

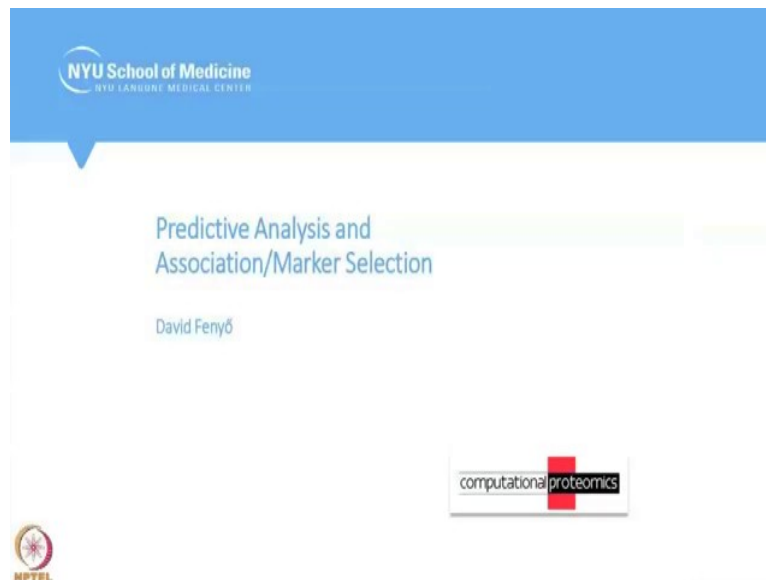
**Introduction to Proteogenomics**  
**Dr. Sanjeeva Srivastava**  
**Dr. David Fenyo**  
**Department of Biosciences and Bioengineering**  
**Indian Institute of Technology, Bombay**  
**New York University**

**Lecture - 46**  
**Predictive Analysis Part I**

Welcome to MOOC course on Introduction to Proteogenomics. Today's lecture by Dr. David Fenyo, will talk about Predictive Analysis. He will provide a detailed information in supervised machine learning, how to link with predictive analysis. Dr. Fenyo will briefly discuss various parameters which are important for how to train a model, how to test a predictive model.

He will also talked to us about how predictive analysis can be used in treatment of cancer, especially in taking decision of treatment strategies. Dr. Fenyo will also talk about image classification and how the image classification could be used for the skin cancer diagnosis. So, let us welcome Dr. David Fenyo for today's lecture.

(Refer Slide Time: 01:14)



Now, we are going to talk about machine learning and specifically about predictive analysis. So, and what that you heard Mani's lecture where he talked about unsupervised and supervised machine learning, so but we are going to talk about today is purely supervised. So


that means, that you need a set of labeled data. So, I mean Mani gave a very short introduction and I will plan to go a little bit deeper and give you more details about this. So, the what I would like to with you learn this morning is first of all how does one train a model.

(Refer Slide Time: 02:07)

**Learning Objectives**

**How to train a model:**  
Gradient descent  
Regularization  
Feature selection  
Selection of machine learning method

**How to test a predictive model:**  
Overfitting and underfitting

 2

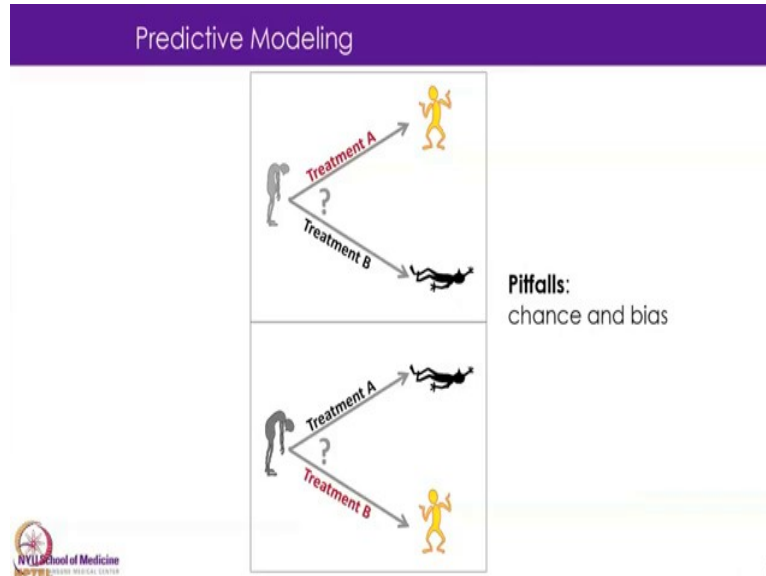
And there we are going to talk about gradient descent which is a way to method a quite general method to find the parameters of the model. Then we are going to talk about regularization which is a method to protect us from over fitting. And then let talk about all these terms in detail. The other thing we talk we will talk about is feature selection. So, one thing that in proteogenomics is that we measure a lot of things.

We measure, let us say tens of thousands of transcripts, maybe 10,000 proteins, maybe 30,000 phosphorylation. So, it is a lot of measurements on different molecules. But most of these will not be relevant to let us say predicting what happens to the tumor. And the; so, what we want to do is focusing in on the important parts and that is what why we do a feature selections. We select out the genes that are important to and it closely related to what we want to predict.

And so, then we will just briefly touch upon that, but people have developed a lot of different methods to machine learning and lot of different approaches on how to do this. And so we will talk a little bit about how to choose the right method for the problem that you want to solve. So, that is another thing, that is quite important. Then there a very important thing is that then after we have trained our model we need to test it. We need to evaluate how good it

is and how well it generalizes. So, that is and there we are going to talk about overfitting and underfitting.

(Refer Slide Time: 04:32)



So, I showed the slides in two days ago and so, this is one example of predictive modeling. So, when this for example, the surgeon cuts out the primary tumor, we can we analyze the primary tumor that is a measure we do RNseq and proteogenomics and then we want to from that measurement a build the predictive model that can tell the oncologist which combination of drugs to give to cure the cancer. And this and this will of course, depends on the both individual and the type of tumor that they have.

So, this just shows in one example in top the treatment A is what we want and, but for the individual in the lower panel we want the treatment B works much better. And this, and because currently as you probably know very well that is not the case. I mean there is some standard of care that is given to everyone and it is only in a few specifications where we can make this decision.

But of course, the hope is that by doing research and produce proteogenomics in the future we will be able to make these kinds of decisions in a more general way. So, what you probably read in the newspapers that machine learning has improved a lot. So, people have been working on machine learning for several decades, but the last few years its really exploded and things work much better than it has in the past.

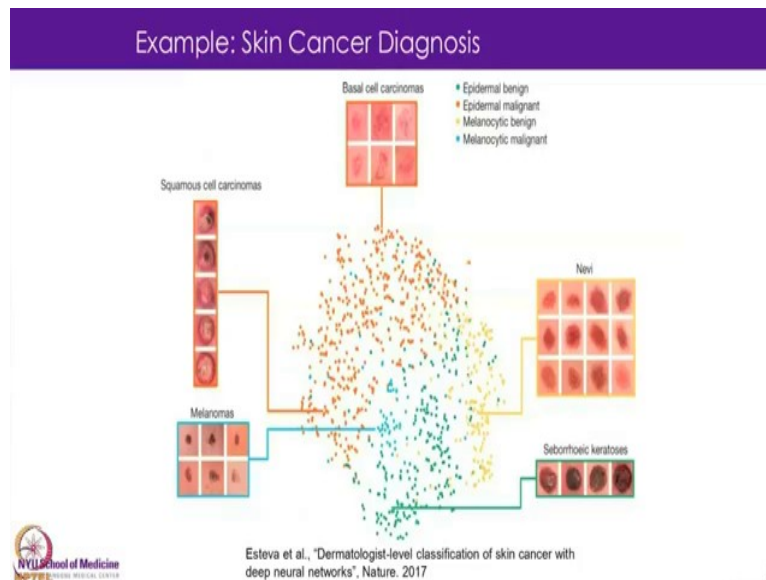
(Refer Slide Time: 06:30)



So, one thing that has been for example, very successful is image classification. So, the like Google and Facebook they have a lot of images, so they have they put large efforts into automatically annotating these images and classifying them. And it is actually amazing how well it works. So, and as you see these are just this is one big data set that is off the news for this training, it is a very large variety of the images, but also how easy it is to see what is in the images. So, that is; but this has really become.

So, for several years they had competitions on who could develop the best algorithm to look at these images. But now they have actually given up on image classification competitions because it works too well, there is not worth doing much more. So, they look at more complicated problems. But this is the general and of course, this we can apply in our field and for example. So, there was a nature publications last year on the skin cancer diagnosis.

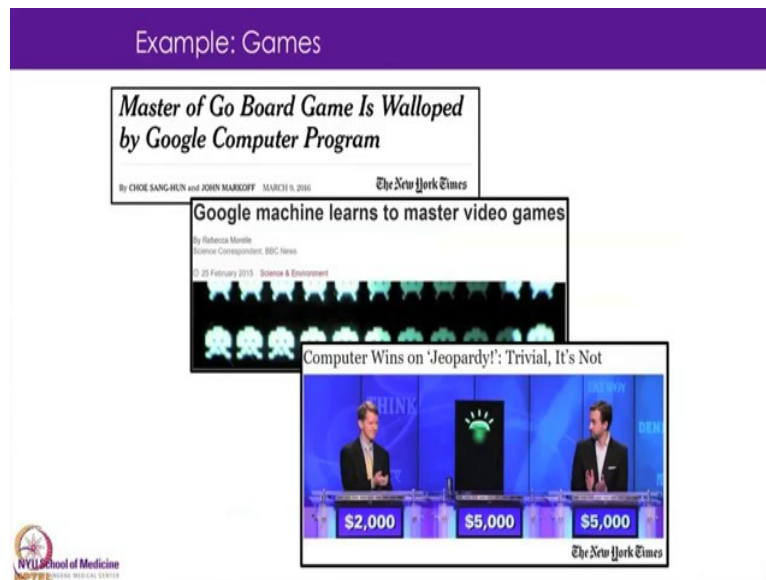
(Refer Slide Time: 07:54)



So, what the authors did was they had cell phone images of mould on people's skin and then they built had pathologist I mean dermatologists look at these images and classify them if they were benign or they were cancerous. And then they built, so they and they collected a lot of images. So, it was I think close to 130,000 and then built the model with that and then they could show that it actually their machine learning model worked better than at least average pathologist.

So, that is and that is quite incredible. I mean you can imagine the implications if you can just you worried about some mould and you take a picture of it on your cell phone and upload it to some web service and then you get an answer back right away what with high accuracy.

(Refer Slide Time: 09:04)



So, another thing that has been very successful is teaching algorithms to play games. So, quite a while ago the chess game machines were became very good at playing chess and then for and became better than any human, but still for a while there if you had the collaboration between an algorithm and a person that was better than any algorithm on its own. But that is not the case anymore. Now, the human does not add anything extra in chess.

So, and then more complicated games like Go has also become an jeopardy. The advantage in games which we do not have in our case is that if you can have different slightly different algorithms play against each other. You can in general you can generate as much training data as you want because in our case we have a certain number of tumors that we analyze and if we analyze more tumors that is more expensive, so that is limited.

But in the game case it is only if you have large computers you can have the algorithms play against each other and learn from these playing. So, if it is in principle generating any amount of training data. So, that is something that people are trying to do in proteogenomics also. But its dangerous because there we have to than if we want to generate more data, we have to model or if you have to build some kind of model how our data behaves and so that of course, then the algorithm will probably mainly learn what we think the data looks like and not really, not anything real.

(Refer Slide Time: 11:19)

Example: Language Translation

## The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

By GIDEON LEWIS-KRAUS | DEC. 14, 2016 | **The New York Times**

NO. 1:

Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai "Ngaje Ngai," the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

NO. 2:

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called "Ngaje Ngai" in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.



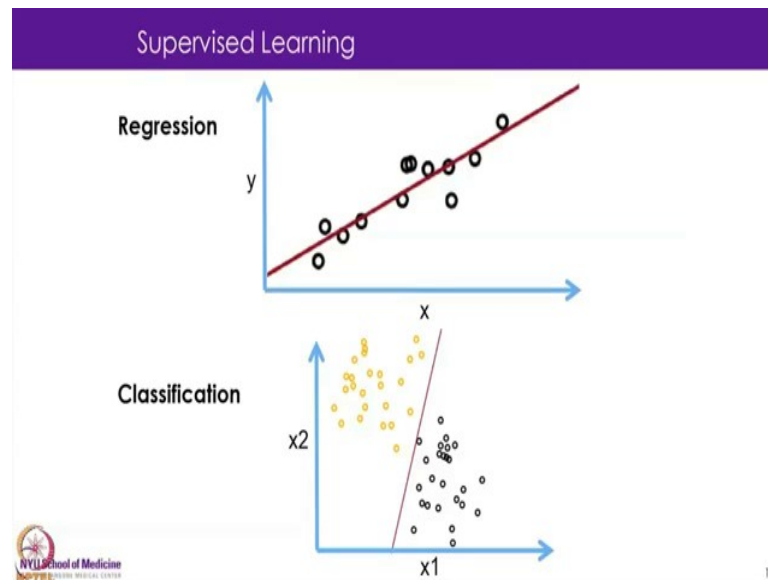
7

Another example that from the general thing is language translation. So, this was an some in the New York times 2 years ago. So, this is a passage from Hemingway's, Snows of Kilimanjaro and so the one of them is the Hemingway original, the other one is was a Japanese translation by a person I mean by an author from English to Japanese and then taking the Japanese translation and translating it back using Google translator. So, which is which.

So, it is mainly the dead body of leopard that is the main and then there are maybe some other nuanced things, but there is only one small grammatical error. So, it is I mean this is quite amazing. So, I am just showing these general examples because as an inspiration that we should do the same for proteogenomics, to be able to do these kinds of things.

But of course, as I said before the advantage in all of these cases both with the image analysis translation and with games is that there is a very large data set that is been labeled. So, that is really what we need. And unfortunately, our datasets are usually limited and not and we would always want them to be larger to be able to achieve things like this.

(Refer Slide Time: 13:14)



So, let us look a little bit more at the details of supervised learning. So, as one I have already as money already mentioned we have two main things, one is regression other one is classification. And so what supervised learning is we build the very, we have a very general model with lots of parameters; the into this model there is no biological knowledge, it is just a very generic and we are going to look a little bit at what is.

So, just a generic model that can pretty much approximate any type of function and then we want to learn the parameters that are that best fit that. So, we are going to look first at a regression and then what the regression is, is that we have some variables that we measure we call them x usually. So, in this case for illustration purposes we are the only show one x axis and then we want to predict what the value y is. So, x would be for example, the level of a transcript that we measure with RNAseq or the level of a phosphorylation site that we measure with mass spectrometry.

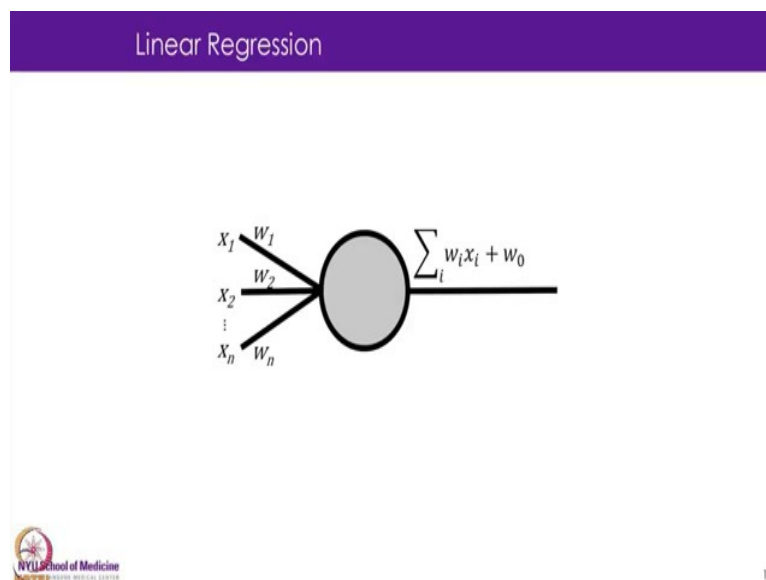
But with all data we have actually many measurements. So, even though I only show one measurement, you should always imagine that there are 10,000 axis or 100,000 axis. It is very difficult to imagine what that happens and also it methods that work on low dimensions become it things behave very different when you go to high dimensions. So, what the regression is that the pretty much try to in this case when we have one x and one y, we try to find the function that describes relationship. It is quite straight forward that way.



And in a classification we try to find the boundary between two classes. And so, in this case there are two measurements or  $x_1$  would be let us say the level of one protein,  $x_2$  the level of another protein and then we have for example, the yellow circles could be that patients that have long survival and the black once the patients that have short survival and we want to find the boundary, so we can classify, when we done the measurements will this we can answer the question will this patient survive for a short or a long time.

And then if we go back to the regression case therapy would the  $y$  could be instead how many months will this patient survive. So, we are going to start with linear regression.

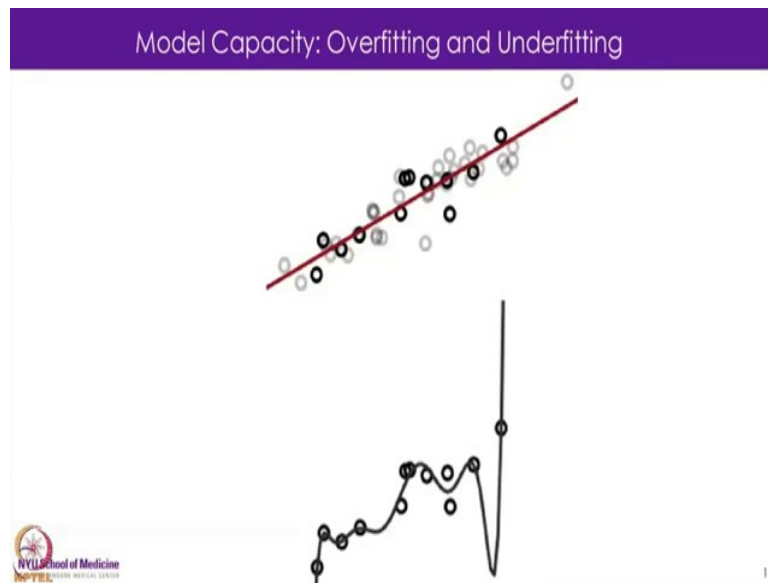
(Refer Slide Time: 16:40)



And so here the axis, so we can have many axis those are different measurements that is we do a quantitative measurements and then for each of them we have a different weights. And so, what output  $y$  is we take each measurement and multiply its weight and sum that up and add a constant. And then, so that you recognize that if we have just one  $x$  as a linear regression case.

And so now one thing to point out that we can also have an orbit arbitrary function of  $x$ , we do not just need we can we do not that limit ourselves to just using  $x$ , but we can for example, have a polynomial as that has an inputs as  $x$ . It is still linear regression because its what linear regression it is linear in the  $w$  in the parameters that is we are learning. So, one thing is that we have limited data, we can build these models arbitrarily complex then. So, what we can, but we have to choose how complex to make them.

(Refer Slide Time: 18:06)



So, in this case if we have these data points, it could be pretty reasonable to draw just have a linear regression, a very, not a very complex function and that could work well. Then, but one could also have a much more complex function and then we would have the lower case.

Student: Can you please give an example of some proteomic data for the linear regression.

So, maybe the output could be we want to predict how many months a person will survive and then we the input would be the several different in proteins the levels of them and then we can use that as the to predict and then what we want to learn all the weights.

Student: So, the natural, the right hand data will get us for how long that person will survive.

Yes.

Student: What all what; so, what will be  $x_1$ ,  $x_2$ ,  $x_3$  and so on.

So,  $x_1$ .

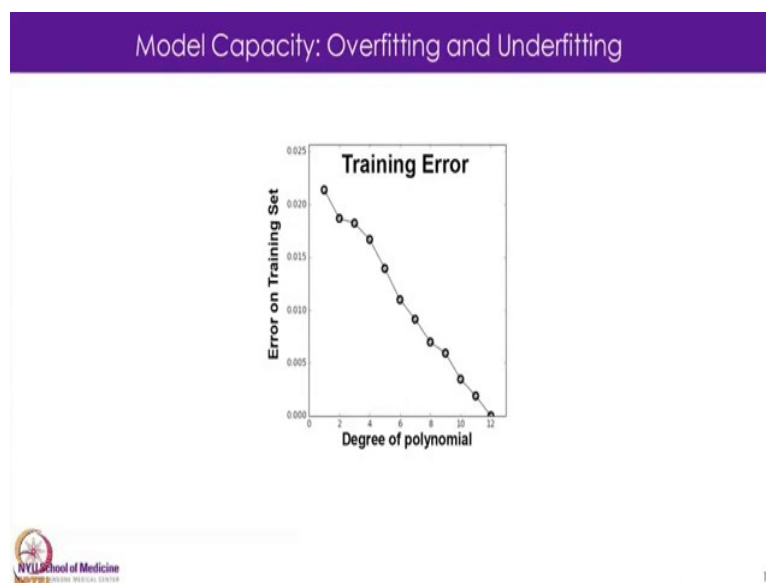
So,  $x_1$  would be one protein,  $x_2$  would be another protein and so on. So, we would and then we are going to talk about some let us say we measure the levels of 10,000 proteins, but that is a lot of parameters. So, we can we probably do not have enough samples to support such a complex model. So, we are going to talk about how to select which proteins are important a little bit later.

So, we have these two cases. So, which one is right correct? Who wants to please? Who thinks the top one is correct? Please raise your hands. The answer is that there is no way to tell, because you if you only have your training data there is no way to tell, I mean yes I agree the top one is more likely that is what we more would guess, but it is just a guess. It is really we do not know.

We need to collect more data and so we need to always train on one dataset and then tests our model on a independent data set. So, let us say that we measure more data and these are now, so the black ones are the same as before. So, we have the gray ones new independent test set that will be now we can say that they trust the linear regression.

But for example, if for some reason this would happen when we do our then we would choose the more complex model, but the main thing is that the its really when you only have your training data sets there is no way to tell how good it is and so that is probably the most important thing from my lecture here today.

(Refer Slide Time: 21:36)



And another way to show this is here now we were on the, this is the same data set that we those 12 points I think it was. If we and we see that then we increase the degree of the polynomial, so we increased the complexity, we can make the error go down in this case to 0, when we have the same degree polynomial as we have number of data points.

So, this is, so one thing about the error that on and you probably familiar with this that we have two of course, both for training and for testing choose a function that we in training minimize and then we in testing we used to evaluate. And for linear regression anyone remembers what we use as the this loss function? It is the sum you take the for each data points the distance to the line and then you square the error and then you take the sum of the square of the errors. You remember that from high school may be. No.

Student: standard deviation!!

No, sum of square errors.

Student: Average of momentum.

Not. I mean usually not when you can take the average it does not matter, but you can just, it is really usually just the sum of the, you take each error for each data points and then you sum them up. It is very, it is simpler than that. It is just taking each error, taking the square and then adding them up. That is the most common, right. It is that is the SD, square deviation.

You can do the mean, but you do not even have to do the mean just the square deviation. So, the sum of the square. You have to sum them, but you do not have to take the mean. So, it is very simple and you will you all know this I am sure and so . So, going back to here, so we know that in a training set if we make our function complex we can have the error go down to 0, but of course this is meaningless because we just have made an overly complex fit to all the sort of noise that is in our training data.

(Refer Slide Time: 24:36)

Model Capacity: Overfitting and Underfitting

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann

WVU School of Medicine  
NOTEL

17

So, because of this a long time ago John Von Neumann said that if you give him four parameters you can fit an elephant to any data and with 5 he can wiggle his trunk. So, meaning; so, what he meant was just this, that if you train on, if you evaluate your model with your training data that is not meaningful and of course, this was a long time ago, so he had much less data.

So, he was worried about 4 parameters, nowadays when people built deep learning model they have hundreds of thousands of parameters and worried much less than John Von Neumann.

(Refer Slide Time: 25:23)

### Points to Ponder

- A predictive model should be build in such a way that it gives a significant outcome without over fitting or under fitting of data.
- The capacity of a model describes how complex a relationship it can model. You could expect a model with higher capacity to be able to model more relationships between more variables than a model with a lower capacity.



MOOC-NPTEL

IIT Bombay

(Refer Slide Time: 25:33)

### Points to Ponder

- Image classification has got a great success and some of the model is very robust which can be used in different diagnosis. E.g.: Skin cancer.
- Regression and classification are both related to prediction, where regression predicts a value from a continuous set, whereas classification predicts the belonging to the class.



MOOC-NPTEL

IIT Bombay

I hope today you learned how supervised machine learning, and regression, and classification plays a role in predictive analysis. Dr. Fenyo also showed how predictive analysis could help a risk in cancer diagnosis and found to be superior when compared to pathology based diagnosis. We also learned that how over-fitting or under-fitting plays an important role in model capacity.

Finally, we understood that the capacity of a model describes how complex a relationship it can model. You could expect a model with higher capacity to be able to model more

relationships, between more variables than a model with a lower capacity. In the next lecture, Dr. David Fenyo will talk more about predictive analysis, giving more emphasis on training a model and testing a model.

Thank you.