NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture 21

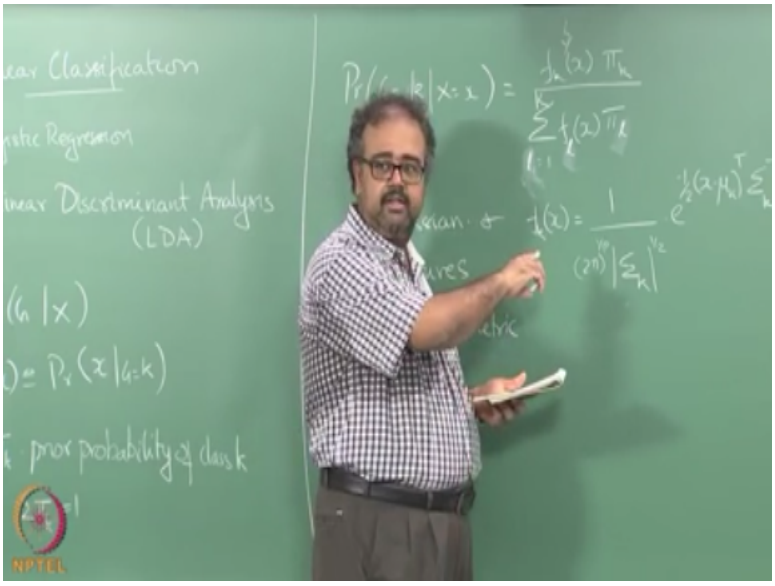
Prof. Balaraman Ravibdran
Computer Science and Engineering
Indian institute of technology

Linear Discriminant Analysis I

Yeah so we started looking at linear classification and so last class we looked at logistic regression right so you remember the assumptions that we made for logistic regressions we assume that the log arts cane modeled as a linear function right. So each the individual the probabilities were given by sigmoid functions right but the log odds was assumed to bilinear right that is where we started off with right and that gave us a linear decision boundary right correct.

So the separating surface between two classes ended up being linear if people remember what log or to us probability of class one divided by probability of class zero the log of that assume that was linear right so that's the assumption we made in logistic regression so today we look at another one of those discriminate based approaches so we already looked at two one was linear regression on an indicator variable right. The second one was logistically aggression they look at the third popular classifier.

(Refer Slide Time: 01:35)



And also known as Lda unfortunately in machine learning there are two very popular algorithms both of which are abbreviated as LD a so this was the older one okay linear discriminate analysis there's also something called latent directionally allocation which we will not get into maybe not mostly not right and which talks about a completely different approach to modeling distributions okay it is nothing to do with classification it is more on modeling distribution that's also sometimes abbreviated as ld8so be context-sensitive when you use LDa.

Right so if you remember so we are really interested in the probability of a class given the data point right we are really interested in the probability of the class given the data point and you can get this using Bays rule if you have the probability of the data point given a class and what is probability yes we can fake that what else you need probability of the class right so probability of the data point given class times probability of the class divided by probability of the data point .

Right so what we will do is we will start by making assumptions on probability of the data point given the class is K okay. so these are also known as class conditioned densities the class conditioned density of the data point apologies for that so I am going to denote by F K of X probability of given that the class was K okay. This is the class conditional density right and I am going to assume that ∂ K is the prior probability of class K.

Right so we assumed that all data points belong to some class or way other so the problem that's going to be one so now I can write right so we have been using guilt walk so that makes sense that is why I told you do not need the probability of the data right I can always fake that way

saying that since the data has to belong to some class right so I can just sum over all the classes I will get the probability of the data.

So they essentially marginalizing over class and I get the probability of the cater okay and now depending on the kind of assumptions we make for our FK the form of FK right depending on the kind of assumptions we make for the form of FK we will get different classifiers. right so some of the most popular assumptions about FK are that FK is causing if K is Gaussian right both Lda and a related method called QD a any guesses what QD is quadratic discriminate analysis right both of them assume that the class condition density FK is given by a single multivariate caution.

It is given by a single multivariate caution right you could also assume that the class conditional densities come from shirts so instead of a single Gaussian right you assume that there are multiple Gaussian which jointly generate the data for you right so people are familiar with the concept of a mixture distribution yes no okay very simple it is like let us do a little jig you here suppose I want to model this following distribution over univariate data right.

So this single dimension right and they axis actually tells me the probability of seeing something right but what I want to do something like that can you think of a parametric form that will give me this kind of a distribution looks a little daunting right can you come up with like a closed form expression for this it looks little daunting right but if you think about it I can look at another Gaussian like that and I can suitably way that two of them and I can combine their distributions right.

So the combined distribution will look like it has two peaks alright so this is essentially the idea behind mixture distributions so if the form of the distribution I want seems rather complex right and I want to have a simpler functional form for the represent for the distribution I can think of writing it as a combination of several simpler distributions right.

So likewise suppose my positive class look like this right so how will it look like in in a 2dsitting right so let us think of it so my data looks like this hey this is the positive class there's another class that comes to here okay so if you think about it this there is more data points here right and then there are more data points here and there is a slight region of min lesser density in between the two right.

If I try to model this as a single Gaussian right and if I use any kind of maximum likelihood estimate where will the peak go it peak probability will be somewhere here who is obviously incorrect right so like where's the peak probability will be somewhere here for the negative class which is obviously incorrect I suppose to that if I say that okay they're both the positive class and the negative class are created way too Gaussian search right then the mixture of this so I can have one Gaussian which has a peak somewhere here other girl shear which has a peak somewhere here likewise one for this and one for this and then I can combine them using some kind of weighting mechanism.

All right so this is what we mean by mixture distributions giving you class densities okay so you can think about this I mean I can have more arbitrarily complex kind of distributions here and then instead of having two Gaussians I can say okay I am going to have ten and also in they need not be Gaussian say could be other functional forms but the more complex the forms I take the harder it is going to become solving thisproblemso this FK of ex.

Is this alright so FK if you remember it's the probability of x given G right given that class is K right so this is a probability that given the classes case so likewise so ifI have two mixture of two Gaussians here that'll be the probability of the data point given that the class is X and here they are given that the class is whatever right so that is what we are mapping here so mixtures are fine if we still want to stay in a parametric space right well then yeah so that is a hard problem right so usually you take some guests from whatever knowledge you have about the domain right or you can do some preliminary experiments you can try Torun some kind of rough clustering by varying the number of clusters and trying to see whether you can decide on the number of mixture components alternatively.

You could do nonparametric methods they are more complex but in the last 5 6 7or so lots of tools have been developed to be able to handle these kinds of nonparametric reasoning okay. so nonparametric is actually slightly misleading it does not mean that it does not have any parameters okay it only means that it has an unbounded number of parameters okay so it does not mean it doesn't have any parameters it is just that I do not fix it a prior like we are fixing the mixture component I'm saying okay the Gaussian spur class right I fixed it okay when you fix these things we call them parametric and nonparametric methods typically can add parameters if the data needs it right you can start off with just one Gaussian okay.

And then figure out oh no I need more okay then I can add another gosh here I can add another Gaussian ands on so forth right so that is essentially what nonparametric methods bias right the ability to grow the number of parameters needed if it is supported by the data the data warrants it right so obviously we'll have to be very careful about doing things like over fitting the data but there are other ways of adjusting for it so I like I said in the last five years a lot of powerful techniques have come up for nonparametric reasoning but I'm not going to cover any of that in the class like I keep reminding you people is animator to ml course right if at all we do an advanced topics in machine learning course then we will probably cover some of that right there hoping to hire a few more faculty members who can start taking all of these courses it was the most popular.

So I do not fix the bond a priori I do not say that okay you can use only three Gaussians per class right so you can keep adding more Gaussians if the data warrants it that's what I meant ban bounded so I don't put the bound a priori and obviously I mean that is always a physical bond but in the modeling sense I don't bound it apriority I don't say that oh you have to use ten Gaussians yeah .

So the most popular of assumptions that people typically make on FK is sometimes called the naive Bays assumption we will deal with this separately and just putting it out here to just tell you that all of this come under the same class okay so the name weighs assumption is essentially to factor my class condition density along each dimension assuming given the class one dimension does not influence the other dimension.

Right so if I have two dimensions here X1 and X 2 so I will say that I can write the probability of x given K as probability of x1 given K times probability of x2 given K that is a very strong assumption if you think about it and I'm saying given the classic x1 is independent of X - if I don't know the class it look like there is some dependence between x1 and x2 but given K I'll ask you mix 1 is independent of X - so this is essentially called the naive Bayesassumption because looks like a very simplistic assumption right looks like Avery naive assumption.

To make about the data so it's called Navy base and it turns out to be powerful in many settings we will come back to Navy base separately in one of the later classes and so right now I am

going to go back to Gaussian right so I am going to assume that Hey so it looks earlier that's the Gaussian distribution okay so you have root 2 Pi Λ okay.

So that is Λ so if I'm looking at aunivariate Gaussian I will write the variance here looking at multivariate Gaussian this is the covariance right this is the covariance matrix this is multivariate Gaussian right and here again this is what X _ mu squared by Λ right so this is the people must be familiar by now when I say X _ mutranspose X _ that is actually this square right in the vector sense and then by Λ becomes Λ Inverse this is the covariance matrix.

so this is the called the multivariate Gaussian right so the univariate Gaussians are familiar about is for this kind of scenario so the multivariate Gaussian will capture these kinds of scenarios right so where the input rate of dementia itself is 2 and now I have to have a Gaussian that is actually jutting out of the window and the board yeah okay I said fine tip flower okay with that.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in