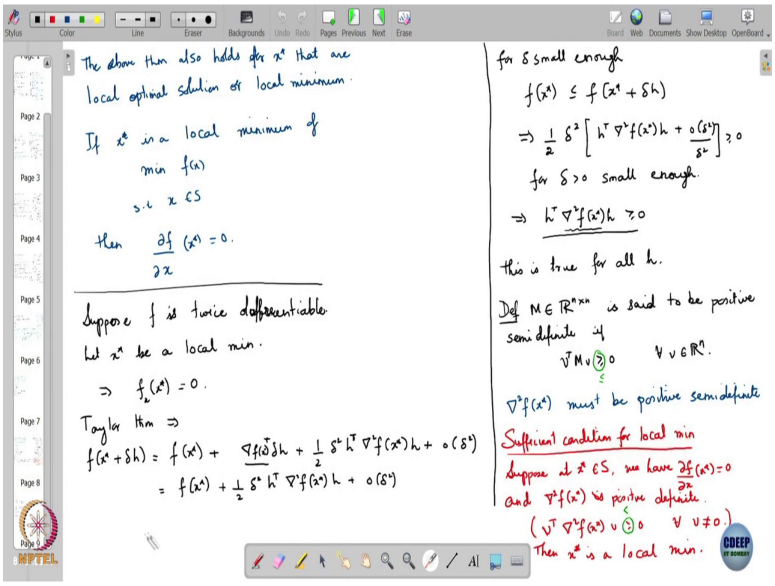
Optimization from Fundamentals
Prof. Ankur Kulkarni
Department of Systems and Control Engineering
Indian Institute of Technology, Bombay

Lecture – 5A
Least square regression

We will; welcome everyone, we will continue with our study of optimization I had some questions from the previous class which I want to address here. Some of you are asking me about; the difference between what is a sufficient condition and necessary condition.

(Refer Slide Time: 00:44)



The above thm also holds for x^* that are local optimal solution or local minimum.

If x^* in a local minimum of

$\min f(x)$

s.t $x \in S$

then $\frac{\partial f}{\partial x}(x^*) = 0$.

Suppose f is twice differentiable.

Let x^* be a local min.

$\Rightarrow f_x(x^*) = 0$.

Taylor thm $\Rightarrow$

$$f(x^*+\delta h) = f(x^*) + \underline{\nabla f(x^*)^T}\delta h + \frac{1}{2}\delta^2 h^T \nabla^2 f(x^*) h + o(\delta^2)$$
$$= f(x^*) + \frac{1}{2}\delta^2 h^T \nabla^2 f(x^*) h + o(\delta^2)$$

for δ small enough

$f(x^*) \le f(x^*+\delta h)$

$\Rightarrow \frac{1}{2}\delta^2 \left[h^T \nabla^2 f(x^*) h + \frac{o(\delta^2)}{\delta^2} \right] \ge 0$

for $\delta > 0$ small enough.

$\Rightarrow \underline{h^T \nabla^2 f(x^*) h \ge 0}$

this is true for all h.

Def $M \in \mathbb{R}^{n \times n}$ is said to be positive semidefinite if

$v^T M v \underset{\le}{\boxed{\ge}} 0 \quad \forall v \in \mathbb{R}^n$.

$\nabla^2 f(x^*)$ must be positive semidefinite

Sufficient condition for local min

Suppose at $x^* \in S$, we have $\frac{\partial f}{\partial x}(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite.

$(v^T \nabla^2 f(x^*) v \overset{<}{\boxed{>}} 0 \quad \forall v \ne 0.)$

Then x^* is a local min.

So, let us just recall this we had concluded that if x star is a local minimum of a of an optimization problem like this o where you have minimizing a function o f x over an open set S ok.

Then it has then and if x star is a local minimum of this optimization problem then it has to be that the derivative of f at x star or the gradient of f at x star they must be equal to 0, this is this has to hold. So, this is what we call a necessary condition this is means that if x star is a solution then it must satisfy this. Now the we also got a stronger necessary condition which said that the Hessian of f at x star must be positive semi definite ok.

So, the condition is written here in blue. So, it said that if x star is a local solution then the Hessian of f at evaluated at x star must be positive semi definite, this is also a necessary condition. So, if x star is a local minimum then the Hessian must be positive semi definite this is what does condition say.

There is a condition which is which goes in the opposite direction which says that if such and such condition is satisfied then x star is a local minimum, that is what we mean by a sufficient condition. So, that is what is written here in red; so a sufficient condition for local minimum.

So, suppose x star is point in S and again we are in the same set in S is an open set f is a different twice differentiable function. And suppose we have that the gradient is of f at x star is equal to 0 and moreover the Hessian is positive definite which means that; V transpose Hessian times V is great strictly greater than 0 for all V naught equal to 0 right.

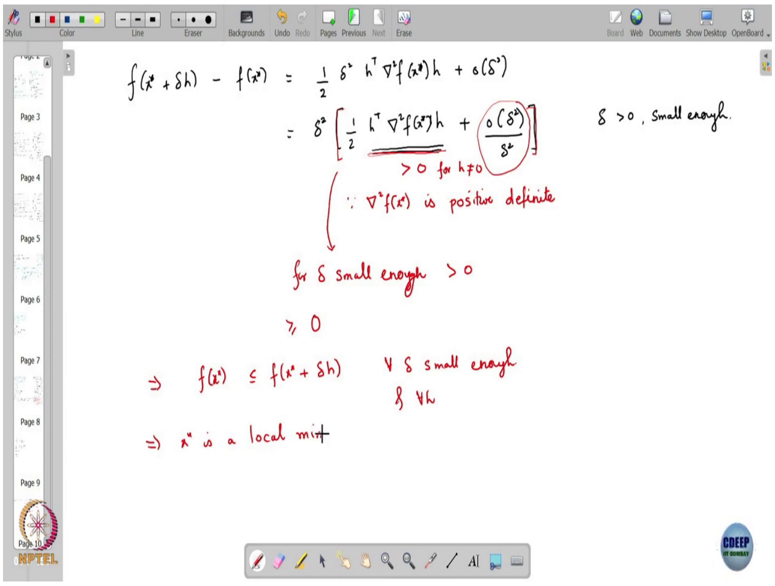
So, if the Hessian is positive definite then x star must be a local minimum ok. So, these guarantees that x star is a local minimum ok, but every local minimum may not satisfy this.

There are local minima that that violate this which are that violate this, but they must satisfy surely the other necessary conditions, which means that that the derivative should be equal to 0 and the Hessian should be positive semi definite ok. Also a quick outline of how the proof for this sufficient condition for local minimum would work remember by Taylors theorem we had argued this equation here, f of x star plus delta h where delta is some small positive quantity h is a direction in r n.

So, f of x star plus delta h can be written in this sort of form; f of x star plus half delta square h transpose the Hessian of f at x star h plus something that is small o of delta square. Now how did we get here? We got here by using that the first order term the linear term in h; actually 0 because the gradient is equal 0 right.

Gradient of f at x star is equal to 0. So, this term vanished and you are left with only this term. Now if you look at the difference between the left hand side and the right hand side ok.

(Refer Slide Time: 04:25)



$$f(x^* + \delta h) - f(x^*) = \frac{1}{2}\delta^2 h^T \nabla^2 f(x^*) h + o(\delta^2)$$

$$= \delta^2 \left[\frac{1}{2} \underbrace{h^T \nabla^2 f(x^*) h}_{>0 \text{ for } h \neq 0} + \boxed{\frac{o(\delta^2)}{\delta^2}} \right]$$

$\delta > 0$, Small enough.

$\because \nabla^2 f(x^*)$ is positive definite

for δ small enough > 0

≥ 0

$\Rightarrow \quad f(x^*) \leq f(x^* + \delta h) \quad \forall \delta$ small enough & $\forall h$

$\Rightarrow$ x^* is a local min

If you look at the difference between the left hand side and the right hand side let me. So, if you look at the difference f of x star plus delta h minus f of x star this quantity is equal to half of delta square h transpose del square f at x star h plus something that small o of delta square.

So, you have this you look at this difference, now we know that when delta is small enough, what must happen?

See remember we are arguing in the opposite direction we now want we want to say that if the Hessian is positive definite then x star must be a local minimum ok. So, which means that we want to say that when delta is small enough the this quantity the quantity on your left hand side here the quantity on your left hand side must be greater than equal to 0 for delta small enough ok.

So, the way to see that is to just simply take delta square common here, you have left within half h transpose this plus this quantity which is small o of delta square divided by delta square this is true for all so for delta positive hence and small enough.

Now we know that when h is not equal to 0, when h is not equal to 0 this quantity this underlined term this term is strictly positives, why? Because at because the Hessian at x star is positive definite this is positive for h not equal to 0 this is because the Hessian is positive definite.

Now, because the Hessian is positive definite this term becomes strictly positive and what do we what can we say about this other term the term that have circled. It is a term because it is the numerator is small o of delta square, I divided that by delta square. What that means, is as its it is quantities as small o of delta square is a quantity that upon dividing by delta square also goes to 0 right. So, small o of delta square divided by delta square goes to 0 as delta becomes as delta goes to 0.

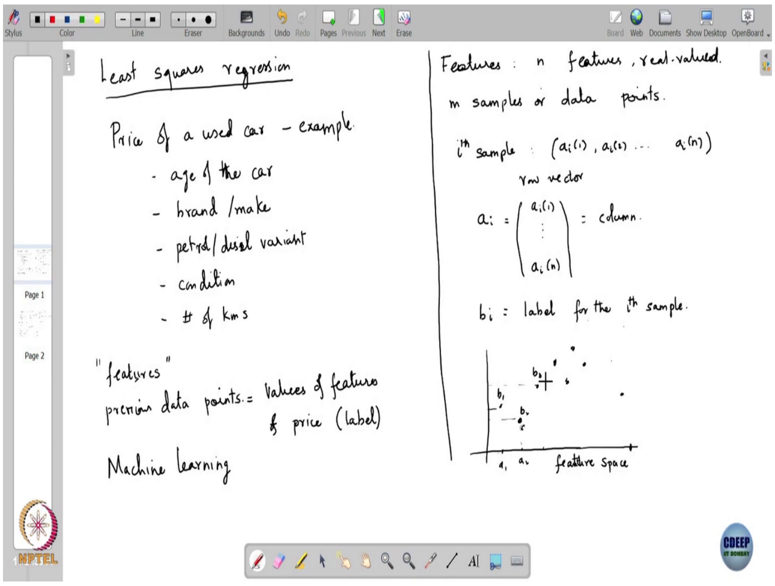
So, as delta becomes small this quantity will become eventually smaller in magnitude than the first term it goes to 0. So, eventually it has to become smaller in magnitude than the positive, if whether this term is positive or negative does not matter to us eventually in magnitude it is going to become smaller than the first term.

So, consequently whatever is there in the bracket here is going to add up to something positive for delta small enough right. So, this is this bracket for delta small enough small enough its bracket is positive and outside I have a delta square which is also positive right.

So, in short what have we got? We got that well if as a consequence we got that this quantity this difference is definitely greater than; this difference is definitely greater than equal to 0, which means that ok. So, that gives you that which implies that in x star is a local minimum ok. So, this is guaranteed that x star is a local minimum ok.

Now let us do a couple of applications of this, first set of applications are going to be where it is evident that the optimization problem is an actually an optimization over an open set. In the second case what is going to happen is the you will it is not very evident that it is an optimization over an open set it has to be mathematically transformed in that sort of way ok. And that will also lead us to our next class of optimization problem ok.

(Refer Slide Time: 09:53)



Least squares regression

Price of a used car – example.

- age of the car
- brand /make
- petrol / disel variant
- condition
- # of kms

"features"

previous data points. = Values of features & price (label)

Machine learning

Features : n features, real-valued.

m samples or data points.

i^{th} sample : $(a_i(1), a_i(2) \dots a_i(n))$

row vector

$$a_i = \begin{pmatrix} a_i(1) \\ \vdots \\ a_i(n) \end{pmatrix} = \text{column.}$$

b_i : label for the i^{th} sample.

So, the first category is this problem of what is called least square regression. So, the problem in this kind of, the question in this kind of problem is that you have a bunch of; you have a bunch of inputs or you can say independent variables which affect an output. And we want to infer a relation between input and output and the relation is not always exact we the inputs and outputs are related to each other possibly in a noisy fashion.

But we do not have a model for the noise, we do not have the may probably many other unmodeled things elements in the problem which are which are affecting the output. So, consequently we do not have a perfect relation.

What we want to infer is a relationship? That will be good enough for predictive purposes. So, it may not be good it is not necessarily good for not it is not enough that we that the relationship is matches very well on the data that we have.

But rather it our goal is actually that it match it should perform well on unseen data on in order to and give us good predictions on unseen data. So, I will explain this with an example, suppose we want to we are talking of say price of a used car, just is an example. So, price of a used car, now used cars come in all kinds of varieties.

So, but we can think that maybe the price of the used car depends on a few a few input variables such as for example, the age of the car, such as the brand or make of the car, it depends on whether it is a petrol or diesel variant [noise,] it depends on the condition height of the car, depends on the number of kilometers it is it has run, what has been it is running etcetera ok.

So, these are in the language of machine learning what are called features; features are what I was calling input variables. So, these are what are called features and the price of the used car is what we want to infer from it. So, if you want to infer using the features.

So, if I the goal is that I give you a new; I give you a sample car which has sample used car which has all these features. It has a certain age tell you it's brand, I tell you whether it is a petrol or diesel car, I tell you it is condition, tell you the number of kilometers it has run and I want you to predict, what the price is going to be? Ok.

At what price will they sell in the market? This is this is the problem and you, now of course, you cannot do this in without any data. So, I will give you some previous data ok. So, I will give you previous data points ok. A previous data points what do these previous data points correspond to? They correspond they give you a list of features, they give you values for features and they tell you the price and the price at which the car is sold ok. This price is what we called a label.

So, you are previously given your given some previous values of the features and you are given also the values of labels corresponding to that ok. And now what we what the goal is that using all this data you come up with some way, some predictor for a new sample; new sample does not mean a brand new car a new sample of an old car of a used car and unseen sample.

Which possibly has a different values of features and use that to predict the price. This is this sort of problem is a is a sort of a basic problem in machine learning ok. Now what we what we let us introduce a bit of notation here these features I will it is denote features are going to be denoted as in the following way. So, I have suppose so I have suppose n features; n features include you know things like this and these I will assume they take values in a continuous space.

So, where these either or type features or brand type features which are which are discrete. Let me now let us ignore them for the moment let us assume that these are n features taking values in the continuous space. So, this could be age of the car some way of measuring condition the number of kilometers it has run and so on ok.

So, these are features that n features that are real valued each of them is the real number. So, and I am going to give you m samples or data points. So, the way this is we will denote this since there are n features. So, what we will do is we will put stack all of them into a vector. So, the for the ith sample the values of the features that you that you have let us denote them in this sort of way ai of 1 dot dot dot.

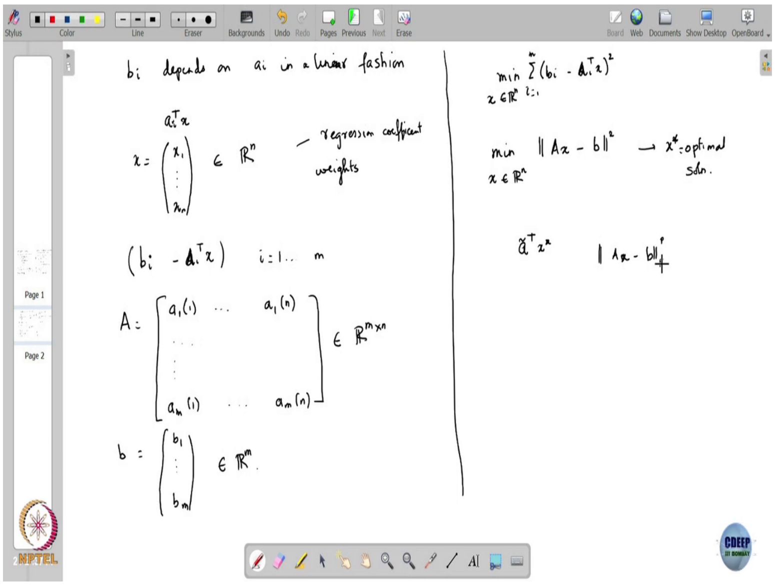
So, ai of 2 dot dot dot; ai of n this becomes a row vector like this ok. The column vector like this ai is will just be the column vector from by this. So, this is the column vector the price or the label that we have got for in the ith sample is denoted bi and this is also an a real number. So, label now what you what one typically has also in addition to this is a biased feature, I will ignore that for the moment.

But so, for simplicity this let us look at this sort of question you have you have previous data points where the this is my feature space the space in which my ai take their values. So, I have a say a 1 here and corresponding b 1 here I have an a 2 here a 2 suppose and b 2 and a 3 and b 3 b 3 etcetera.

And so you get this sort of you have these you have these m samples ok. Is everyone understands this there are m samples that is the m is the number of dots that I have here like this of this kind n is the is the length of the feature vector ok.

Now what I would like to do is somehow come up for a new unseen point. So, here is a new point in the feature space which I have never seen before I would like to know what should be the right value of the price or the or the label for me correct. So, if I want to do this kind of prediction what I need to do is build a model using what is using the data that I have. And so what we will positive there is we will hope for linear relationship ok.

(Refer Slide Time: 19:56)



b_i depends on a_i in a linear fashion

$a_i^T x$

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ – regression coefficient weights

$(b_i - a_i^T x) \quad i = 1 \ldots m$

$A = \begin{bmatrix} a_1(1) & \cdots & a_1(n) \\ \vdots & & \\ a_m(1) & \cdots & a_m(n) \end{bmatrix} \in \mathbb{R}^{m \times n}$

$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$

$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} (b_i - a_i^T x)^2$

$\min_{x \in \mathbb{R}^n} \| Ax - b \|^2 \rightarrow x^*$: optimal soln.

$\tilde{a}^T x^*$ $\quad \| Ax - b \|_p^p$

We will let us pose it in this sort of way let us pose it that that b bi depends on ai in a in a linear fashion. Now what this mean is, so let us let us suppose that you compare the true value ai a true value bi with a with the value of a linear function of the features which is denoted like this ai transpose x ok, ai transpose x.

So, this is a linear function on a linear function of the feature. So, the coefficient here x are this vector x 1 till x n. These coefficients if I as I vary the coefficients I will get a different value for ai transpose x and I am hoping to match get coefficients in such a way that when I these coefficients in such a way that when I that when I input a new set of features here unseen set of features it will give me a good it will give me a good approximation or a good prediction for the for the feature or the label ok.

Now, the trouble is if you look at the sort of would figures that I do here data the way it is because there are so many unmodeled elements there is clearly no need not really be a linear relationship between the label and the feature ok. So, what one tries to do is then you think of a cost function, a cost function or the loss function.

We try to we look at you look at the mismatch or look at the residue like this look at the residue like this and we say well we want to look at we would like to this we would like to find an x such that this these residues I goes form one till m, these residues are minimum are as low as possible ok.

This is this is what is called a training problem. So, you would look you try to find xs and thereby define this function define this linear function through those xs. So, you have try to find xs so that this residue is as small as possible. So, a popular a popular way of doing this is to do is to simply write it like this. So, let us write a as this matrix I will. So, this is this puts together all the observed features of the m data points let us me write b also as this column vector.

These are the observed labels of the m data points. So, a is now a matrix in Rm cross n b is a vector in Rm and what because that need not really be a linear relationship what we will try to do is find an x such that; these residues are the some of these residues is minimized ok.

So, now, the residues can take any sort of signs, so it does not make sense to just simply add them. So, what we; so you need to take a norm of them. So, let us is what I am taking here is the I am taking the 2 norm I am squaring the residues and summing all of this over the m sample. So, I am looking for an x in Rn. So, belongs to Rn remember that minimizes this particular loss function ok.

Another way of writing the same thing is to simply write that this is norm of this vector, this vector Ax minus b the l 2 norm of that vector is what we are have here alright ok. So, what this problem does is it gives you the optimal set of the optimal xs are x s have different names they are called regression coefficients in some language, in some communities they are called in weights by some.

What matter is basically they define for once they x the optimal x has been found once the x star, the optimal x. So, let us call that x star the optimal solution. So, if once the optimal x star is found all I in order to predict the in order to predict the price for an unseen sample, what I need to do for an unseen sample say suppose; if I find a new sample a tilde. What I need to do to predict the price for that is to simply do a tilde transpose x star ok, this is my prediction, this is the idea behind yes

Student: No

No there are there are so different norms have different properties and different shapes. So, I this goes for more and more into data science; I was not prepared to discuss it right now.

But I will give you a general idea the when you look at points like when you get a bunch of data points like this not. Some of them can also be have out what can what you can call outliers. So, a data point like this which really is well out of the trend right. So, the general hypothesis in this sort of work is that be the data is comprises of a trend plus a bit of noise and we want to capture the trend, but not the noise.

But there will occasionally be an outlier which come because noise is random there will be occasionally be an outlier like this. And now what that kind of outliers can do is that it can have an adverse effect on the way on your choice of x star because of the way you are training your algorithm right.

So, on the way you have defined the because of the way you are because of the way the x star has been found you that that it can skew the choice of x star little bit. So, that to the point where now the noise starts paying a bigger role than the trend.

Now what as consequence what you would one way of getting rid of that is. Instead of hand picking what I am saying that well this is an outlier and that is an outlier you probably do not know what is an outlier to begin with. Instead of doing that, you try to look for a different loss matrix itself.

So, instead of the Euclidean norm look for an take a different norm. So, for example, you can take an l p norm. So, now, you can raise all of these to the instead of a power two you can raise all of these to the; to some power some power. So, this is the l p norm there are other there are other types of loss function that penalize large deviations much more than they penalize small deviations ok.

So, that the though you can took where you can choose one of them. So, this goes more and more deep into, what you want? What the purpose of the study is? What the nature of the data is and so on. But this is just a simple illustration that I wanted show you.