Artificial Intelligence for Economics

Prof. Adway Mitra

Artificial Intelligence

Indian Institute of Technology Kharagpur

Week – 03

Lecture - 11

Lecture 11 : Decision Tree and Random Forest

Hello, welcome to this lecture on the course of Artificial Intelligence for Economics. I am Advay Mitra, an Assistant Professor in the Indian Institute of Technology, Kharagpur. So, today we are beginning our lecture 11 of this course, the title of which is Decision Trees and Random Forest. So, like in the last lecture we discussed that concept of clustering which is an unsupervised learning. Now, from today we are moving to supervised learning. So, basically we are considering how to learn from data that is given data or which are basically which is basically past observations, how we can draw insights from past observations and use them in future to make our future decisions that is the like the main aim of this for the this set of 5 or 6 lectures which will be dealing with.

So, the which concepts which we are going to cover in today's lecture first of all discriminative features. So, we in the last lecture itself we talked about feature vectors that is every observation is going to be represented as a vector of certain features which we will choose. Now, among these features which of these features are discriminative that is they can be helped for classification purposes. Then we will discuss the concepts of entropy and information gain and based on this we will come to the algorithm for classification and regression trees and finally, we will also discuss the concept of random forest.

So, let us start with a simple application which is rooted in economics. So, like this is something which of we often come across when we are using various these retail websites such as Amazon and so on. So, let us say we want to buy a camera. So, when we are buying a camera we look at different features or different that is we may in we may initially give some kind of search string in the retail website based on which we will have a lots of suggestions. So, now, we look at any one of those suggestions and we have to decide whether we will buy it or not.

Now, the if you see it from the perspective of the search engine, then the search engine will of course, want to make sure that you as a customer buy the products that it is

recommending and you are also satisfied with it. Because if you are not satisfied, then you will probably not come back to the retail engine for the for at the future shopping. So, it wants to satisfy you. So, now how do you how does the retail engine understand whether you are satisfied or not I mean it is not a local shopkeeper to whom you can just go and complain that the camera you gave was bad. So, then how do you do it.

So, then how do we let as customers how do we let the retail engine know whether it did well or not in its recommendations, the answer is through this kind of feedback. So, the feedback rating is like it helps the retail engine or the to understand how satisfied the customers are with their service, so that they can improve their service. On the other hand they will try to do their best by trying to give you the products which you will which you will like. So, you for a given product for any given product it will have to first guess whether you will like the you as a customer are going to like that product or not or rather if you were to give a rating to that product how much rating would you give it out of 5. So, it will it needs to the retail engine needs to predict what your possible ratings for this product will be.

If its guess is that your you will give a rating of let us say 4.5 to that product, then you will it will probably show it to you high up in the listing of products. But if it predicts that you are going to give it a rating of let us say 3.1 or 3.2 then it may not show you the product or even if it shows you it will be towards the middle or towards the bottom of the list right.

So, now how does it how can it guess how what rating you are going to give to that product. So, it probably your rating probably depends on the features. So, then it has to understand which are the feature that you like most. So, for this it will have to delve into the past shopping history of you as a customer. So, let us say that the like you are a camera like a camera like a camera buff who have purchased a large number of cameras from this same websites.

So, for each of the cameras which you have earlier purchased its features are stored in the in your purchase history and the ratings which you have given they are also stored. So, now the machine has to analyze your past purchase and rating data to analyze which are the kinds of cameras to which you are getting high ratings. So like for example this camera you gave a rating of 5 while this camera you gave a rating as 2. Now there are many differences in the features of this camera their companies are different their colors are different their resolutions are different video frame rate price everything is different. So, then we out of all these features which are the features that is determining whether you will give a high rating or a low rating to the camera.

Is there only one feature on which your rating depends or are there multiple features on

which your rating depends. If like in either case which are the features. So, like the aim is to understand the features which are important to you as a customer. So, if we now if we do not have customer specific data then of course, like it will have to make a like a generalization over multiple customers. So, they will have to like they will consider the rating data and I mean the purchase and rating data of customer.

many customers and from that it will have to make a general statement about which kinds of features are identified by customers in general or it may even try to identify certain segments of customers. In the previous lecture on clustering we had mentioned one application known as the market segmentation. So, like it may now different sets of customers may have different kinds of behaviors while rating the products that is because they may have different kinds of choices. There may be one group of customers who are interested in feature 1, another group of customers who are interested in feature 2 and 3 and so on and so forth. So, like it may make sense to like segment the customers by some clustering algorithms in some way and then build a model specifically for each set of customers.

But anyway, so let us say that for one customer or for a set of customers, this is the rating data. So, then the question arises that which are the features on which the rating depends. So, let us go feature by feature. Let us say the first feature is the company. So, the company there are two possible companies in the past cameras that have been purchased C_1 and C_2.

So, like if we see the rating performance for company C_1. Let us say that total 100 cameras have been rated by this user and like out of which 54 were from company C_1 and 46 from company C_2. So, this is the we see the breakup of the different ratings which the user has given. So, we see that like the cameras from company C_1 seem to be mostly getting low ratings like 1, 2 or 3 very few of them get ratings like 4 or 5. while for company 2 or company C_2 it seems to be just the opposite.

So, we may be wondering whether company itself is the determining factor. So, it may be that this customer has some for some reason has some bias towards this company the company C_2 and it gives high ratings to the products from this company. Or it may also be that there are certain features which are common to company C_2. So, and those features are like by this customer. So, company name is just a proxy for those features either possibility is there.

And now let us consider another feature the color. So, let us say that there are among the cameras that have been purchased in the past there are mostly two colors black and white. Out of the 100 cameras 70 were black and 30 of them were white. So, now if you look at

the break out of the ratings given to the black cameras. or to the white cameras we do not see any particular trend that is it seems to be fairly even.

So, if we look at the company feature of a given camera we see some signal that it is does seem that like cameras from one particular company seems to be getting more ratings than the examples from the other company C_1. But in case of color there does not seem to be any signal like this. So, now if we want to just calculate the probability. So, let us so there are 5 ratings are possible. So, let me just say that the first 3 ratings 1, 2, 3 are called low and the 4 and 5 they are called as high ratings.

So, let us say I want to calculate the probability that given that the particular example is from company C_1 what is the probability that it will get high rating. So, I can calculate it based like by interpreting probability as the relative frequency that is the frequentist view of probability. So, we just simply consider how many cameras belong to company C_1 that is 54 and out of them how many of them had high rating. So, the answer in this case turns out to be 5 plus 6 equal to 11. So, 11 by 54 which is around 0.

2 that is the probability of getting a high rating for those examples which are for company C_1. But for the examples from company C_2 the probability of getting high rating we can calculate in the similar way and that turns out to be 25 by 26 by 46 which is like 0.55 which is definitely much harder. So, if we like so, like we can so, we can understand that just by looking at the company feature there is a significant difference between the probability of getting high rating. So, if our task is to predict the rating I mean the this predict whether the rating is going to be high or low then the company seems to be a good feature.

Because if the company is C_1 then the probability of getting high rating seems to be very less But when if it a company C_2 then the probability of getting high rating seems to be considerably higher. But if you consider say price in fact, you want to see whether the price is less than 300 or not. So, in that case the probability of getting a high rating for the if the price is less than 300 turns out to be something like 0.36 in again you do the same kind of analysis you identify that. There are 70 cameras whose price was less than 300 and among them 25 of them had got high rating.

So, this probability you calculate to be 0.36. While for the remaining cameras which are most more costly than 300 among them the probability of getting high giving high rating you also turns out to be 0.4. So, there is very less difference in this case between 0.

36 and 0.4. Whether the price is less than 300 or not does not seem to be making much of a difference in determining whether the rating will be high or low. Now instead of 300 let

us consider another threshold for the price let us say 500. So, now we find that there are 90 cameras whose price was less than 90 sorry less than 500 out of them 35 got high ratings which is like 0.4 or 40 percent. While the more expensive cameras which are even more expensive than 500 only 2 percent sorry only 20 percent of them that is 2 out of 10 they got the high rating.

So, now in this case again there is a clear difference of in the probability of high rating between this feature and this feature that is whether the price is less than 500 or more than 500. But the second number that is this 0.2 is a little bit tricky because it has been calculated based on only 10 examples. So, like when we are considering or when we are calculating probabilities based on relative frequency there is a concept known as the law of large number which like basically says like which which is the basis of this frequency statistics. Now, that comes or that becomes applicable only when the number of examples is high enough.

So, in this case we can argue that 10 examples is not high enough. So, what we are seeing out of 10 examples that is if out of 10 examples only 2 have been given high ratings that may be simply an aberration. If instead of 10 if we had let us say 50 examples from this which cost more than 500. the maybe we would see a much more than 20 percent chance of getting this high ratings right. So, although the their stars seem to be a difference between the probability of high ratings in the case of this feature, it is a big on weak ground because it is calculated on the basis of very few examples.

Now, the question is so, what the different features we are seeing that in some cases there is a difference between the probabilities of high ratings in some other cases there is not such no clear difference between them. So, that means, that is some of the features we can say that it is more discriminative that is like by looking at that feature we can like clearly distinguish between the possible distributions of the final rating or the final target which we are trying to predict. So, we may be interested in quantifying the discriminativeness of different features. So, one such measure of discriminativeness is called entropy which is calculated like this. So, the let us say like the different values of the output with they have different probabilities $p_1, p_2,$ etcetera for a particular value of the feature.

So, now so, there it defines a probability distribution. So, let us say k values are possible for the output variable like if it is ratings then 5 values are possible. Now, we are like they all of them have some particular probability which as we saw earlier we can calculate as them those probabilities as relative frequencies from the data. Now, using those probabilities that is $p_1, p_2, \ldots, p_k$ we can calculate the entropy using this formula. It turns out that for those probability distributions where the p's are all close to each other

that means, what that means that no class is significantly more probable than the others.

In other words it is like the prediction is in a confused state. So, in this case like we it will turn out that h is the value of the h which is entropy is quite high. On the other hand like if we consider a situation like this where the probabilities of the different values of the output variable like it is a highly skewed distribution that is there is one value which is highly probable and the other values are very less probable. In that case that means, there is a higher surety that is something like this kind of a situation right that is in like we see that one class is or this or this either of this. So, it means that in this case the probability of low is very quite high and the probability of high is quite low in this case it is just the opposite.

So, like we can say that in this case the entropy will be quite low. So, like for this distribution that is 0.9 0.1 the entropy works out to be 0.

5 which is quite low. In the first case the entropy works out to be 1 which is significantly higher. So, that means, like if the distribution is if the predictive distribution is more sure of itself is confident of itself that is there is one particular value of the output which has high probability in that case the entropy will be low. But if it is not sure of itself that means, all the possible values have seem to be having more or less comparable probabilities in that case the entropy will be high. So, then entropy seems to be a good measure deciding whether a particular feature is discriminative or not. To be more precise a discriminative feature is some like if I consider a discriminative feature, then I will get a more confident classification or prediction of the output variable.

So, it is entropy will be low on the other hand if I consider a less discriminative feature in that case the induced entropy is going to be high because the prediction will not be very confident right. So, now, let us see the entropy like how to calculate the entropy. So, now, let us so like we consider the quantity known as the information gain what is information gain? Suppose like if the original data set when all the examples were together. So, like some of them had different class levels. So, based on those class levels irrespective of their features we can calculate what is known as the original entropy like in this case for example.

like if you just calculate the I mean the all the 100 examples. So, you can see that out of the 100 total number of examples which got the rating of 1 is 21. rating of 2 is 24, rating of 3 is 18, rating of 4 is 20 and rating of 5 is 17. So, like in the overall data set without considering any feature the probability of getting a particular rating let us say 4 will turn out to be 20 divided by 100 that is 0.

2. Similarly, I can calculate the probabilities of all the other ratings also. that is the

original based on those we can calculate the original entropy. Now, since more or that is all these numbers or all these probabilities of the different possible values of the rating they are comparable to each other.

So, we just saw 0.2, 0.18, 0.24 and things like that they are all numbers which are comparable to each other. So, the original entropy that way is expected to be quite high. Now, suppose I split the data set on the basis of any particular feature like which we have considered so far. So, we will see let us say two sets of the examples one having one value of the feature and the other value the other value of the feature. Now, within those two sets of the features I will calculate the entropy and hopefully the in this case the if the feature is discriminative then in one set one type of level will dominate in the other set another kind of level will dominate.

So, that is the in these two subsets which are obtained after splitting the original data set the entropy hopefully will be quite low because the prediction will of the label with respect to that subset will be quite confident. That is one probability one class will have high probability the others will have low probability that is when the entropy is low. So, original entropy was high after the splitting the entropy is less. So, the difference is known as the information gain.

So, this is the formula for information gain. So, original entropy minus split size 1 times split entropy 1. plus split size 2 times the split entropy 2. So, what is split entropy 1? So, as I said like we split the data set into 2 or more parts like on the basis of a feature. So, that is like I can either consider 2 possible values of the feature or I can even consider 3 or 4 or 5 or more values of the feature also. So, accordingly I divide the data set into those many subsets and I calculate the entropy of each of those split parts of the data set and I also multiply that with the size of the split.

So, now why is this important? So, you will remember in this case I this particular thing I was not I was not giving much importance because it was only on the basis of 10 examples. So, that the size was low. So, that is why I include the size also of the of the split also in the information gain formula. So, even if there is one split for which the entropy is very high or very low, if the its size is very small then I will not give much importance to it. So, now in this case we see that the data set we are considering if we split it with respect to company C_1 or company C_2, then we get a significant that we get a decent reduction of the entropy that is the we get a high information gain.

So, compared to the original entropy the split data sets they seem to be having significantly low entropy like the original data set had an entropy of 0.66 the split data set is having like entropy like at least this one is having a significantly lower entropy of 0.51.

But if I consider the splitting criteria as whether the price is greater than 300 or not, then the information gain turns out to be 0. That is there is absolutely no change in the entropy from the original data set of all 100 cameras to these two split data sets or one of them may have 70 cameras and the other has may have 30 cameras.

Similarly, on the basis of the splitting with respect to the whether the price is more than 500 or not. In this case, we see that there is actually a reduction of entropy for this split 2 where we are considering only 10 cameras. But since that is only 10 out of 100, so the this split size is making sure that. the this the reduction of entropy obtained from this part is actually not a very significant reduction that is from like the original entropy is 0.

66. If you consider only those 10 examples their entropy is 0.5 which is significantly lower, but because I am multiplying it with the size of the split which is quite small this is not really bringing down the overall entropy by much from that is the information gain which I obtained is just 0.01. So, we see that compared to these this feature that is the company feature this seems to be having a more information gain. So, I can consider that to be a more informative feature and I can split according to that. So, what is the algorithm then I choose the feature which provides the most information gain and I split the data set according to this that feature.

So, that feature has in this case has two values C_1 and C_2. So, I consider the examples for the value of C_1 and I separately consider the examples with the for the value of C_2. Now, if you consider this subset of examples for C_1, we see that among them the low rating dominates. So, if I like if a future example comes for which I am I know its feature and I know that its belongs to its feature is I mean its company is C_1. So, then with already with high confidence I can predict that it is going to get a low On the other hand if it is which company is C_2 then I bring it here. So, like in the subset of examples where the company is C_2 I see that 26 out of 46 have high ratings.

So, we can say that there is a decent chance that the any new example which is of company C_2 will also get high ranking. So, I will note that in this case I am more confidence that is 43 out of 54 in this case I am less confidence that is 26 out of 46. Now, like just these 100 examples if I which are used for this training if those 100 examples are again used to like I mean if we try to predict their levels according to this algorithm we will achieve an accuracy of 69 out of 100. So, now like instead of stopping at just one feature I can actually continue splitting these datasets further based on the remaining features. So, initially I had the 100 cameras the full dataset I split I identified that the company C_1 is I mean the company is giving the biggest information gain.

So, I split it into two parts one for company C_1 another for company C_2. Now, again this

set of 54 examples I may want to split according to some other features. So, again I calculate the information gain with respect to the other features apart from company and then maybe I find some feature for which I get a good information gain. So, I split it further according to this. This one also I may do the same exercise that is among all the remaining features I see how it can be split and it may turn out that there exists a split which gives me information gain. So, I just so, like every split I just look for information gain and if I do get a decent amount of information gain.

Now I may have some threshold about what how much information gain or reduction of entropy I can consider as decent. So, if that is the case then I split the data set further if not I just stop there. So, this is the decision tree algorithm it has two parts the training phase and the testing phase. In the training phase we identify the feature that results in maximum information gain and we accordingly split the data set. Now, in the in each split data set I identify if any feature can result in further information gain and if yet then you stop further you split further and if no that is no feature is able to give you any further information gain then you stop it there.

Now, so you get leaf nodes in the decision tree which cannot be split any further. So, like as is obvious in this from this figure each also these are the leaf nodes of the tree. Now, each leaf node has like represents a split of the population and it has like some examples in it. Now, on the basis of those examples which are present in that leaf node we can calculate the prediction for that leaf node whichever class level is most frequent or which are the mode of modal class level of that node we use it to classify any further.

So, once this decision tree has been built. So, let us say a new test example comes whose company is C_2 and its price is 350. So, first I check the according to a decision tree I check the company. So, I find that its company C_2.

So, I bring it to this intermediate node. Next, I look at its price. So, there are two options either that is less than 500 or greater than 500 its price is 350. So, less than 500. So, I come to this leaf node. Now, in this leaf node I find that there are 40 examples 25 of which have got high ratings. So, I am 25 by with confidence of 25 by 40 which is about 62 percent confidence I predict that this camera is going to get high ratings and so on and so forth.

So, this is a decision tree algorithm its advantage is that it is easy to interpret it is easy to classify a test time you just have to look at one the different features one after another and predict the level. Its disadvantage is that no optimal solution is known and this idea of splitting the population on the basis of information gain it is just a heuristic. And, it can like it is not known that if this will create the best possible decision tree. And, if the

tree keeps on growing deeper and deeper like it may cause over fitting also that is that is to say the data the model which the algorithm which you will learn the decision tree which you learn may be doing very well on the data for which you trained it. But, if you give data from outside the training set then you may not get good performance that is what is known as over fitting.

Now, the same thing we have this thing we have considered for classification, but for it can be extended for regression also. So, supervised learning has two parts classification and regression. So, regression we see when the say when the target variable is a real valued as opposed to a categorical valued. So, let us say we want to predict the average rating which can be any real number between 1 and 5. So, in this case also I can use the decision tree the only thing is that in that case we cannot calculate the entropy because we are like there are infinite number of choices.

So, I cannot calculate the probabilities of each of them. So, instead what I do is I go for variance. You just like entropy basically gives you a measure of disorderliness same is given by variance. So, low variance means more orderly population. So, that is so like we our split criteria may be instead of information gain or reduction of entropy we may change it to reduction of total variance across the different things. Now and one more thing which we can do is instead of considering a single decision tree we can consider an ensemble of decision tree that is build a set of classifiers which are based on slightly different training criteria.

So, in between this case what we can do is like we are doing this based on so many features. So, let us say that all the observations have let us say 100 features. Now, for eachwe define let us say we will train 10 different models, but for each of the models we will focus on a different subset of features. Say model 1 will look at some some 15 randomly chosen features, model 2 will look at some other 15 randomly chosen features and so on and so forth. So, each models may be slightly different from each other because they will be focusing on different sets of features.

So, now we have got different models which are called as the decision trees. So, given any new example this is the first tree will predict one level for it, second tree will predict another level for it, third tree will predict third level for it and so on. So, now all the trees will make some predictions you simply take a vote you see that each level is getting predicted how many times by the different trees and accordingly you make the prediction. So, this is the ensemble based decision tree which is also known as a random forest. What is forest? A forest is a collection of trees.

So, in this case let us say like 8 features are there. So, this is the input. So, you have three trees like this. The first tree focus on these four features, the second tree focus on these

four features, the third tree focus on these four features. And let us say they all have their own class predictions. Say the first one predicts the class level to be C, the second classifier second tree predicts the class level to be B, the third one predicts the class level to be C.

So, C you can say has got two words, B has got one word. So, you predict it as C. So, now why is this useful for economics? So, it is useful for consumer behavior analysis where we may want to make customer specific make predictions as we already saw in the examples which we started off with. Then there is market basket analysis which where the task is to analyze how different items in a market can or in a can be merged together that is instead of like selling individual products if you can somehow merge products together and sell them that may be beneficial to the customer because they will not have to search the things together. But then for that you need to predict that which sets of items are usually purchased together.

So, that is much market basket analysis. So, that is can be solved in this process. Then loan default prediction. So, suppose a bank has many customers applying for loans. Now the bank needs to look at different features of that customer and predict whether they will be able to repay that loan or not. So, decision tree provides them a nice sequential approach in which they can look at different features of the customer in a sequence and then finally, come to a conclusion whether the loan can be given or not. Similarly, fraud transaction analysis or public policy impact assessment also we can like be the broad idea is that you can look at a sequence of the features to come up with the final prediction.

So, to conclusion is that like in any classification problem some features are more useful than others in predicting the target values and entropy or information gains are used to evaluate the discriminative power of the different features. Now, decision tree specifies a sequence of features to be examined based on which the classification can be done. The decision tree can be extended for regression as well by using variance instead to identify the feature discriminativeness. And finally, we can consider an ensemble of decision trees which vote that is each decision tree looks at a few features and makes a prediction you like you consider them as votes by the decision trees and make the final prediction based on that. So, this is known as a random forest and this random forest is like even though it is quite simple it is considered to be a very powerful and successful classifier in many practical tasks.

So, with this we come to the end of this lecture number 11. In the coming lectures we will discuss more algorithms related to supervised learning. So, all of you please stay well see you soon again bye.