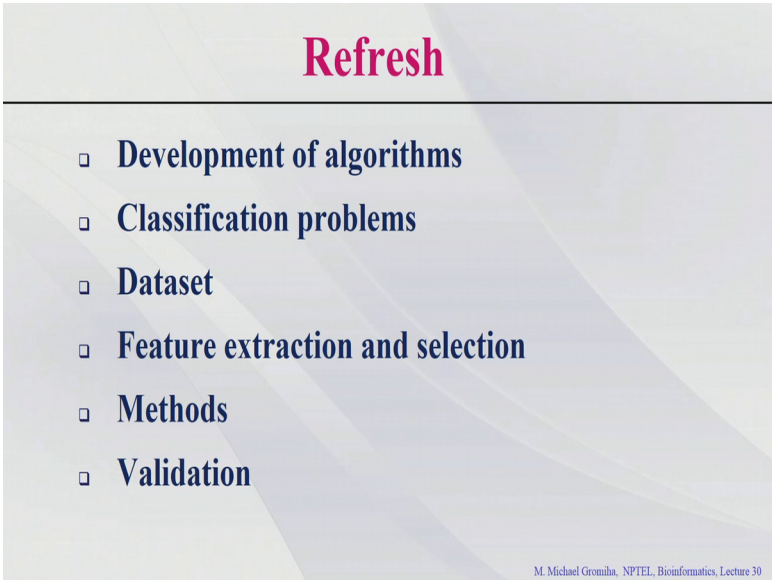
Bioinformatics
Prof. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 30a
Applications of bioinformatics I

In this lecture, we will discuss about the applications using machine learning techniques. For example, last class, we discussed about various problems right for example.

(Refer Slide Time: 00:30)



How to classify different types of proteins, what are the different classification problems?

Student: Classification real.

Right DNA binding or not or the transmembrane proteins or not right, transmembrane helical proteins or beta-barrel proteins or, driver mutation or passenger mutations, also we have image processing we will discussed whether this can be a malignant or it is a benign right. So, we discussed various aspects in the classification problems. Then again also we discussed about the other real value problems. So, for example, if you take solvent accessibility, either you can classify is buried or exposed or you can predict the exact values there is 15.5 angstrom square so on.

So, then we discussed about the various aspects, we have one has to consider for developing any algorithm right what the various important aspects we have to consider?

Student: (Refer Time: 01:21).

First point is dataset. So, dataset is unbiased non-redundant right and reliable dataset right with the different various range of this values. Once we have the dataset then we can derive the features. So, we have to be cautious and it should be aware that the features what you extract, that should be applicable to what the; your problem right. So, if you for example, if you take about the DNA binding proteins, so you should know there are some bias in the positive charge residues.

Likewise you have to choose the features right there should be applicable to your problem. Then once we select the features, then we need to extract and as well as the choose; what are the important features. For example, a few features are very closely related for example, the correlation is 0.95, then we introduce a bias because we are using the same feature 2 times. So, we need to select the features which are important as well as which can explain or which can discriminate or distinguish different types of proteins.

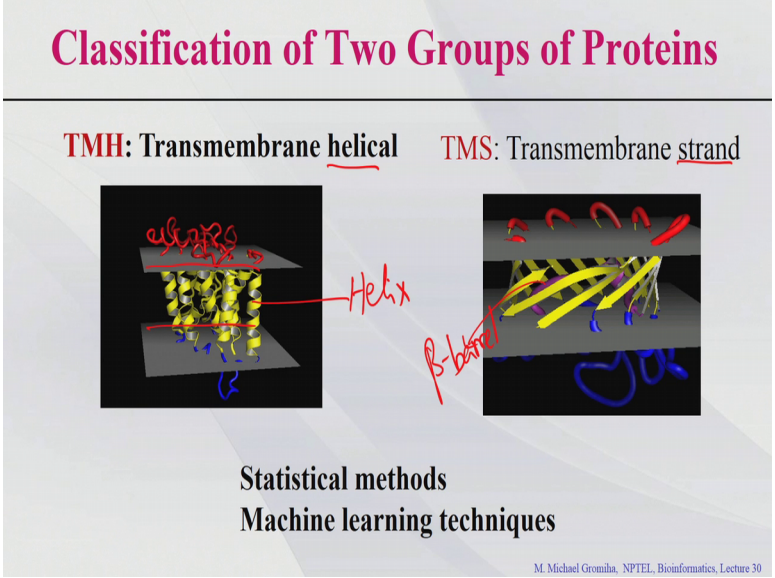
Then once we select the features or when we extract the features, then we use the different methods like the statistical methods or machine learning techniques right to discriminate or to classify the different types of proteins or different applications. Once we make the method then we need to assess we use various measures to assess the performance what are various measures?

Student: Accuracy.

Sensitivity, specificity, accuracy, ROC right precision, recall. So, we used various measures to assess the performance of the method. We have to be sure that accuracy is only that not only the measure right we need to check up or the sensitivity as well as specificity, to see whether the method is performing well or not. Then the various validation procedures like we consider back check or the self consistency or n-fold class validation or Jack-knife or split sampling and so on. So, we can validated all these things. Then once if we are satisfied with your method and the performance, then you can develop your web server right for the applications. So, this can be use for others.

So, among all these classifications today I explain one of the applications right mainly with the statistical methods as well as machine learning techniques, to classify 2 groups of proteins right. This will a give you more details what are the various aspects a we have to consider for developing any algorithm right on any applications.

(Refer Slide Time: 03:47)



Here I mention 2 different types of proteins, like one is the transmembrane helical proteins, and the transmembrane strand proteins. We already discussed in the previous classes see about the transmembrane proteins right; what are transmembrane proteins?

Student: (Refer Time: 04:03).

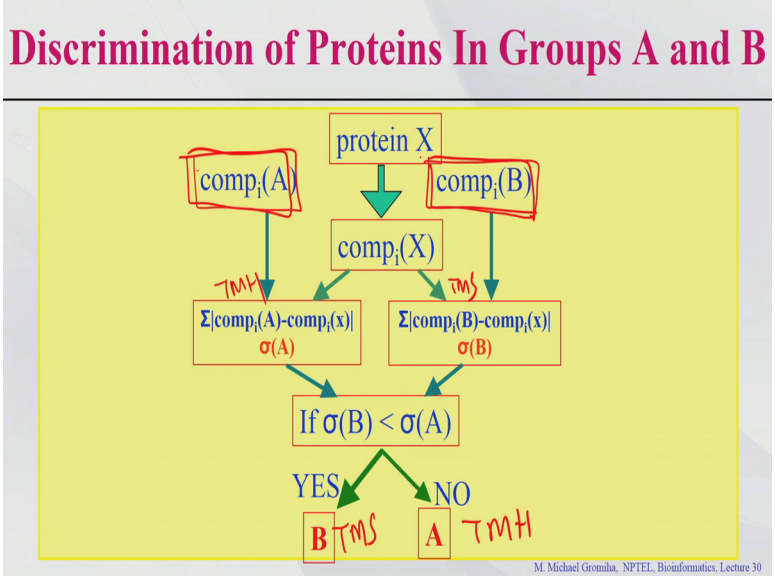
Right proteins which are embedded in the membranes right the biological membrane. So, we can see this is the membrane part right, this is embedded into membranes. So, here you have discussed 2 different types of membrane proteins right this one you see the inside the membrane, this is helical right. So, in this type of proteins we call as transmembrane helical proteins. We second type here if you see this is the protein which contains the beta barrel right. So, this proteins we call as transmembrane beta barrel proteins or transmembrane strand proteins right. So, abbreviation I used TMH and TMS right. If you have a pool of proteins, can we identify the proteins like which belong to the helical ones or which belong to the strand ones?

So, there are various methods to do this I will explain 2 different methods, mainly the statistical methods and the machine learning techniques how to do that right, what do we first step we have to do?

Student: Dataset.

Dataset right first we have to agree a dataset. So, if we have a datasets then maybe you can derive the features from the data set right. So, here first I explained about the method what I already discussed earlier. So, when we have dataset this is your standard dataset A and B right you can derive the features.

(Refer Slide Time: 05:16)



For example, here I show the amino acid composition right. So, if we calculate this amino acid composition, let us continue take that to this node. Then for any protein x right if I calculate the composition compare with the class A for example, TMH are compare the class B this is TMS, and see the deviation if the deviation is less in the case of B, this is the transmembrane strand proteins very less in A then we say TMH protein right that is fine ok.

So, for getting this information right to get the standard values, we need to have a standard set up data right. So, in this case we have to get a data set. So, we can have various a resources to get the data, we discussed about various datasets various what are different databases we discussed?

Student: PDB.

PDB is for.

Student: Structure.

Protein structure for the DNA sequences.

Student: Genbank.

Gen bank or (Refer Time: 06:21) protein sequences.

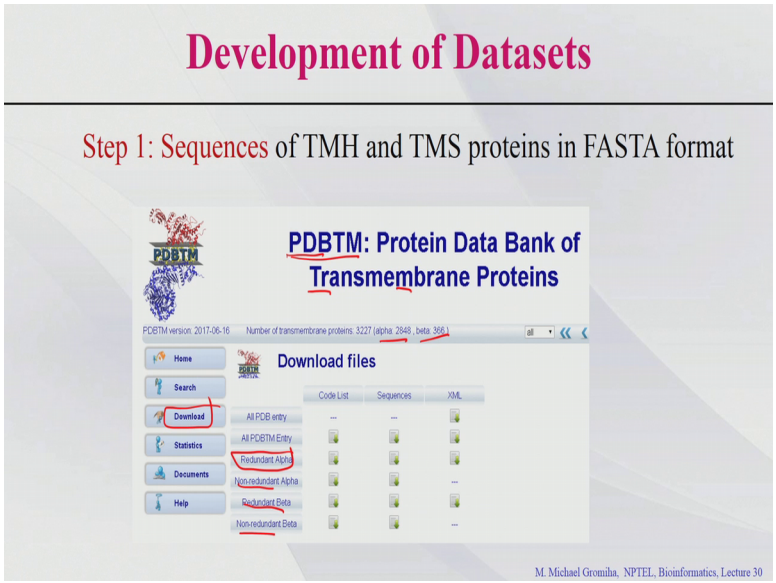
Student: UniProt.

That uniprot right like a for a stability data.

Student: Protherm.

Protherm. So, I have discussed various data bases, here our aim is to classify transmembrane helical and strand proteins right. You can get the data from protein data bank also there are several databases, which are specifically mentioned for the transmembrane proteins right TM topo or PDBTM and so on.

(Refer Slide Time: 06:47)



So, here t m PDBTM it is you are updating quite frequently given there is a recent update right. So, we use the PDBTM, this is the protein data bank the same PDB protein data

bank of transmembrane proteins right this is how the abbreviate the PDBTM. So, this is a data base you can see different types of transmembrane proteins either helical type or the beta barrel type, also it has various options to extract the data based on the number of helices, number of strands right also this we have developed your program to identify the regions where the residues which located in the membrane or extracellular space right and so on.

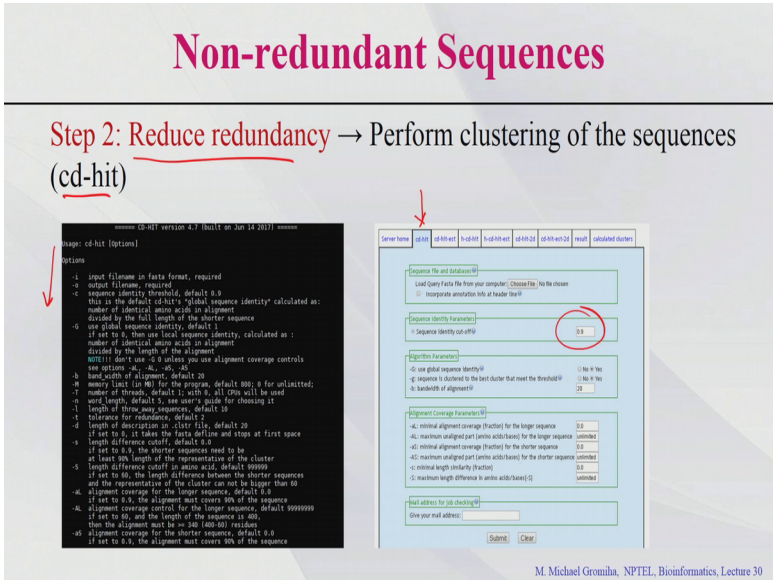
So, you also have the option to download the data based on the non redundant datasets for the sequences as well as the structures. So, currently if we have transmembrane proteins right this is the about 3000 alpha as well as about 300 beta barrel proteins. So, when they go the download option right. So, there you can get the redundant alpha and the non redundant alpha right also they redundant beta and non redundant beta right.

So, first we take a the all the redundant ones, if we get the redundant data then we need to get the non redundant ones right or you can get the non redundant, but here they use a specific cut off, either 30 percent or 40 percent right. Sometimes we like to have a specific the cut off as per our desire. In this case we need to download the redundant data; that means, you will get all the data once we get the all the data then what we have to do?

Student: redundancy (Refer Time: 08:27).

You have to take a non redundant dataset right. Now if you go to PDBTM we will get the all the sequences of transmembrane helical proteins and TMS proteins right that is fine dataset is done. Then we have to remove the redundancy right that is redundancy we will discuss about various methods.

(Refer Slide Time: 08:41)



What are the various methods we discuss earlier to reduce redundancy?

Student: (Refer Time: 08:51).

Clustering you can see cd hit, blastclust.

Student: blastclust.

On the poises.

Student: (Refer Time: 08:56).

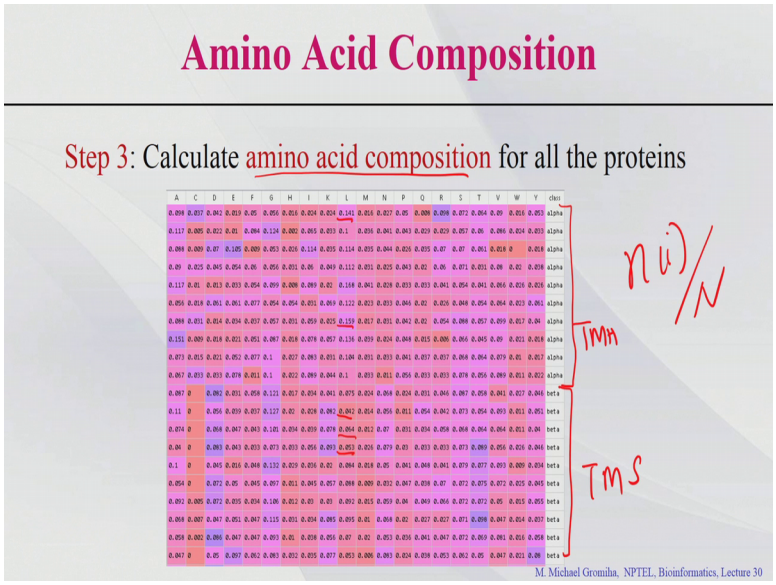
Various methods we discussed right. So, if we say cd-hit right they will have the lot of options also we discussed about the comment what we have to give right, the cd hit to get the non redundant sequences, we can specify the (Refer Time: 09:05) size you can specify the redundancy right or say when we (Refer Time: 09:09) input files and output files you can get that. So, these are the various options available in the cd hit right also we can see the net/online also we can reduce the redundancy using cd hit. So, you will ask for the sequence identity cut off right the word size. So, then you can get the non redundant sequences.

So, for example, if you want to 30 percent redundancy, you can use the 0.3. So, you can get the non redundant sequences. First we get the sequences for this case we use PDBTM and once we download the sequences then reduce redundancy that at any cut off. For

example, if 30 percent cut off we can use cd hit to get the sequences with 30 percent cut off. So, now, we need to think about the features; what the various features we can derive from the sequences and which features will be meaningful and applicable to this type of proteins right. One of the features you can think about that is hydrophobicity because if you see the transmembrane helical proteins, the transmembrane regions are enriched with the hydrophobic residues. So, you can expect that these residues are highly hydrophobic nature.

Then this hydrophobicity you can see in the composition right because if you see the transmembrane region, this regions are enriched with leucine, alanine, valine right; as well as the aromatic residues. So, composition could be also be another important feature to classify these 2 types of proteins right. So, let us see how to get the composition how to calculate the composition?

(Refer Slide Time: 10:39)

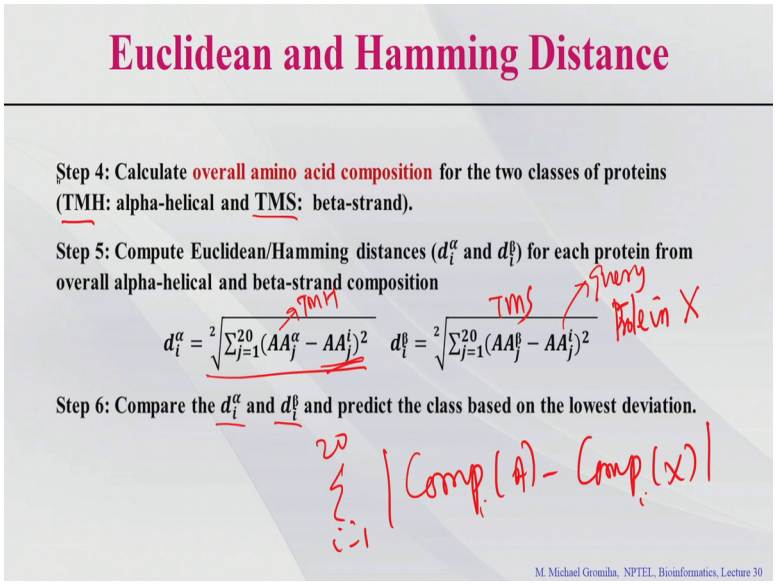


Student: (Refer Time: 10:43).

Right take the n of 'I' for 20 residues divided by N, N is the total number of residues in a protein right. So, here I show the data for the different amino acid residues, you can see this is the different 20 amino acid residues we have the values. Here if you see this is for the data for the; are alpha the TMH proteins right, here we have data for the TMS proteins. If you compare the values we can see some residues are dominant in the alpha proteins and some of them in the TMS proteins.

For example if you see the leucine this is the 0.14 or 0.15 right, but in the case of beta you see this is even sometimes this is very less 0.04, 0.05, 0.06 and so on. Like there is some other residue other residues right for example, if it is the valine or tryptophan you can see the difference right we discussed earlier right about the composition were the different proteins. So, you can take the composition, one could be an important feature to classify theses 2 type of proteins right. I will show the results right what is a prediction accuracy or performance if you use the composition as the feature. First we talk about the statistical methods right. So, now, we can get the overall composition right take for example, if you have the 500 proteins or 1000 proteins, take all the proteins and you can take the overall composition right for the in the case of the helical proteins as well as for the a TMS proteins, fine.

(Refer Slide Time: 12:04)



Once you have the composition, you now take the new protein. Protein X. We take the protein X, then you can calculate the deviation right you can see either the Euclidean distance or you can do the hamming distance right. So, what is hamming distance how to calculate the hamming distance?

Student: The difference between the composition of.

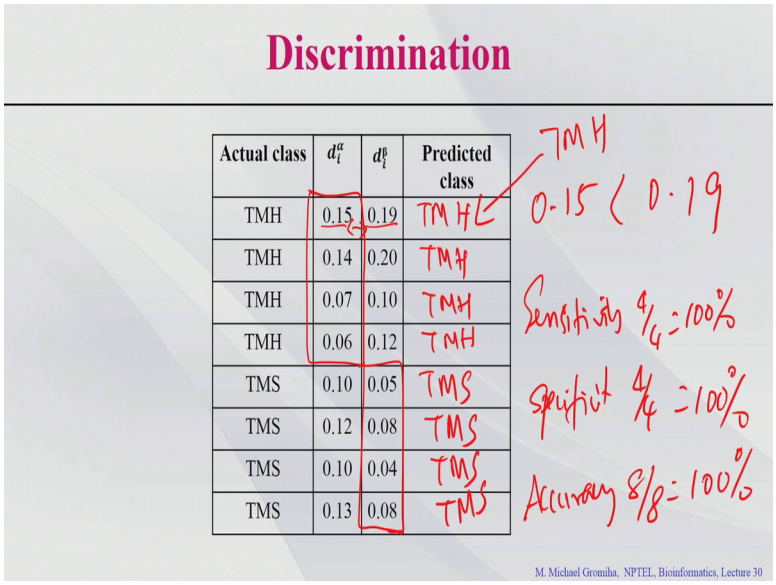
Difference right.

Student: (Refer Time: 12:42).

Composition of A minus composition of X right. So, you can see this is the sigma i equal to 1 to 20 like if you take the set of proteins. The composition of the whole proteins A and for any X right i equal to 1 to 20 right you can take this i. So, you can do at 20 residues then take the absolute values of the difference, you can get that. But in the case of the Euclidean distance you can use this equation right square root of i equal to 1 to 20 right. For the case of the all of all TMH proteins right. This is TMH right this is for any protein. For your query protein.

Likewise you can do for the TMS and your query protein here. This is your query protein right get the difference square root of the difference and sum up and take the square root then you will get the d alpha i and d beta i. When you compare these two d alpha i and d beta i right either you get obtain from Euclidean distance or you obtain from the hamming distance then it tell you whether your protein is of TMS or the TMH

(Refer Slide Time: 13:52)



For example, if this is the value you can see the actual class. So, alpha it is 15 this beta it is 0.19 it is 0.15, 0.19. So, what is the predicted class?

Student: Alpha.

TMH right this is TMH second one.

Student: TMH.

TMH third one.

Student: TMH.

TMH forth one.

Student: TMH.

TMH, this one.

Student: TMS.

TMS, TMS, TMS, TMS now because this is less here, here this is less right compare these 2 numbers and see which one is less right. Then you get the predicted class because 0.15, this is less than 0.19. So, this is TMH. So, we put the as it is in the class TMH. So, if you do this. So, you can see this is perfectly matching. So, we can see the all the examples right data examples you can get the full results right you can calculate you assess with the sensitivity, specificity and accuracy right. So, what is the case sensitivity here?

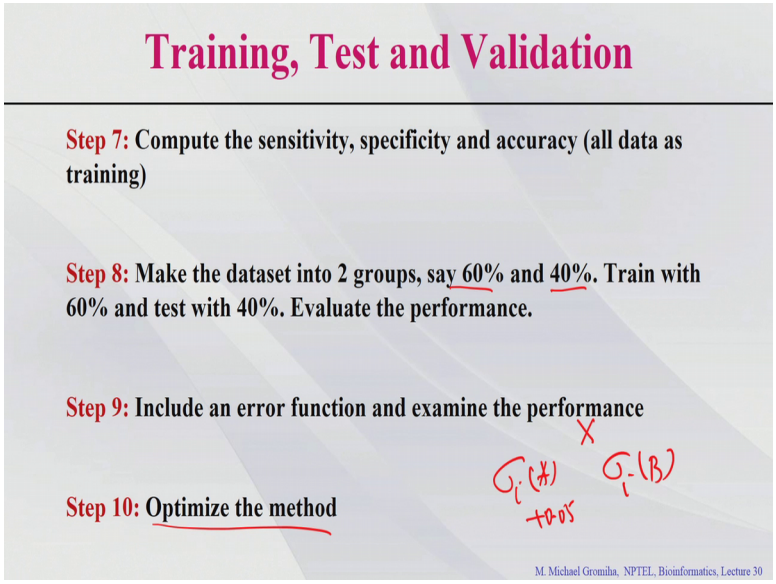
Student: 100 percent.

100 percent right. So, this is 4 by 4 this is equal to 4 by 4 this equal to 100 percent specificity is equal to100 percent accuracy.

Student: 100 percent.

8 by 8 this is equal to 100 percent right fine. So, this is fine. So, you can do this.

(Refer Slide Time: 15:27)



So, here we use 8 data see 8 data right likewise you can use all the dataset training, and you can evaluate the performance using training. Then you make it as 2 groups first you say 60 percent. 60 percent we use to get the composition for the TMH and the TMS, then remaining 40 percent we use to test right use the same values. So, we can a get the 40 percent of value data, each protein this 40 percent you can test and then see whether these proteins are also correctly identified as TMS or the TMH then you can evaluate the performance.

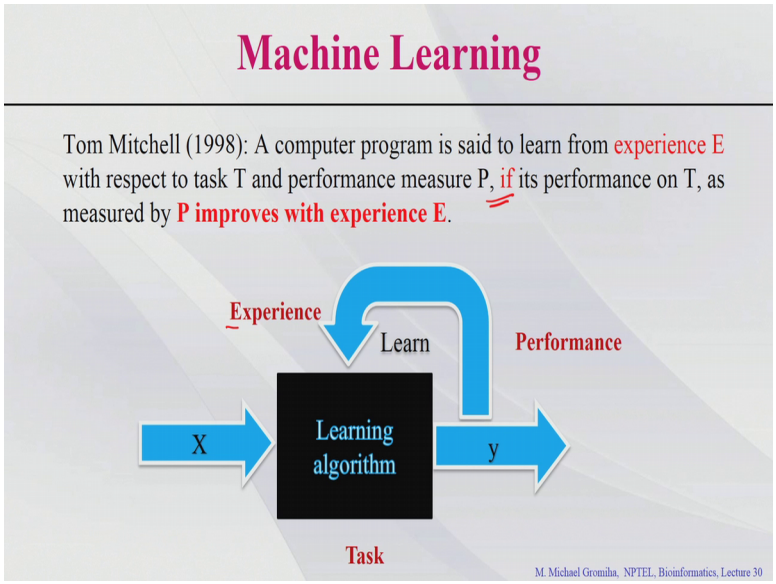
Then sometimes if you have a minor deviations for example, 0.1 or 0.05, see for small deviations the performance will be less. In this case you can add another function instead of you can compare the sigma i of A and sigma i of B right with any X right you can add a error function for example, this is a 0.05. So, then a increases you change a error function and you can optimize your method show with the highest performance right you can do that, in this case you can get the method with the better performance then just using the a difference between the sigma A and sigma B with the protein X right then you can do that.

So, this is statistical method you can do with that right this is we have to deliver features and we have to use your features by yourself, and you have to assess the performance, and the error function also you have to choose and you have to tune the parameters and optimize the method. So, likewise this is another type of techniques that is machine

learning. So, in this case you do not have to worry about all these things. If you collect the set of features the machine learning will learn your input data as well with respect to the output, and this will choose the important features and also it will train in such a way that your input will be mapped with output with the highest performance.

Last class I will show a discuss a little bit about the machine learning techniques. This class I will go for little bit more details or for additional details you can take a class even in the computer science courses.

(Refer Slide Time: 17:51)



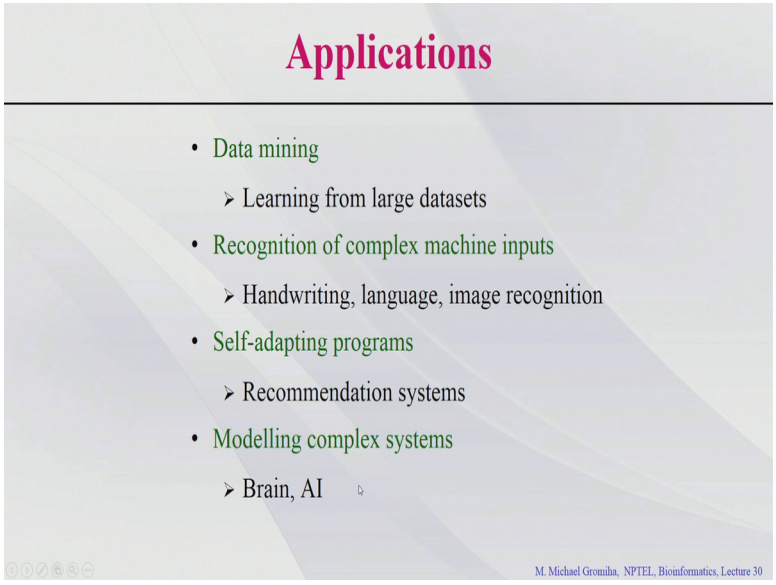
What is X here? X is your input. So, y is your output, here it is learning algorithm. Your task is you have to discriminate these type of proteins in X right into y whether it is TMS or TMH right. So, you how do you know whether this algorithm your program learn from this experience? So, if you said to learn from the experience right this experience you can see E right for any task what is our task here? Classifying the transmembrane helical and transmembrane beta barrel proteins right.

So, with respect to a performance; performance what are the various performance we discussed sensitivity, specificity, accuracy and so on right. If we the (Refer Time: 18:34) learns only if it is a performance right on this task right measured by the any P right it improves with the experience E. If it gets a information right based on the experience the learning algorithm, if it increase the performance. Then we say that the program learns right from the experience E with respect to the our task T that is classifying

transmembrane helical and strand proteins, with respect to the improvement in the performance P for example, accuracy right.

That means it learns only if the performance increases for our own task. So, what are various applications in machine learning? So, machine learning is wide range of applications not only in biology, in all aspects of in the system right. So, lot of applications.

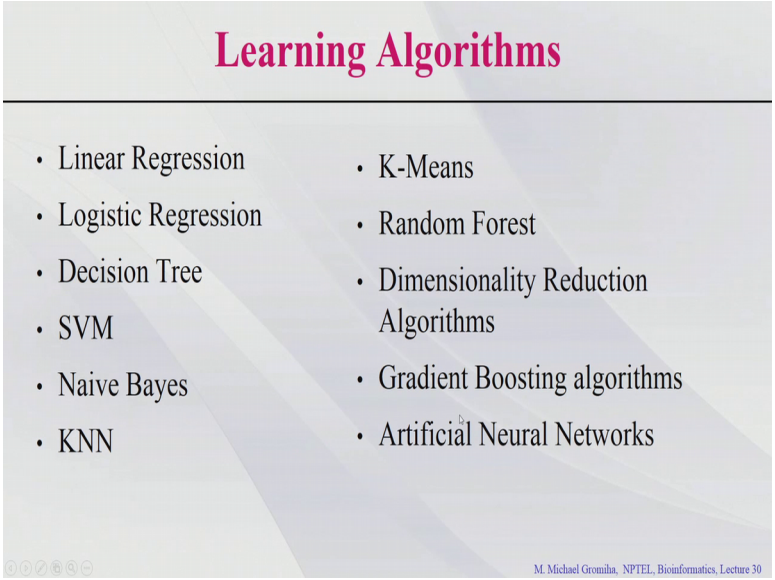
(Refer Slide Time: 19:22)



For example in the data mining, you can learning from very large datasets right maybe you can take in physics or the chemistry or the engineering sites whatever. So, it learns from very large dataset and you can derive the features to get what the performance we required.

For example if you see the recognition of complex machine inputs, like handwriting, language and image recognition right. It can read these images and then you can correctly tell that this is what this mean right different handwriting languages. Then we also in the self adapting programs recommendation systems as well as modelling complex systems right for example, see you see brains. So, you have you can see the several neurons right you can use for example, the artificial intelligence right you can model this complex systems right using a different a machine learning techniques.

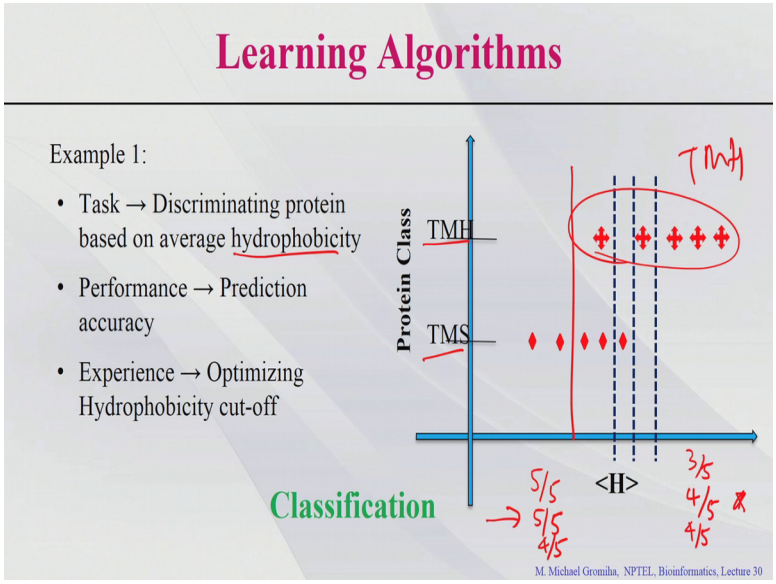
(Refer Slide Time: 20:13)



Then what are the various learning algorithms right there are various learning algorithms right for his machine learning. For example, linear regression or logistic regression , decision tree this will classify this data, if A is more than B. So, it took can be this one then compare with the D. So, if this A is less than D then it could be the group another group; likewise it compare various features then finally, this will make a tree right and finally, you classify these are the conditions are met then your protein right this is the kind of transmembrane helical proteins right they can do that.

SVM a support recommendations right this I will explain soon, and Naïve Bayes and k nearest neighbours, k means clustering, random forest right, artificial neural networks and so on. So, various machine learning algorithms are available in the literature and each algorithm right as their own advantage and disadvantages depending upon the problem we choose, as well as the input parameters right what you use as well as you and your a final output what you require right.

(Refer Slide Time: 21:20)



So, the depends on your problem we can choose different types of algorithms. So, here I give few of the few examples now the first task is. So, we have to discriminate the proteins right with the any feature. For example, here I used a very simple feature that is hydrophobicity right. This is the only one value right using this one value, how can we discriminate 2 types of protein right. Earlier we discussed in the case of TMH and TMS right helical proteins are enriched with the hydrophobicity.

So, let see how to use this hydrophobicity whether we can classify or not. So, we calculated the different proteins here I show an example with 10 proteins. This 5 they belong to TMS and this 5 belong to TMH.

How to discriminate right for example, we can assign the H values right you can set any threshold, for the threshold you can see whether we can do better performance or not right put a line here. So, how about the prediction performance of TMS?

Student: (Refer Time: 22:24).

5 and 5 right all the 5 you can predict, because left side you see the TMS and out right side you can see this is the TMH, here TMS is 5 by 5, for the TMH.

Student: (Refer Time: 22:41).

3 by 5 right then we can try we can set different threshold for example, you put this threshold, now what will happen?

Student: (Refer Time: 22:50) 5 by 5.

This 5 by 5.

Student: 4 by 5.

4 by 5 now which is better now?

Student: New one (Refer Time : 22:56).

This is better right second one is better right. Then we try another threshold here now what will happen?

Student: 4 by 5, 4 by 5.

4 by 5.

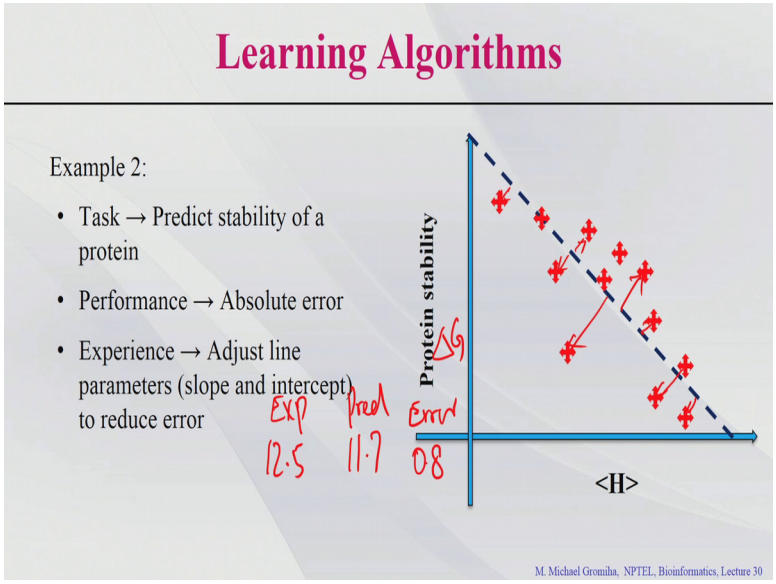
Student: 4 by 5.

4 by 5. So, then you would be 3 threshold which one is the best.

Student: (Refer Time : 23:10).

Second one is the best right. So, we can use your different thresholds and you can optimize, you see the performance and which one is the best to attain the best performance right. So, you are doing by (Refer Time: 23:22) we are doing. So, in the machine learning it will take all this efforts and it will tell you this is the best choice right you can go with the highest performance. So, here the performance is a prediction accuracy and experience is optimizing the hydrophobicity, when they learn the data. So, it will finally, optimize this is the best performance you can get right. If you put one more line here, again this because the performance is vast now in this case you can optimize this is the optimize one you can go ahead with that.

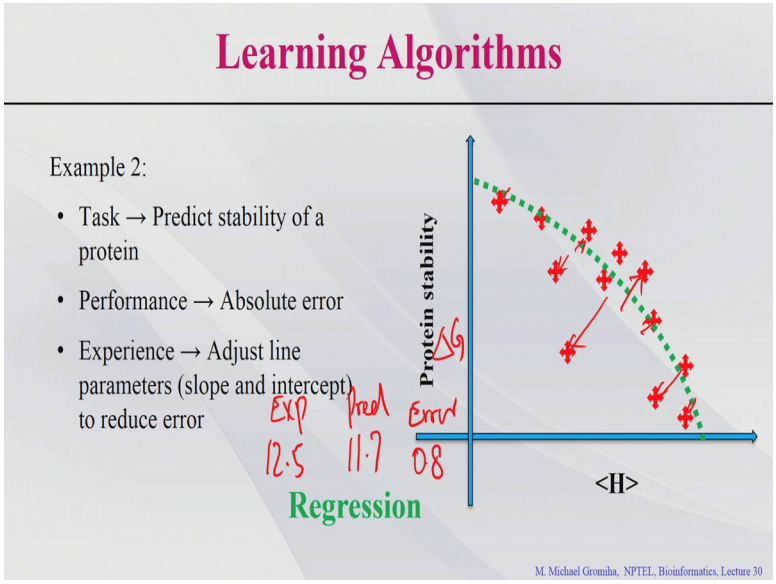
(Refer Slide Time: 23:54)



So, now, we can use a this in classification problem. Then if you go with the real values for example, if you talk about the protein stability how far the value changes with H with delta G right with this H hydrophobicity this delta G we get different points. So, now, if the task is. So, predict the stability of a protein. So, we if H is known whether you can predict the stability right. So, and how to optimize the method? So, how to perform check the performance. Absolute error for example, if the actual value is 12.5 and the predicted one is 11.7 what is the error? 0.8 right this is the experimental this is predicted this is error right fine.

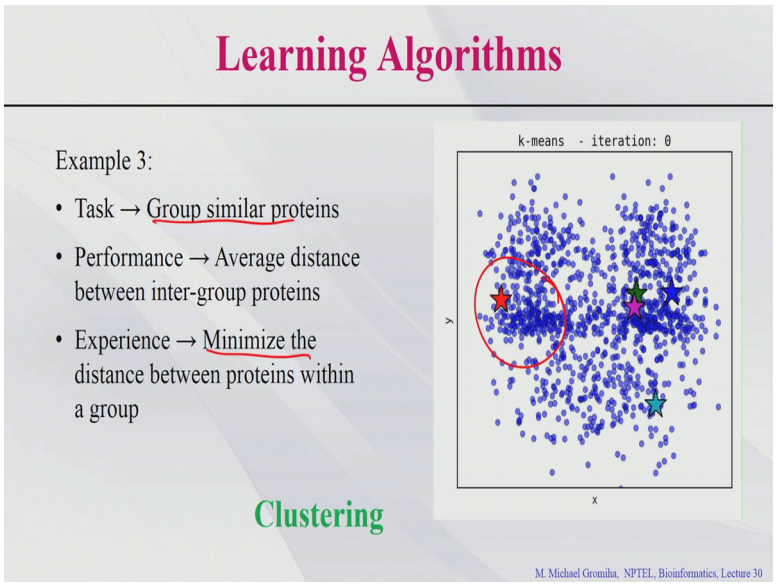
So, and how to reduce the error? For example, if I put a line here right here what is the average error. So, from here this here is a difference, here is the 0, this difference this difference here this is the difference see all the difference from this line right take the average you will get the number is it possible to make another line or different other functions for example, like this right.

(Refer Slide Time: 25:10)



So, the machine learning it will trying this data and make this line in such a way that the MAE should be less. It is a mean of all this errors should be less right this is how it works in the case of the absolute values.

(Refer Slide Time: 25:28)

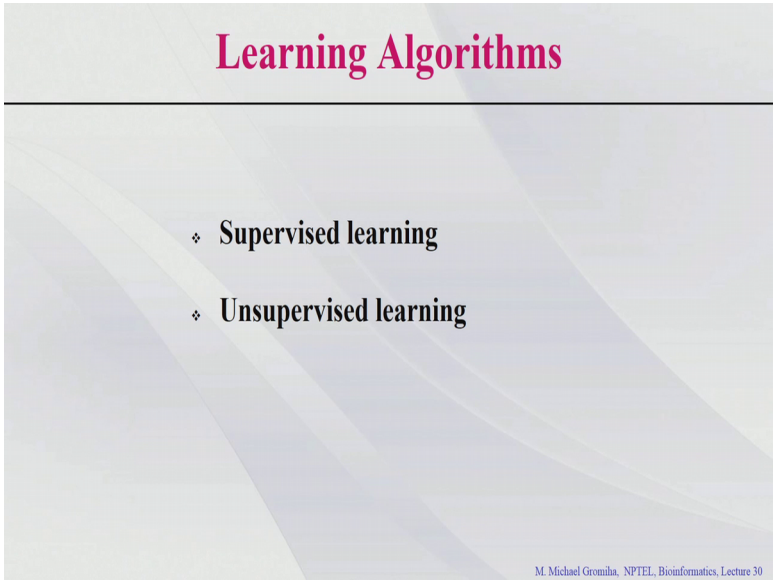


So, this is example three. So, if you take the similar type its many proteins, the task is to group similar proteins right. You can see there are similar types of proteins right how to classify the similar types of proteins. So, how to assess the performance? You can see the average distance right between the inter group proteins and this that should be minimum.

So, if you see this one. So, you can see this group is the best fit because within their group they difference is less now, if you go to other group the difference is more.

In this case this is the group we can fit with the optimally less distance. So, you can minimize the distance between the proteins for example, if this is the cluster right the proteins within this cluster, they have less difference right they have a similar perform values.

(Refer Slide Time: 26:17)



So, the 2 different types of learning algorithms right we discussed different types of learning algorithms, one is supervised learning and unsupervised learning. What is the difference between supervised learning and unsupervised learning?

Student: (Refer Time: 26:29) previous data is known then it is.

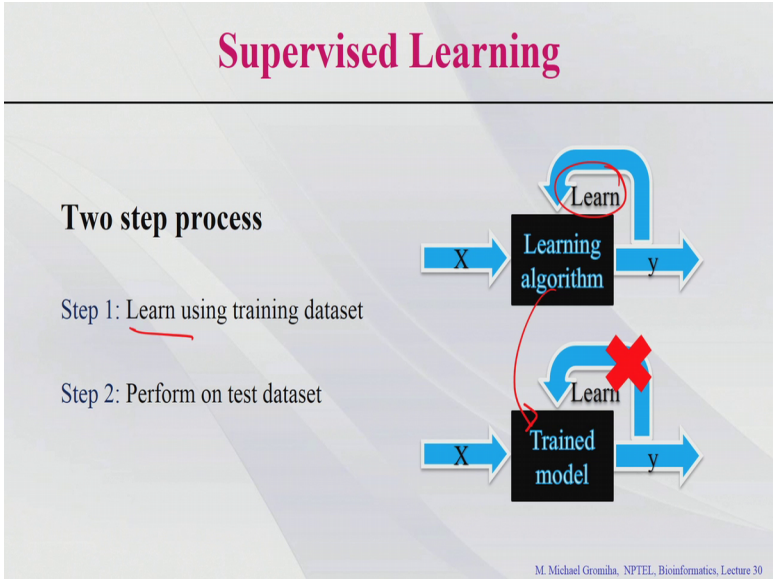
Right if you label the data. So, if your result is known, then we can learn as a supervised learning say this will have the experience right with the known output. So, you can use the label data. So, we can try in your model try to develop your model, in the unsupervised learning.

Student: No information.

Right you can (Refer Time: 26:54) know the labeled ones. So, you can learn by yourself right and then see this group should be the probably the similar group right like for

example, the clustering right it will do by themselves. Supervised learning it will be guided by the labels right this will tell you this is the good way to do that, but this not good way to do that because we know the data. So, you can see the difference. So, then we can optimize the methods right with respect to the levels. And I will explain some of these methods; supervised learning it is a 2 step process in the first case you have to learn using any training dataset right.

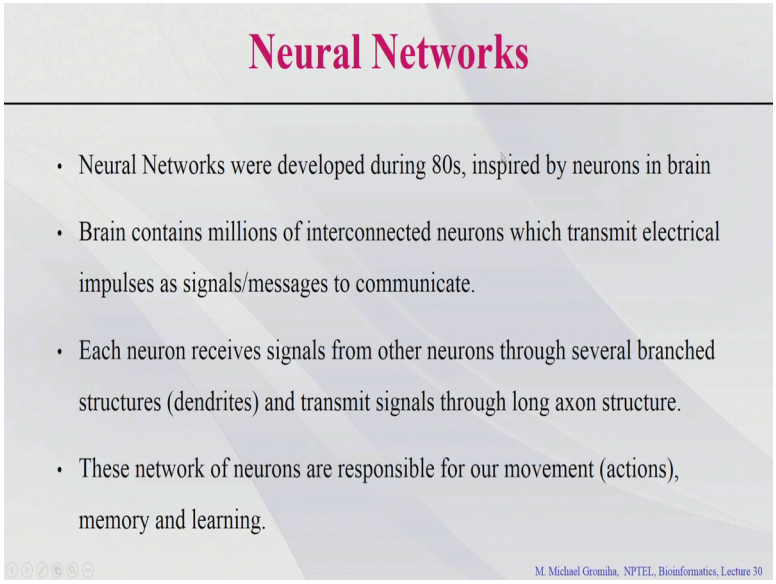
(Refer Slide Time: 27:27)



For example, in the first case this is the your input and this is the output. So, with respect to your output it will map the input data right by learning the result form the output right.

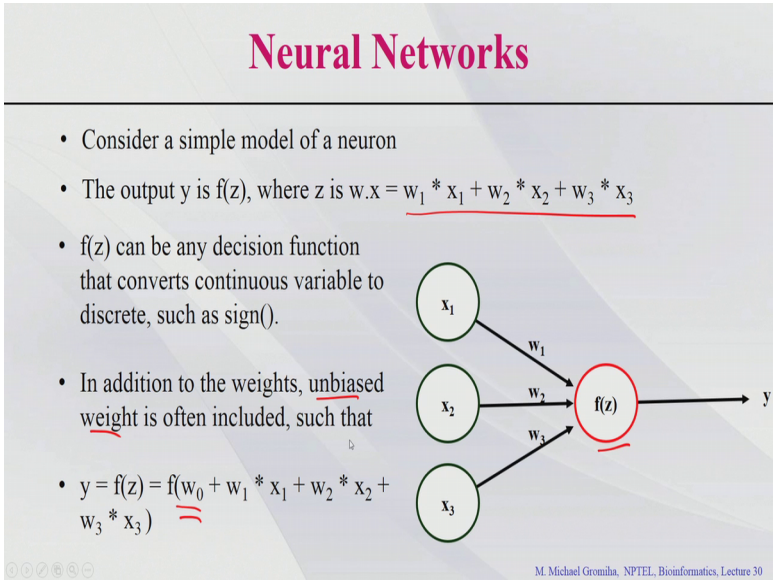
So, you can learn and finally, optimize then we use a same algorithm to test. Now here this no learning because now just we have the model this is the model is already trained. So, no more learning is you take any X input, it should give the output right then you can compare the performance right this is the 2 step process. So, different methods available for the supervised learning right. So, one is the neural network, this is widely used right if you see the history of this neural networks, they are developed early in eighties right merely inspired by the neurons in brain right.

(Refer Slide Time: 28:12)



If the brain contains millions of interconnected neurons, which transmit the electrical signals or the messages to communicate. So, what happens in this case each neuron receives signals right through a several branch structures right and transmit the signals through a long axon structure. So, this network of neurons they are responsible for our movement for example, actions, memory as well as learning and so on. Likewise you can use this type of information right in the different classification problems right for example, if you take a simple model of a neuron, this is a kind of regression in 1d.

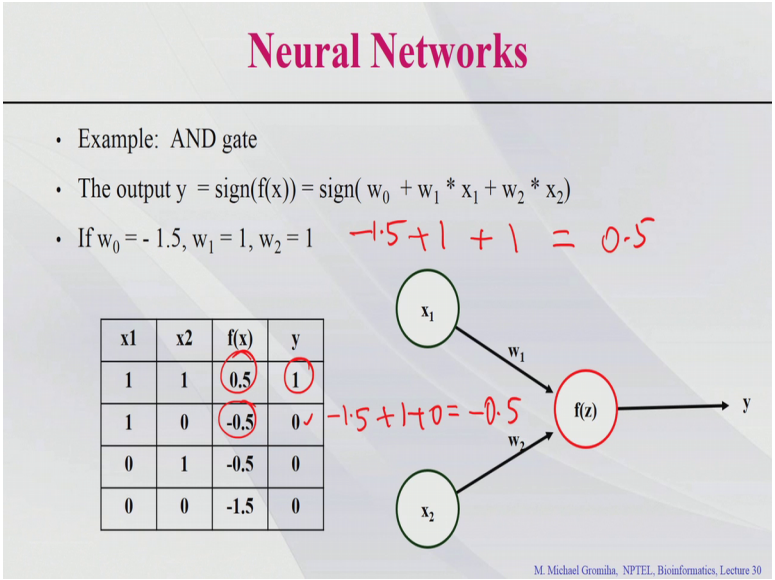
(Refer Slide Time: 28:57)



So, if see if you take the output is f of z right this is connected by different inputs for example, x, x2, x3. So, as saying with different weights, it consider this w1, w2 and w3. So, the function f of z this you can write right in terms of the weight as the inputs. So, you see this equation right w1 into x1, w2 into x2 into w3 into x3 right. This is can be any decision function right. So, depending upon your sign. For example, this is a plus then this is TMS and its minus it is a TMS right. You can you do any sign function of sign right, to assign well this is a which class fine. So, these are the assigned weights with along this assigned weights we can also use the unbiased weights like there for example, this is the w0. So, in this function.

So, in this case you can rewritten this function y you say f of z this is equal to the same right plus this f of this w 0 this unbiased weight right like the constant in the a Regression equations.

(Refer Slide Time: 30:05)



So, now, for example, if you have AND gate. So, output is y right. So, you have the sign this is equal to w0 plus w1 into x1 plus w2 into x2. So, we have x1 and x2. So, for example, if you take a weight w0 equal to minus 1.5, w1 equal to1, w2 equal to 1. So, x 1 equal to one, x2 equal to one, what is a f of x? F of x you can use this function this is w0, this minus 1.5. w1 into x 1.

Student: (Refer Time: 30:41).

Plus 1 w2 into x 2.

Student: 1.

Plus 1 this is equal to.

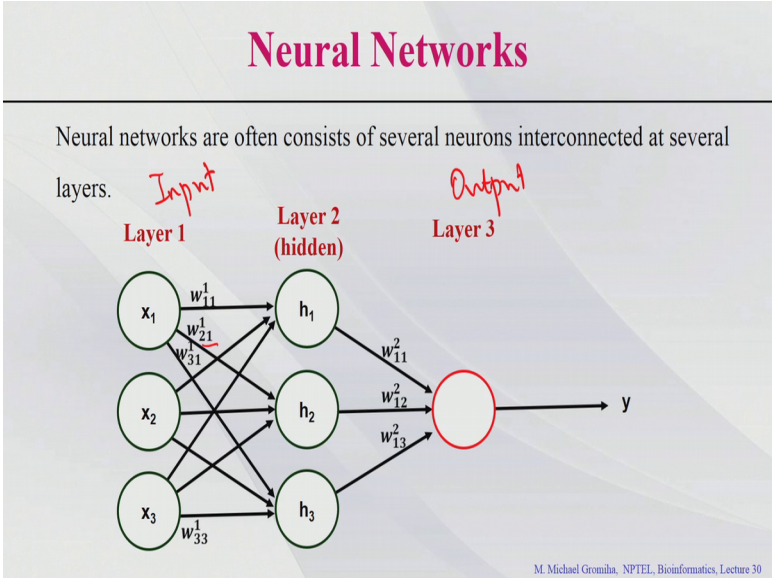
Student: 0.5.

0.5. So, you get the 0.5. So, you can see a sign minus or plus this is minus or plus is plus. So, this is equal to one right and if you take second example, if x1 equal to 1 and x2 equal to 0. So, in this case minus 1.5 plus 1 plus 0. So, this is equal to.

Student: minus 0.

Minus 0.5. So, this is this case minus 0.5, this sign is minus. So, this is equal to 0 right next 2 classes both are minus. So, both are zeros right you can see with respect to the sign you can decide whether this is a TMH or this is a TMS.

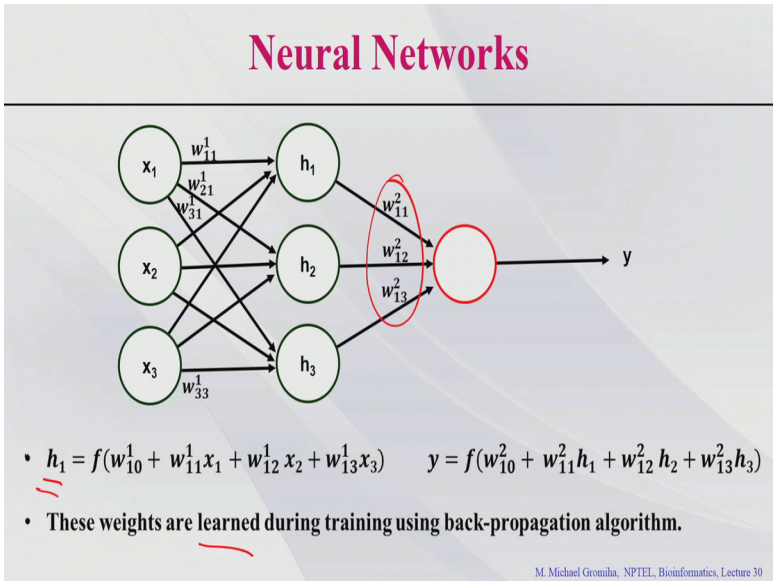
(Refer Slide Time: 31:29)



So, how the construct networks right it has the different layers. For example, if the layer one this is the input layer right and this another layer, the layer 3, the output layer. In between you see layer 2 this is hidden layer. So, here this will optimize the all this weights. So, because this was the interconnected networks x1 and h1 are connected by w1 one this layer one. So, likewise x1 and h2 are connected with this is a 2 and this is 1 this I put 21 and this is from layer 1. So, you put layer one. So, weight of 21.

Likewise is completely interconnected. So, one x1 is connected with the h1, h2, h3, x3 is also connected with the h1, h2, h3 and x3 is also connected with the h1, h2, h3 then see the output. So, from this hidden layer it go to the output. So, in this case weight 11 or the 12 or is 13 right or is three. So, when we do all these things finally, you have a decision function this will tell you where this is which class this is the TMS or it is the TMH.

(Refer Slide Time: 32:35)

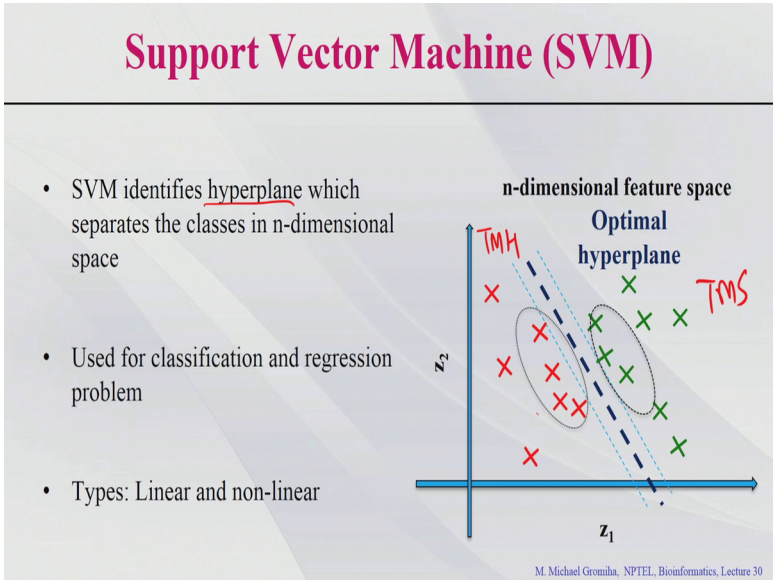


How to do that we can see the h1 this is given as this function I discussed earlier. So, here the 10 right this is the kind of constant you have this one x1, x2, x3 with the weights w11, w12, w13 and so on right for a h1 likewise h2, h3 you can write.

Then y you can write a function based on this information right w11 and 12 and 13 with respect to the hidden layers right. So, we can you have to take the function. So, how to get the weights right how you have learned the weights. So, they learned during the training using back propagation algorithm. So, you get the output then go back and adjust the weights and see the performance and do it again. So, finally, using the back propagation algorithm. So, it can learn and finally, optimize this weights right in the case of neural networks and once is optimize, then you can use the model for any prediction right that is fine. Then the second methods I will discuss about support vector machines right.

What is a support vector machines how it works right say if for example, if you see set of data.

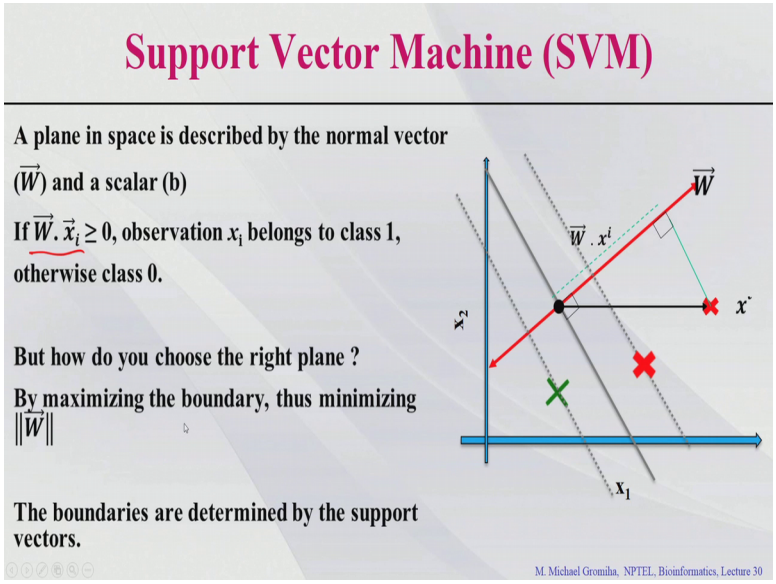
(Refer Slide Time: 33:43)



Support vector machine right identifies the hyperplane which separates the classes in n-dimensional space. For example, in this 2 sets of data how to identify the hyperplane right we can we can plane right for example, if you do it here right. This will clearly separate this is the group one or for example, you take the TMH and this is TMS you can easily classify right.

This line can be linear or this can be non-linear. So, anything. So, then we have see the distance. So, how much the width right this 2 have this a plane such a way that the width should be the maximum right between the 2 groups of classes for example, TMH as well as a TMS. So, if you see this you can have make this support vectors. So, what are the lines they are very close to line. So, these are the 3 cases in a TMS and here these are the four cases that is a very close to this line right then we can decide these support vectors to define the width right how to do this?

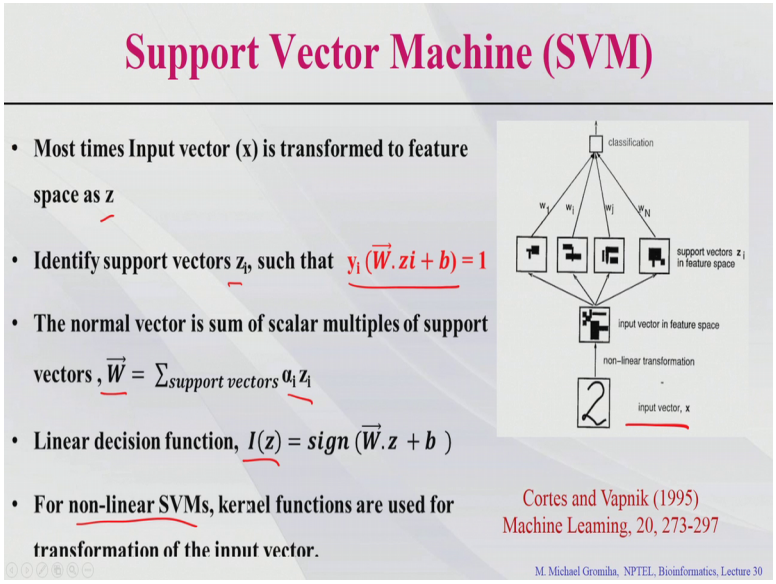
(Refer Slide Time: 34:51)



For example, if you take any points right any plane this space right here, this is the your hyperplane right. So, take a normal vector. So, this is the normal vector say W vector right and then if you if W dot x i more than equal to 0 you can take this observation belongs to class one.

See the here you can see TMH right otherwise you say TMS then the problem is how to choose this plane right that is a problem. So, in this case we have to maximize the boundary, in this case you can minimize this a vector W vector. So, that you can see the right plane it is move the planes, and then you can see the right plane, How do you define the boundaries right you can see the support vectors, a support vectors help right to determine the boundaries for assign this a boundaries right in this case you can do the different use hyperplane.

(Refer Slide Time: 35:49)



So, how it works for example, many times input vector its here, the input vector right this is transformed to the feature space like z space.

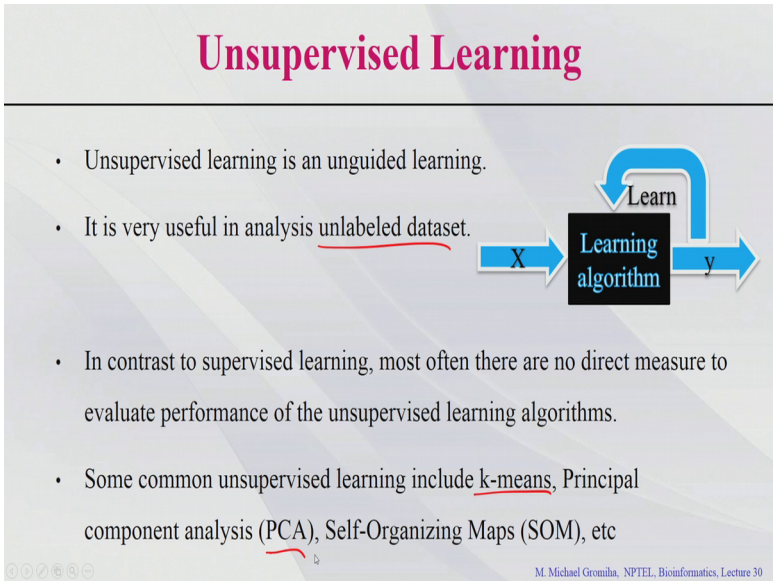
Then you can identify the support vectors of the zi such a way that this is yi of this a w dot z i it is this is the your support vectors right plus ‘b’ a constant this is equal to 1. So, if you define the support vectors such a way that. So, yi of this w dot zi dot product plus b equal to 1 right. Then take the normal vector it which is the sum of the scalar multiples of the support vector w, you take the this alpha is constant zi is the support vectors, and take the sum of all the support vectors right you get the w right this will give you the a normal vector right. Then you have the decision function i of z, this will tell you with respect to a sign right of W dot z right plus b if it is a plus and it is a minus you can define different classes.

In the case of linear ones, but the take the non-linear SVMs right the use various kernel functions right for the transformation of the input vector right. So, you can listen lectures afford by the computer science department, for more details right and they derivations for different machine learning techniques right. So, we discussed about 2 types of supervised learning what are 2 types of supervised learning?

Student: SVM

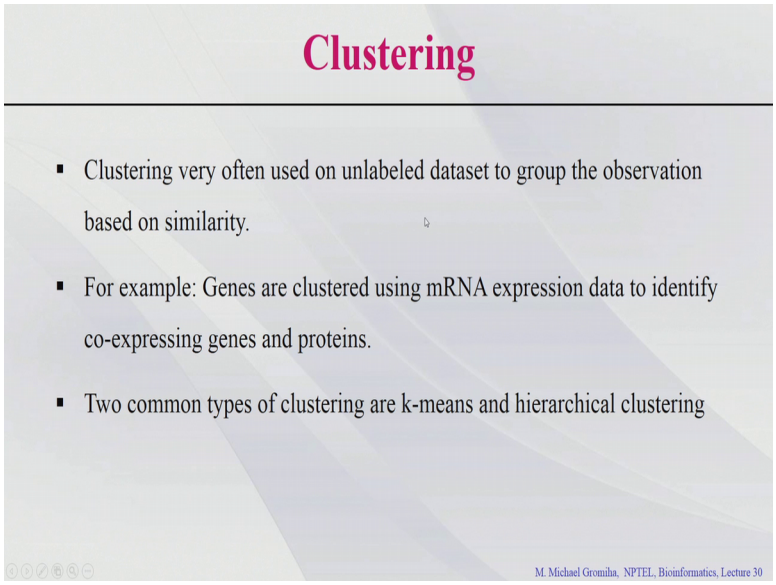
Neural networks and support vector machines right now we can see the unsupervised learning. So, it is an unguided learning. So, it is useful mainly this is unlabeled dataset.

(Refer Slide Time: 37:26)



So, if you drawn a output, we can for in unlabeled dataset you can use that right. In contrast to super supervised learning here the no direct measure to evaluate the performance in the case of unsupervised learning algorithms, but for example, some of this a methods of k means clustering or principal component analysis, and self-organizing maps and so on these are the various methods, let see how it works.

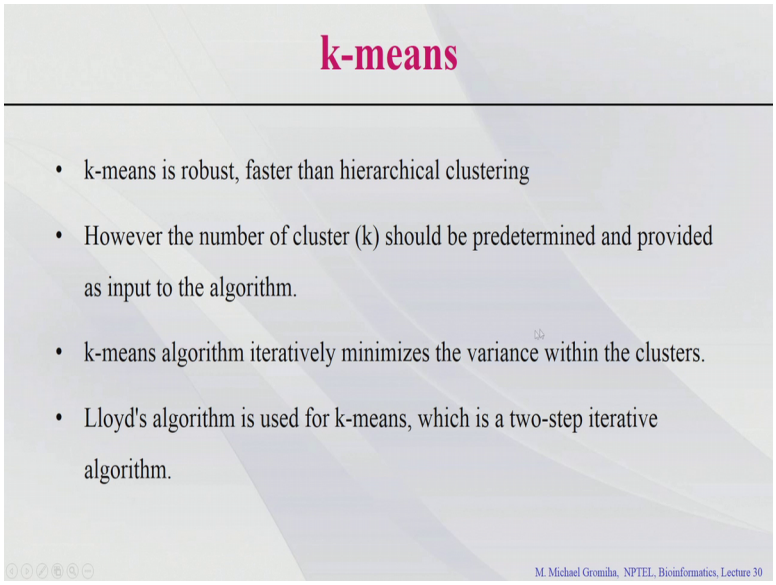
(Refer Slide Time: 37:50)



So, we take clustering, this is one of the unsupervised learning this is the very often used for unlabeled dataset to group the observation based on similarity. For example, if you have a group of 100 proteins right and see the hydrophobicity your feature. So, all the 100 proteins you know the value of hydrophobicity right then you group the proteins right with which are very close in the values, then you move this numbers here and there. So, that each group the average difference between the hydrophobicity values are the minimum right likewise you can cluster.

For example genes are clustered the mRNA expression, to identify the co expressing genes and so, are the proteins. There are various types of clustering techniques right generally normally this one is the k means clustering as well as the hierarchical clustering. So, now, we will see how this clustering techniques work. So, take the k means clustering it is robust and it is faster than the hierarchical clustering.

(Refer Slide Time: 38:43)



Already we discussed about k-means clustering with the another application right for which application we discussed k-means clustering.

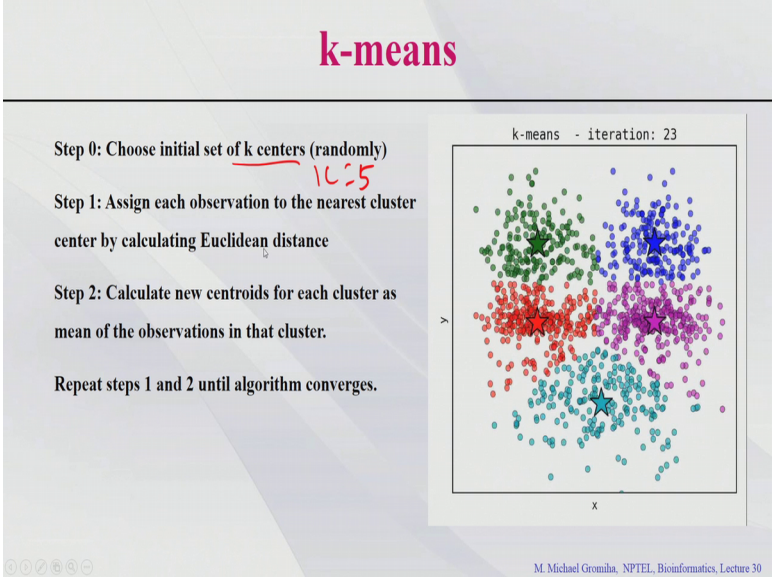
Student: Grouping.

Yeah grouping this a non redundancy sequence sequences right when you use the cd-hit, we discussed about the a clustering technique right. So, for in this case the number of clusters should be predetermined. We require to tell we require 4 clusters or 5 clusters or

how many clusters we need. We took group of 100 proteins if we have 2 clusters here it is a symbol though you can make is (Refer Time: 39:16) symbol cut off, I discussed a earlier. So, in (Refer Time: 39:19) space you cut off you can move cluster right if we take number of clusters. So, kindly we can group based on the values.

What does they k means clustering do? So, iteratively minimizes the variance within the clusters right. So, mainly the Lloyds algorithm is used for the k means which is a 2 step iterative algorithm. So, I will tell you how to do that.

(Refer Slide Time: 39:37)



So, if you see it is a group of data right how to cluster? First see the initial set of clusters k centers right how many clusters we use choose here? 1 2 3 4 5 different colors right you can show in different colors 5 different clusters, we set this k is equal to 5 right. Then assign each observation to the nearest cluster by calculating the Euclidean distance.

For example if you see composition, you can calculate the a Euclidean distance for the different proteins and see what is a values right. From this you can see what are the a nearest clusters which can form in one cluster. Then get this centroids of each cluster as mean of the observations. So, if you have a 10 proteins in one cluster. So, you can get the average right this will the mean of the cluster from that we can see what is the difference and what will happen, if this between this another cluster.

So, iteratively it will move here and there right finally, when it is optimized these are the proteins in this cluster right when the algorithm converges, then we will decide this is the 5 different clusters 1 2 3 4 5 right these are the proteins which are belonging to each cluster this is fine.

So, once it converges then you can see this is the cluster, we can a use these clusters to get the repeat steps of data right for any of your sequences or any datasets.