the exam, which go like this. Suppose you have been given a set of training examples and they go like x 1, y 1, x 2, y 2 till x n, y n. So, this is your given training set.

It has n examples and these are points from the two dimensional real space. So, we are going to fit one single straight line to this data. Find the equation of the line that best fits the data and minimizes rather in the sense that, it minimizes the squared error; all right. So, we are going find that straight line, which minimizes the squared error and thus it is a best fit to the data.

So, the first step in this problem is to estimate the squared error. So, number 1 step is estimation of squared error. So, this is called squared error, but this is actually sum of squared error or you can also consider mean squared error. This does not change the solution, but it changes the shape of the error function a little bit. So, the squared error function J will look like it is equal to summation i equal to 1 through N y i minus f of x i squared; where, f is your linear regression hypothesis.

(Refer Slide Time: 04:17)



$$J = \sum_{i=1} [y_i - f(x_i)]$$

where $f(\,)$ is your linear regression hypothesis.

Here, $f(x) = mx + c.$

$f : \mathbf{X} \rightarrow \mathbf{Y}$

$$\boxed{y = mx + c := f(x).}$$

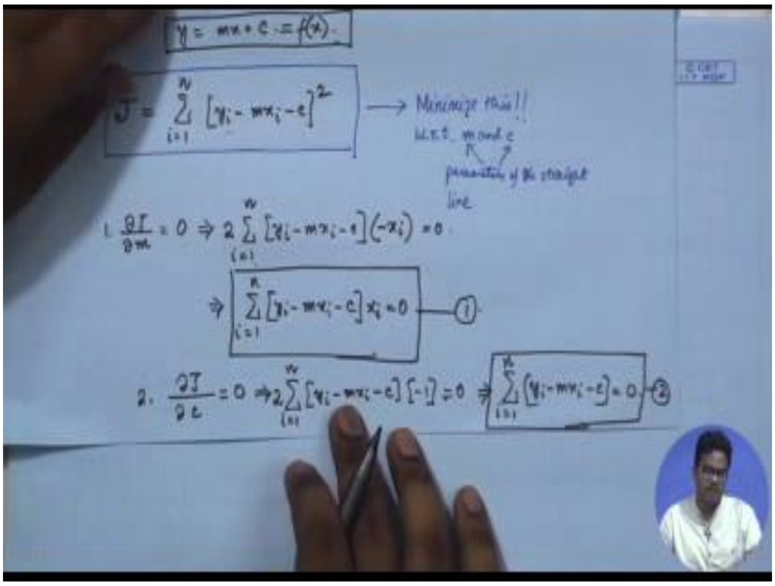
$$\boxed{J = \sum_{i=1}^{N} [y_i - mx_i - c]^2}$$ → Minimize this!! w.r.t. m and c

parameters of the straight line

Now, as is very clear from the name that, f will be of the form; think is of linear regression; f x is going to be of the form m x plus c. So, rather I can write it as f maps from x to y. So, let this x be little different. So, this is the set of x’s to the set of y’s; all right. So, then the definition is y equal to m x plus c. So, this is f x; all right. So, this is the equation of the line that you are trying to figure out; all right. So, what we will do is we will rewrite the error.

So, the error is equal to 1 through N – y i minus m x i minus c whole square. So, this is going to be the error function that we will try and minimize. So, objective is to minimize this; of course, with respect to m and c, because these two are the parameters of your learning algorithm of the straight line, which you are trying to tune and adjust to the data. So, how do we go about solving it? We find the derivative of j with respect to m and c and set them to 0.

(Refer Slide Time: 06:09)



$$\boxed{y = mx + c = f(x)}$$

$$\boxed{J = \sum_{i=1}^{N} [y_i - mx_i - c]^2} \longrightarrow \text{Minimize this!! w.r.t. } m \text{ and } c$$

parameters of the straight line

$$1.\ \frac{\partial J}{\partial m} = 0 \Rightarrow 2\sum_{i=1}^{N} [y_i - mx_i - c](-x_i) = 0$$

$$\Rightarrow \boxed{\sum_{i=1}^{n} [y_i - mx_i - c]\, x_i = 0} \text{ --- } (1)$$

$$2.\ \frac{\partial J}{\partial c} = 0 \Rightarrow 2\sum_{i=1}^{N} [y_i - mx_i - c][-1] = 0 \Rightarrow \boxed{\sum_{i=1}^{n} [y_i - mx_i - c] = 0} \text{ --- } (2)$$

So, first we do del J del m equal to 0; all right. So, this is going to be - del J del m is going to yield you to i equal to 1 through n; then, y i minus m x i minus c times minus x i; right? So, you first find the derivative of this entire quantity. So, you first take the derivative inside the summation sign; then, you take the derivative of this quantity with respect to the quantity inside the square and you get this to times the quantity inside and then you will take the derivative of the quantity inside this thing in terms of m – by chain rule. And, what comes out is minus x i.

And, you are trying to set this thing equal to 0; which amounts to summation i equal to 1 through n. So, we can send this minus n and 2 to the other side. So, we can divide both sides by minus 2. And, we end up with this thing. So, it is just y i minus m x i minus c times x i equal to 0. So, this is going to be our first equation. So, I will keep this. So, let me try and make both of these things visible; oops; all right, let us keep it this way for continuity. So, now you can see; all right. So, you have the error function, which is j; you have the first equation.

And, the second thing is we are going to find out the derivative of; let me write here; del J – del – this is the second thing, c equal to 0. So, what is this going to yield us? So, again we go back to this. So, first, we have this summation – i equal 1 through n; then, we have y i minus m x i minus c. We can actually get rid of this, because we have this. So, this is what we are trying to take the derivative of. So, del J del c is equal to again twice – 2 comes before twice of this time minus 1; right? Because the derivative of this quantity with respect to c is going to be minus 1, which is coming here, so this quantity will be set to 0 in order to minimize; and this again boils down to yes this boils down to this quantity, summation i equal to 1 through n y i minus m x i minus c equal to 0, because we can send this minus 2 to the other side and we are left with this quantity. So, this is the second equation.

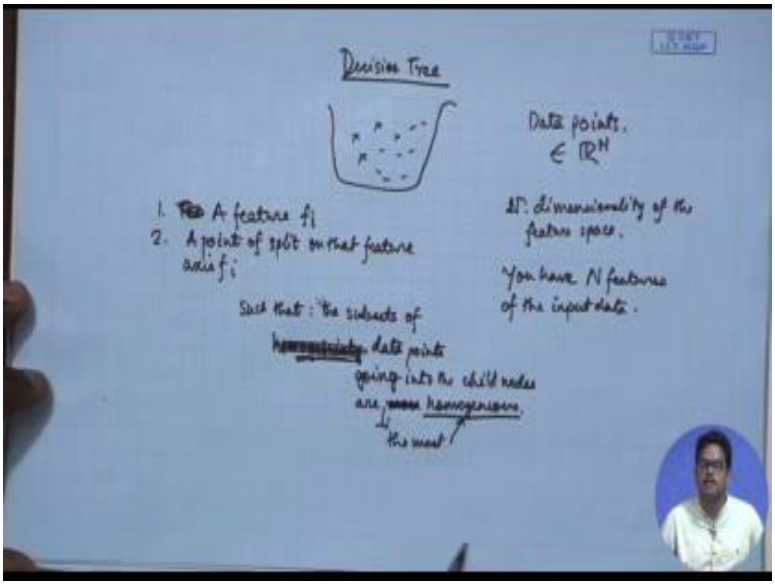
Now, what you can do is; yes you have these two equations and you can plug in the values of x i's and y i's; all right. So, say you have been given this set, right? You have been given this set of values x i and y i – 1 through n. So, you are going to put these values in these two equations. So, this x 1; once you will have one expression – one of these terms for i equal to 1, so you put the values of x 1 and y 1 over here. And, over here as well; and, use to make it some over all those values. So, you get now one equation.

The second equation comes from here. If we put the same values, x – the values of x i's and y i's in this equation; and, you get another. And then, you simultaneously solve these two equations and you get the result; all right. So, it is going to be pretty easy, pretty straightforward. And, I hope that there will be no problem in solving this kind of questions. So, this is how you find the equation of a best fit line given a set of data points.

Now, the next and another kind of question that you can face from this section is like asking about the expression of squared error should be; and, it is very easy; the one we wrote over here. And, this is the expression of the squared error. So, you have to choose from one of the examples maybe. And, if say some other kind of (Refer Time: 11:35) in the exam if difficult question comes; and, you have been asked, it has been specified that, the error is something different from squared error like cross (Refer Time: 11:46) You may look up the web what it really is.

So, in that case, the expression of j will change. And, such kinds of error functions are applicable for different kind of like different kinds of applications. They are tuned to certain applications. And so, if a different kind of error function is given, then you have to evaluate it this way and then you have to take the derivative with respect to the different parameters of your model and then find out the values by solving simultaneous equations.

(Refer Slide Time: 12:26)



Decision Tree

Data points.
$\in \mathbb{R}^N$

N: dimensionality of the feature space.

You have N features of the input data.

1. A feature f_i
2. A point of split on that feature axis f_i

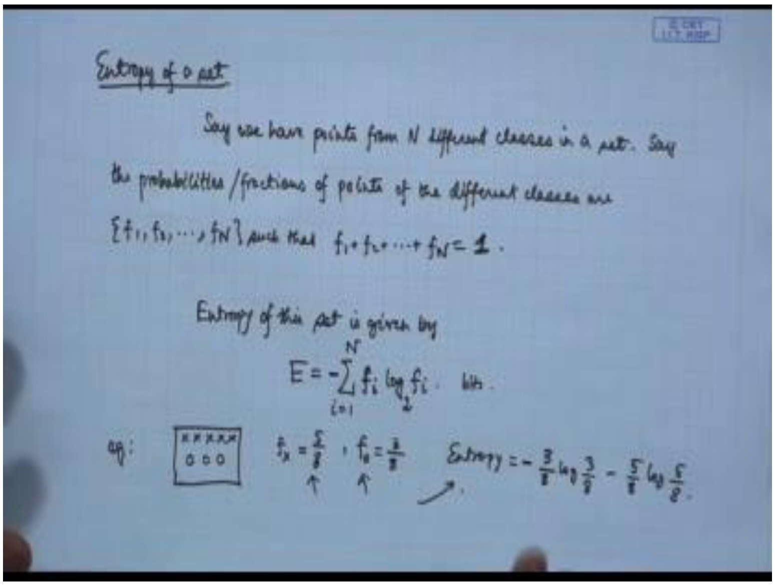
Such that : the subsets of ~~homogeneity~~ data points going into the child nodes are ~~more~~ homogeneous.
the most

The next topic that we are going to take up today is decision tree. So, decision tree is nothing but a set of nested if else conditions, which take one huge dataset, say we have this bucket full of data points. And these data points are from in different dimensions; right? So, our data points – these are points belonging to say R raised to the power N. From an n dimensional phase, so you have space. So, you have n is the dimensionality of the feature space.

Or, in other words, you can say that you have n features of the input data. So, how do you learn the decision tree? At every step, you are going to choose number 1 a feature; and second, a point of split on that a feature f i, I would say; on that feature axis f i; all right such that homogeneity; geneity – n e i t y, whatever. Such that the subsets of data points going into the child nodes are more homogenous. So, there should be a method of calculating the homogeneity of a set in this context.

So, at every step, you have to first estimate for which feature we have the; I would say the most possible as homogenous as possible; all right. So, you have to choose that particular feature and that particular split on that feature axis such that the homogeneity of the subsets being produced as a result of that is maximized. So, there should be a method of quantifying the homogeneity of a set and or homogeneity or purity of a set. And, one of those metrics is entropy. So, let us look at the expression of entropy.

(Refer Slide Time: 15:48)



Entropy of a set

Say we have points from N different classes in a set. Say the probabilities / fractions of points of the different classes are $\{f_1, f_2, \dots, f_N\}$ such that $f_1 + f_2 + \dots + f_N = 1$.

Entropy of this set is given by

$$E = -\sum_{i=1}^{N} f_i \log_2 f_i \quad \text{bits}.$$

eg: $\boxed{\begin{matrix} \times\times\times\times\times \\ \circ\circ\circ \end{matrix}}$ $f_\times = \frac{5}{8}$, $f_\circ = \frac{3}{8}$ Entropy $= -\frac{3}{8}\log\frac{3}{8} - \frac{5}{8}\log\frac{5}{8}$.

Entropy of a set as you have studied earlier, let us say we have points from n different classes in a set. And, say the probabilities, if we calculate the probability by the frequency definition of probability, then this amounts to; this is the same as saying that, let the fraction or fractions of points of the different classes are say f 1, f 2 this way until f N. So, let us say that, these are the different fractions of points; all right. Such that f 1 plus f 2 dot dot dot till f N is 1.

So, this is also the probability of a particular class in that set. So, we have a huge mixture of points of n different classes; and, this is the class distribution. These are the fractions of the points from different classes. So, how do you calculate the entropy? So, entropy of this set is given by E equal to summation i equal to 1 through capital N; of course, a negative sign f i log f i.

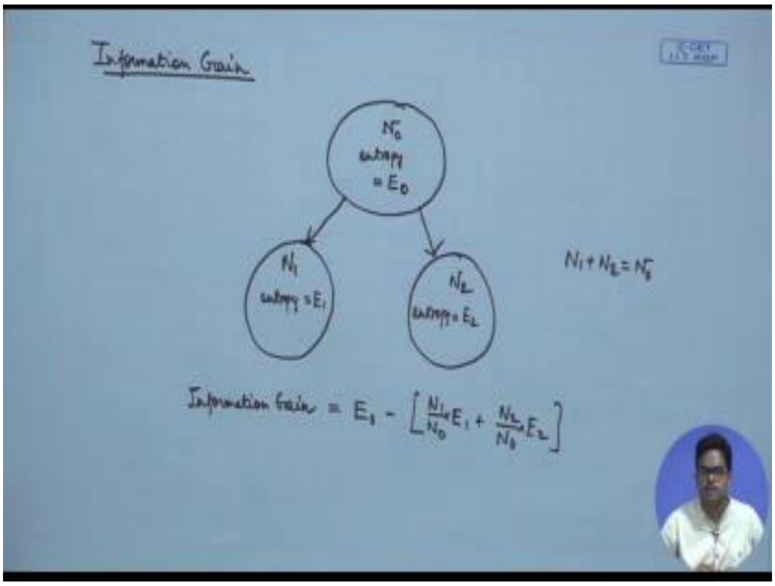
Now, the base of the logarithm decides the unit of the entropy. And, popularly, we have a log base 2 in the definition of entropy. And then, the entropy is quantified end units of bits. So, given a set of say N different classes, all right? Say we have this for example; we have a set, which looks like this. See there are five points of class crosses and say three points of class circles. So, f cross is going to be how much? So, it is going to be 5 by 8; right? And, f naught or f f – you know circles, is going to be 3 divided by 8.

And, entropy is equal to summation or just I will write it is going to be 3 by 8 log of 3 by 8 minus 5 by 8 log of 5 by 8. So, whatever is the answer? So, first you have to estimate the probabilities or the fractions. And, if you are estimating from frequency definition, then it is equal to the fraction of points. And then, you are going to calculate the entropy. So, you can find this kind of questions in the exam in which you will be given a set of points and you have to calculate the entropy of that set; pretty easy right? So, this term as you can say as you can see that, this term is going to be minimum and it can be proved when the set is a uniform one; so you have the same number of members from every single class; and in that case, the entropy is maximized; and, because there like complete randomness – uniform distribution.

And, entropy is going to be zero when there is a just point from one single class. So, the fraction of that particular class will become 1. So, it is 1 log 1. So, log 1 is .0. So, you are going to have zero entropy. And, in course of training a decision tree, our motivation is to keep splitting the dataset into fragments, into subsets until we have close to zero entropy. So, we should be able to say for certain that, given we are at a particular leaf node, all right; the class distribution is almost certainly just 1.

We are completely sure that, the point belongs to one particular class. So, you keep start pushing the unknown sample from the top of decision tree and you end up at a leaf node. And, that particular leaf node should stand for one particular class. So, when the example lands up in that particular leaf node, we can say that, yes, this example belongs to that particular class, which that leaf node was corresponding to; right? So, this is the notion of entropy.

(Refer Slide Time: 21:06)



Information Gain

Information Gain $= E_0 - \left[\frac{N_1}{N_0} E_1 + \frac{N_2}{N_0} E_2\right]$

And, in connection with entropy, there is another quantity, which is called information gain. So, assume that, we had an initial set of capital N examples. Now, these kinds of questions are going to come in exam. And, in this particular demonstration, I am not going to use numbers; I am going to just use symbols, so that you can remember the expression and kind of like use it to quickly solve problems in exam.

And, also to get a feel of what is really happening. So, say we are starting with capital N examples; and, in the initial entropy – entropy of this particular set is equal to say E 1. Or, rather let us say E zero or this is N 0. So, initially, we have in our set, N 0 samples and the entropy of the set is E 0, which was calculated in the way that I just demonstrated to you, using this definition. So, the entropy was calculated. And, this is what we have.

Now, we chose a particular feature axis and a particular split on that feature axis; and, we ended up producing these two subsets. So, the subsets are N 1 and N 2 large. So, N 1 plus N 2 is equal to N 0. And, the entropy is E 1 and E 2. So, the information gain is going to be calculated as E 0 minus N 1 by N 0 into E 1 plus N 2 by N 0 into E 2. So, this is the formula for calculation of information gain. And, this quantifies how much randomness has been reduced or how pure the subsets become as a result of this split.

And, at every step of decision tree learning, we choose the feature axis and a split on that feature axis, which maximizes the information gain. So, this is one of the criteria of decision tree learning. So, in the exam, you are going to find questions in which you will be asked that, which particular feature is the best to choose and in the context of like it for which feature maximizes the information gain and thus is the best choose.

So, you have to try out for every single feature given in the question and calculate the information gains associated with them. And, thus you can like figure out which feature is the best and the one which maximizes the information gain is the best to choose. Or, you will be given this kind of a scenario and ask to calculate the information gain. You will be able to do it; right? So, this concludes the tutorial of this week. And, the assignment will be released this Sunday and an announcement will be made in the forum. This tutorial video will also be made available along with the notes.

So, the deadline will be the Thursday after this week. So, one and a half weeks after the start of week 2, the deadline will be set. So, all of those will be announced in the portal. And, best of luck; wish you can solve these kinds of questions in the exam quite comfortably.

Bye-bye, see you next time.