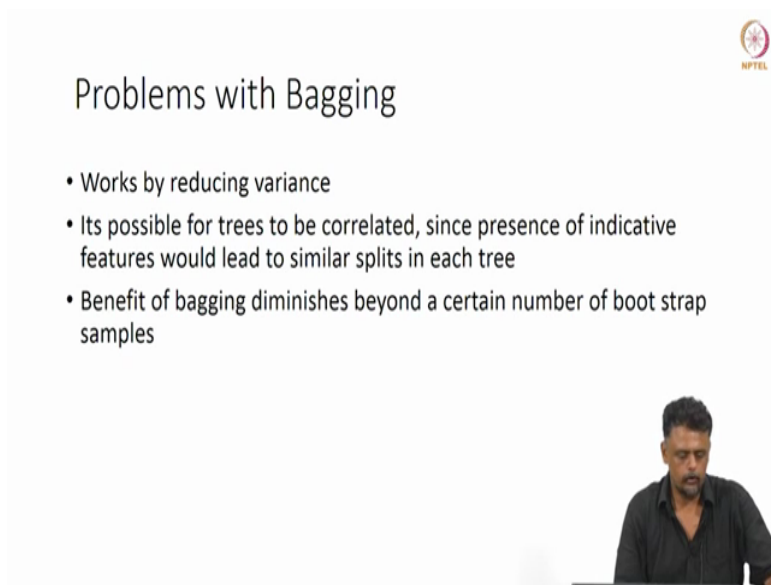



**Machine Learning for Engineering and Science Applications**  
**Professor Dr. Ganapathy Krishnamurthi**  
**Department of Engineering Design**  
**Indian Institute of Technology, Madras**  
**Random Forests**

Hello and welcome back, so in this video we will look at random forests so which is based on binary decision trees.


(Refer Slide Time: 00:22)





### Problems with Bagging

- Works by reducing variance
- Its possible for trees to be correlated, since presence of indicative features would lead to similar splits in each tree
- Benefit of bagging diminishes beyond a certain number of boot strap samples



So we saw that a bagging can be used or bootstrap aggregation can be used to improve the variance in when you are using binary decision trees this is by training a bunch of decision trees using bootstrap samples of your training data, ok. So because binary decision trees tend to over fit having a very large number of those the variance radius is reduced and it performs and improves generalization performance, ok.

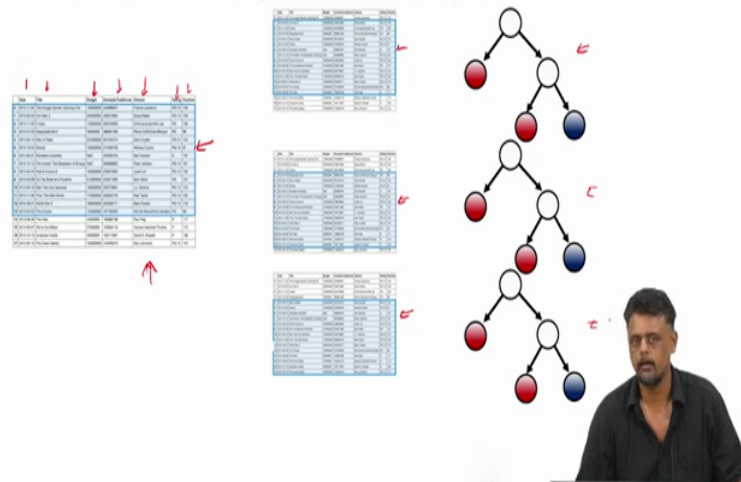
However the problem with bagging is that it is possible for trees to be correlated in the sense that if we have a very strong couple of strong features which always split first or the most are the most what do you call the best information gain is obtained by splitting on those features then no matter how many trees you average over that they are correlated, so beyond a point bagging will not reduce the error in your predictions, ok.

So this is primarily again to reiterate because there will be some more few indicator features or a bunch of strong features which lead to maximum information gain and averaging core correlate variables will not help reducing the variance.

(Refer Slide Time: 01:36)

## Random Forests

Grow decision tree from multiple bootstrapped samples



So are not a deal with problem we do something called random forests as I guess the name implies it is again a bunch of the decision trees but there is a difference from how we do with compared to bagging, so with bagging we grow decision trees from multiple bootstrap samples so this is here is our training data and the blue highlighted areas what we are choosing as the training data as part of the entity entire data set and the blue region is where we choose the training data from.

As we saw in the previous lectures this is just the movie recommendation database so the date of the movie, the title of the movie, the budget so you have a date, title of the movie, budget the domestic total gross, director of the movie, the rating in the runtime on given and so we are just choosing this the subset highlighted in blue as our training data ok, so what we do is we select with bagging we just selected with replacement no random samples from the training data.

So that is basically we bootstrap the training data to produce so if you have one set of training data we produce a let us say in this case three sets of training data from the original training data, so as you see the blue box keeps moving around inside this white table which is the actual total number of data sets available to you and with each bootstrapped training data sample you will fit a tree, a binary tree and we saw and we will, will not go into details but we saw last video how we for a per given test data set we will run the test data set to all the trees that we have trained using the bootstrap data and it will just average or in this case maximum voting for classification, an average for regression, ok.

(Refer Slide Time: 03:31)

## Random Forests

Grow decision tree from multiple bootstrapped samples

The slide shows a decision tree structure with nodes and branches. Some nodes contain red and blue circles, representing different classes or outcomes. The tree is surrounded by several data tables, likely representing bootstrapped samples. A blue line highlights a specific path through the tree, and red arrows point to specific nodes. In the bottom right corner, there is a small inset image of a man speaking.

So for random forests again it is the same principle that we will grow decision trees from multiple bootstrap samples with the exception that so with the exception that even the features will be chosen at random, so let us say we have let us say little this particular tree here which is chose a trained using this particular data set a bootstrap data set highlighted in blue we will not use all the features, so at the root node I have cut out this these two features here these two columns of features ok and when you come to the particular node here again I have cut out to some of the features so let us say this I left out the title feature again and grow the decision tree, ok.

So the difference between this bagging and random forests is that from a very top level view point is that when we grow these multiple decision trees from bootstrap samples not only do we bootstrap the there training data set but we also only choose a subset of the features available to us at every node in your tree, in every tree ok, so what happens so this kind of reduces the correlation between the trees that we train ok, so that is differentiable behind random forests.

(Refer Slide Time: 04:47)

## Random Forests



- Modification to Bagging
- Grow decision tree from multiple bootstrapped samples (Total M data pts)
- For every tree at every split consider only a random subset of features (Total N Features) i.e. say D features where  $D < N$  ( $D = \sqrt{N}$ )
- Learning on a random subset of features reduces the correlation between trees



So it is a modification to bagging wherein we grow decision trees from multiple bootstrap samples let us say we have M data points we will have bootstrap samples M data points and for every tree at every split we will consider only a random subset of features, so that say you have total of n features we typically use small n I will use capital n here that is fine, let us say D features and typically D is square root order of square root of n ok, so at every node in the decision tree in every decision tree you only choose a subset of the features.

So the learning on the subset reduces the correlation between the trees ok, the disadvantage behind using a tree like a random tree this way like is basically building a random forests from trees which are trained on bootstrap samples as well as randomly chosen features is that it (diff) it becomes difficult to interpret ok because every node even though we even if you grow all the trees to a particular depth every node might be split based on a different variable.

So it will be very difficult for us to interpret which variable gave the best information gain and things of that sort ok, however there is a huge improvement in performance compared to bagging because we are decorrelating the trees that we are training ok, so we typically would train hundreds of trees and as we saw in bagging for a regression output we will consider an average of all the trees outputs of the our trees and for a classification output we will do maximum voting, ok.

So this is a brief note about random forests again the fundamental principle is a decision tree just that how we train them and how we interpret the results or the only difference, thank you.