NPTEL

NPTEL ONLINE CERTIFICATION COURSE

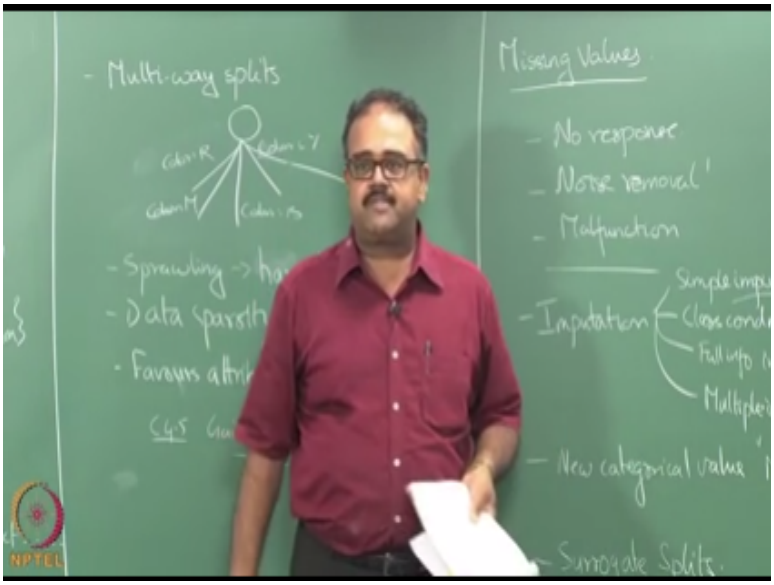
Introduction to Machine Learning

Lecture 45

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

Decisions trees – Missing Values, Imputation, Surrogate Splits

(Refer Slide Time: 00:15)



Missing values right, so what I mean by missing values? Some classes that are a different thing we come we will talk about class will much later. Suppose I have again let us go back take a real scenario right, you are filling in some survey questionnaire and then I am going to take all the data from you and then I am going to build a decision tree that allows me to predict whether you will buy a machine in my or buy a new computer from my shop or a new TV from a shop or something right and then the next somebody comes in and I am supposed just look at you and say okay he says he is going to buy a computer he is not right.

So I should be able to classify people like that but then when I fill in the survey I would just not answer some questions. So i will have a data point right so i am assuming that x is down from RP right or whatever the space that I am drawing x_i from but for each x_i assume I already know the values x_i to x_{ip} and so far we have never talked about the case where some of these might be unknown, some of this might be missing they might be missing for a variety of reasons right. So one could be right there could be a no response in a surveyor something right the other could be could be due to noise removal.

What do I mean by that? I look at my patient record data I find that somebody has a temperature of 223, so obviously some noise there right do I make it 22 or 23 when or 22.3 right nothing seemed straight right, I will tell you what scale it is in but still the still does not seem right so what do you do just remove it okay. So let us just assume that the nurse did not record the temperature of this patient right, so you remove noise from your data you might lose some attributes hi everything else about the patient I just do not know whether he was running a fever or not when he came into my clinic right.

So likewise you could just not have recorded it right that is equivalent to no response, so that guy messed up might have come with alike a bleeding right hand with it is just hanging off a wrist or something and you are not going to say hey first get this temperature and put it in there I do not want any missing values right, so this thing says it might not just get recorded you know so those kinds of things are there is an equal until no response, all right so anything else yeah exactly that is what is a malfunction right.

So it just that you might be recording sense of data from somewhere and the sensor just turns off for a while it may be it over heated or something went wrong and just for just a while you do not see any of this data being recorded, so the variety of reasons why you could have missing values in your data in fact if you work with real data right more often than not you will have significant missing values. In fact when I work with some data I have had cases where people have given me data where some attributes were missing in more than 80% of the data point's right.

So what you do in that cases remove attribute itself okay, you mean I shall not worry about the attribute because you are not going to be able to use it in any practical setting right so we just removed a trigger itself, but in other cases if it is missing only in like 5% or 10% of the data points you do not want to throw away that data point like throwing away 20% of the data point is

still a big thing right and yes and you do not want to remove the attribute also because it is available in 80% of the data points and you do not want to throw the attribute away you do not want to throw the data points away right.

There are two things you can throw the column of a can throw the row way right if it is missing in more than 80% you can directly throw the column away, but it is somewhere in the middle right summer small numbers then you do not know what to do throw the column or throw the attribute throw the rope, do not both exactly. So there are lots of different ways of handling this missing, very the statisticians have studied these new ones right so they have come up with many techniques for handling missing values and why am i bringing it up while we are talking about decision trees and not other classifiers.

Because there are some techniques which is peculiar tradition trees which are not available for other classifier. I am going to talk about all of these right so that I mean in general also you could use some of these techniques the first one all right so we will give it a fancy name called imputation. That imputation is essentially filling in a value for the missing attribute right, so how do you fill in the value for the missing attribute, you the mean the simplest thing is to do the mean you could do a regression on the attribute right.

So you could regress on the attribute in fact what is the best way of doing regression on the attribute? You should do it in a class conditioned fashion, use the class also because you are talking about the training data here right, and so use the class also as in part of your regression or part of your averaging. So what you can essentially do is okay to take all the data points that are of class 1 and use those to predict the value of the missing attribute for in that set of data points right suppose I have like a hundred thousand data points and let us say thousand or of class 1and of which 4 of them are missing some attribute 3.

Let us take those thousand data points and I can i will do a regression and predict for those 4 data points i will use the remaining 9996 data points as my training data and fit the curve and now I can predict what that one missing point is, so why is this kind of conditioning on the class useful? You use that feature glass, so if there is any kind of variation right the correlation between that feature in the class right this will help me preserve it right. If I am going to do this across the entire 100,000 data sets I lose the correlation I will lose the effect.

At least for these attributes it will get polluted right but this way I will be able to retain it, do not lose anything you do not do anything by doing it this way if that is correlation you actually preserve it if there is no correlation you do not lose, anything sorry Victoria was asking just saying what if there is no correlation do not lose anything by doing it this way right. So this is imputation the different ways of doing imputation you can use the mean you can use the class condition mean right you can use regression for doing the imputation and there is something anymore complicated technique.

That something called multiple imputation on using the regression for doing the imputation is also called full information imputation 8, only the statisticians have been at it for awhile right, so they have all kinds of the full information imputation is it because you are using all the known attributes, for predicting the unknown attribute right we are doing the mean you are only using that attribute in the same attribute in other data points and multiple imputation is a little weird thing. So what you do is you use all the data that you have right and setup a probability distribution over the missing attribute values right.

Like I said I have 996 data points in which that attribute is not missing right I will use that and figure out okay for if for red what is the probability for blue what is a probability for green what is a probability for discrete see there for continuous values i have to pick some distribution, let us say I pick precaution and else okay I will find what is a mean and the variance of the Gaussian that will predict the missing attribute value okay. Now what I do is I draw samples from this distribution and use those samples to fill in the missing values.

I will get one data set again other set of samples and fill in the missing values i will get another data so that is called multiple imputation. So i can create multiple copies of the data point by repeatedly sampling from this distribution and in some cases this has much better variance with much lower variance than using some of the other method. So even though this entails significantly more computation okay, so imputation is one that is another handle this I just introduced a new value for the variable right and I will call it missing.

Why would this be useful? exactly so there might be some kind of systematic reason for which the data goes missing and if I instead of trying to somehow guess what the value should be if I actually pay attention to the fact that it went missing right that would be useful, so I did not see who said that okay yeah, so infact it is actually a very practical practically useful thing it because

quite often the reason it goes missing is that is a specific reason for it, and you can in fact the fact that it is missing might be predictive of and you know.

So how likely is my patient to recover the temperature reading is missing, so those kinds of things, so use something called surrogate splits, so what is a surrogate split? okay so surrogate splits actually a server slightly different function, it works for imputation and this can be used during training itself right but the circuit splits thing we typically use during testing you can also use it during training is suppose. The basic idea is this for every attribute right that I have I will try to pick another attribute okay, that tends to split the data in the same way right.

Suppose let us say again let us take the same example I have 100,000 data points I split on attribute say 3 can I get two groups right this has says 70,000 data point this as another 30,000 data points. I split on some another attribute let us say four okay again I get two groups one has 68,000 data points other has 32,000 data points and not only that it turns out that the intersection of the 70,000 and 68,000 is something like 65000, on the intersection of the other two is something like 25,000.

So essentially three and four give me more or less the same splits, we are finding correlation we are not really reducing it here, so we are finding correlation what we do is if we have selected attribute 32 split on our tree and then we suddenly find that attribute 3 is missing in the data point we just split on an attribute 4 and behave as if we split on attribute 3 and go on. So this is what it means a surrogate right it is like putting proxy right, so I have attribute for can put proxy for attribute 3 and then he just continued working with your tree right.

So that is essentially what circuit splits up and it does is it exactly finds, that right it actually looks at correlation between the attributes and tries to exploit that okay, so as you can see that imputation right and adding this new categorical values could work with any kind of classifier that you are working with right. As long as you have a way of handling categorical attributes it is just one more value that you are handling, while the surrogate split something very specific to trees right likewise we are going to look at fragment which is also something very specific to trees.

So this is a little subtle so what I am going to do is the following right, so I come to a point I am going to make some query I am going to make this x3 <5 that is a query that I am going to make

it is a variable x3 <5, that is a query i have to make at this point in the tree right and what do I find my data point does not have x3. That is for categorical drive talking about categorical attributes, so it will be like okay I am going to RM V YG missing like that that or if i am going to do it in two subsets I will put missing into one of the subsets right.

But suppose this is there x3 <5 so what do i do x3 is missing right what I do is I look at that all the data points for which x3 was not missing okay, I will see what fraction went down this way let us say oh 0.6 went here 0.4 went here, so what is 0.6 all the data points that did not have x3 missing 60% of those we are < 0.5% of those had $x_3 > 0.5$, so now what I do is I am looking at one data point right one data point that came here I am going to split it into two right, so it is going two point six of the data point is going to travel down the left and 0.4 of the data point is going to travel down the right.

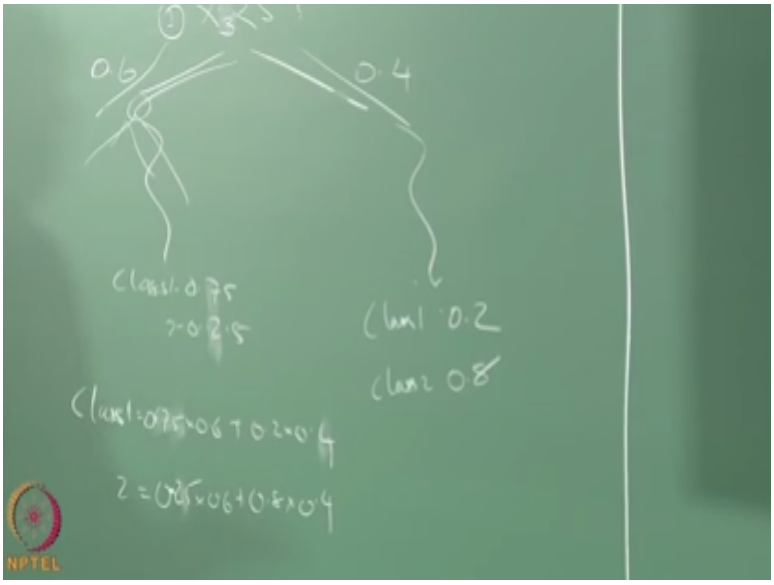
So what I do is it is essentially I am actually letting the data point travel all the way reach a leaf right and the leaf is going to make some prediction, so some probability it is class one some probability risk class to some probability is class 3 right. So the 0.6 part will make one prediction the 0.4 paths will make one prediction, I make a weighted combination of the two predictions and I output that as my final is it is seems like quantum mechanics name. So let us say I go down I finally reach here and I say its class 1 with probability I do not know 0.6 and class 2 with probability 0.4 and this one winds down somewhere.

And I say this is class 1 with probability 0.2 and class 2 with probability 0.8 right, so overall think that i will report is the probability of class 1 is okay, that makes sense no this is maybe this is a bad choice, do that make sense I am using it only once right. So what is the meaning of saying 0.6 of the data point goes down this side is essentially I will go all the way down and I will say that finally I will use the 0.6. I am not using the 0.6 anywhere here and these telling orders the semantics of saying 0.6 goes down this way.

So the reason we are carrying this weight along is at some point further down the line if I have another missing attribute and I decide to split it I have not been splitting one I will be splitting only 0.6 right, so this can get weird right, so I can have a data point which has multiple missing attributes traveling down more than two paths it will reach multiple leaves and then I eventually combine all the leaves so this is called when I am calling this fragmenting method right.

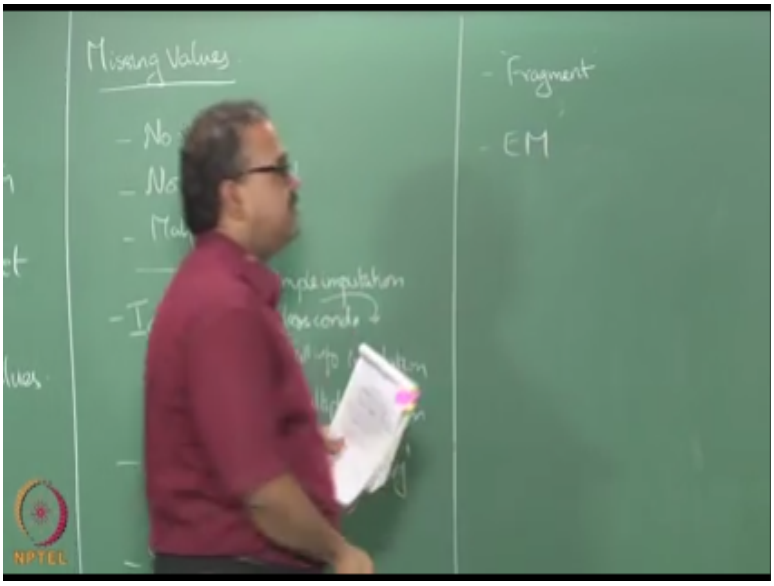
Again this is pretty unique to trees so if you think about it this is somewhat similar to doing multiple imputations. Of course the whole idea is to use training data to make a prediction on this data point right the whole the subject is predicated on using the behavior of other data points to predict output.

(Refer Slide Time: 20:27)



Of new data right, so it should not matter right.

(Refer Slide Time: 20:34)



So the last way of handling missing values is something called am expectation maximization right and it is going to keep cropping up all over the place as we go along but we will do it we will actually formally do with deal with expectation maximization much later. So just be aware that when we look at EM this is one of the applications of am okay and Link missing values I am not going to get into this is a pretty involved thing and in fact if you think you have been having difficulty with any of the concepts we have covered so far in the class you will not seen anything yet right so am is the one thing which everybody struggles with when they look at it first time so we will come to that later.

IIT Madras Production

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

www.nptel.ac.in