NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture 43

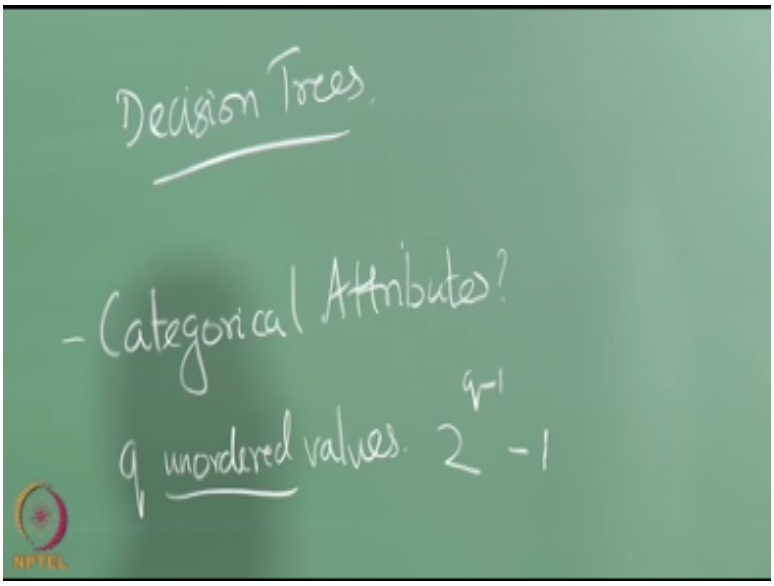
Prof. Balaraman Ravindran
Computer Science and Engineering
Indian institute of technology

Decision Tree – Categorical Attributes

We will continue looking at decision trees so like I said there is a little bit more about trees that I wanted to look at and then we will actually do a, an example today of how to construct tree, so starting from data set I will actually constructed decision tree right okay, so we already looked at a couple of issues with regard to addition trees what are the what are the things we looked at a well how will you we need to talk about that yet.

So when we look at how will you pick a cell a splitting attribute right what is the splitting value in the splitting attribute, so how large you should go a tree right and how do you prune it okay these are the issues that we looked at right whether several other questions that we could ask right, so when we talk about splitting attributes and split points inherently we are assuming that our attributes are continuously right. So that we can talk about a split point right so what happens if I have categorical attributes.

(Refer Slide Time: 01:24)



So no categorical attributes things that take some discrete values right, so it could be things like color red, blue and green, right or it could be things what you normally would believe our continuous variables like age right but for a variety of reasons they have been recorded as discrete values young, middle-aged old right, so most surveys and things like to if you look at it when you answer these things you know they do not ask you for an exact age they ask you are you lesser than 25 or in between 25 and 34 or greater than 35 or things like that right.

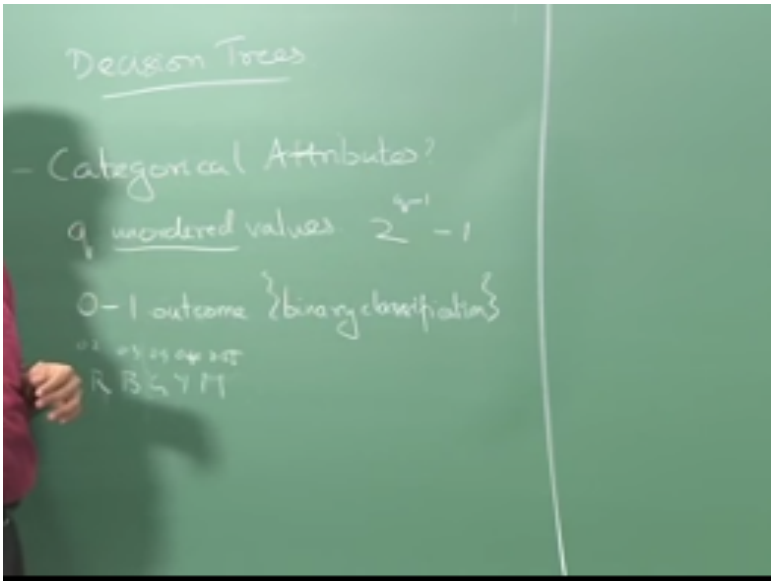
So somehow it is discretized either by for reasons of anonymization or for convenience or whatever you end up discretizing the values right, so you quite a lot of circumstances in fact especially if you are involving with the involved with medical domain right or with as kind of marketing kind of a domain right you will end up having discrete attributes right, so in which case what is the meaning of a split point, yeah I put this binary yes right suppose it is color right, so we think about it if I have q right so that is a key part here think of q unordered values right age itself has some kind of ordering in it.

So I can still fake it you know I might have only 3 different age entries but I can think of it young or middle-aged and old right young and middle-aged or old right it does typically usually does not make sense to look at young and old and middle-aged you know because it is I mean I can say you can fake it you can think of it as a ordered attribute and then you can do this but suppose it is unordered right.

So essentially what you would really have to do is think of splitting the values into two subsets, so like I said let us say take color as an example right I might have to put okay red, blue and yellow into one group and green and I do not know give me a few more color names that is about how I how much I know magenta, purple okay so all of this into another group right and so like that right I have to figure out some way of splitting it into two okay, so how many possible splits are there like that.

I have q values to 2 power? good yeah, right so that many possible combinations is not really not going to be feasible for me to go over all of them in order to pick the split point right, so that is exactly what we were doing right if you remember algorithm from the last class you are actually going over all possible split points and we said there are only finitely many such possible split points because we only have to look at n of them right but now you have to look at even though I have only n values I mean n data points for training and I potentially have to look at 2^{q-1} split points right.

(Refer Slide Time: 05:37)



So that is not going to be feasible so there are 2 ways of handling this actually there are 3 ways of handling this the first one is if you have a we have a 0 - 1 outcome basically that is what people would call a binary classification problem if you have a binary classification problem you can do one clever trick what is it that you can do any ideas no I do not want to explore the number of attributes right I am making it some something very restricted here right I am looking at binary classification problems.

So says something that you can think of that you can do here, so what exactly are you looking at when you are trying to find a split point what you are trying to do is trying to make sure that your prediction right when I do it on one half versus other half is more accurate and the prediction I did on the data as a whole before the split right so that is exactly what you are looking at from the slip point trying to find a split point such that it is more accurate than the other right.

So what you can do essentially here is you can pick one of the classes let us say you pick class one right let us say I have 5 predictors or I am sorry not yeah I have 5 predictors, so predictors or this the unordered values right, so let us say I have 5 values for a particular thing let us say colors right red, blue, green, yellow, magenta right as I say I have 5 colors, now what I will do is I will take red okay I look at all the data points that have color red okay I will see what fraction of them or class one right.

Then I will take all data points that have color blue I will see what fraction of them class one likewise for the other 3 colors it make sense so far I look at each color figure out what fraction of

that color that data points having that color or of class one now I will arrange them in some order ascending order let us say of this probability when I say fraction it means what fraction that is a probability that a data point having color red will be class one right from the training data I shall ascend arrange it in ascending order of this probability.

Then I will just treat it like any other ordered variable and then I will split right does that make sense, so why does this help us think about it a little bit right suppose I suppose I have put it in some order right let us say that. So red has 0 .2 % of the data having class 1 right let us say suppose something like this no need not be why should they be I am just looking at each fraction of the data points with color red that were class 1 more than one color no this one attribute that says color of the data point okay whatever is it the one attribute that says color of the data point right.

And that way that attribute can take 5 values red, blue, green, yellow or magenta right suppose it is taken color red I look at what fraction of those data points that have color red or of class one right suppose I find that there are 10 data points that I have color red and two of them are of class 1 then it is 0.2, so obviously so these numbers do not have to sum to 1, right because they are only for that right, now we can tell me what is a good place to split this okay before somebody asked me a question about what if it is exactly 0. 5 okay.

There you go oh 0.2, 0.3, 0.4, 0.45 and 0.55 see how we go about doing the lot of one thing good point yeah so yeah, so you know how to do this come on pick up pick a thing and tell me what you know the Gini index and you know the you know Gini index or information gain or something like the right all of you know that all miss classification error let us use miss classification error as the splitting criterion, so for me to find an optimal split I do not really have to consider R and Y going to one part right B, G and M going to the other side right.

So it will either be here or here or here or here or here I mean that is a really bad attribute to pick if it is here right, but it will only be left to right all right, so these are the only subsets I need to consider I do not have to consider all of the other subsets right it does not make sense you can intuitively see this here right, so since this fraction of the class one keeps going up right, so either you break here or here or here or here, so that is a heuristic for this right in fact with a little bit of thing you can show that for two classes right.

You will get the same optimal split right by using this method as you would get by exhaustively searching through all the splits I am seeing too many puzzle looks we did this decision trees day before yesterday if remember decision trees okay, so split points if I have categorical attribute split points are going to be like subsets of the values the attributes can take right, so a split points would be okay do I consider our red and green to one side blue, yellow, magenta to the other side that is good potentially a combination right.

So in this case I am saying you do not have to worry about all possible subsets all you need to do is after you have done an arrangement like this okay wherever you choose to split depending on the criterion you are using so wherever you choose to split right so everything to one side will form one subset ever thing the other side will form another subset and these are the only subsets that you need to consider while you are trying to find the optimal space you do not have to consider all the 2^{q-1} subsets okay fine.

IIT Madras Production

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

www.nptel.ac.in