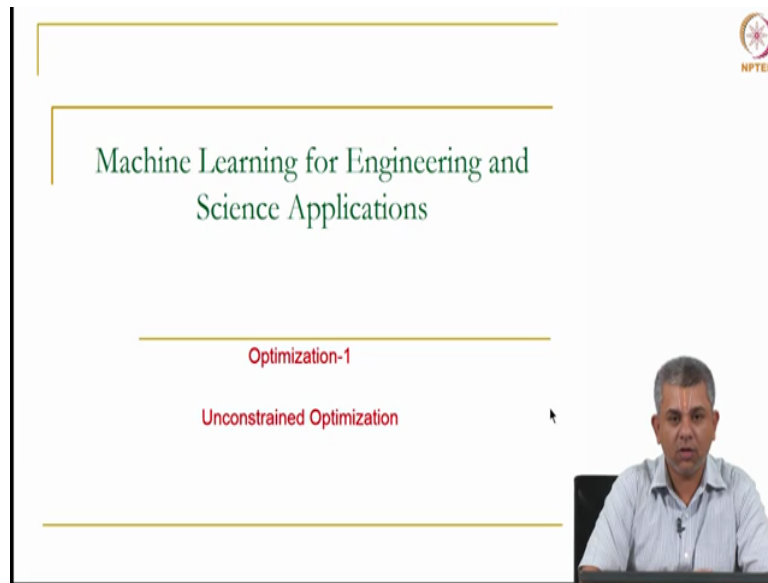


Machine Learning for Engineering and Science Applications
Professor Dr. Balaji Srinivasan
Department of Mechanical Engineering
Indian Institute of Technology, Madras
Optimization – 1 Unconstrained Optimization

(Refer Slide Time: 0:15)



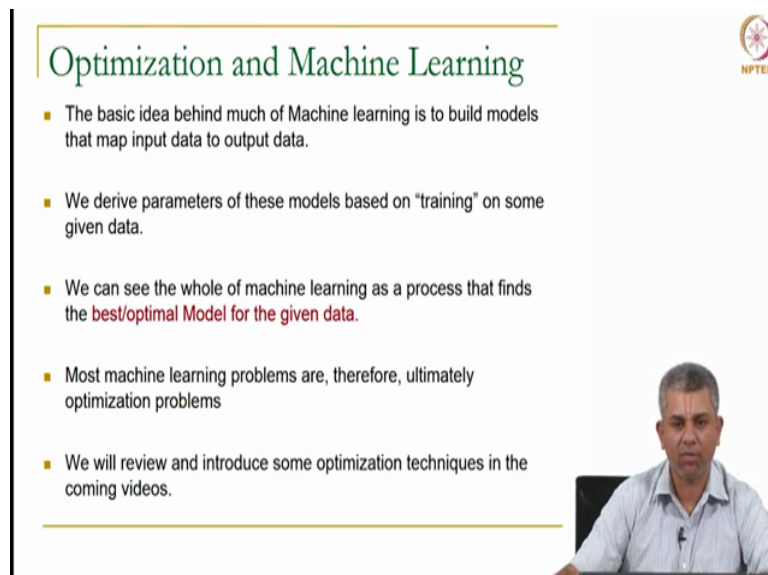
Machine Learning for Engineering and
Science Applications

Optimization-1

Unconstrained Optimization

In this video we will be looking at some beginning of optimization specifically unconstrained optimization.

(Refer Slide Time: 0:21)



Optimization and Machine Learning

- The basic idea behind much of Machine learning is to build models that map input data to output data.
- We derive parameters of these models based on "training" on some given data.
- We can see the whole of machine learning as a process that finds the **best/optimal Model for the given data**.
- Most machine learning problems are, therefore, ultimately optimization problems
- We will review and introduce some optimization techniques in the coming videos.

So the relevance of optimization machine learning is very very high, as we saw in the first week the basic idea behind most of machine learning is that you want to build models so data

models that input they take some input and map it to output data. Now usually (what the) our maps depends on certain parameters and the way we improve our models as you will see in the next week is based on something called training that is you give more and more data and try and improve your parameters.

So usually you would like to know how much does your output change depending on the parameter, so you have some quantity which is a vector quantity that is changing based on some other vector quantity. So most of it most of our machine learning is dependent on finding out the best or optimal model for some given set of data so most machine learning problems can usually be rewritten as optimization problems.

So what we will be doing in the next series of videos is to try and introduce as well as review some of you will be familiar with some of these ideas already so we will try to introduce some optimization techniques in the coming videos.

(Refer Slide Time: 1:38)

Optimization $f: \mathbb{R}^n \rightarrow \mathbb{R}$

- The general optimization task is to maximize or minimize a function $f(x)$ by varying x .
 - The function $f(x)$ is called the **objective function** or **cost function** or **loss function**
 - The function $f(x)$ maybe a **scalar** (single objective) or a vector (multi-objective)
 - In this course (and most of Machine Learning) we deal only with a single objective. That is, $f(x)$ is a scalar.
 - However, x is, in general, a vector.
 - Therefore, $f: \mathbb{R}^n \rightarrow \mathbb{R}$
 - For example, $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2$. Here, $f: \mathbb{R}^3 \rightarrow \mathbb{R}$
- It is possible to reduce all optimization problems to minimization problems.
 - That is, all problems can be written as find x that minimizes $f(x)$
 - Any maximization problem can be written as minimization of $-f(x)$
- We denote the solution to the problem as $x^* = \arg.\min f(x)$

Handwritten notes on slide: $f(x) = x^2 + 1$, $\min f(x) = 1$, $\arg\min f(x) \rightarrow x=0$, $\min f(x) \rightarrow$ Minimal value of f , $\arg\min f(x) \rightarrow$ That x which results in $\min f$, optional minimization

So typically what you try to do in a general optimization task is to maximize or minimize a function once again like in the previous videos this function can be something that takes in a vector and gives out a scalar. So this function the one that you are trying to maximize or minimize is called an objective function or a cost function or a loss function. So this terminology is used interchangeably.

So the function usually in a general optimization task can either be a scalar and this is called a single objective optimization problem or you can have f itself as being a vector. So in that case it could be a multi objective vector and in this course we are going to restrict ourselves

to this case to the single objective optimization problem is of even that is an involved problem.

So we will be only dealing with that and this is actually true of most of practical machine learning anyway we try and define a cost function or an objective function which is a scalar for itself, x in general remember is a vector and typically we are going to deal with the case where f goes from \mathbb{R}^n to \mathbb{R} . So an example of such an f for example could be f of x vector which is 3 dimensional, so x_1 square plus x_2 square plus x_3 square, here f is going from \mathbb{R}^3 to \mathbb{R} .

Now even though the general optimization task is to either maximize or minimize a function, we will typically talk only about minimization because all optimization problems can be called as minimization problems, why? Because if it is a maximization problem you simply minimize minus f of x , so whenever I will be talking in the next few slides as well as in the next video I will only be talking about minimization because maximization is a trivial change by simply changing the sign.

Now here is some notation the optimal solution let me write the right word optimal or the minimal so we will be writing that as x^* this star denotes optimal. Now notice the term $\arg \min$, \min of f of x would simply mean minimal value of f , $\arg \min$ of f of x is that x which results in minimum of f . So just to give you an example if f of x is let us say x square plus 1 then minimum of f is 1, but $\arg \min$ of f is what value of x gave you the value of f equal to 1 this is x equal to 0. So we will be using this notation quite often, $\arg \min$ is that argument or that value of x which gives us minimum of f of x .

(Refer Slide Time: 05:02)

The slide is titled "Optimization – Scalar x" and features a graph of a function $f(x)$ versus x . The graph shows several local extrema: a local maximum, a global maximum, a local minimum, and a global minimum. A red box highlights a local minimum. To the right, a smaller graph shows a local maximum and a saddle point. Handwritten notes in red include $f(x) = x^3$ and "Saddle point". The NPTEL logo is in the top right corner.

- We will look at the **unconstrained problem**. That is, find x that minimizes $f(x)$ with $x \in \mathbb{R}$. That is, no constraints on x .
- It can be shown that any local extremum will have the property $f'(x) = 0$
 - Such points are called **stationary points** or **critical points**.
 - The stationary point may be a (local) minimum, maximum or saddle point

Exercise: Prove this using Taylor Series
- If $f''(x) > 0$, it is a local minimum
- If $f''(x) < 0$, it is a local maximum
- If $f''(x) = 0$, it could be a saddle point
- The absolute lowest/highest level of $f(x)$ is called the global maximum/minimum

<https://www.nptel.org/...>

So here is a quick review of scalar optimization, so as you remember if you have some function f of x versus x it is in general going to be a curve and you are going to have various minima for now we are going to look at the unconstrained problem, unconstrained problem means there are no constraints on x there are no limits on x , we are looking at x belonging to the whole of the real line, we will look at the constrained case in the next video for now in this video we are only looking at the unconstrained problems, so we will assume x can go from minus infinity to plus infinity, okay.

So in such a case you could have some global minimum and global maximum and you could also have a local minimum and local maximum that is locally if I just put a box here all the values around the local minimum are greater than the minimum local minimum but this might not be the global maximum or the global minimum.

Now it can be shown that both these extrema we are not going to show it but both these extrema whether it is local minimum of local maximum all of them will have the property that $f'(x) = 0$ in the unconstrained case. So these points are called stationary points or critical points. So the stationary point as I have just shown could be a local minimum or a local maximum or something called a saddle point.

Now how do we figure out whether it is a local minimum or a local maximum? Typically you look at the second and higher derivatives we will look at just the second derivative case here. So if $f''(x)$ if the second derivative is positive for example here in such a case it is a local minimum. For example if you look at the slope here you will see that the slope of the

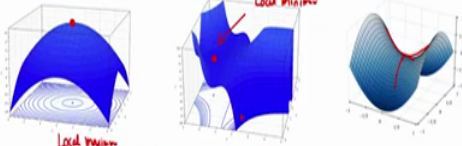

slope okay so as I move away from here the slope increases which means that it is a minimum here.

So here the slope is 0, here the slope is positive so that is why $\nabla^2 f$ is a local minimum as an exercise you can try and prove this, this is an optional exercise for those who are interested you can try and prove this using the Taylor series. Similarly if $f''(x)$ is less than 0 then it is a local maximum once again it has the same idea. Now it can happen that your $f''(x)$ is actually 0, in such a case it is called a saddle point this for example is if you look at $f(x) = x^3$ around $x = 0$ this is precisely what happens.

Now what is happening here? This is like the shape of a horse's saddle as we will see in multiple dimensions also. You will see that in this direction there is an increase, in this direction there is a decrease, it is sort of the combination of this curve and this curve. So from one side it looks like a local minimum and from another side it looks like a local maximum. So this happens when $f''(x)$ is also equal to 0 and in such a case it could be a saddle point. All of you are familiar with the notation of global maximum and minimum this is the absolute maximum or the absolute minimum that you will get over all of space.


(Refer Slide Time: 8:40)

Optimization – Multivariate x

- In this case the **unconstrained optimization problem** is to find x that minimizes $f(x)$ with $x \in \mathbb{R}^n$. That is, there are no constraints on x .
- Since x is now a vector quantity, we need to evaluate the gradient $\frac{\partial f}{\partial x} \equiv \nabla_x f$
 - It can be shown that any local extremum will have the property $\nabla_x f = 0$ → $\frac{\partial f}{\partial x_1} = 0, \frac{\partial f}{\partial x_2} = 0, \dots, \frac{\partial f}{\partial x_n} = 0$
 - Such points are called **stationary points** or **critical points**.
 - The stationary point may be a (local) minimum, maximum or saddle point
- The type of critical point is decided by the nature of the Hessian $H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$
 - If $H_{i,j}$ is **positive definite** it is a local minimum → p.d. → All eigenvalues of $H > 0$
 - If $H_{i,j}$ is **negative definite** it is a local maximum → n.d. → All $\lambda_s < 0$
 - If $H_{i,j}$ is **indefinite** (i.e. neither p.d or n.d) then it is a saddle point

https://en.wikipedia.org/wiki/Matrix_and_eigenvalue#In_Matlab/Parabola.png



So now let us look at the multivariate case. In this case you are trying to find out f of x once again which is x that minimizes f of x , but x now belongs to \mathbb{R}^n instead of simply belonging to \mathbb{R} now it belongs to \mathbb{R}^n , once again we are looking at the unconstrained problem there are no constraints on x . Now as we saw in the derivative and gradient slides since now x is a vector

quantity we will now have to evaluate instead of simply df/dx , you have to now evaluate the gradient of f .

So in analogy to what we saw earlier any local extremum, so for example here this is a local extremum in this case the gradient will be 0, remember this is the 0 vector which means $\frac{\partial f}{\partial x_1}$ will be 0, $\frac{\partial f}{\partial x_2}$ will be 0, so on and so forth if it is an n dimensional vector x $\frac{\partial f}{\partial x_n}$ is actually going to be 0. Once again these are called stationary points or critical points and like in the 1 dimensional case you could have a local minimum, local maximum or a saddle point.

So some examples are given here, this is a local as well as a global maximum, here for example is a local minimum which is not a global minimum because there are values lower than this going on and this is the example of a classic saddle point, in one direction it is a local maximum and in another direction it is a local minimum so that is what a typical saddle point looks like.

Now how you find out whether this is a local minimum, maximum or whether it is a saddle point now depends on the Hessian rather than the simple second derivative remember for vectors the generalization of a second derivative is the Hessian, okay. So as we saw in the previous slides Hessian now is a matrix, now unlike before I cannot simply say Hessian is positive that has no meaning because it is a full matrix.

So when Hessian is positive definite, you might remember this from the linear algebra slides what does positive definite mean? Positive definite means all eigenvalues of H are positive. So this is not even positive semi definite you have to have all values of H being actually positive. So if that is the case and remember since the Hessian was a symmetric matrix we are guaranteed to have real eigenvalues so that you can talk about this meaningfully.

So if the Hessian is positive definite then it is a local minimum, if the Hessian is negative definite which would mean all eigenvalues are less than 0 then it is a (local minimum) sorry local maximum and if the matrix is indefinite, what is meant by indefinite? It is neither positive definite nor negative definite. So some eigenvalues are positive, some maybe negative or some even if they are 0 then it is a saddle point, thank you.