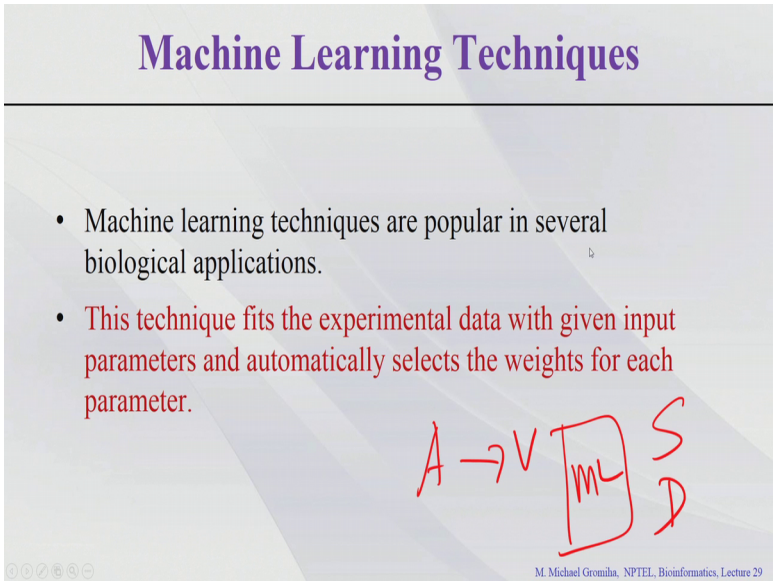
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 29 b
Development of Algorithms II

Now, we have the dataset ready; for the dataset, we derived some features. Output we know. For example, if we make a dataset; if we take the stability, this will tell you this stabilizing or destabilizing. In this case we can use various methods, it can be statistically methods or the machine learning techniques for example, we take machine learning techniques.

(Refer Slide Time: 00:40)

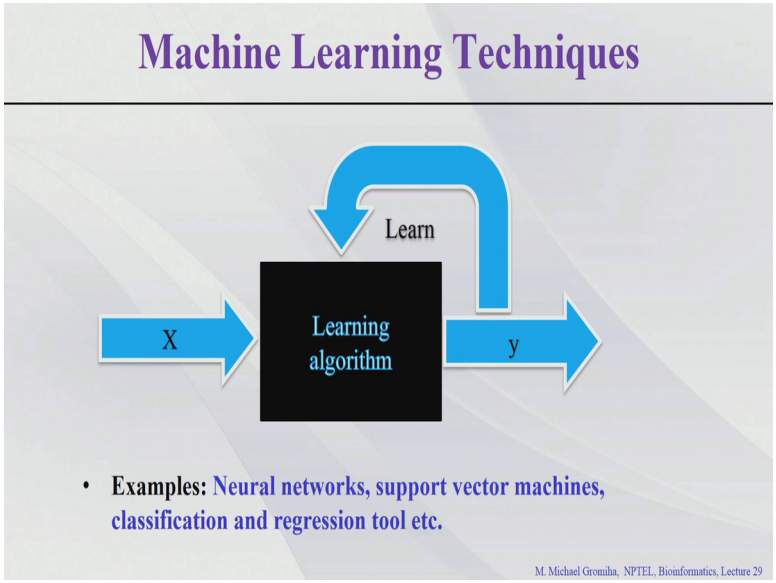


What is a machine learning technique? This will fit the data for example; a machine learning techniques fit the experimental data for stabilizing or destabilizing. So, this is the stabilizing or destabilizing. Here you have the features.

For example, alanine is mutated to valine; they take this mutated residue, position and some other information. For example, where this is located based on secondary structure or based on the accessibility surface area; they take all this information and here you can see the machine learning technique. And fits some of the parameters which you give the input and the map with this output; so, they will give the mapping between the input and the output.

So they are popular in several biological applications because machine learning techniques work far better than this statistical methods; some of the performance also very high.

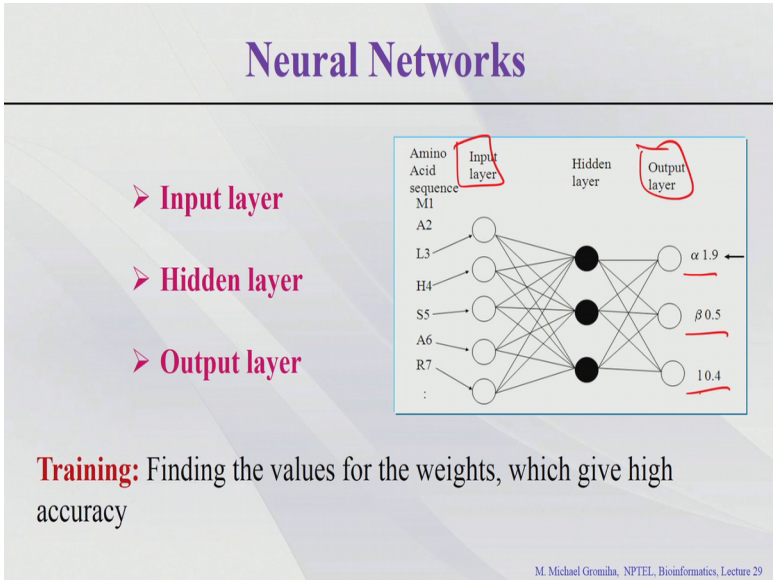
(Refer Slide Time: 01:33)



So, how it does? For example, this is the X this is the input and Y is the output. So, if you use the learning algorithm; this is your machine learning techniques. For example, neural network, say support vector machines or the classification and regressions tools and so, on.

So, with this why it will learn? And change all the weights right and then again put into this machine learning again. So, then with the new weights they train this X data and finally, give the output. So, you repeat it again and again finally, unless you get the best performance.

(Refer Slide Time: 02:09)

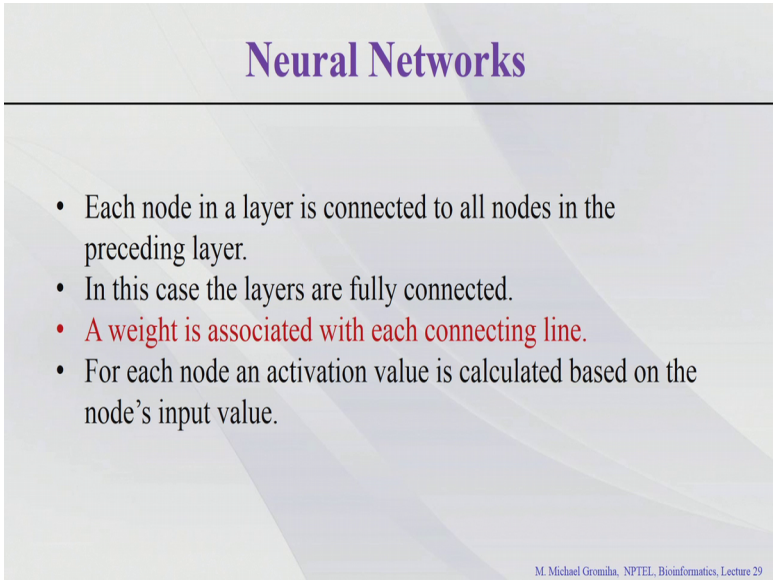


This is how the machine learning techniques work; so, one is the neural network, I will just tell you the basics.

So, here they have three different layers one is the input layer and this is the output layer. This is the what we want; this is the what we give, so that is the input hidden layer. In this the hidden layer they are connected with this inputs layer as well as the output layer this is the completely connected networks. So, they assign some weights and based on this weights finally, the output layer will tell, whether this belongs to any particular class.

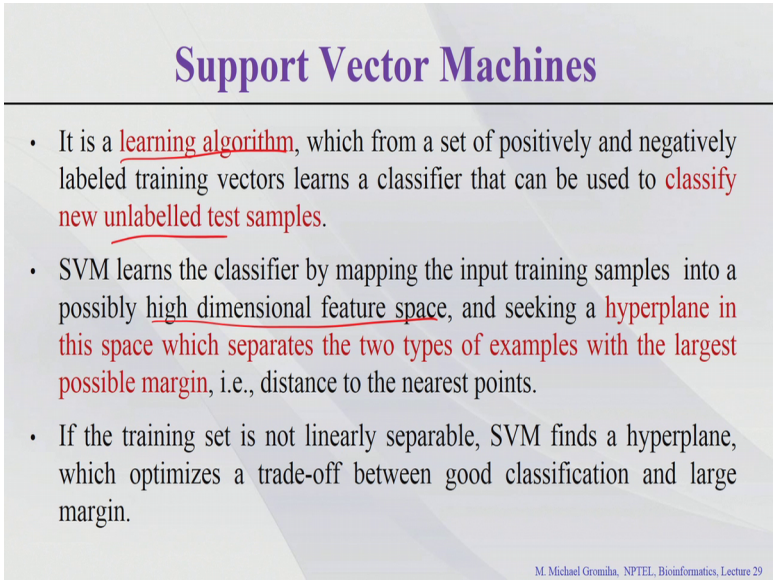
For example, here this is alpha or beta or loop; so if we depending upon the numbers we get they will assign.

(Refer Slide Time: 02:51)



this can assign helix or this is strand or it is a loop and so on. So, this how it works; now we can see the each node is connected to all node in the preceding layer; this is the connected networks, it is fully connected. And you assign the different weights and finally, based on this function you can see this is in which class.

(Refer Slide Time: 03:06)

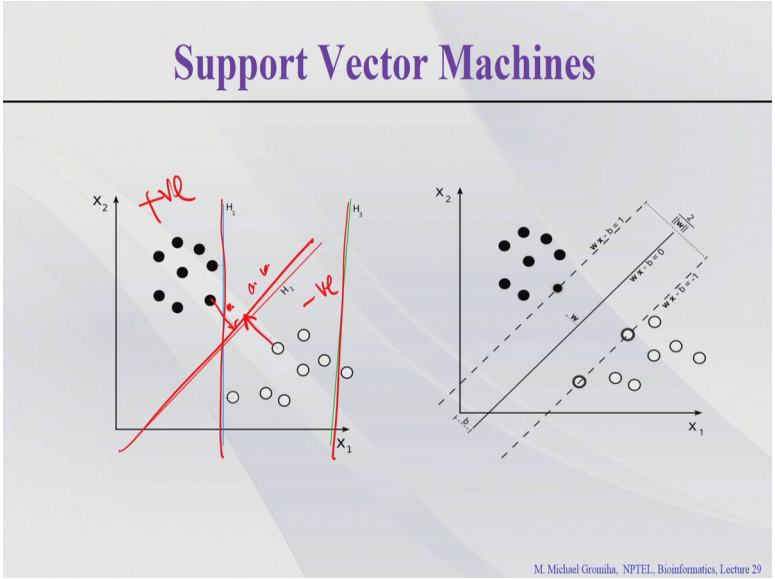


This is another one that is called support vector machines. So, I will give you the details with the equations in the next class. So, it also a very popular algorithms widely used

algorithm; so it is a kind of learning algorithm, which has two set of values for example, positive negative labeled vectors.

They can classify; this positive negatives and using this information, they can classify the unlabeled test samples. The SVM this learns the classifier by mapping the input training sample in the possibly high dimensional feature space and seeking a hyper plane which will separate the two examples whether it is a positive or negative.

(Refer Slide Time: 03:49)



For example, if you see here this is the dot this is for example, if you say positive and this is negative. So, there are various ways to set the hyperplane for example, if you put the hyperplane here; can it separate correctly? No, because this is a only one and all other on the other side; for example, if you put this line hyperplane one this you can separate? Yes. So, if you have another line here; which one is better? It is 1 and 2; both can separate; which one is better?

Student: 2.

This one; why this is better?

Student: Because the distance

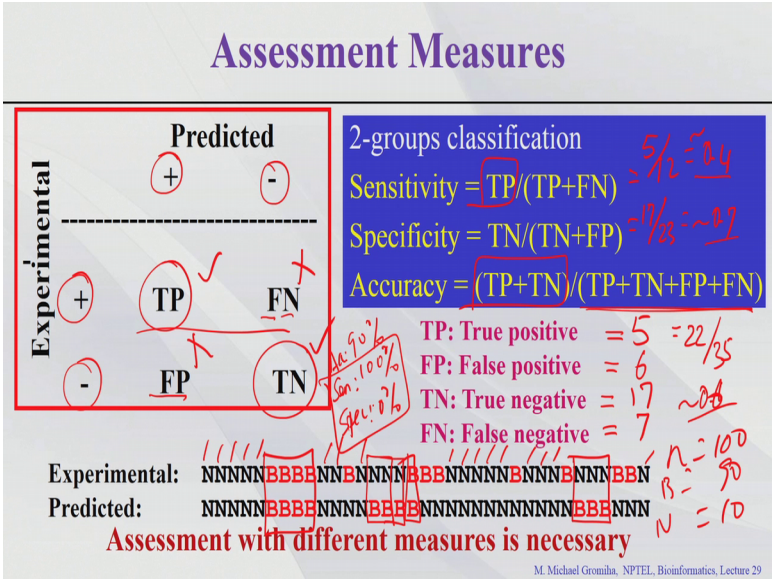
Distance; if you see the distance from this one and this one and this one; so this fine. So, in case if the margin is wide then even if you something happens here or here; in this

black ones, then you can discriminate correctly. So, we have this vectors; so make them margin as well as possible and make other things as one group as possible.

So, this is how the support vector machines; you can use straight lines or you can use any type of non-linear fits you can do to classify the different two different classes. So, now, we have the data; so we have the data sets, from the data sets we derive features and the features we give in the machine learning technique.

Finally, we can get the output now the question is whether the output we get is reliable or not? How far the output matches with your experimental known data? For example, if you have 100 data; you can know it stabilizing or destabilizing and the output we get 100 results. Then how far they are correlate? How far they coincide with each other? Is a kind of two state problems? So, we can have two groups; one is you can see one is positive group, one is negative group.

(Refer Slide Time: 05:41)



If this is a experimental is positive and the predicted is also positive then we take this as true positive. For example, a mutation stabilize a protein and you predict the stabilize the protein; so, both are same. So, in this case is positive because we taken or considered stabilizing as positive that is true. So, this is a true positive for example, if the experimental it is negative, but we predicted as positive; this false positive because it predicted as positive, but that is not correct that is wrong; so, this way is called false positive.

The experimental is plus and the predicted is minus; so this is not correct this false and predicted as negative; so this is false negative. Then if it is negative experimental and predicted somewhere minus; then this is true because both are correct where both are in this case it is negative; so even in this case it is true negative. So, we can access the values; the experimental and the predicted into four groups true positive, true negative, false positive and false negatives.

Among the four which two are the correct one? Which where the wrong ones?

Student: True positive and true.

These two are correct and these two are false predictions. So, in this case based on these measures we can define a different terms called sensitivity, specificity and accuracy. Sensitivity is the one which you can deal with the positive dataset, specificity is with a negative dataset. So, sensitivity we can see this true positive divided by true positive plus false negative; rest of are all the positives. Then specificity you can say true negative divided by true negatives plus false positives.

So, you can calculate sensitivity and specificity; accuracy you can calculate, this is a total number of instances and total number of positives. Because this is the correct one to any to TN is also the correct one; this is a total number of samples. For example, if you have 100 samples and 70 is correctly predicted, accuracy will be? 70 percent. So, I give one example here; so, here we have the two different sequences; one is the experimental and the other one is predicted.

So, you can see B's, you can say binding and this case a power positive; B is a power positive and N is negative. How many true positives? True positives means what is my true positives?

Student: Positive is predicted.

It is positive predicted as positive how many true positives?

Student: 5.

4, 5; how many false positives predicted as plus 2.

Student: 3.

This also right 3.

Student: 3.

3, 6 true negative; 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and the false negative 1, 2, 3, 4, 5, 6, 7; if you do this what is sensitivity? True positives is 5; 5 divided by true positive plus false negative plus true that is all right.

Student: 12.

5 by 12 equal to 0.4 approximately you can (Refer Time: 09:48) approximately; 0.4. Then specificity; true negative is 17 plus false positive that is 23; 0.7, 0.8; 0.7 its right. So, that is a accuracy true positive plus true negative 22; divided by 35; this is approximately 0.5, 0.6; 0.6.

So, now we can see that; so it is biased with the good to predict the true positives or exclude that true negatives. If you see this is point for sensitivity is 0.4; specificity is 0.7, which one is higher? Specificity. So, in this case you can better to get the true negatives then the true positives; here sensitivity is only 0.4. So, this will tell you if the accuracy you can see 0.6, but sometimes the accuracy is biased depending upon the datasets for the unbalanced datasets.

So, I another example I tell you I will show you if there are 100 residues; 90 are binding and 10 are non-binding; there is 100 residues, N equal to 100, binding equal to 90 and non-binding equal to 10. And you program identifies the all the 100 as binding; what is a accuracy?

Student: (Refer Time: 11:29).

90 percent because 90 are correct; out of 100, 90 is correct. So, in this case the accuracy is 90 percent this method is good? It is not good, because if you see the accuracy; this is good. Because 90 percent 90 percent very good; you can use it, but if you on other hand if you calculate sensitivity and specificity; this accuracy what is sensitivity? 90 are binding. So, all the 90 are predicted correctly; so, that will be 100 percent; so, what is specificity?

Student: 0 percent.

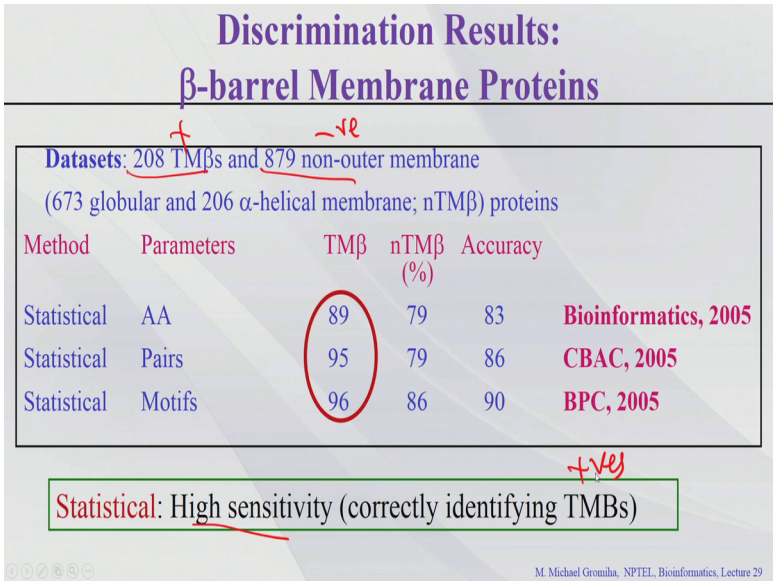
0 percent because nothing is predicted; so, if we see these two numbers, then you can see that these are not only good methods because it is completely useless method because specificity is 0. So, in this case if you develop your method you have to make sure that your accuracy lies between sensitivity and specificity. And there is a balance between sensitivity and specificity and another way is you can get the average takes sensitivity plus specificity by 2.

In this case what is a number? 50; so, in this case it is a just random because two states classification, randomly if you get the values you will get 50 percent right this is completely. So, another measure use here a ROC; Receiver Operator Characteristics; in this case you can calculate from this area under the curve, this I will explain now.

So, likewise you can calculate the different methods to see whether your method your method is good or not. So, not just with our based on accuracy it will equal sensitivity or specificity or accuracy or you can get the balance between the sensitivity and the specificity to get the performance.

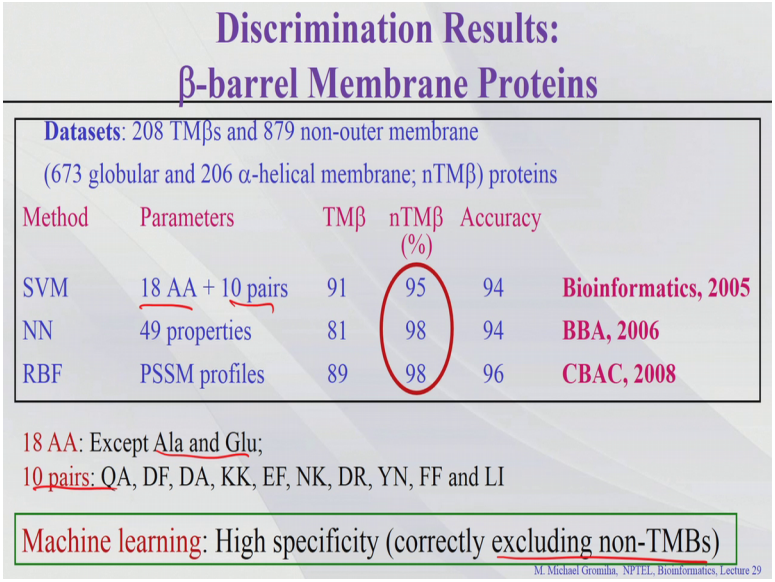
So, here write one of the results I show here; so, if we take statistical here we use statistical methods and machines learning techniques. Statistical methods I discussed earlier in the classes; we have we use a composition and get the new protein; this you compare with the group A and group B and based under we decide.

(Refer Slide Time: 13:29)



So, if we see this one here the data set is 208 positives and this is the negative data sets; this is plus and this is the negative data set. And most of the statistical methods; they try to identify correctly these positives. So, this is high sensitivity; it is correctly identifying the positives.

(Refer Slide Time: 13:48)



So, on the other hand if we use some machine learning techniques it is the same data set. If you use here it is biased with these negatives so, here 18 means the ammonia acids excluding these two 10 pairs which are given in this pairs. So, if we do this here; this is a statistical methods is less here you can see the specificity is high, you can see excluding non TM betas.

So, from this one you can see that whether you use machines learning or statistical methods. Here the statistical methods predicts the positives correctly and the machine learning, it excludes in negatives correctly. So, why the machines learning excludes negative correctly?

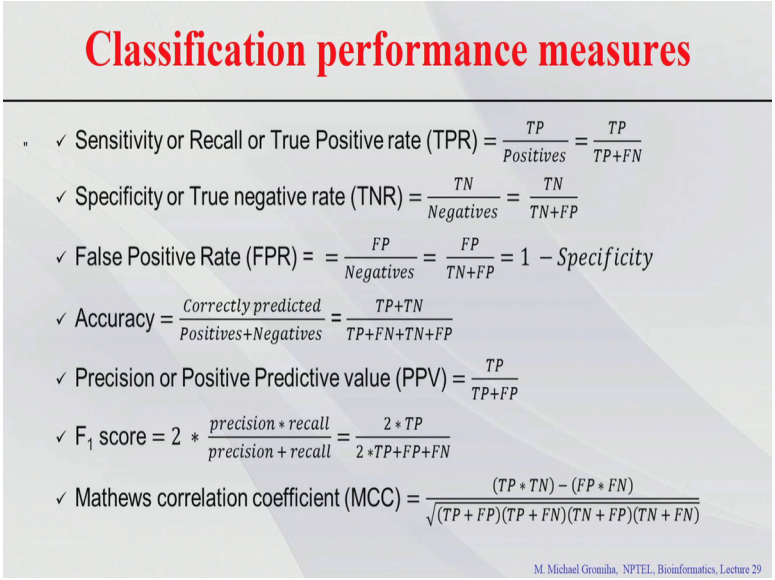
Student: (Refer Time: 14:29).

Because mainly because the dataset because here you look as a bias this is 200 or this this 900. So, this is more than four times higher, so to improve the performance they will bias with these non negative set. So, this is why they improve the performance, so you

have to be sure that that is a balance between these two. Otherwise if what is high with this a negative sets or the positive set then the reliability is very less.

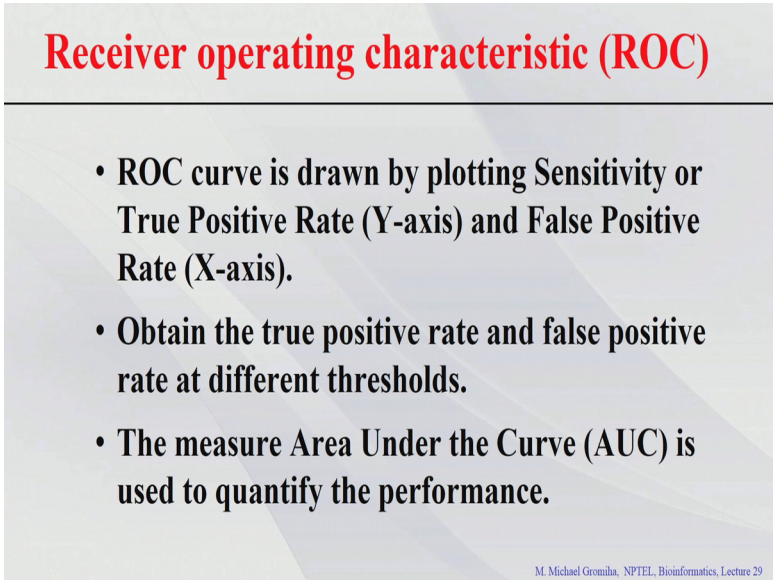
So, that need to check with the accuracy along with sensitivity and specificity; so, these are other performance.

(Refer Slide Time: 15:04)



For example the false positive rate or the precision, F-score; Mathews correlation coefficient this is also you can use to measure the performance of your method. So, this another one is currently is widely used because if you go for any classification problems; now immediately will ask what is ROC? What is the AUC?

(Refer Slide Time: 15:24)

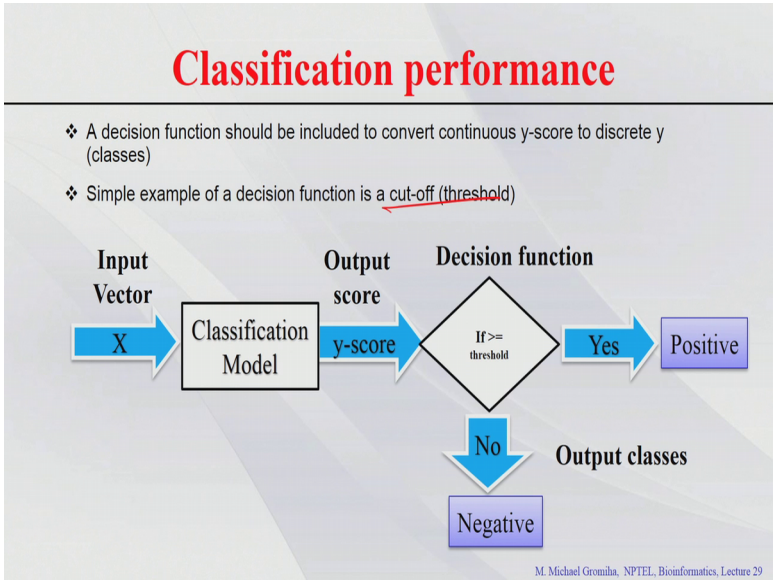


So, here this is the receiver operator characteristics, this is a curve; by plotting the sensitivity or the true positive rates against the false positive rate.

You can write the false positive to the X axis or you can versus the true positive rate in the Y axis. You can draw and then see how far your curve goes; I will show the curve just right now and what is the area under the curve. So, for this case you need two data not just two data, two sets of data. You need the true positive rates and the false positive rate at different thresholds; not just one threshold we get one number. So, we need to change the threshold and get this set of data; using this set of data you can make a curve.

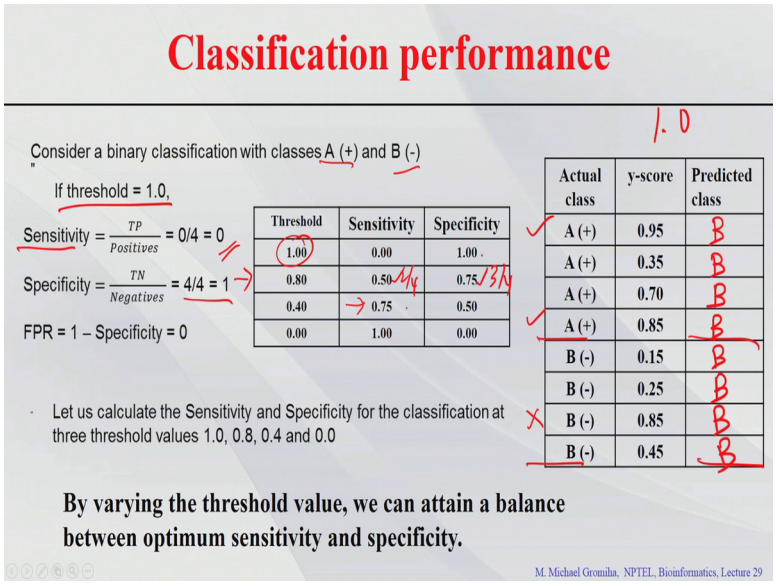
So, then if you make a curve then you can see the area under the curve; that is called AUC, this will estimate the performance of your particular methods I will just explain.

(Refer Slide Time: 16:16)



So, for example, you can see this X is your input and you do any classification model and finally, you get the output. Now you take any decision function, then if this is a greater than or equal to your threshold for example, 0 or 1; then we can take this yes is positive and if it no then this is negative cases; you can see the output class. This is based on your decision function; you can take any cutoff to decide whether this is the type X or type y; a 1 or 0.

(Refer Slide Time: 16:47)



So, now we take the binary classification for example, we have the A as positive and B as negative. Because the threshold is 0; threshold is 1; for example, threshold is 1; then what is this class? For example, it is more than 1, this is A; if it is less than 1; it is B. So, in this case what is the class?

Student: (Refer Time: 17:09).

This B; so, everything is B this B, B, B, B; why everything is B?

Student: (Refer Time: 17:17).

Because everything is less than 1, if the threshold is less than 1; the threshold you put 1; similarly less than 1, then this is case; everything is B. Now you can calculate this sensitivity, what is sensitivity true positive by positives? But no A's are correctly predicted; so it is 0 but specificity in how many B's? In the data set 4 B's; how many B's are correctly predicted?

Student: 4.

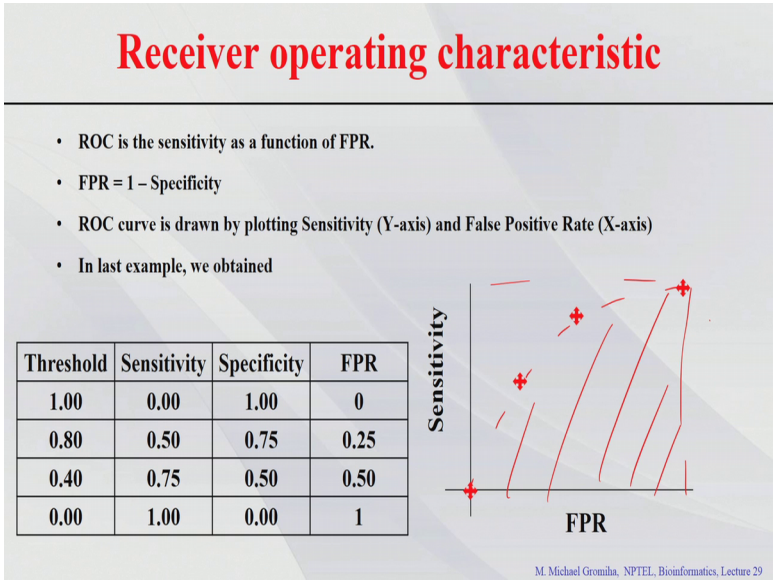
All the 4 are correctly predicted. So, 4 by 4 this is equal to 1; so, if you take the threshold 1; sensitivity is 0, specificity is 1. So, we take the false positive rate; this as equal to 1 minus specificity. So, specificity is 1; so, this is equal to 0; for example, if you take the threshold is 0.8; what is sensitivity? How many true positives if it is 0.8? If it is 0.8; this is correct and this is also correct; so, 2 by 4; this is 0.5.

In this case this is wrong; so 3 by 4; this will 0.75; this is the 2 by 4; this is 3 by 4. Likewise, if we take the 0.4 sensitivity is 0.75 and specificity is 0.5; specificity is 0, just other way around of the threshold 1. So, this is equal to 1 and this equal to 0; so you can get the different threshold, you can get the sensitivity and specificity.

So, now you can varying threshold value; we can see a balance; if we see a threshold 1 is completely biased in one side or the threshold 0 other side. But if you make the intermediate thresholds; sometimes it is better in the specificity or sometimes if it is in the sensitivity. On the other hand, specificity is also better; a sensitivity is 0.75, but specificity is also good. If we see this 2; if you take the average here we get 0.5, here we will get .615; now this is better.

So, you can change the threshold and then see how their performance varies.

(Refer Slide Time: 19:27)



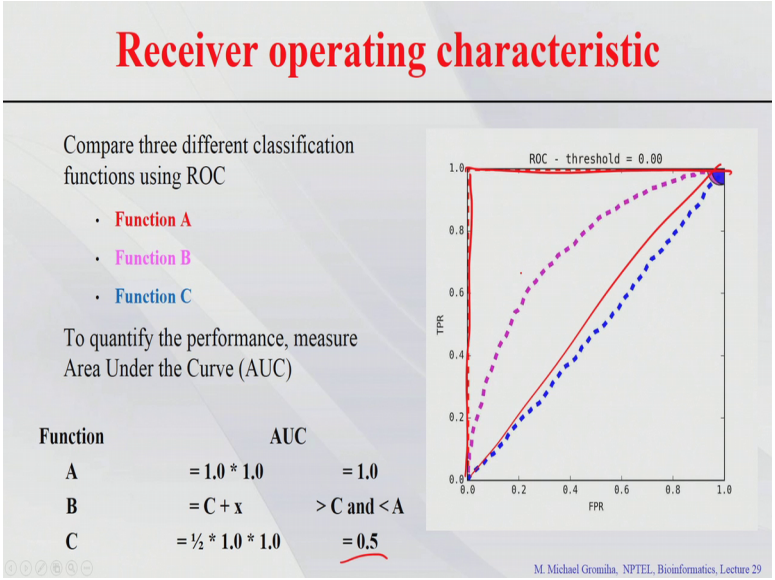
So, we can see the ROC is the sensitivity; as a functions of false positive rate, just I mentioned; what is false positive rates?

Student: 1 minus.

1 minus specificity is that is also right. So, now, we can see this is the data I showed earlier different threshold you can get the sensitivity specificity and false positive rate. So, where is false positive rate and the sensitivity? What is the value first plot here; first one? Where? This place. So, we can see here because 0; 0 this place. And the second one; 0.5 that is 0.25 and third one is a 0.75 and this is 0.5 and the last one this is 1 and 1 if you get this curve.

So, here now this is the complete graph; so, here this a place where we can see the curve; you can see how much this is equal.

(Refer Slide Time: 20:27)



This is another example, what is a minimum value the AUC will have? What is the minimum value the AUC can have? If you make different curves like this what is the maximum value it can take; if your curve goes like this?

Student: 1.

Or the curve goes like this; so, I will show the example. So, for example, you three different functions; so, three different colors here, so you can see how it goes. So, for example, with the red one; you can see this completely straight. So, it occupies the full space; so, for the red one what is the AUC?

Student: 1.

Which one? Because this is the curve; the complete space that is 1. If we take the blue one, this completely random because both the values are the same; in this case it is 50 percent of this full square. Now 50 percent means this and this is equal to 0.5; like a square this is just 50 percent; so this will be 0.5. And if you take the magenta one; function B; so, what will be the value?

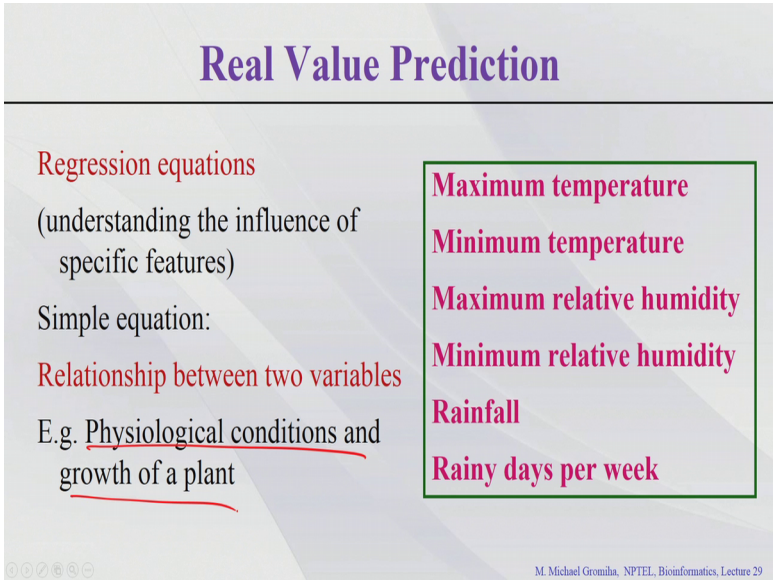
Student: Between.

That is anywhere between 0.5 and 1; so if you look at this it is very close to here this is for example, you can see 0.7 or something. You can say this value as 0.7; so, from these

numbers we can tell this is the balance between sensitivity and specificity or from this numbers we can say whether your method performance good or not; for any classification problems, so these are the classification problems.

Now, next one is the real values for example, if we have the binding affinity of 0.2 kilo cal per mole and the stability of 1.2 kilo cal per mole; how can we do that? So, in this case; we can do the regressions equations; so, the Y axis you can say this is your experimental data and the X axis you can see the various important features. For example, if you want to have the physiological conditions and the growth of a plant; this in agricultural field it is very important; if you make/plant; so everyone likes to grow.

(Refer Slide Time: 22:26)



But, we cannot grow apple here; so, likewise we need to know which conditions this plant can grow and which conditions the plant can give the fruit. So, there are some of the various aspects; I listed some of them and for example, what is a maximum temperature at that place? What is a minimum temperature? For example, some plants require high temperature; some plants require very low temperature. So, in this case if you plant which requires high temperature; so, if you do it in the cold countries, you can get it will not grow.

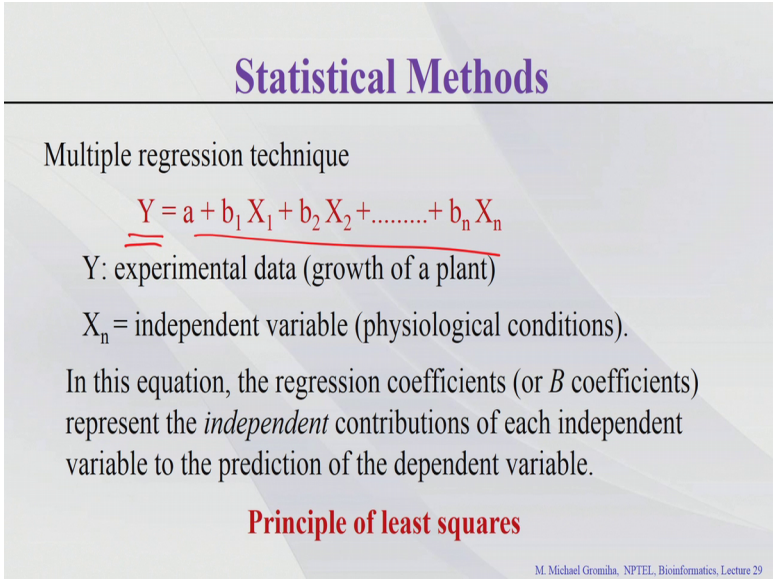
So, that is a maximum relative humidity and the minimum relative humidity or you can see the rainfalls; what is the approximate rainfall? And you can see rainy days per week.

Now if you go to the agricultural department, you can get this information. So, any local place we can get the information.

With this information, you will also tell that if you do it in June to September; this is a case and the yield is 70 percent or yield is 100 percent. Then we know the conditions and we know the yield; accordingly you try to plant in that conditions. Either you have to artificially you have to maintain the temperature or you have to think; this is the season where we have the temperature in this range and we can do.

This is the reason sometimes it fails; we expect mainly June and July we will get rain. So, this will plants, but if we do not get rain; then we do not get the fruits; we do not get the yield. Likewise we can use this techniques; over this is your experimental plant.

(Refer Slide Time: 24:00)

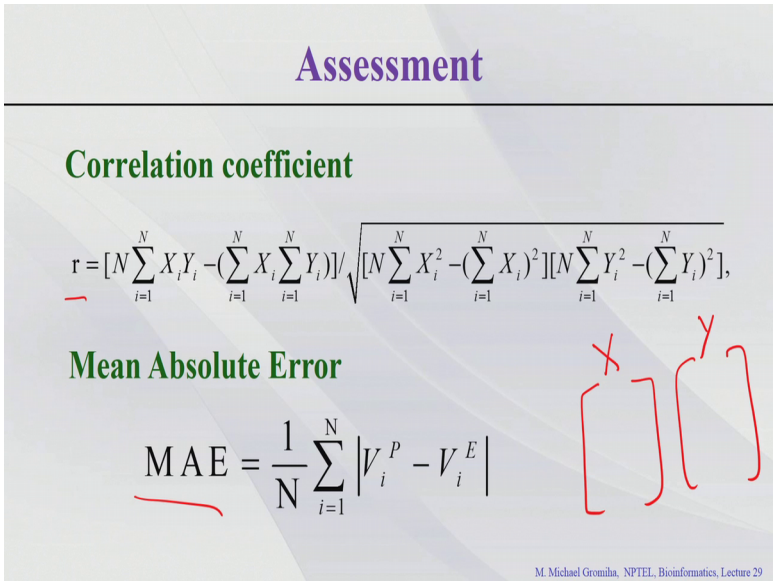


Growth of your plants or yield and here these are your variables; how this growth depends upon the different other factors like the temperature or humidity or rainfall and so, on. So, we can use the principle of least square and we discussed earlier in the previous classes. So, I do not want to explain more; so, use this one and then finally, I get the coefficients.

You set the coefficients, then you can see probably if this is the temperature in the humidity then what will be the yield? At least if you have 100 percent; if you get 80 percent then we can try. Because the worst case how much it can go because for we

expect rain, if it rainfall is less than what we expected up to which conditions; we will get the profit. If you very worst case then we do not do that; this is how we can do this.

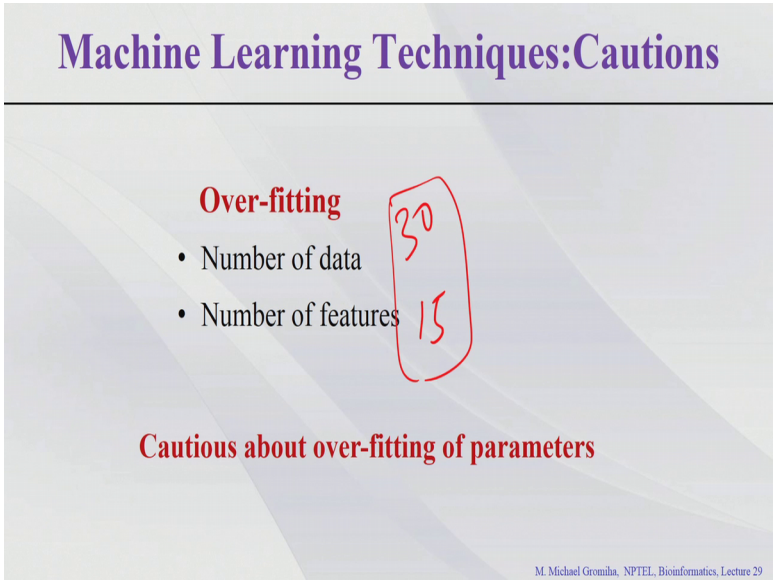
(Refer Slide Time: 24:55)



So, now how to estimate the reliability? So, we can get the correlation between the two variables for example, here we have this experimental values the growth of the plant or here this is your predicted ones. Then how far we can with this features; what is the correlation? Which highly correlated or less correlated highly correlated means you can use this equation; for giving this the growth of new plant with this particular conditions and this mean absolute error this also we discussed in the last class.

So, this you can get 1 by N; a sigma positive predicted minus experimental, take the absolute values and take the average, then we will get the mean absolute error. So, mean absolute error should be high or less? Less. Correlation coefficient should be high; if here the high good correlation and should go the less error, then we can say your method works fine.

(Refer Slide Time: 25:49)



Now, the next question is; so, can we need when you make the machine learning techniques, we have to check the number of data as well as number of features. Because last time in the QSAR also we discussed; so, what is the rules for QSAR with respect to data and the variables?

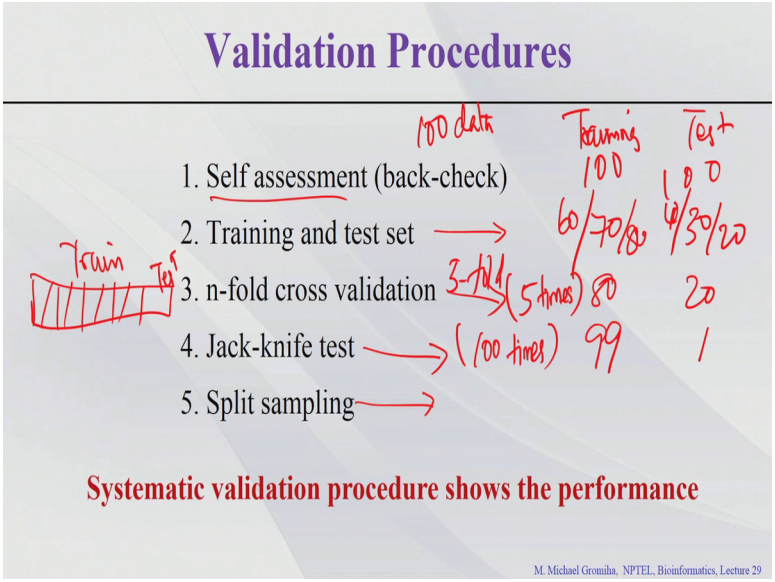
Student: Data is also be five times more.

At least more than five times and essentially if you take the; ideally if you see three to five features data are more than that. So, if we see the number of data for example, we have the 20 data or 30 data and if we do it fifteen features. For example, here 30 data and we fit with the 15 features; we will get good results, we will see the accuracy, sensitivity, specificity, ROC and get very high values, but this will work for the new data because we do not work because here you have see that you get 30 data your are fitted with 15 features.

So, we can cause over fitting because even if 1 parameters can take care of 2 data; all the 15 parameters can take care of all the 30 data; accuracy will be 100, but new one comes if you will not fit within any of these two values, then completely that is totally wrong. This is the reason why we need the less number of features and more number of data. In this case, this small number of features could account the more number of data. Even if the new one comes because the probability is very high; so, in this case it has high probability to capture the information from the new data and you can give better results.

So, there is a reason why we need to have the less number of features as well as more number of data; for doing any of this techniques to avoid over fitting. Now here we do the validation, what different validation procedures we need to do? Because systematic validation is important, usually if you do try in with the same data set; we get good results, then people be happy my accuracy is 97 percent or 98 percent. So, my method is fine; this is what we do, this self assessment for example, you have 100 data.

(Refer Slide Time: 27:54)



So, in the training you take the 100, you derive the parameters and the test you take the same 100. So, expect high performance and low performance? High performance, why high performance? Because these 100 data will capture the important information and then you predict with the same set. So, whatever you capture there then you can apply to their same that you can do the tests; we will give a good results and second one you can take the training and test sets.

For example, take some part for example, 70 percent; 70 training and 30 test. So, then you capture the features only for 70 data. Already we discussed about data sets its complete non redundant; in this case with 30 or the also non redundant not related with the 70 and in this case even if the features capture from this 70 are able to predict the 30; then we can trust if any one comes is possible to capture. Then you change this a training test 80; 20 or 60; 40 and see even if the less number of training, if we can predict well in

the more number of test; that means, whatever the feature we get from this training that is reliable and that is important for predicting this a test set of data that is very important.

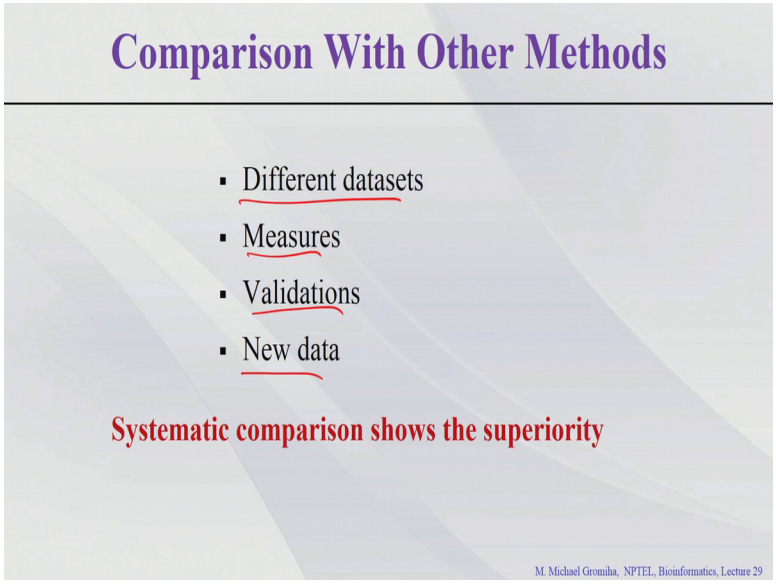
So, we need to have different sets 70, 30 or 80, 20 or 60, 40, you can do different sets of a training and test. And the third one you can do this N fold cross validation; that means, if you have 100 data you made into for example, if you take five fold cross validation. So, we divide into five groups; so you can see into five groups, 1, 2, 3, 4, 5 and train 4 and test. So, 80 you train and 20 you test, but do it for five times so that each one will be at least one time in the test.

In this case, if that is not the case something is training there is the test, but it is not if the other way around. In this case here if we do the cross validation at least one time; all the sample will be at least ones in a test. The more rigorous one this a jack knife test in this case if it is a 100 training data take 99 to train and 1 to test I repeat 100 times. So, then you do it for 100 times then see whether the perform after that we get all the output. So, you get the performance compare with the experiments how are we can do it in the jack knife test.

Next split sampling this is a random randomly choose some set of data for training and get the features and using that to develop a model and use the model for predict the remaining ones. You can take any randomly take this split and you can do that for example, if we take this 65, 35 and so on; any numbers you can take. So, now if you do this right for example, if you have good results then the question is whether your method is better than other methods or not?

If we develop any method you have to show, your method is better otherwise immediately they will ask; so what? So, you got something; so, what is the importance? why do you need to use your method? So, you have to show in different ways that to get data is the method is better; then you need to check the literature and see what are the other reports available literature. If yours is a first one, then it is little easy because you can get this novel this is a first report. So, we could get the actual accuracy of up to that much percentage; if already reports are available in the literature.

(Refer Slide Time: 31:52)

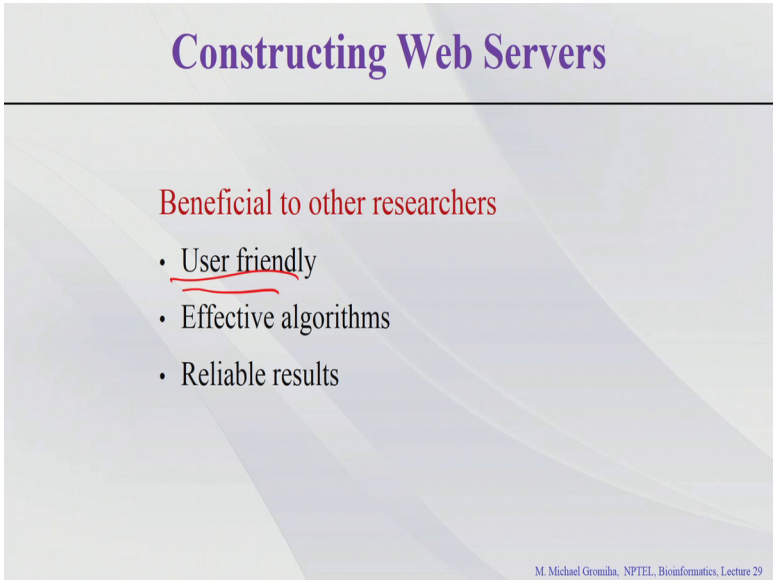


Then you have to compare on different aspects or you can use different data sets; you can measure the different measures, different validation procedures and different new data. If you have new data set and your method can correctly predict and other methods are not able to predict, then you can say that if you require any new data or method will perform better than other method in the literature.

So, you have to compare with other methods different ways to comparison; either you compare with the same data set what they use; mainly for the test or you can select the data which are currently available in newest data; that you never used, they also never used and you compare; common data set. Otherwise, sometimes we do not know which datasets they used for training, but in this case your training and their training will be different. So, you have to make sure that your some common data set to compare with the different methods.

Then once it is done your method is performed the best then we can try about some applications; how do the others know that you have a method, in this case you can develop a server.

(Refer Slide Time: 32:57)



So, you have the applications that should be user friendly. So, a new satisfied with their method and compare with other methods in the literature. So, for the applications you have to construct a server so that others can also use your methods and see whether your method is reliable and how to compare their data also with your method. So, with the beneficial to the other researchers; you develop a server that should be user friendly because if anybody access your server, already we discussed some of the database.

What are the features consider to develop a database. Likewise if the other researchers like to use your server that should be user friendly so that they can get the results quickly and what they want they we need to give provide the data. Or you can do effective algorithms, put the server and this case you can give the reliable results so the others can use your data set and that will be beneficial to other researchers in this speed.

So, we discussed various aspects; two different types of problems, what are the two different types of problems?

Student: Classification.

Classification problems and the real value predictions then what are the various aspects you need to consider; The first one is?

Student: Dataset.

Data set you have to very careful about developing the data set that is very important very sensitivity. Once data set is done and then? Go over the features depending upon the problem you are working on say you have to derive the features. Once you have the features in data sets and then develop a method; you can put in the machine learning or statistical methods so that you can predict the output, then you need to access.

So, various measures to access the performance; what are the various measures? Sensitivity.

Student: Specificity.

Specificity accuracy.

Student: ROC.

ROC curve right area under the curve. So, from that you can access the performance of your method. Then you have to validate. So, N fold cross validation or jack knife test, other say split sampling we could any cross validations methods and see whether its perform better or not. Once everything is fine then go with the comparison with the other methods literature. So, see how far your method could perform better than other methods in literature and then you make a server.

In this case everybody will can use several users right they can use your dataset, your server so that they can predict for near values and compare how far our method is better than other method in the literature. Then they can use it because all the bioinformatics problems why we do it because we have some potential applications. For example, for the drug design or different work; like I mentioned earlier like the cancer mutation and different things aspects.

So, your methods works fine and if you have a server then they can use this as their inputs and this can be used for doing any experiments. So, we will give the first end information for designing various experiments; they are useful for experiment list also to design their own experiments based on the data obtained from any new methods.

Thanks for your kind attention.