# Data science for Engineers
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture- 44
Performance measures

In this lecture, we will talk about typical performance measures that are used once a classifier is built. This idea is useful for all kinds of classifiers and in fact, some of these ideas could be used to benchmark different classifiers.

You will see a result like this,

(Refer Slide Time: 00:46)
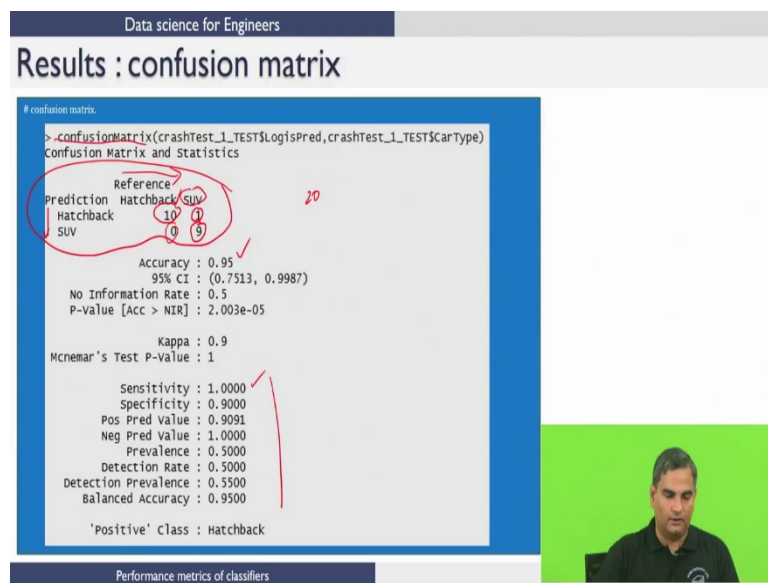
when you run, r code for any of the classifiers that you are going to see as the teaching assistants describe how to do case studies and generate this table. The intention in this lecture is the following; since, you are going to see a result like this, for most classification problems. What I want to do, is I want to really, look at all of these, terms here. So, for example, there is accuracy, what does sensitivity means, specificity means positive predictive value and so on mean.

So, I am going to, first, describe what these mean in this lecture and once you understand this, in all the future lectures, whenever you look at this, you will see, how well your classifier is doing. So, let us look at, the first thing on this slide, which is something like this here. So, this is a case study which you will see later where, based on certain

attributes of a car, you are trying to classify whether it is a Hatchback or an SUV.

So, that is the kind of example problem that we have here. So, really the two classes are Hatchback and SUV. And this table that you see right here is what is called as the confusion matrix. So this is the result that you typically get. So, the way to interpret this is the following. So, if you go down this path this is what the classifier predicts for a given data and this direction is the actual label for that class.

So, for example, these is a result of one classification algorithm and let us try and see how we interpret this result. So, if you notice the total number of data points that were used in this classification algorithm can be easily found out by summing all the four elements here. So, the number of data points that were used in this algorithm are 20 and then we could interpret each one of these numbers. So, this number basically is a prediction that a car is a Hatchback and this basically says, that car was truly a Hatchback. So, these are all right predictions of Hatchbacks.

So, the prediction was Hatchback and the car was also Hatchback. Now, if you look at this number, because we are going across this row, the prediction still remains to be Hatchback. So, the classifier predicted this car to be hatchback, however, the reference is really SUV. So, this is a wrong classification of an SUV as a Hatchback. Now, if you go to the next row and then look at this, you would see that the prediction is a SUV and the reference is Hatchback. So, this 0 means there was no car which was a hatchback, which was predicted as an SUV. So, that is why you get a 0 and if you go to this number, the true car class is SUV and the prediction is also SUV. So, 9 SUVs were correctly predicted to be SUVs. So, that is what this means. Now, we are going to de ne some terms as we go along and we will come back to the same matrix to show you how these calculations are done and show you what the meaning of these definitions, the these are so that, whenever you look at the results of another classification algorithm, you are able to kind of judge whether the algorithm did well or not and so on.

So, the previous table when we put this in this form, remember when I said reference in the previous table, this was a true condition. So, this was the true condition, this was Hatchback, this was the true condition, this was SUV and so on and this is the predicted. So, in this down, you get prediction, this is the truth. So, if you take this right here, the true condition is positive. The predicted condition is also positive. So, this is what we call as a true positive result.

So, notice, something important. This positive, the second word actually relates to the prediction and the first word refers to the truth or non truth of the prediction. So, if you say true positive, it has both information about the prediction and the actual condition also the true condition, because if I say it is positive prediction and that is true then the actual condition should have also been positive. Now, if we come to this here since, we are on the same row, the prediction still remains positive, however, the first word says it is a false prediction of positive. So, the true condition is actually negative.

So, this is a mistake of the classifier and this is a success of the classifier. Now, if we go to the second row, the second row the predictions are all negative because predicted condition negative and if you look at this right here. The prediction is negative, but this prediction is false. So, the truth is actually the condition is positive. So, this is also a mistake or a failure of the classifier. Now, if you come to this right here, since we are on the same row, the prediction is negative and we also know that this prediction is true.

So, the true condition is also negative. So, this is again a success case. So, when we think about this this way, then basically if we have

only the diagonal elements and the o diagonal elements are zero then, we have a perfect classifier in some sense. There are no mistakes that have been made by the classifier. Now, in statistics this is called the power of the test, this is called a type one error and this is called the type two error.

(Refer Slide Time: 08:15)

Now, based on those four numbers, there are many ways of measuring the performance of a classifier. So, before we do that let us make sure, that we summarize all that we said in the last slide in, in this one slide. So, we are going to use the notation TP for true positive TN for true negative FP for false positive and FN for false negative. As I mentioned before if we say TP the prediction is positive and that the true condition is also positive. So, this is correct identification of positive labels. TN then the prediction is negative that is the truth. So, correct identification of negative labels F P P, the prediction is positive and it is false. So, incorrect identification of positive labels and FN is that we predict as negative, but that is an incorrect identification. We can also see that the total number of samples that we have worked with has to be = TP + TN + FP + FN, because any label, we can classify it to one of these four possible outcomes.

(Refer Slide Time: 09:33).

So, the first definition is how accurate the classifier is. So, this accuracy is very simply defined by, in all the samples how many times did the classifier get the result right. So, we had, true positive, true negative, false positive, false negative. So, I already told you that the first letter tells you the truth or the success of the classifier. So, in these four cases the true positive and true negative are the success cases and these are failure cases. So, accuracy would be true positives + true negatives divided by the total number of samples. So, this gives you how many times did I get, did it right or how many times did the classifier get it right. Now, because N is a sum of all of these and we notice from the last slide that these are the o diagonal elements and I said the best classifier is one which has 0 o diagonal elements. So, when N is the sum of all these four, if these both are 0 N will become TP + TN. So, accuracy = 1 or the maximum value that accuracy can take is 1. So, this is an important measure that people use, to study the performance of classifiers.

(Refer Slide Time: 11:03)

Now the other definitions, there is a definition for sensitivity, where we want to find out how effective the classifier is in identifying positive labels alone. So, in the four cases again, let us look at it true positive, true negative, false positive, false negative. So, the classifier has effectively identified a positive label only if the identification is positive and that is the truth. So, this is when the classifier identified positive labels. So, that is goes into the numerator. So, we want to find the effectiveness in identifying positive label. So, what we are wanting is, of all the positive labels that were in the data, how many times did my classifier correctly identify positive labels?

So, the new denominator has to be the total number of positive labels in the data. So, this is a positive label, this is the actual condition is also true here because this is negative and true. Here the prediction is positive, but the actual, actual condition it is false. So, this is also a negative label. Here, the prediction is negative, but that is wrong. So, the actual condition is positive. So, if you want to just take the total number of positive labels in the data that will be TP + FN. So, if you divide TP divided by TP + FN, then you get what is called sensitivity. Specificity, on the other hand is the effectiveness of classifier to identify negative labels and using the same logic, you can quite easily find that the specificity will be TN by FE + TN, because this, these are the total number of negative labels, correctly identified among all the negative labels. So, true negative will be negative labels and false positive will also be negative labels, to that is the true condition of these will also be negative. So, you get this ratio to be this, you can quite easily notice that the values of Se sensitivity and Sp specificity will both have to be between 0 and 1 and the best result is when both are 1.

So, these are two other measures that people use. And there is also another measure, which is balanced accuracy, which is sensitivity + specificity by 2, that is an average of sensitivity and specificity. We will come back to this, because these are in some sense both things that we should look at and I will tell you strategies, where, you can get sensitivity be 1 always or specificity to be 1 always, but clearly, will also tell you that those will not be the most effective classifiers for us to use.

(Refer Slide Time: 14:14)

Then there are other measures, there is measure called prevalence, which talks about how often does this condition actually occur in our sample. So, how many positive labels are there, totally in our sample. So, true positive is a positive label and false negative is also false positive label, because, the prediction was negative, but that is false. So, the true condition is actually positive. So, that divided by the total number in the sample space, it will give you the prevalence. Now, positive predictive value is the following. If the classifier identified several labels are as positive, what proportion of this is actually a correct result is what is. So, if all of these are identified as positive by the classifier, there is a proportion of this, which is correct. The proportion of correct results in labels identified as positive is what is called positive predictive value. Similarly, if a classifier identifies several samples as negative, the proportion of correct results within this is what is called negative prediction value.

So, this actually, is something that is used quite a bit for example, in medical community and so on. So, basically you might understand that, that this is very important right. For example,, if you go, do a test for dengue and if you get a positive result, how likely is that result to be

correct is given by, this kind of number and so on. So, these are important other measures that one could use.

(Refer Slide Time: 16:05).

Then there is something called a detection rate, which is defined as true positives divided by total number of samples and detection prevalence, which is true positive + false positive divided by N. So, these are easy to calculate and there are interpretations for this. The last number that we are going to talk about is what is called a K number or K statistic, which basically in some sense benchmarks whatever result you get with a random chance based classifier.

So, this is little more complicated than the other measures. So, I am just going to de ne this and then going to show you the formula for this. So, what you want to know is this K gives you the observed accuracy, whatever the classifier gives you as a result - what accuracy, you would expect for a classifier, which is designed based on this notion of random chance, divided by 1 - expected accuracy. So, this is the definition of K. So, what you want to have is this observed accuracy to be larger than expected accuracy.

(Refer Slide Time: 17:22).

So, if this is a slightly more complicated formula. So, if I have abcd defined as true positive, false positive, false negative, true negative, then the calculation for K is this. You do not have to do this calculation. This comes out of the code, but just so that you know, what these numbers are.

(Refer Slide Time: 17:50).

So, let us go back to the same example, that we had and then look at, these numbers for, for this example, this example; we talked about hatchback and SUV. Clearly, we are not talking about a positive label, a negative label here. However, if you want to use, these measures, you have to make one of these a positive label and the one, a negative label, you could make either one positive or negative, but whenever, there is a result like this. That is shown in, in R, then the, the first one is the positive label and the second one is the negative label. In fact, you can see that here positive class = hatchback. So, this is the positive label.

So, now let us look at this and then see whether we can do all these calculations for this example. So, let us look at this number. We will see, what this is. So, the prediction is hatchback and the truth is also hatchback. So, this is true positive. Now, here the prediction is hatchback, but the truth is SUV. So, this is a false positive and here the prediction is a SUV and the truth is hatchback. So, this is what I would call as false negative and here the prediction is SUV and the truth is also SUV. So, we will call this as the true negative. So, true positive = 10, false positive = 1, false negative = 0 ,true negative = 9. So, this is what we have this now, let us go through the formulae that we had before and then see whether all of this fits in. So, the accuracy is the number of times we got it right. So, in this case, if I sum up all the diagonal elements, which will be true positive + true negatives.

Those are the number of times, we got this right. So, the numerator for accuracy will be 19 divided by the total number of samples, which is 20. So, you see that 0.95 answer for this. Now, when we look at sensitivity, we said sensitivity is defined as true positive divided by true positive + false negative. So, this is how we de ne sensitivity. So, this is going to be = 10 divided by 10 + 0. So, you get 1 here. Similarly, you can verify the specificity to be 0.9. Now, let us look at the positive predictive value. So, the positive predictive value is one where, of all the labels that were identified as positive. How many of them were actually true positive. So, that is what this would be.

So, this would be true positive divided by true positive + false positive. Because we are talking about all the labels that are identified are as positive that is, this number of which how many are true. So, if you look at this then, we will substitute this here. So, the true positive is 10 and false positive is 1. So, 10 divided by 11 and you will get this number 0.9091. And the negative predictive value would be the number of times that the negative prediction is right. Among all the predictions for as negative. Very similar to the positive predictive value and you can use the formula that I shown before to use this. In fact, this negative predictive value will be true negative, by true negative + false negative.

So, again of all the labels that are predicted as negative, how many are actually correct? So, if you do this calculation. So, true negative is

9 divided by true negative, 9 + false negative 0. So, 9 by 9, which is 1, which is what you see here. The other ones prevalence detection rate, detection prevalence and, balanced accuracy are very-very simple calculations based on the formulae, that we have in the slide, of this lecture. So, this kind of, gives you an idea of how to interpret the results that come out of, a R code, for a classifier. Now, also you notice that, this K value is here and based on the complicated formula that I showed you before.

Now, if one were to ask a question as to, what are good values for this, then that I, where a little bit of subjectivity comes in. There are applications, where, you might say, sensitivity is very important or there might be applications very much, say specificity is little more important than sensitivity and so on. So, it is kind of application dependent and depending on what you are going to use these results for, that is something that you should really think about, before finding out which, which of these numbers are important. From your application viewpoint, nonetheless, what we wanted to do was in one lecture kind of, give you the calculations for all of these. So, that, it is a handy reference for you, when you work with the case studies, in R. One last curve, which is seen in, in many papers and reported is this curve called ROC, which is an acronym for receiver operating characteristics and this was originally developed and used in, signal detection theory.

(Refer Slide Time: 24:01).
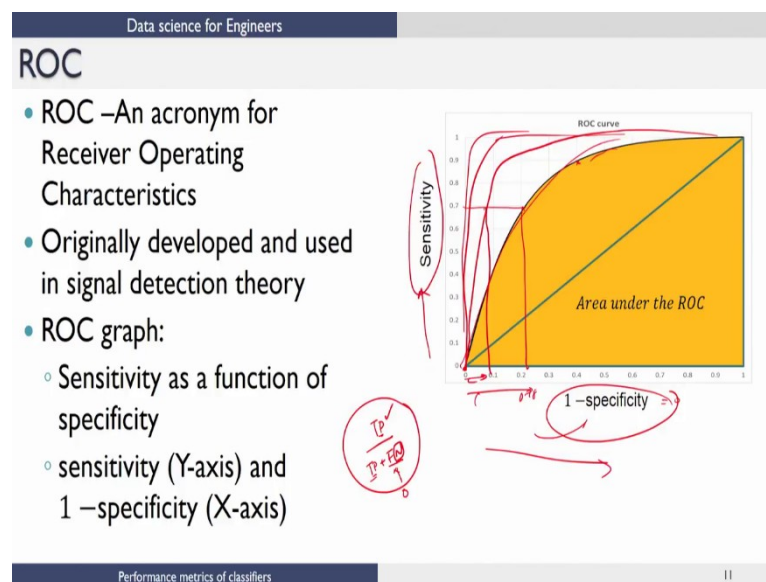
What this ROC curve is, is, a graph between sensitivity and 1 - specificity. So, it is important to notice that this is 1 - specificity. So clearly, we know that the best value for sensitivity is 1 and the best value for specificity is also 1. So, ideally you want both of these to be

one that would mean that you know, as you go, along this sensitivity curve.

So, for example, if you take a particular sensitivity, this is the ROC curve, this right here, this is ROC curve. So, if you take a particular sensitivity. So, you push your sensitivity to be more and more let us say, you go to 0.7 then the, specificity is, let us say this is 0.22, something like that. So, let us say, the specificity is 0.78, because 0.22 is 1 - specificity. So, the way to think about this is the, the best specificity point, is actually this right, because 0 1 - specificity is 0 would tell me specificity is 1.So, this is best point for specificity, but it is the worst point for sensitivity, because sensitivity is 0. Now, as you try to push your sensitivity to be more and more, then if you are sitting on this curve, you are going away from the best specificity point.

So, this happens to be the best specificity worst sensitivity and as you push this sensitivity more and more, you go further away from your best specificity point. So intuitively, if you want a good ROC curve, then what you want is the slope here to be, something like this. So, as I go away from sensitivity, I do not want to lose too much specificity. So, if the curve becomes something like this, it is better, because at the same 0.7, if you notice, I have given up only this much. Whereas, for this curve, I have to give up this much and if it is even sharper, then you give up less and less.

So, this curve kind of benchmarks different classifiers so, that is the most important thing to remember. Another thing to remember is, if you told me that I want the best sensitivity, I do not care about specificity, then that is a very trivial solution. The reason is, remember, how sensitivity is defined? Sensitivity is defined as TP by TP + false negative. So, this is how sensitivity is defined. So, how many times do I get true positive divided by true positive by + false negative. Now, think about this, if I want to make this one, which is the best number for sensitivity, then my strategy is very simple, I will do no classification. Every label, I will simply call it positive. Now, if I do that then let us see what happens to this. Notice an important thing, I said this is what the classifier predicts and this is the truth or the falsity of what the prediction is. Now, if I come up with a strategy, a classifier, which simply says positive for every label, let us see what happens to sensitivity. So, you will have true positive on the numerator divided by true positive, and this false negative will be 0 and the false negative will be 0, because negative speaks to the prediction, but I have a classifier, where I am never going to play it like negative.

So, this will be 0. So, it will be true positive by true positive sensitivity will be 1. Similarly, if I come up with a classifier, which does nothing, but says everything is negative label without doing anything, then that classifier will have specificity value of 1 right. So, if I want to get a sensitivity value of 1, I simply come up with a classifier, which does nothing but says everything is a positive label

and if I want, a specificity of 1, I come up with a classifier which says every label is negative without doing anything. So, both of these classifiers are useless.

So, there has to be some give take and that given take is what is shown by this curve right here and if you take a normalized area and then say this area under ROC is this yellow portion, then you would notice that better and better ROC curves are things like this. So that means, as the area under the Roc curve goes closer and closer to 1, I am getting better and better classifier designs.

(Refer Slide Time: 29:47).
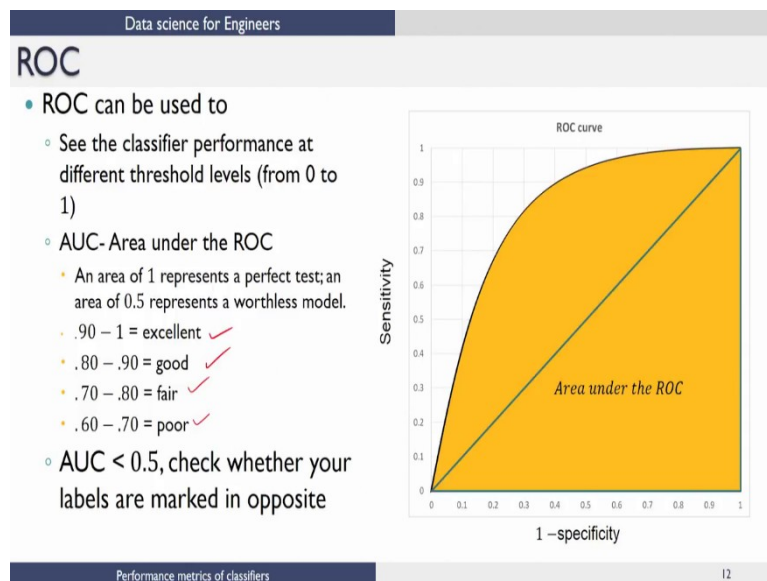
So, in general, if the area under the curve is between 0.9 to 1, we would call that an excellent classifier and similarly, definitions for good, fair and poor. If actually the ROC curve goes below this line, then there is some serious problem. So, you might want to go and check whether there is anything wrong with your data and so on.

(Refer Slide Time: 30:20).
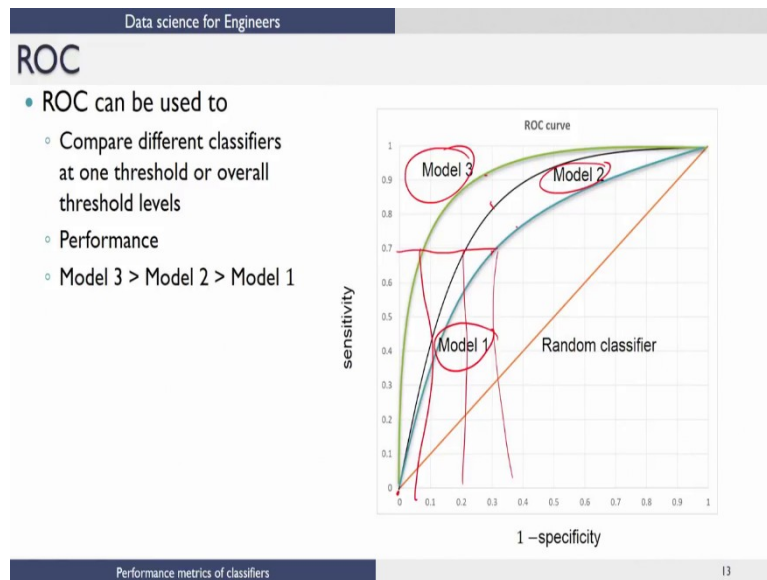
So, this is a for a single classifier, but if you have several classifiers that you are using, then I would say this classifier model one. This is classifier with model 2 classifier, model 3, then this classifier is better than this Classifier is better than this classifier, because if you take at any sensitivity level, if I go across the amount I give up in specificity, because this is the best specificity point for classifier 3 is less than this, is less than this. So, if I have to pick this to get the same sensitivity, I have to give a lot more of specificity. So, that is a key idea, when you try to benchmark different classifiers in terms of their performance.

So, I hope, this gives you an idea of, how you can benchmark the performance of various classifiers and how to interpret numbers that one would typically see with, with the confusion matrix and so on. So, this is an important lecture for you to understand so that, when these case studies are done and when the results are being presented, you will know, how to interpret them and understand these results, thank you very much.

In the next lecture, after, an case study on logistics regression is presented to you, I come back and talk about two different types of techniques. One is called K means clustering, the other one is really just looking at neighborhood and doing classification in a very nonparametric fashion, which is called the K nearest neighbor approach. So, I will talk about both of these, in later lectures.

Thank you.