NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

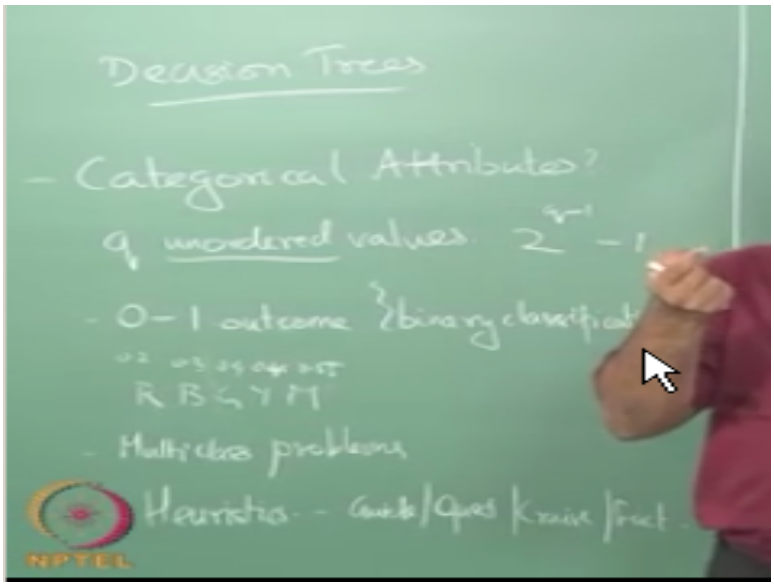
Lecture 44

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

Decision Trees - Multiway Splits

So what about multi class problems many, many classes not just 0 and 1 and I have like five classes for echo design labels from right what about multi class problem for multi-class problems this kind of simplification is actually not possible right so we have to end up using some heuristic or the other right so typically what people do is they end up doing some kind of very rough clustering on the values that attribute can take right and then try to define split points based on that and yeah so I am not going to go into the details of the heuristics I mean if, if at all you are going to use this and you will probably be using a packet but will be good to read up some of these.

(Refer Slide Time: 01:17)



If you are interested as you can imagine as soon as you enter heuristic territory right everyone can have their own favorite heuristic so there are many that have been proposed in the literature is guide quest Clues fact the one says annoying things in many of this machine learning data mining literature as people sometimes go out of their way to come up with pronounceable acronyms.

So people have clustering algorithms called chameleon imagine how much work they must have gone to produce chameleon as an acronym and so in fact I think in fact what you essentially end up doing is you use some kind of indicator variables right for each of these right this is something I think you suggested that right this is an indicator variable for each of these dimensions and then they try to do some kind of dimensionality reduction on that I try to pick a discriminating direction right.

And then project on to that and then use that dimension for splitting suppose I want to spit on color right I will not do it on color so I will create 5 variables okay which is essentially one variable or color is red one variable for color is blue one variable for color is yellow and one will for color is magenta but I will not use those as Boolean variables right and I will try to find some kind of a projection from this 5 five dimensional space on to a single dimension and then flick that single dimension as a continuous dimension and try to do my projection on it essentially ends up doing some kind of clustering instead bring some kind of clustering on that one dimension.

You talk about clustering little later but you but you know what Clustering is I already told you what the problem is in the very first class okay the other approach to doing this is to is to do multi very multi-way splits, so what do I mean by that if okay if I decide to split on color or I have to evaluate color in so splitting it into two groups I will split it into 5 groups in our case in our example because they are 5 values color can take I will split it into five groups right so in my decision tree instead of always looking like this will suddenly start looking like that.

So what are the problem with multi way split so why do not we use multi-way splits all the time too much computation in what way not each of the class right why are we determining the split point for each of the class talking about an attribute that describes the data right this is that some confusion people are having here when it when I talk about categorical attributes I am talking

about attributes of the data other than the class label the class label will always be categorical right if it is continuous then it becomes a regression problem right.

But then the values that are describing the data itself you normally assume that X comes from RP right I was telling you that that need not be the case right if suppose I am filling out a survey form you in stop filling in a rage or something will going to say less than 25 or something right so in such cases how will you test on that variable right how will I split on that variable that is a question we are asking so in now instead of saying that you see less than 25 and between 25and 35 will go left and greater than 35 and greater than 45 will go right.

Instead of saying that and say okay this will be less than 25 this will be between 25 and 35 this will be between 35 and 45 will be greater than 45 or something then splitting it all the ways in one go really does not it is a little bit more computation because when you are computing the score of each attribute that you have to do some additional work but it is not too much okay what is bad is it no yeah but moving you always remove the whole sub tree right yeah so interpretability becomes a casualty right.

So because if you are going to have multi-way splits becomes harder to interpret so the tree becomes very sprawling all right so remember as one of the biggest advantage of decision trees is that they are easily interpretable now if I am going to say okay there is a ten way split and then you have to go down the 10 way split and go down further then it becomes harder to interpret right.

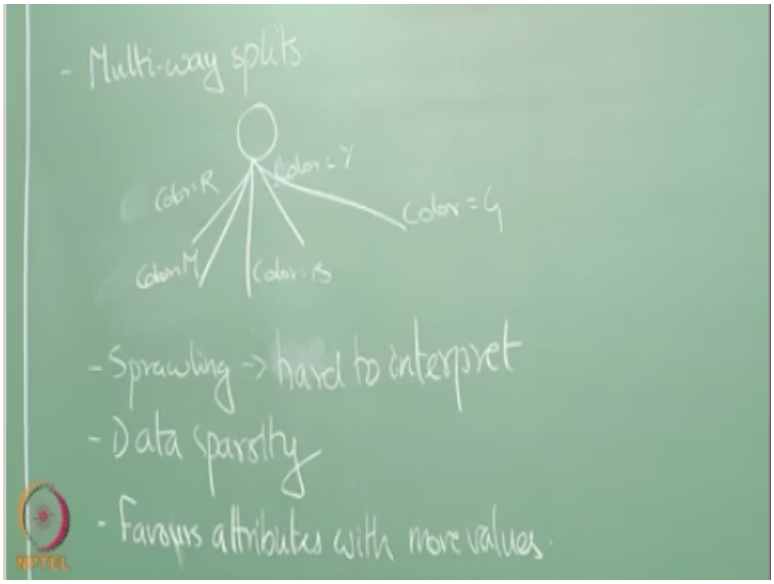
So like if she was saying you might lose insights right that is essentially saying that some amount of interpret interpretability is lost but there is another problem with having sprawling trees yeah, so variance is more but the related problem to that right is the fact that if you do this multi-way splits the amount of data that is available might come down drastically right, so each path might have suppose let us say Magenta is an Rare color right so here that has nothing to do at this 0.55 okay magenta might be a rare color right it might be just only 10 people in my million customer database ever have magenta color shirts okay.

Right but then 55% of them might be positive I do not know see that that has nothing to do with it right so how predictive it is of the positive class is nothing to do with the size of the population there right but the problem is I will only have 10 people here on which to make further decisions

right so if I am going to do this multi-way splits I run into data scarcity problems very quickly does it make sense I mean I know I really cannot ask you questions and exams or things like that with all of these things more like practical guidelines for you to when you actually start using these algorithms.

What are the things you should be watching out for right it should we are using decision trees you should make sure that you are not running out of data points very quickly if you run the some branch in your tree becomes sparse quickly right then it becomes harder for you to trust the trick okay and this is related to the variance question because we are making decisions based on very small number of data points then naturally the variance is going to be high here decision branch that is what I say.

(Refer Slide Time: 09:19)



So if you want me to actually fill in some things here there are no two choices when you pick and that is the whole point I am eliminating the whole question of splitting again picking a split point right so at the color attribute I will say a color not is sorry color equal to R you go that way like that, so this is essentially how your free will look up, so this how it is going to look like yeah exactly so this is though so see you remember you compute the quote-unquote the utility of splitting on a particular variable right.

So you pick a split variable and then you find optimal split point in that split variable and then you look at what is the least quality whatever you can achieve rate we look at squared error we looked at entropy in whole bunch of other things, so you essentially look at that so here instead of looking at the best possible split point once you pick an attribute you split on all the values attribute can take and then compute the measure whether it is squared error or entropy or whatever it is you can compute the measure.

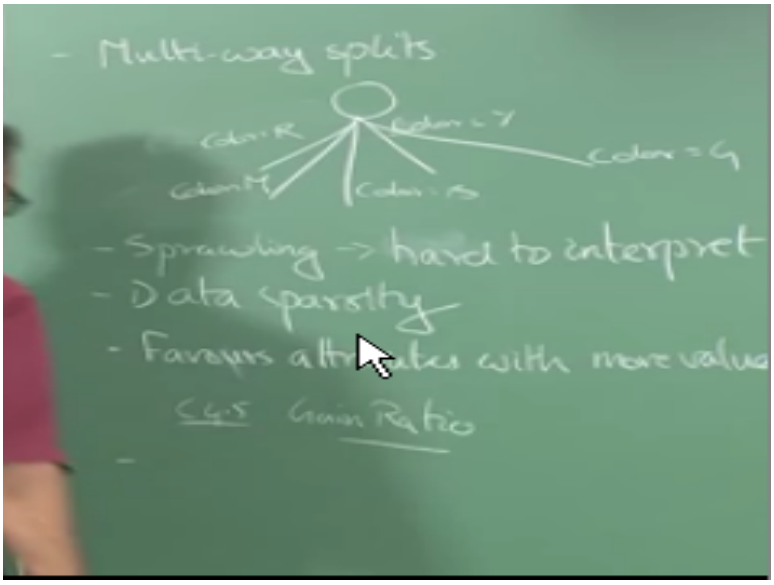
So all the measures we talked about where in contingent on the split being binary right essentially you just looked at R_1, R_2 for simplicity sake but I could have had R_1, R_2, R_3 , R_4, R_5 and I was computing the expression if you do not and if you cannot relate to what I am saying this flip back if you actually were taking notes you would know what I mean right, so we have any way looking at the squared error measure right I wrote down as R_1 and R_2 where R_1 was it is lesser than J and where R_2 is greater than J or something or S whatever those split point.

But here we do not there is no choice of what is an optimal split point here once you pick an attribute you split on all the values it can take so it becomes prowling so that leads to so if we are doing this multi-way splits it is green natural please favor attributes with more values, so let us say I have color and then I have something else like a tan color has 5 values and ages like 15 different bins I have split age into right so when I split on color I will split into 5waybranch when the split on age I will split into 15 way branch right of course there can be exceptions but I would more likely to find pure leaves when I split into 15 then when I split into 5 right.

I split it into 15 ways and more likely to find leaves that are pure and if I split into 5 ways I am less likely to find leave set of pure right so just pure in the sense they have the same class right so this kind of multivariate tends to favor attributes with more values right, so that is not necessarily the best way of doing the splits because you might not be generalizing properly later

right, so for this people use all kinds of tricks, so they are very popular decision tree algorithm called C 4.5 which uses something called gain ratio.

(Refer Slide Time: 13: 37)



So people recall information gain as you spoke about in the last class it is related to entropy right the information gain thing so information gain tells you how much less information you need right by splitting on a particular attribute for encoding the class labels right, so what again ratio says is hey forget about the fact that I have this way this variable suppose I split the data into 10 ways randomly how much information would again vs. splitting it into 10 ways based on this attribute you see the defense I take the data split it into just randomly split it into 10 groups right or I take the data and split it into 10 groups based on this attribute okay.

So that ratio is what I will use so if I can just figure it out arbitrarily split the data into 10 groups under out of still gain the same information as spitting on this attribute then I do not want to split on this attribute it is no better than random right and heaven forbid ration is less than 1 I really do not want this right, so the ratio should be higher than, so that is what I will be looking for so I can instead of using information gain I will use gain ratio in likewise you can order this for any of the attribute as any of the measures that you use that you can always adjust it for random splits. So that is essentially what we end up doing right.

So you gain anything special about expressive power for the tree we are doing multi-way splits as a tree become more expressive in the sense that can it represent more functions than you could with binary splits know whatever I do you do it is multi-way spits I can do a recursive binary splits and I can achieve that not adding to the explicitly it just avoids the question of picking a split point right that is not a trivial thing okay.

You have to come up with all kinds of heuristic to split bit points no pick split points but still if you can okay the recommendation is to avoid multi-way splits and stick with binary splits but in some cases just easier to do this especially if the number of ways in which you will split is small enough if you know if you are not going to split it into 20 different things or 50 different things right you can still do multi way splits like 5 or 6 should be fine.

IIT Madras Production

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

www.nptel.ac.in