NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

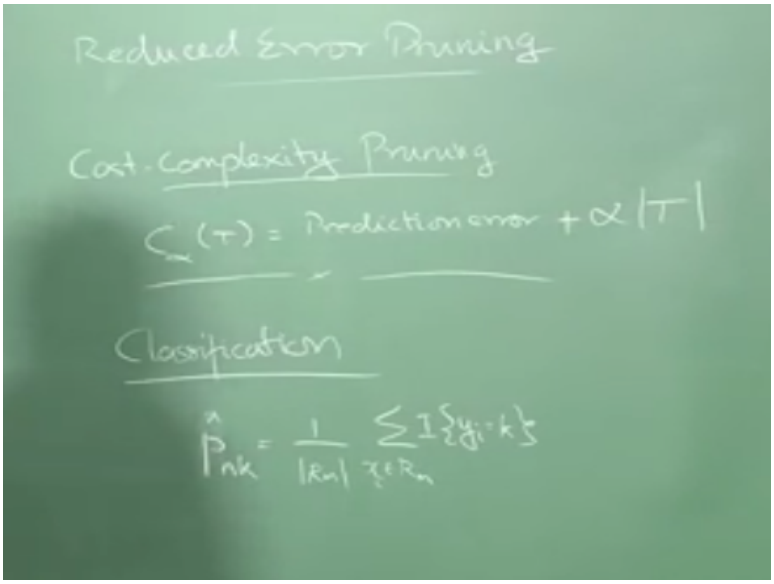
Lecture 42

Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Decision Trees for Classification-
Loss Functions

So look at the probability of a data point in M region, belonging to class K which is P_{MK} right not talking politics but so you estimate that by there is no counting the number of data points of class k and region m and dividing it by the total number of data points is fairly safe further this is how I do the prediction right so what about how do I grow a tree to do classification it is exactly the same as this except that I do not use square error right can I use square error why not exactly.

(Refer Slide Time: 0:33)



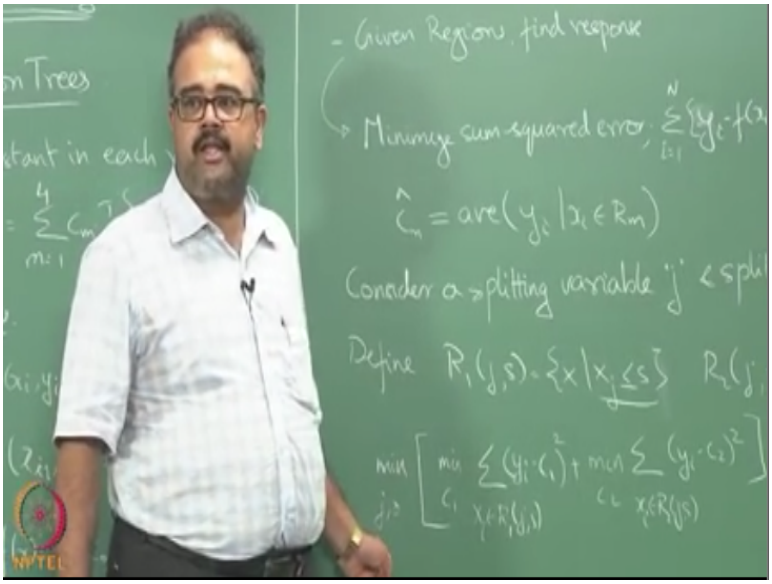
So it depends on how I encode it right I mean earlier in the linear regression you are kind of faking it by encoding it as indicator variables or whatever right so here I am going to have actual outputs I can still look at the distance of the prediction vector to the indicator variable vector

right and try to do that right I can still do that but there are better ways of doing it right so the first thing I can use is the miss classification error.

So denote by K(m) the class label that I am going to assign to the entire region M just like we did the $\hat{C}(m)$ as the, the response that I am going to assign for the entire region M so k of M is the rest for the class label I am going to assign for the entire region M okay that just say arg max of this okay. so now the Miss classification error is I am going to count all the data points in RM which do not have K(m) as their label right that is a Miss classification right is all those data points the label I will be outputting K(m) right for all the data points in RM.

I will be outputting K(m) has the label so all the data points in RM which do not really have k of M has their label or misclassified right and divided by the total number of data points that gives me the average miss classification error is there some way to simplify this 1- $P_{m\ K(m)}$ okay so because the fraction of data points that will be correctly classified or $P_{m\,K(m)}$ right.

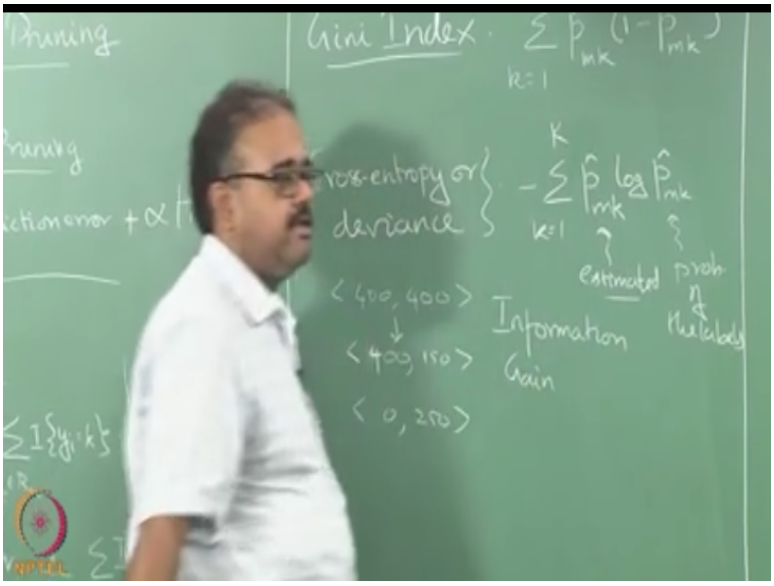
(Refer Slide Time: 04:35)



Because the two label is K(m) I am outputting K(m) there will be correctly classified right so the fraction that you be misclassified is 1- $P_{m\,K(m)}$ okay so that is a Miss classification error so how do I use this essentially I plug it in here right find the split point and splitting variables such that the Miss classification error in each of the regions is minimized at the sum of the Miss classification error in each of the regions many ways.

So remember that this has a very specific solution for minimizing so you do not really have to do this minimization is a fixed process right as soon as you find the region you just take the most abundant class in that region and set that as the class label for the entire region okay right is it clear I will do the how you used to miss classification error rate.

(Refer Slide Time: 06:18)



So the next thing we would like to look at this so one of the downsides of not being able to say anything very theoretically formal about decision trees is that it leads to dogmas so there are two camps of people who are very sure that this is the right way to do decision trees right they, they just keep fighting each other right and there are two very, very popular measures for doing classification using decision trees okay.

So the first one is called the Gini index ok so the Gini index was actually originally proposed by economists to look at disparity of wealth right so let us look at the wealth distribution in a population okay so are there more rich people than poor people or their lot more poor people than rich people I mean how does a disparity of the distribution of wealth okay that is essentially introduced that so in that in some sense you can roughly see that right.

So are there more class one data points than class two data points or anything else suppose I have K classes okay in this particular region are there more lot more class one data points than 22 k if I am able to split my regions like that then I am doing something good right because I can output

the class level as one and I will have less error correct so if I am able to split region says that the class distribution is actually skewed within that region.

Then I am doing something good and if the class distribution is uniform within that region then I am doing something bad that because that is not a good region because whatever class table I output I am going to have a lot of error but if the class distribution is skewed in favor of one class over the other then I can output that class in fact the ideal leaf would be so skewed.

There is only one class present click so the skewness measure is what I have to look for and the more skewed the data is the better so the Gini index is actually more popularly given by this form so I do this for each region so this is for a single region I do this for all regions so the other popular measure is cross entropy or deviance but it is more popularly known by the name I will give it to you in a minute right.

And this is given by this expression this looks familiar to you guys Shannon's entropy kind of thing races cross entropy where is the cross part you have P hat $_{mk}$ and P hat $_{mk.}$ there so why do they call it cross entropy okay it turns out that that they see the true output label distribution that you have right from the data that is given to you right and this is what you do for estimating this is the estimated label distribution.

And since you are using an unbiased estimator for the probabilities you end up actually estimating the true probabilities so that is why it is called cross, cross entropy this, this is supposed to be the that is the estimated okay and since you are anyway just counting the number of labels of each class and then dividing it and doing this so it is essentially end up with the same thing okay.

So the first one is the output label distribution this is the estimated one and so if you end up with the same thing right so another way of thinking about it is if you look at the prevalence of the labels in the data and I give you100 data points right essentially if I am going to randomly pick a data point and look at the label right so this is the probability of seeing label k correct so going back to your ideas of Shannon’s entropy so if I have if I have a sequence of 100 things I have k possible symbols that can occur right.

And this gives me the number of bits I need to encode these k symbols given the relative frequency of those symbols right if I had not done the splitting right if I had not split into M

regions right if I had kept the data as a whole I would have required some number of bits to encode the output level that make sense suppose let us look at it this way so I have my data so there are 400 data points of each class.

I will require some amount of bits to encode this right half, half the entropy is I mean the probability is half and half I will need some amount of thing to encode this suppose I split it up so that I get I get two regions one gives me 400,150 other gives me 0 and 250 that is how many bits do I need to encode the output variable here none right always the big improvement.

I do not need any bits for encoding the variable here and here I will need some but that certainly be less than this because we know half of the worst case right so in, in terms of the number of bits that I need for specifying the label I have some improvement when I do the split when I go from 400, 400 when you go from there and I get these two splits the number of bits I need has come down right.

So I have gained some information by doing this split right so how much information have gained? sorry right so the original entropy minus this quantity gives me the amount of information I have gained right so sometimes this is also known as the information gain criteria because of that right so either you, you minimize the cross center of PR you maximize the information gain let us information gain is essentially some constant minus this so that is information again.

Therefore you maximize the information gain or minimize entropy so again the process is very simple you for every feature J you try to find that split point S such that this or this is optimized right one of these three things but the most popular or actually the Gini index and the cross entropy so one thing I want to point out so when you are splitting this into two things right and then I have to find out the overall cross-entropy are devious right.

So what I need to do is so the entropy of this will be weighted by 250 the entropy of this will be weighted by 550/800 right both of these cases so I will have to have some kind of weighted combination of the code or the Gini index whatever it is I have to have a weighted combination of the Gini index of the individual partitions or the deviance of the individual partitions so I have to be careful about that just do not add the M up okay.

You have to use the weighted combination so for this it is fine because it is per region yeah so again you have to be we have to make sure you are combining it appropriately right yeah there is only one output will come right only one symbol is present there is only one symbol present you do not need any bits to encode it because that is only symbol this present class one will not happen so, so 400, 400 means class one there are 400 data points class 2 there are 400 data points 0 to50 means class 1 there are zero data points class 2 there are 250 data points.

The symbols I am talking about are the classes right here there will be no occurrence of class 0 and only class 2will occur okay so one again one other caveat we are using this for classification you are doing cause complexity pruning right almost always you are supposed to use the Miss classification error because that is eventually what you are trying to optimize so you grow that tree with whatever error measure you want but when you prove the tree use the Miss classification error.

Because at the end of the day I am going to evaluate you based on the Miss classification error not on the Gini index or information gain or anything and these are in some sense they are relative measures we are good for comparing one feature against the other right but the final performance measure is only miss classification error right so use that when you are doing the protein ok so I will stop here.

IIT Madras Production

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

www.nptel.ac.in