

Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture 15

Principal Component Analysis and Regression Applications in Agriculture (Contd.)

Welcome friends to this NPTEL online certification course of Machine Learning for Soil and Crop Management and this is week 3 and we are at our lecture 15. So, this is the fifth and final lecture of week 3. And in this week we are talking about the Principal Component Analysis and Regression Applications in Agriculture.

So, in our previous four lectures of this week, we have discussed, what is principal component analysis, what are the, how principle component analysis can be utilized for dimensionality reduction, how to calculate the principal components and then what are the score plots, what are the loading plots and what is biplot PCA biplot.

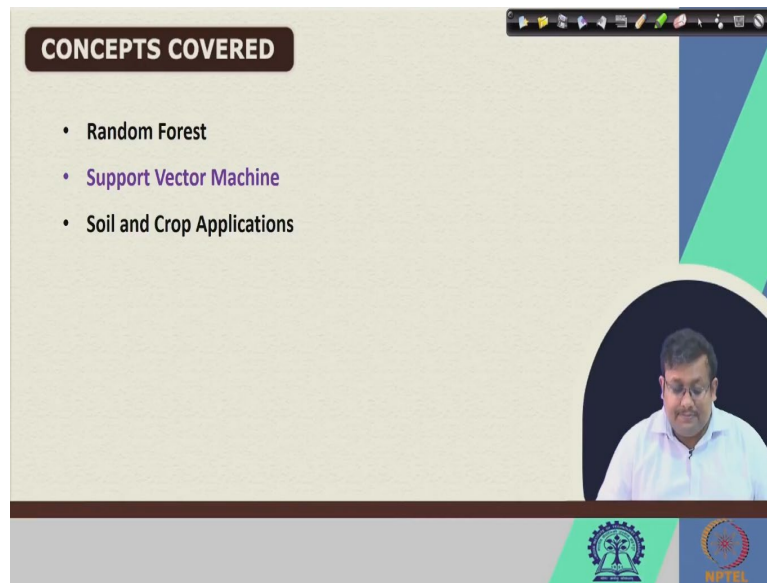
We have also seen, what is p c screeplot and from the pc screeplot, how we can calculate the number of important principle components. And based on the number of principle components, how we can go with the smaller, small, short, smaller dimension, I would say, for representing the features into pc base, space base.

And also we have seen the principal component regression, how to use the, what is the principal component regression, how we can use the pc's for principal component regression, we have seen. We have also seen partial least squares regression, how it differs from ordinary least squares regression of multiple linear regression and principle component regression, how we can calculate the latent factors or latent variables of PLSR, we have also seen.

And then we have also seen the the application of PLSR for different crop and soil properties. In our last lecture, we have discussed about the classification and regression tree, which is one of the important machine-learning algorithm. We have seen how to build a CART algorithm and then what is the recursive partitioning, what is pruning we have seen.

To fit the, to avoid the overfitting, we have seen the different types of impurity measures like entropy then Gini Index and how we use them for recursive partitioning in CART we have also discussed. And ultimately we have seen some examples of crop and soil application of CART.

(Refer Slide Time: 03:06)



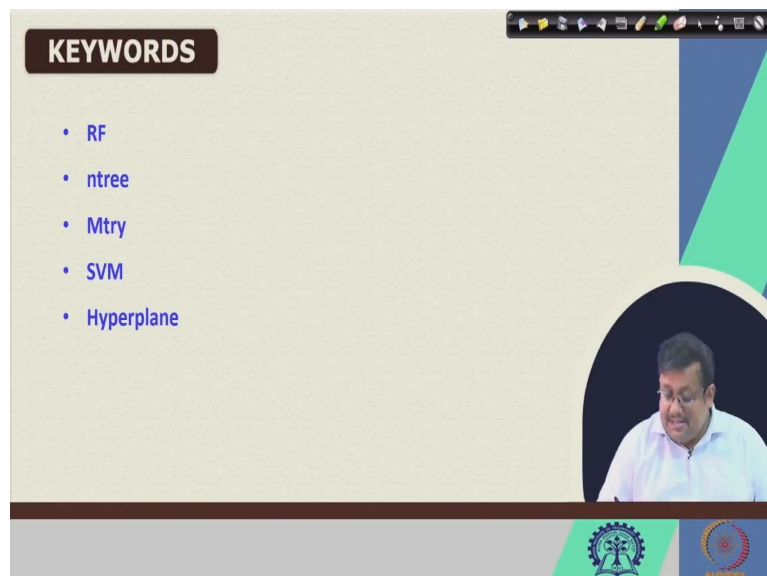
A slide titled "CONCEPTS COVERED" with a list of topics. The slide has a light beige background with a dark blue header and footer. A circular inset in the bottom right shows a man in a white shirt. The footer contains logos for IIT Bombay and NPTEL.

- Random Forest
- Support Vector Machine
- Soil and Crop Applications

So, today in this lecture, we are going to discuss two widely used machine-learning algorithms in soil and crop based research and these are Random Forest and Support Vector Machine or in other words Support Vector Regression.

So, today we are going to discuss in this lecture random forest, what is random forest, what are the features of random forest, how it differs from decision tree, what are the advantages of random forest and also we are going to discuss, what is support vector machine or support vector regression. And then we are going to see some examples of soil and crop application of both these algorithms.

(Refer Slide Time: 03:35)

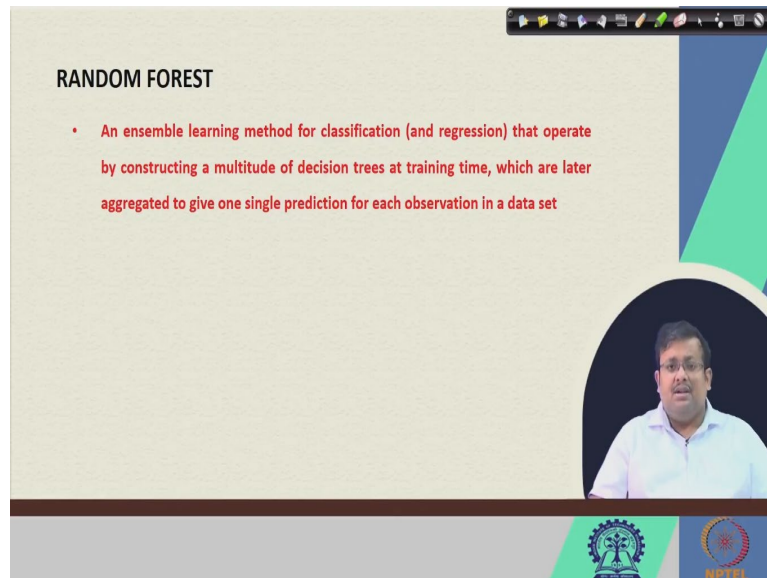


A slide titled "KEYWORDS" with a list of terms. The slide has a light beige background with a dark blue header and footer. A circular inset in the bottom right shows a man in a white shirt. The footer contains logos for IIT Bombay and NPTEL.

- RF
- ntree
- Mtry
- SVM
- Hyperplane

So, some keywords are there like Random Forest, then ntree, then Mtry, then SVM, Hyperplane we are going to discuss.

(Refer Slide Time: 03:44)



RANDOM FOREST

- An ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time, which are later aggregated to give one single prediction for each observation in a data set

So, let us see what is a Random Forest. Random Forest is basically an ensemble learning method for classification and also regression that operates by constructing a multitude of decision trees at training time, which are later aggregated to give one single prediction for each observation in the data set.

So, what happens in this random forest algorithm, we generally grow multiple trees, hundreds and even thousand number of trees, just like we have seen the discussion on the trees, so, these individual trees are being grown in the random forest algorithm. And then we either take the majority of the classified classes from those individual trees, by voting as our final prediction, or in case of regression, what we do, we take the average of the prediction from the individual trees.

So, you can see by using a multiple weak learners, we are getting a strong learner. So, random forest is a very very robust and I would say it is a very widely used method in the machine-learning domain and in many domain of science, this application of random forest has been widely accepted.

(Refer Slide Time: 05:19)

WHY RANDOM FOREST?

- Random sampling of training data points (bootstrapping) when building trees (ntree)
- Random subsets of features considered when splitting nodes [mtry, \sqrt{M} for classification and $M/3$ for regression]
- Forms lots of decision trees with random selection of samples and random selection of features.
- Provides the class of dependent variable based on many trees
- Random trees
- Many random trees=Random forest

Photo Credit: Skitterphoto from Pexels

The slide includes a video inset of a man in a light blue shirt speaking. At the bottom, there are logos for IIT Bombay and NPTEL.

So, this random forest, since we are also using this random forest algorithm extensively in soil and crop studies, we are going to discuss this in detail. So, how this, the question comes to our mind, why we call it a random forest? Remember that using the bootstrapping; we are doing the random sampling from the training data set.

Suppose, we have 300 points in the data set, so, we do the bootstrapping or bootstrap samples from the 300 data set. So, it is a random sampling of the training data points for building the trees and the number of trees which varies from hundreds to thousand, we define in r , by this parameter called ntree, ntree stands for the number of trees.

So, for building this individual tree, we take a bootstrap sample from the original data set, and this bootstrap sample, as you know it is a random selection with replacement. So, this is one level of randomness in this whole algorithm, sampling, random sampling of the training data points. Now, once we have sample randomly, then in the tree, we use the random subset of features, even there are features.

For each sample we have suppose, for all the samples suppose, we have 100 features or 200 features. So, again from the feature space also, we are selecting a random subset of features for splitting the node. So, this is basically denoted by this term in r called mtry. So, generally for classification problem it is a square root of M and in case of, if M greater, capital M is the total number of features, it will be square root.

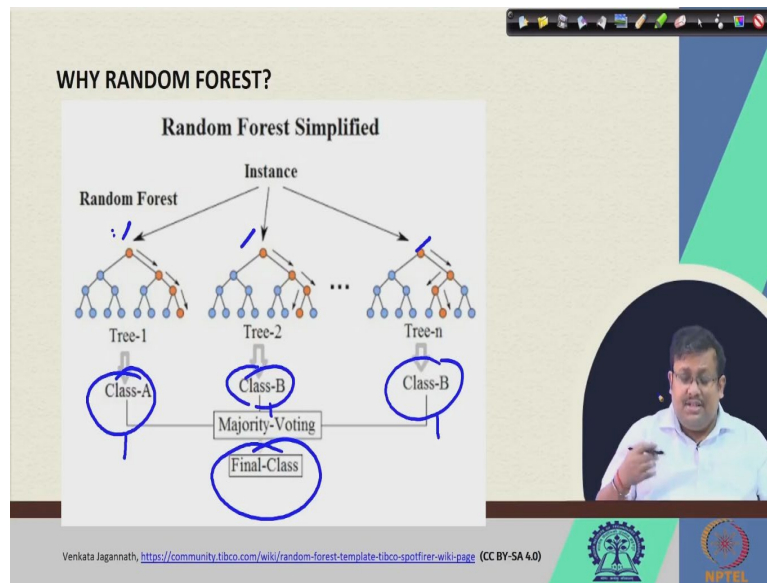
Generally, in case of square root of M in case of classification problem and in case of regression problem, generally, we take M by 3 number of features randomly. So, you can see here, we are, in two stages, we are sampling randomly. For first of all we are randomly sampling from the training data points for building the each of the tree and we are building number of trees.

And also we are, for building this individual tree, we are selecting random subset of features for the, while we are splitting the nodes. So, what happens? So, basically we forms lots of decision trees with random selection of samples and random selection of features. So, it provides the class of dependent variable based on the many trees and in case of regression trees, it is a random forest regression.

It gives us the prediction for these individual trees and then we take an average for that. So, you can see these trees are random trees, because we are sampling randomly bootstraps samples for building a tree and then the feature which we are using to for splitting the nodes, are also randomly selected. So, that is why this tree which are being formed are known as the random trees and when there are multiple number of random trees, hundred to thousand number of random trees, we call it random forest.

Just like in a forest you can see multiple hundreds and thousands of trees, so similarly here in random forest algorithm also, we are building these trees. And since these trees are random then thereby, therefore we are considering this as a random forest. So, this is why we call this algorithm as random forest algorithm.

(Refer Slide Time: 09:09)



So, a simplified representation of the random forest, you can see here, suppose, this is the data set and from this data set, we take the bootstrap sample and fit the number of trees, here, you can see this is tree 1, tree 2, tree 3 and up to tree n. So, n number of trees we can build. And suppose if it is a classification problem, where our target variable is a categorical variable, then you can see, our final class is suppose A and final class is B from tree 2 and for example final class for tree n is class B.

So, after we get the prediction of the classes from these individual trees, we take a majority vote. And the majority vote based on this majority voting, we select the final class. In case of regression tree, what we do? We take the prediction for these individual trees and then we average it. So, that will be the final outcome for an unknown sample. So, this is the prediction. So, this is how this random forest algorithm generally grows.

(Refer Slide Time: 10:14)

WHY RANDOM FOREST IS WIDELY ACCEPTED?

- Most of tree can provide correct prediction of class for most part of the data
- The trees are making mistakes at different places
- Strong fundamental concept: a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

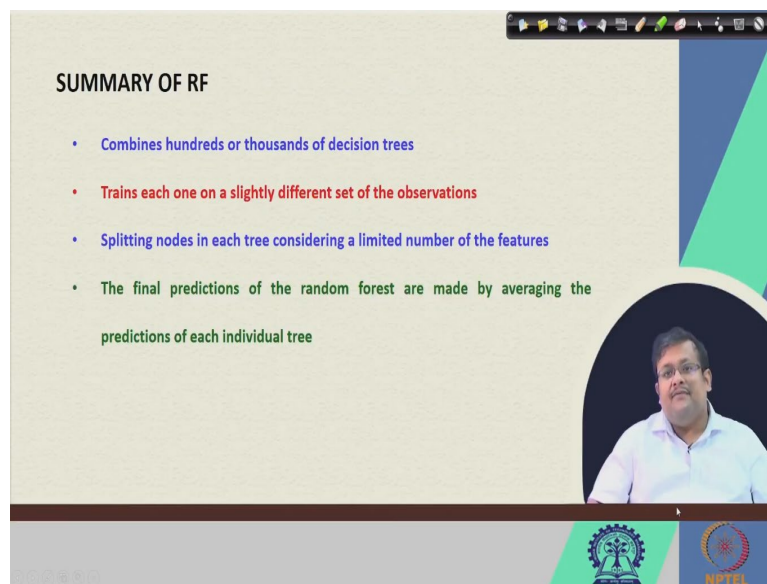
Now, why random forest is widely accepted? Why it is more robust than the single decision tree? There are three reasons. First of all most of the tree can provide. There are two assumptions, major assumptions; first of all, we assume that most of the tree can provide correct prediction of class for most part of the data.

This is the first assumption of random forest, the second assumption of random forest is, the trees are making mistakes at different places. First of all again, most of the tree can provide correct prediction of the class for most part of the data and the trees, if they are doing also mistake, they are doing the mistake at different places.

Since there, they are being built by taking the random samples and also using the random features, so that is why the ultimate, even this weak learner, ultimate, when we are getting the values from all the individual trees then it becomes much more robust and strong. So, strong fundamental concept is there, when a large number of relatively uncorrelated models or trees operates as a committee, it will outperform any other individual constituent models.

So, here we can see, here, large number of uncorrelated models are operating as a committee and then ultimately it is outperforming all other models which are available. So, this is how the random forest model is widely accepted is more robust than other models and it has also very good prediction performance also.

(Refer Slide Time: 11:59)



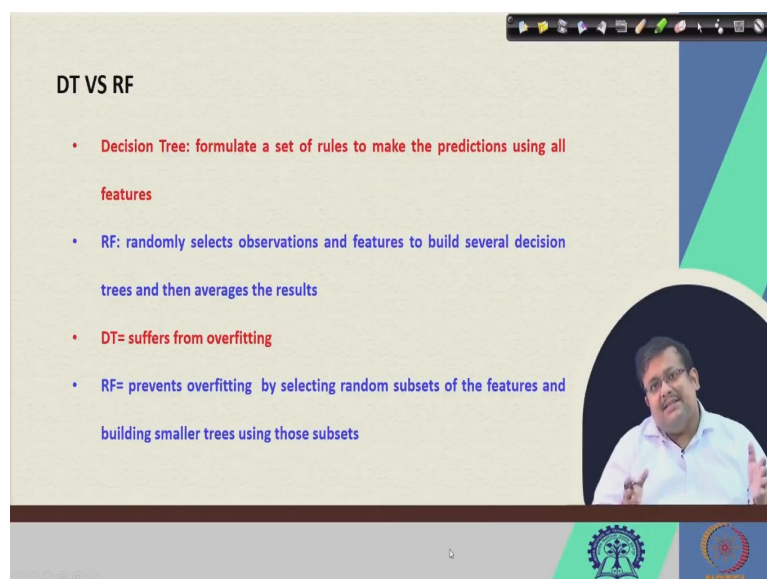
SUMMARY OF RF

- Combines hundreds or thousands of decision trees
- Trains each one on a slightly different set of the observations
- Splitting nodes in each tree considering a limited number of the features
- The final predictions of the random forest are made by averaging the predictions of each individual tree

The slide includes a video inset of a man in a white shirt speaking. At the bottom, there are logos for IIT Bombay and NPTEL.

So, what is the summary of random forest? The summary of random forest is, it basically combines hundreds and thousands of decision trees. It trains one on a slightly different set of observation, because we are taking the bootstrap samples and then splitting nodes in each tree considering a limited number of features and then the final prediction of the random forest are made, when it is a regression problem, the final prediction for the random forest are made by averaging the prediction from the each individual tree. So, this is the summary of random forest.

(Refer Slide Time: 12:34)



DT VS RF

- Decision Tree: formulate a set of rules to make the predictions using all features
- RF: randomly selects observations and features to build several decision trees and then averages the results
- DT= suffers from overfitting
- RF= prevents overfitting by selecting random subsets of the features and building smaller trees using those subsets

The slide includes a video inset of a man in a white shirt speaking. At the bottom, there are logos for IIT Bombay and NPTEL.

Now, what is the difference between a decision tree and the random forest? So, decision tree generally formulates a set of rules to make the predictions using all the features. We have seen that in our previous lecture. How they are setting up the rules based on the combination of the feature and their values. Whereas random forest randomly selects observations and features to build several decision trees and then averaging the results.

So, of course, in case of single decision tree that suffers from over fitting, however the random forest prevents overfitting by selecting random subset of the features and building smaller trees using the subsets. So, this is the difference between a decision tree and random forest and this is why random forest is more robust than that of a decision tree.

(Refer Slide Time: 13:32)

RF: PROS AND CONS

- Can be used in both classification and regression: Versatile
- Can measure the relative importance of the used features
- Hyperparameters are easy to understand and produce strong prediction performance
- RF prevents overfitting
- Large number of trees: slow to train RF model

So, what are the pros and cons of random forest? So, this random forest can be used for classification, both, classification and regression problem and also this random forest can select the relative importance of the features or variables or predicted variables in any classification or regression problem. And then the hyperparameters are easy to understand and produce the strong prediction performance.

And it also prevents the overfitting, as I have discussed in our previous slide. However, there is one drawback in random forest that large number, when you grow the large number of trees, that sometimes slows the model. So, this is one of the drawback of random forest and model. But, overall random forest is widely used in as a machine-learning tool in soil and crop domain of agriculture.

(Refer Slide Time: 14:42)

The slide is titled "SUPPORT VECTOR MACHINE (SVM)". It contains three bullet points:

- Constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification or regression
- A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin)
- The larger the margin, the lower the generalization error of the classifier

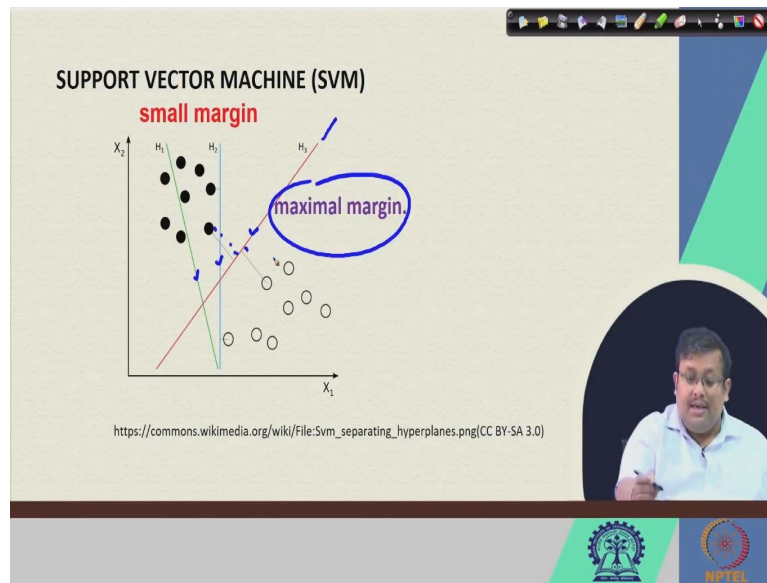
The slide also features a video inset of a man in a white shirt speaking, and logos for IIT Bombay and NPTEL at the bottom.

So, let us see another important, very important machine-learning algorithm called Support Vector Machine or SVM. So, what is SVM? So, SVM is an algorithm which constructs a set of hyperplanes in a high or indefinitely, indefinite dimensional space or infinite dimensional space, which can be used as a classification or regression. Again just like random forest, just like classification regression tree, we can use this SVM for both classification and regression.

And when we do the regression, we specifically call it SVR or Support Vector Regression. So, what happens in this SVM? So, this SVM basically tries to create a hyperplane in multi-dimensional data set for classification regression. And how this hyperplanes are being done? Hyperplanes are being drawn to maintain the largest distance from the nearest data points. So, largest distance from the nearest data points of any class.

So, they are called the functional margin. So, as again, the hyperplane the, a good separation in case of hyperplane is achieved that has the largest distance to the nearest training data point of any class, I will show you in our next slide. So, the larger the margin, the lower the generalization error of the classifier. So, this is how, this hyperplanes are being drawn in case of support vector machine.

(Refer Slide Time: 16:29)



So, here you can see here, two features are here x_1 and x_2 and these are the different types of soil, different for example two different types of samples. So, we can build hyperplane in different fashion. We can build a hyperplane, we can build in hyperplane here, we can also build hyperplane here.

However, this one is being selected because that this hyperplane is creating the or showing the largest distance from this nearest samples belong to a specific class. So, you can see here, we are maintaining the maximal margin using this red hyperplane, however for all other possibilities we are not getting this maximal distance between the samples belong to two different classes.

So, the idea, the basic idea for sub SVM is to draw this kind of hyperplane, which can divide or which can maintain the maximum margin between the samples belongs to different groups. And these samples which are present along the boundary of the margin are known as the support vector, ok. So, this is how the name comes the support vector name originates.

(Refer Slide Time: 18:07)

SUPPORT VECTOR MACHINE (SVM)

$f(x, w, b) = \text{sign}(w \cdot x - b)$

• denotes +1
• denotes -1

Support Vectors are those datapoints that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin

This is the simplest kind of SVM (Called an LSVM)

- The objective of a SVM= to find a hyperplane in an n-dimensional space that distinctly classifies the data points.
- Support Vectors: the data points on either side of the hyperplane that are closest to the hyperplane

<http://www.cs.cmu.edu/~awm/tutorials>

Logo: IIT Delhi, NPTEL

So, here you can see in this picture this is the, this is again the hyperplane and this is the maximum margin linear classifier, is the linear classifier which shows the maximum margin and you can see these are the support vectors. These are the support vectors, these are the data points that the margin pushes up against.

So, here this is the simplest type of support vector machine and we call it a Linear SVM or LSVM. Now, the objective of support vector machine is to find a hyperplane in an n dimension space that distinctly classifies the data points. And what are the support vectors? I have already discussed, support vectors are the data points on either side of the hyperplane that are closest to the hyperplane.

So, you can see here, this is the hyperplane and the either side of the hyperplane these are the closest point. So, that is why they are known as the support vector.

(Refer Slide Time: 19:10)

SUPPORT VECTOR REGRESSION (SVR)

- SVR= supervised learning algorithm that is used to predict discrete values.
- SVR= uses the same principle as the SVMs

<http://www.senspirit.com/artificial-intelligence/machine-learning/regression/support-vector-regression/support-vector-regression-in-1/>

The slide features a 3D plot of a parabolic surface with data points scattered around it. A small inset video shows a man speaking. Logos for IIT Bombay and NPTEL are visible at the bottom.

Now, this is another representation of the support vector regression. Now, support vector regression is a supervised learning algorithm that is used to predict the discrete values and these SVR generally use the same principle as like as SVMs.

(Refer Slide Time: 19:31)

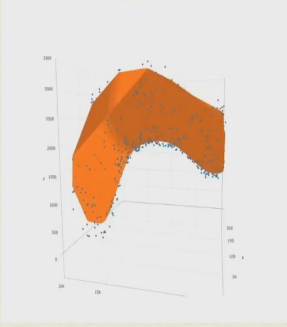
SUPPORT VECTOR REGRESSION (SVR)

- The basic idea behind SVR is to find the best fit line (hyperplane that has the maximum number of points)
- The larger the margin, the lower the generalization error of the classifier

<http://www.senspirit.com/artificial-intelligence/machine-learning/regression/support-vector-regression/support-vector-regression-in-1/>

The slide features a 3D plot of a parabolic surface with data points scattered around it. A small inset video shows a man speaking. Logos for IIT Bombay and NPTEL are visible at the bottom.



SUPPORT VECTOR REGRESSION (SVR)



• The basic idea behind SVR is to find the best fit line (hyperplane that has the maximum number of points)

• The larger the margin, the lower the generalization error of the classifier

<http://www.semspirit.com/artificial-intelligence/machine-learning/regression/support-vector-regression/support-vector-regression-in-c/>






Now, the basic idea behind SVR is to find the best fit line or hyperplane that has the maximum number of points and the larger the margin the lower the generalization error of the classifier. So, this is in nutshell about the Support Vector Regression.

(Refer Slide Time: 19:52)

SVR: PROS AND CONS

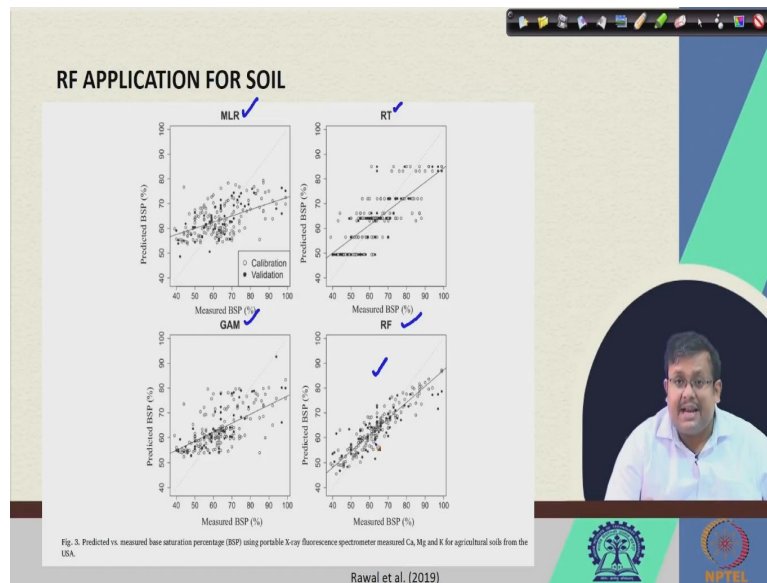
- Robust to outliers
- Easy implementation
- High prediction accuracy
- Not suitable for large datasets
- Underperform when # of features > # training samples
- Does not perform well for very noisy data

Now there are certain benefits and disadvantages of support vector regression. The benefits are given here in blue text and the disadvantages are given here in the red text. You can see that the support vector regression is robust to outlier. It can help in easy implementation of the, it can be easily implemented. This support vector regression algorithm can be easily implemented and it generally gives us the higher prediction accuracy also.

However, there are some disadvantages also. First of all it is not very much suitable for a very large data set and if the data is very very noisy then also it does not perform well and finally it under performs when the number of features are more than the number of training samples. So, we have to take care of this consideration while trading with the support vector regression.

(Refer Slide Time: 20:49)

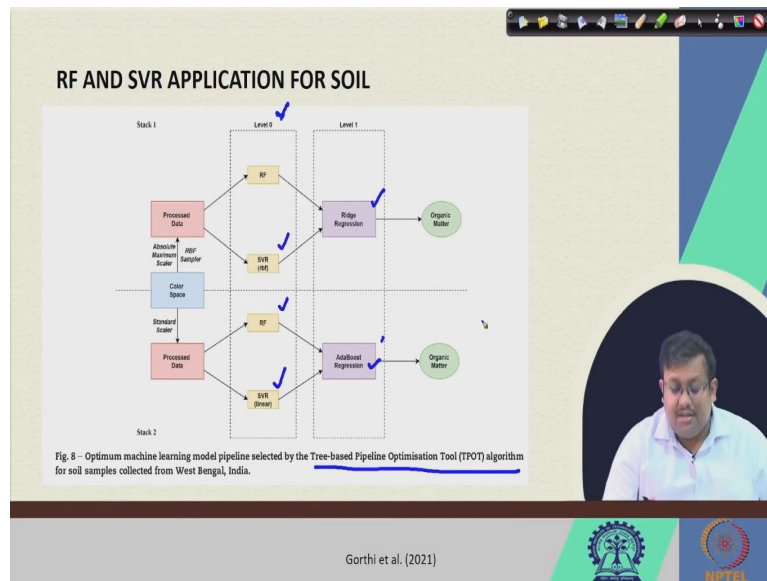


Now, let us see some example of random forest and support vector regression based application for soil and crop. Here in this research, you can see made by Rawal et al in 2019. They have used the different, four different models like here you can see, it is a regression tree model, multiple linear regression model, generalized additive model and random forest model to measure the base saturation percentage of the soil based on the portable x-Ray fluorescence spectrometer.

So, that shows the application and they used the random forest. The random forest gives the best prediction accuracy. So, the random forest shows, visually also, it can be seen that random forest is giving the best accuracy. So, that shows the application of random forest for soil base saturation percentage measurement. Base saturation percentage generally, gives an idea of the important nutrient content in the soil.

So, just like at a (21:55) capacity. So, here we can see, using the elemental data values coming from any proximal sensor, if we can use the support random forest regression, we can predict the best saturation percentage of the soil, non-invasively using this method.

(Refer Slide Time: 22:17)



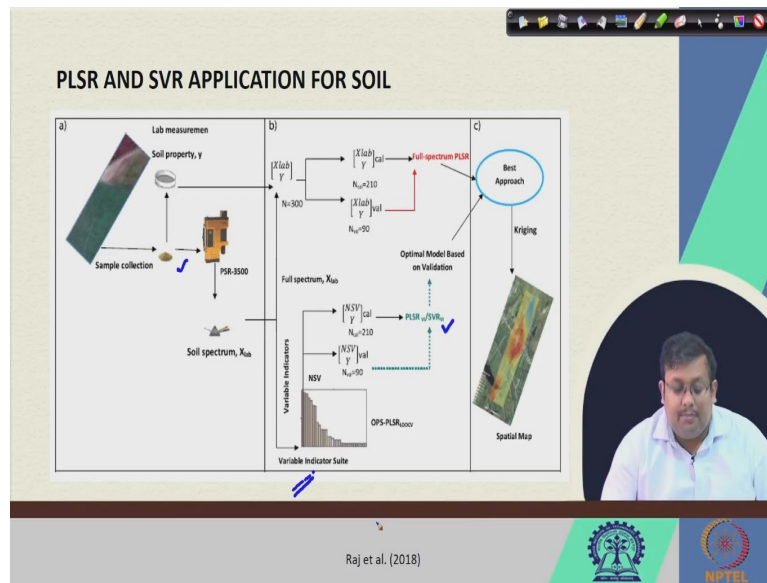
Another application from Gorthi et al in 2021, they have used image derived indices to predict the soil organic matter, using multiple machine learning algorithms in two different levels. And this type of construction is known as the Tree Based Pipeline Optimization Tool algorithm. So, TPOT tool, algorithm tool.

So, here you can see, there are two levels, in the level 0, in the level 0, there are random forest then radial basis function, support vector regression, the Linear SVR, then also random forest. And also there are two types of data, standard scalar and absolute maximum scalar and using the feature space, you can see here, there are two to three different types of model here in the level 0.

And the level 1 they have tried both Ridge regression and also AdaBoost regression. And ultimately they try to predict the organic matter. So, this shows the application of both random forest and support vector regression for image based prediction of soil properties. We are going to discuss this research in details in our upcoming weeks, when we will discuss the image base, image processing and subsequent machine learning algorithms.

But here, we can at least see that using random forest and support vector regression in combination with other machine learning algorithms; we can predict the soil organic matter.

(Refer Slide Time: 24:04)

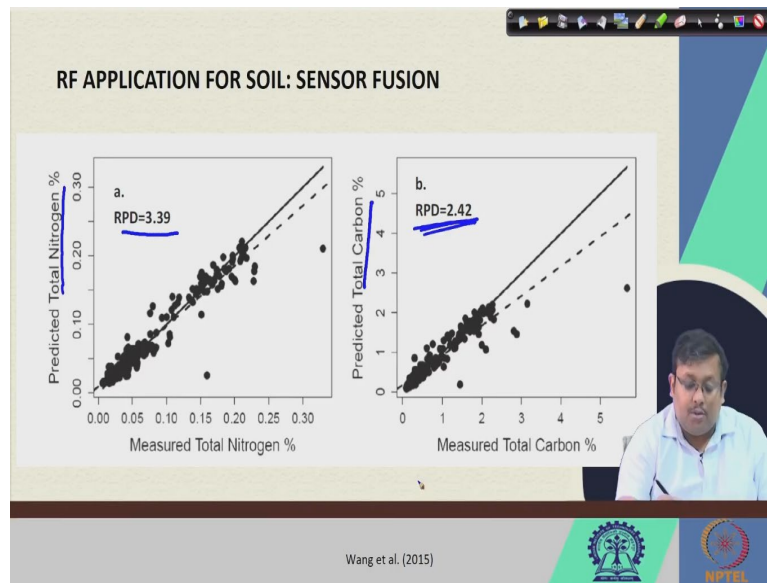


Another example of the application of support vector regression you can see here that using the diffuse reflectance spectroscopy, Raj et al in 2018, they have selected the spectral data, they have gathered the spectral data from the soil and after gathering the spectral data, they have utilized some indicators, they call them variable indicators.

So, using the variable indicator suite, they have selected a number of spectral variables instead of taking the whole number of spectral wavelengths from three 350 to 2500 nanometer, they have selected a certain number of variables or wavelengths from the whole feature space and then they have utilized that for developing the support vector regression model to predict certain soil properties.

So, this shows the application of PLSR as well as the support vector regression for soil property measurement.

(Refer Slide Time: 25:15)

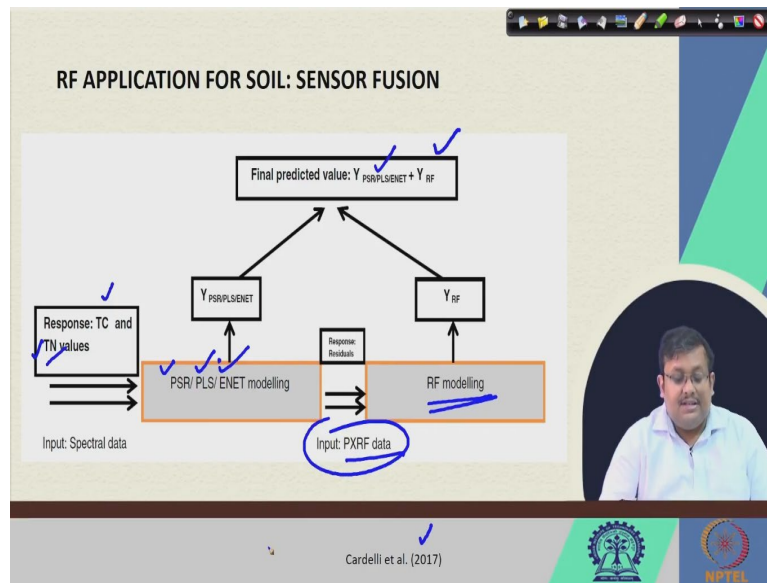


Also we can see another RF application, random forest application for soil sensor fusion. We are going to discuss sensor fusion in our coming weeks, but remember that when we fuse multiple sensors, sometime we get better accuracy or synergistic accuracy than using those sensors individually.

So, here you can see, using the random forest algorithm, Wong et al in 2015, they have predicted the total Nitrogen content by sensor fusion technologies and also they have produced the total Carbon prediction also and they got acceptable accuracy with an RPD value, residual prediction variation, residual prediction variation of 2.42 and here 3.39 and anything which is greater than 2 is showing the stable model.

So, they have produced based on these RPD values they have shown that using the RF and using the sensor fused data, from multiple sensor, it is possible to predict the total Nitrogen and total Carbon content in soil.

(Refer Slide Time: 26:37)



So, this is the construction of the sensor fusion, another research being done by Cardelli et al in 2017, you can see here, that here total Carbon and total Nitrogen that were the target values. So, using the PSR, Penalized Spine Regression or Partially Squares Regression or Elastic Net Regression, they have tried to model this total Carbon and total Nitrogen using the spectral data and they have predicted the y .

And ultimately the residual values were further modeled by random forest using the portable XRF elemental data and the outputs were the Y_{RF} . So, ultimately the final predicted values is the combination of the prediction from both these two algorithms. So, this is how different types of machine learning algorithms are utilized for sensor fusion for predicting certain soil properties.

(Refer Slide Time: 27:42)

RF AND SVR APPLICATION FOR PREDICTING CROP HEIGHT AND MATURITY

Table 3. Grouping of means of mean absolute error (MAE), root mean squared error (RMSE), and Pearson's correlation coefficient (r) between estimated and observed values for plant height in soybean obtained with different machine learning models and input configurations.

Model	Input		
	WL	VI	WLVI
	MAE		
DL	8.32 Cb	10.89 Bb	11.92 Aa
RF	8.38 Ab	8.09 Ac	8.11 Ac
SVM	8.98 Aa	8.55 Bb	8.49 Bb
LR	9.03 Aa	8.65 Bb	8.67 Bb
	RMSE		
DL	10.51 Cb	13.38 Ba	14.58 Aa
RF	10.88 Ab	10.49 Ac	10.51 Ac
SVM	11.77 Aa	11.05 Bb	10.97 Bb
LR	11.78 Aa	11.24 Bb	11.21 Bb
	r		
DL	0.79 Aa	0.77 Ab	0.75 Bb
RF	0.77 Ba	0.79 Aa	0.79 Aa
SVM	0.73 Bb	0.76 Ac	0.77 Ab
LR	0.73 Bb	0.75 Ac	0.75 Ab

Means followed by the same uppercase letters in the same row and the same lowercase letters in the same column do not differ by the Scott-Knot test at 5% probability. DL, deep learning; RF, random forest; SVM, support vector machine; LR, linear regression; WL, spectral bands wavelengths; VI, spectral vegetation indices; WLVI, a combination between spectral bands and vegetation indices.

Table 2. Grouping of means of mean absolute error (MAE), root mean squared error (RMSE), and Pearson's correlation coefficient (r) between estimated and observed values for days to maturity in soybean obtained with different machine learning models and input configurations.

Model	Input		
	WL	VI	WLVI
	MAE		
DL	6.05 Bc	7.45 Aa	7.62 Aa
RF	8.21 Ac	6.09 Ac	6.08 Ac
SVM	7.11 Ab	6.05 Bb	6.39 Bb
LR	7.37 Aa	7.41 Aa	7.41 Aa
	RMSE		
DL	8.01 Cd	10.23 Ba	10.96 Aa
RF	6.98 Ac	6.28 Ac	6.21 Ad
SVM	9.75 Aa	8.25 Bb	8.82 Cc
LR	9.31 Ab	8.41 Ab	8.39 Ab
	r		
DL	0.66 Aa	0.57 Bb	0.54 Cc
RF	0.62 Ab	0.65 Aa	0.65 Aa
SVM	0.60 Cd	0.53 Bc	0.57 Bb
LR	0.51 Ac	0.56 Ad	0.56 Ad

Means followed by the same uppercase letters in the same row and the same lowercase letters in the same column do not differ by the Scott-Knot test at 5% probability. DL, deep learning; RF, random forest; SVM, support vector machine; LR, linear regression; WL, spectral bands wavelengths; VI, spectral vegetation indices; WLVI, a combination between spectral bands and vegetation indices.

Teodoro et al. (2021)

Let us see some crop examples also here, you can see that people have tried or have tried to use the random forest and support vector machine regression for predicting the crop height and maturity. This table shows the crop height and this table shows the prediction accuracy values while for predicting the crop maturity using different machine learning and deep learning algorithms.

Here DL stands for the deep learning method, random foreign support vector machine and linear regression both these cases. So, and also based on different types of multispectral data, they have collected different multispectral data and they have extracted those bands of wavelengths and they calculated some spectral vegetation indices and then WLVI stands for a combination between spectral bands and vegetative indexes.

So, they have calculated some spectral bands and vegetative indices and then they try to predict, incorporate them as predictors, input predictors and they try to predict the crop height as well as the crop maturity, based on different models and they tried there, and they compared their accuracy prediction accuracy. So, you can see the random forest and support vector machines are why so widely used in crop in and soil.

(Refer Slide Time: 29:13)

RF AND SVR APPLICATION FOR PREDICTING MAIZE WATER INDICATORS

Figure 2. (a) Matrix 300 UAV integrated with the Altam sensor serving as the imaging platform used in this study. (b) Altam camera, (c) flight plan for the study image, and (d) the calibrated reflectance panel.

Table 3. List of vegetation indices (VIs) used in the modelling of crop water content and related source references.

Index	Full Name	Formula	Reference
Direct water-sensitive spectral VI			
NDWI	Normalised Difference Water Index	$\text{Green} - \text{NIR} / \text{Green} + \text{NIR}$	[11]
Indirect water-sensitive spectral VIs			
NDVI	Normalised Difference Vegetation Index	$\text{NIR} - \text{Red} / \text{NIR} + \text{Red}$	[18]
NGRDI	Normalised Difference Green/ Red Index	$\text{Green} - \text{red} / \text{green} + \text{red}$	[12]
NDRE	Normalised Difference Red-Edge Index	$\text{NIR} - \text{rededge} / \text{NIR} + \text{Rededge}$	[8]
NDVI rededge	Red-Edge Normalised Difference Vegetation Index	$\text{Rededge} - \text{Red} / \text{Rededge} + \text{red}$	[8]
CIgreen	Green Chlorophyll Index	$(\text{NIR} / \text{Green}) - 1$	[12]
CIrededge	Red-edge chlorophyll index	$(\text{NIR} / \text{rededge}) - 1$	[8]

Table 5. Prediction accuracies of EWT_{leaf} , FMC_{leaf} , and SLA_{leaf} were derived using optimal models based on the RFR, DTR, ANNR, PLSR, and SVR regression models.

Model	$\text{EWT}_{\text{leaf}} (\text{gm}^{-2})$			$\text{FMC}_{\text{leaf}} (\%)$			$\text{SLA}_{\text{leaf}} (\text{m}^2 \text{g}^{-1})$		
	R ²	RMSE	RRMSE	R ²	RMSE	RRMSE	R ²	RMSE	RRMSE
RFR	0.89	1028	313	0.76	0.45	1.00	0.73	0.0004	3.48
DTR	0.73	25.16	7.67	0.65	1.08	1.35	0.7	0.0009	8.16
ANNR	0.64	14.29	4.35	0.34	1.54	1.92	0.68	0.0007	6.60
PLSR	0.74	17.1	5.15	0.45	0.48	0.60	0.6	0.0008	19.33
SVR	0.78	15.05	4.76	0.69	0.70	0.89	0.71	0.0005	18.82

Ndlovu et al. (2021)

Another research was there where Ndlovu et al in 2021, they have used the UAV images and after getting the UAV images they have, these this is the UAV, using multiple integrated with the imaging platform, ok. So, there are different types of camera they have used and they have taken the image by flying the UAV.

After getting those images, they have calculated different types of industries like NDWI, normalized difference vegetation index, then NGRDI that is Normalized Difference Green Red Index, then NDRE, NDVI rededge CIgreens, Clrededge and so on so forth. So, and these are their formulas.

After calculating those, they predicted the maize water indicators, there are different types of water indicators like EWT leaf FMC then SLA leaf, by using different types of algorithms like PLSR then artificial neural network then decision tree, random forest then support vector regression and then they tried then compared their model accuracies.

You can see from here, the random forest regression in terms of R-square is getting highest higher prediction accuracy in case of EWT, whereas in case of FMC again random forest is giving the highest prediction accuracy and in case of SLA also we are seeing the random forest is giving the highest accuracy. So, you can see that why random forest is widely accepted, it is showing higher prediction accuracy for multiple soil properties.

(Refer Slide Time: 31:01)

REFERENCES

<http://www.cs.cmu.edu/~awm/tutorials>
<http://www.semspirit.com/artificial-intelligence/machine-learning/regression/support-vector-regression/support-vector-regression-in-r/>

Cardelli, V., D.C. Weindorf, S. Chakraborty, B. Li, M. DeFeudis, S. Cocco, A. Agnelli, A. Choudhury, D. Ray, and G. Corti. 2017. Non-saturated soil organic horizon characterization via advanced proximal sensors. *Geoderma* 288:130-142.

Gorhi, S., R.K. Swetha, S. Chakraborty, B. Li, D.C. Weindorf, S. Dutta, H. Banerjee, K. Das, and K. Majumdar. 2021. Soil organic matter prediction using smartphone-captured digital images: use of reflectance image and image perturbation. *Biosystems Engineering* 209: 154-169.

Ndlovu, H.S.; Odindi, J.; Sibanda, M.; Mutanga, O.; Clulow, A.; Chimonyo, V.G.P.; Mabhaudhi, T. A Comparative Estimation of Maize Leaf Water Content Using Machine Learning Techniques and Unmanned Aerial Vehicle (UAV)-Based Proximal and Remotely Sensed Data. *Remote Sens.* 2021, 13, 4091.

Raj, A., S. Chakraborty, B.M. Duda, D.C. Weindorf, B.Li, S. Roy, M.C. Sarathjith, and B.S. Das. 2018. Soil mapping via diffuse reflectance spectroscopy based variable indicators: an ordered predictor selection approach. *Geoderma* 314:146-159.

Rawal, A., S. Chakraborty, B. Li, K. Lewis, M. Godoy, L. Paulette, and D.C. Weindorf. 2019. Determination of base saturation percentage in agricultural soils via portable X-ray fluorescence spectrometer. *Geoderma* 338:375-382.

Teodoro, P.E.; Teodoro, L.P.R.; Baio, F.H.R.; da Silva Junior, C.A.; dos Santos, R.G.; Ramos, A.P.M.; Pinheiro, M.M.F.; Osco, L.P.; Gonçalves, W.N.; Carneiro, A.M.; et al. Predicting Days to Maturity, Plant Height, and Grain Yield in Soybean: A Machine and Deep Learning Approach Using Multispectral Data. *Remote Sens.* 2021, 13, 4632.

Wang, D., S. Chakraborty, D.C. Weindorf, B. Li, A. Sharma, S. Paul, and N. Ali. 2015. Synthesized use of VisNIR DRS and PXRF for soil characterization: total carbon and total nitrogen. *Geoderma* 243-244:157-167.

So, these are the references which I have used guys. It is not possible to cover all the machine learning approaches, because there are multitude of machine learning approaches for regression which are being utilized by the soil scientist and crop scientists for predicting soil and crop properties like generalized additive model, like elastic net, like partial, like penalized spline regression, there are multiple types of regressions, which are there.

It is not possible, but I try to cover all the major machine learning algorithms as far as the predictions as far as the regressions are concerned, to discuss like PLSR, random forest, support vector regression, decision trees, we have discussed. If time permits, we will also discuss the others in our subsequent weeks.

So, I hope that you have now some understanding of these different machine learning prediction algorithms regression approaches and also the principal component analysis and how people are using this for different soil and crop studies. So, let us wrap up our lecture here and in the next week we will start discussing about different clustering algorithm. Thank you very much.