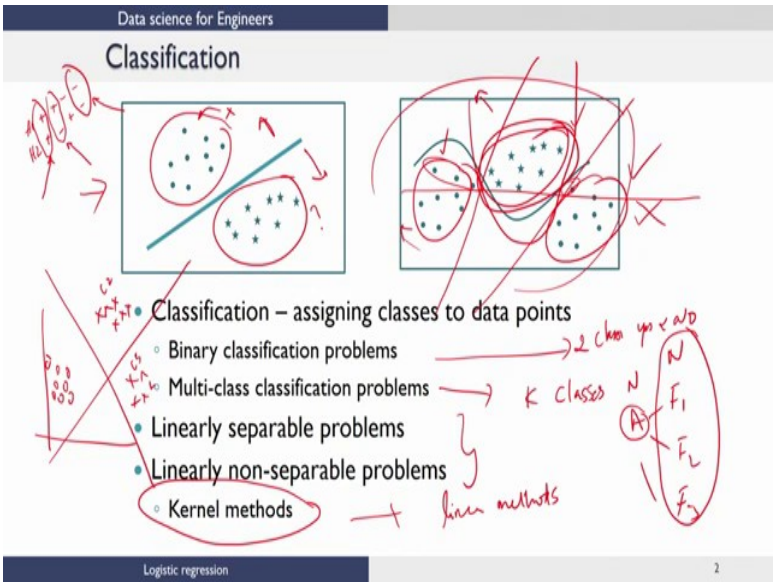
Data science for Engineers
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 41
Classification

Let us continue our lectures on algorithms for data science, we will now see some algorithms that are used in what we call as the classification problems. I will first start by discussing classification problems again. We have done this before, but I thought I will come back to this for this part of the lectures and then I will tell you the type of classification problems quickly and then describe some characteristic that we look for in these classification problems and then I will teach three techniques which could be used in solving classification problems.

In general many of the techniques that I used in classification can also be modified or used for function approximation problems. So, this is something that you should keep in mind, similarly techniques used for function approximation problems can also be modified and used for classification problems. None the less in this series of lectures we will look at these algorithms and then give them a classification flavour and that would come through both the way we describe the algorithm and also the case studies that are used with the algorithms.

(Refer Slide Time: 01:50)



So, let us look at what classification means. So, we have described this before if I am given data that is labelled to different classes that is what I start with, then if I am able to develop an algorithm which will be able to distinguish these classes. And how do you know that this algorithm distinguishes these classes. Whenever a new data point comes if I send it through my algorithm and if I originally had K different labels the algorithm should label the new point into one of the K groups.

So, that is typically what is called as classification problem. So, pictorially we represent here, here we say, here is a group of data, here is a group of data. Now, if I were to derive a classifier, then line like this could be a classifier and remember we have seen this in linear algebra before, I have 2 half spaces and I could say this half space is class 1 and this half space is class 2. Now, you noticed that while I have data points only in a small region this classifier has derived and unbounded classification region. You could come up with classifiers which are bounded also we will discuss that later.

But here now if I get a new point if I get a point like this, then I would like the classifier to say that this point most likely belongs to the class which is denoted by star points here and that is what would happen and similarly if I have a point here I would like that to be classified to this class and so on. Now, we can think of two types of problems. The simpler type of classification problem is what is called the binary classification problem. Basically binary classification problems are where there are 2 classes, yes and no.

So, examples are for example, if you have data for a process or an equipment and then you might want to simply classify this data as belonging to normal behaviour of the equipment or abnormal behaviour of the equipment. So, that is just binary. So, if I get a new data I want to say from this data if the equipment is working properly or it is not working properly. Another example would be if let us say you get some tissue sample and you characterize that sample through certain means and then using those characteristics you want to classify this as a cancerous or a non cancerous sample. So, that is another binary classification example.

Now, complicated version of this problem is what we call as a multiclass classification problem where I have labels from several different classes. So, here I have just 2, but in a general case in a multi class problem, I might have K classes. A classic example again going back to the equipment example that we just described, instead of saying if the equipment is just normal or abnormal. If we could actually further resolve this abnormality into several different fault classes, let us say fault 1, fault 2, fault 3 then if you put all of this together normal fault 1, fault 2, fault 3, now you have a 4 class problem.

So, if I have annotated data where the data is labelled as being normal or as being collected when fault F_1 occurred or as having been collected when fault F_2 occurs, fault F_3 occurs and so on, then whenever a new data point comes in we could label it as normal in which case we do not have to do anything or if we could label it as one of these fault classes then we could take appropriate action based on what fault class it belongs to. Now from a classification view point, the complexity of the problem typically depends on how the data is organized.

Now, if the data is organized, let us talk about just binary classification problem and many of these ideas translate to multi class classification problems. In a binary classification problem if the data is organized like it is shown in this picture here. We call this data as linearly separable where I could use hyper plane to separate this data into 2 sides of the hyper plane or 2 half spaces and that gives me perfect classification.

So, these are types of problems which are called linearly separable problems. So, in cases where you are looking at linearly separable problems the classification problem then becomes one of identifying the hyper plane that would classify the data.

So, these I would call as simpler problems in binary classification. However, I also have a picture on the right hand side this also turns out to be a binary classification problem. However, if you look at this, this data and this data both belong to class 1 and this data belongs to class 2. Now however you try to draw a hyper plane.

So, if we were to draw a hyper plane here and then say this is all class 1 and this is class 2 then these points are classified correctly, these points are classified correctly and these points are poorly classified or misclassified. Now, if I were to come out similarly with a hyper plane like this you will see similar arguments where these are points that will be poorly classified.

So, whatever you do if you try to come up with something like this then, these are points that would be misclassified. So, there is no way in which I can simply generate a hyper plane that would classify this data into 2 regions. However, this does not mean this is not a solvable problem it only means that this problem is not linearly separable or what I have called here as linearly non separable problems.

So, you need to come up with not a hyper plane, but very simply in layman terms curved surfaces and here for example, if you were able to generate a curve like this then you could use that as a decision function and then say on one side of a curve I have data point belonging to class 2 and on the other side I have data points belonging to class 1.

So, in general when we look at classification problems we are we look at whether they are linearly separable or not separable and from

data science view point this problem right here becomes lot harder because when we look at the decision function in a linearly separable problem we know the functional form, it is a hyper plane, and we are simply going to look for that in the binary classification case.

However, when you look at non-linear decision boundaries, there are many many functional forms that you can look at and then see which one holds. For example, one could may be come up with something like this or one could may be come up with things where I just do something like this and so, on.

So, there are many many possibilities. Now which of these possibilities would you use is something that the algorithm by itself has to figure out. So, since there are many many possibilities these become harder problems to solve. So, all of this we described for binary classification problems. Many of these ideas also translate to multi class problems. For example, if you take let us say, I have data from 3 classes like this here. So, these are 3 classes. Now if I want to separate these 3 classes and then ask myself if I can separate this to through linear methods.

Now it is slightly different from the binary classification problem because we needed only one decision function and based on one decision function we could say whether a point belongs to class 1 or class 2. In multi class problems you could come up with more decision functions and more decision functions would mean more hyper planes and then you can use some logic after that to be able to identify a data point as belonging to a particular class.

So, when I have something like this here let us say this is class 1, this is class 2 and this is class 3 what I could possibly do is the following. I could do hyper plane like this and a hyper plane like this. Now then I have these 2 hyper planes, then I have basically 4 combinations that are possible. So, for example, if I take hyper plane 1, hyper plane 2, as the 2 decision functions, then I could generate 4 regions. For example, I could generate + +, + -, - +, - -. So, you know that for a particular hyper plane you have 2 half spaces, a positive half space and a negative half space.

So, when I have something like this here then basically what it says is, the point is in the positive half space of both hyper plane one and hyper plane 2 and when I have a point like this, this says the point is in the positive half space of hyper plane 1 and the negative half space of hyper plane 2 and in this case you would say it is in the negative half space of both the hyper planes.

So, now, you see that when we go to multi class problems if you were to use more than one hyper plane then depending on the hyper planes you get a certain number of possibilities. So, in this case when I

use this 2 hyper planes I got basically 4 spaces as I show here. So, in this multi class problem which is I have 3 classes, if I could have data belonging to one class falling here data belonging to another class falling here and let us say the data belonging to the third class falling here for example.

Then I could use these 2 hyper planes and the corresponding decision functions to be able to classify this 3 class problem. So, when we describe multi class problems we look at more hyper planes and then depending on how we derive these hyper planes we could have these classes being separated in terms of half spaces or a combination of half spaces.

So, this is another important idea that, that one should remember when we go to multi class problems. So, when we solve multi class problems, we can treat them directly as multi class problems or you could solve many binary classification problems and come up with a logic on the other side of these classification results to label the resultant to one of the multiple classes that you have.

So, these give you some basic conceptual ideas on how classification problems are solved particularly binary classification problems and multi class classification problems, the key ideas that I said that we want to remember here are whether these problems are linearly separable or linearly non separable and in the linearly non separable problems there are multiple options one way to address the multiple options is through a beautiful idea called Kernel methods, where this notion of checking several non-linear surfaces can still be solved under certain conditions on the non-linear functions that we want to check using simple I would I am going to call it linear methods and I will explain this later in the course.

So, the idea here is that if you choose certain forms of non-linear functions which obey certain rules and those rules typically are called you know Kernel tricks. Then you could use whatever we use in terms of hyper planes, those ideas, for solving those class of problems.

So, Kernel methods are important when we have linearly non separable problems. So, with this I just wanted to give you a brief idea on the conceptual underpinnings of classification algorithms the math behind all of this is what we will try to teach at least some of it is what we will try to teach in this course and in more advance machine learning courses you will see the math behind all of this in much more detail.

So, as far as classification is concerned we are going to start with an algorithm called logistic regression. We will follow that up with k n n classifier then we will teach something called k means clustering, k means come under what are called clustering techniques. Now, typically you can use these clustering techniques in function approximation or classification problems and I am going to teach these

techniques and use case studies that give a distinct classification flavour to the results that we are going to see using these techniques.
So, I will start logistic regression in the next lecture and I hope to see you then.

Thank you.