Regression and Prediction
Week 4

Topics covered so far

1. Regression: Linear Regression
    a. General Statistical Framework
    b. Linear Regression
    c. Performance Assessment - Estimating parameter means and confidence intervals for prediction

2. Regression: Model Evaluation
    a. Prediction vs Modeling
    b. Assumptions behind Regression
    c. Bias-variance tradeoff
    d. Overfitting and Regularization
    e. Cross-Validation
    f. Bootstrapping

Discussion Questions

1. What is the difference between machine learning and statistical learning?
2. What is linear regression and how does it work?
3. What is multiple linear regression? What are some examples where it could be used?
4. How do you measure the performance of a linear regression model?
5. What are the underlying assumptions in the linear regression model?
6. What is bias variance trade off?
7. What is Regularization? What are its different types?
8. Why do we use Cross-Validation? How does it work?
9. What is the concept of bootstrapping and why do we need it?

Statistics vs Machine Learning

The difference between machine learning and statistical learning is their purpose. Machine learning models are designed to make the most accurate predictions possible, whereas statistical models are designed for inference about the relationships between variables.

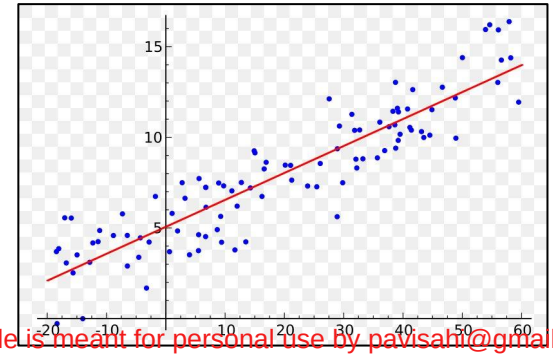
The following table highlights the major differences between statistics and the machine learning point of view:

Statistics	Machine Learning
Emphasis on deep theorems on complex models	Emphasis on the underlying algorithm
Focus on hypothesis testing and interpretability	Focus on predicting accuracy of the model
Inference on parameter estimation, errors, and predictions	Inference on prediction
Deep understanding of simple models	Theory does not always explain success

Linear Regression

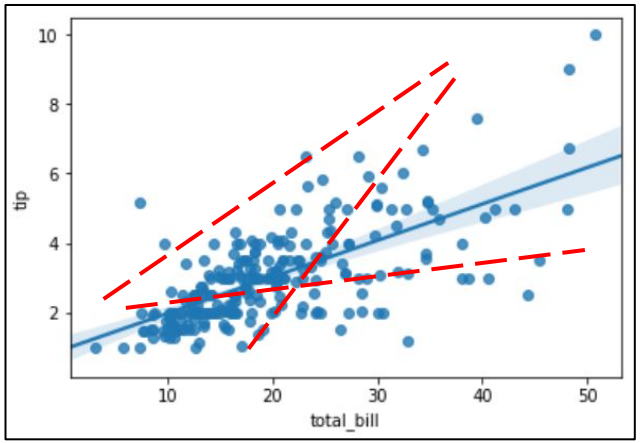
- Linear regression is a way to identify a relationship between the independent variable(s) and the dependent variable.
- We can use these relationships to predict values for one variable for given value(s) of other variable(s).
- Linear Regression assumes the relationship between variables can be modeled through a linear equation or an equation of a line.
- The variable which is used in prediction, is termed as independent/explanatory/regressor, whereas the predicted variable is termed as dependent/target/response variable.
- In case of linear regression with a single explanatory variable, the linear combination can be expressed as:

response = intercept + (constant * explanatory variable)



Linear Regression: Line of Best Fit

- Learning from the data, the model generates a line that fits the data.
- Our aim is to find the regression line that **best fits** the data.
- By best fit, we mean the line will be such that the cumulative distance of all the points from the line is minimized.
- Mathematically, the line that minimizes the sum of squared errors of residuals is called the Regression Line or the Line of Best Fit.

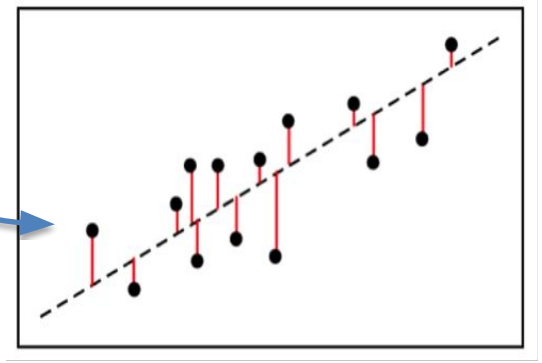


- In the example here, you can see a scatter plot between the total_*tip* amount and the *total_bill* amount.
- We can see that there is a positive correlation between these variables. As the bill amount increases, the tip increases.
- The blue line is the 'best fit' line and those in red are some examples of other lines that are not the 'best fit'.

Regression Example

Obs	Height in Inches, X	Act Weight in Pounds, Y	Predicted Weight $\hat{Y}$	Residual/ Error $e_i = Y_i - \hat{Y}_i$	Residual^2/ Error^2 $e_i^2 = (Y_i - \hat{Y}_i)^2$
1	63	127	120.1	6.900	47.61
2	64	121	126.3	-5.300	28.09
3	66	142	138.5	3.500	12.25
4	69	157	157.0	0.000	0
5	69	162	157.0	5.000	25
6	71	156	169.2	-13.200	174.24
7	71	169	169.2	-0.200	0.04
8	72	165	175.4	-10.400	108.16
9	73	181	181.5	-0.500	0.25
10	75	208	193.8	14.200	201.64
				0.000	597.28

Sum of Squared Residuals



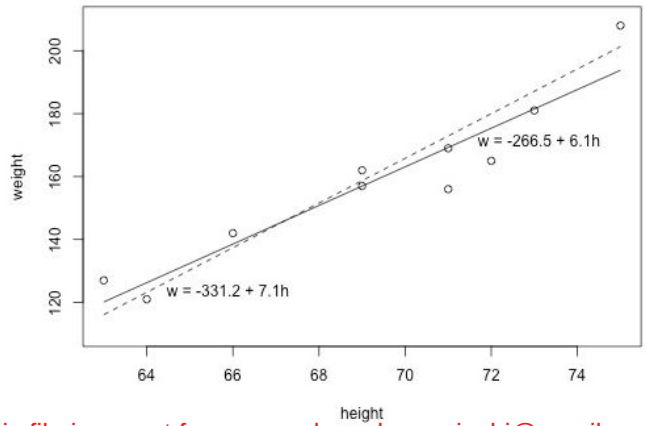
1. Say weight is regressed on height of an individual
2. Linear regression model for the above data: **$\hat{Y}$= -266.53 + 6.1376X**
3. Model is obtained by **minimizing the sum of squares of residuals.**
4. Sum of residuals is always equal to zero.
5. The line will always pass through the centroid **($\bar{X}$, $\bar{Y}$)**

Regression Example – Interpretation of Intercept

- The intercept $\mathbf{b_0}$ is the measure of **y** when **x=0**.
- **Never extrapolate** the model beyond the range of x values.
- When range of x does not include zero, y @ x=0 is not meaningful.
- Simple Linear Regression helps in prediction of y within the range of x values in the data.
- Ex: What would be the weight of an individual having height = 67.5 inches?

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = -266.53 + 6.1376X$$



What is Multiple Linear Regression?

- This is just an extension of the concept of simple linear regression with one variable, to multiple variables.
- In the real world, any phenomenon or outcome could be driven by many different independent variables.
- Therefore there is a need to have a mathematical model that can capture this relationship.
  - **Ex:** Predicting the price of a house, we need to consider various attributes such as area, number of rooms, number of kitchens etc. Such a regression problem is an example of multiple linear regression.
  - The equation for multiple linear regression can be represented by:

    target = intercept + constant 1*feature 1 + constant 2*feature 2 + constant 3*feature 3 + ...

- The model aims to find the constants and intercept such that this line is the best fit.

Regression - Evaluation Methods

R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
• Measure of the % of variance in the target variable explained by the model • Generally the first metric to look at for linear regression model performance • Higher the better	• Conceptually, very similar to R-squared but penalizes for addition of too many variables • Generally used when you have too many variables as adding more variables always increases R^2 but not Adjusted R^2 • Higher the better	• Simplest metric to check prediction accuracy • Same unit as dependent variable • Not sensitive to outliers i.e. errors doesn't increase too much if there are outliers • Difficult to optimize from mathematical point of view (pure maths logic) • Lower the better	• Another metric to measure the accuracy of prediction • Same unit as dependent variable • Sensitive to outliers - errors will be magnified due to square function • But has other mathematical advantages that will be covered later • Lower the better

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - yhat_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}$$

Assumptions of Linear Regression

Assumption	How to test	How to fix
There should be a linear relationship between dependent and independent variables	Pairplot / Correlation of each independent variable with dependent variables	Transform variables that appear non-linear (log, square root, etc.)
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance Inflation Factor)	Remove correlated variables or merge them
No Heteroskedasticity - residuals should have constant variance	Plot residuals vs. fitted values and check the plot	Non-linear transformation of dependent variable or add other important variables
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of independent or dependent variables

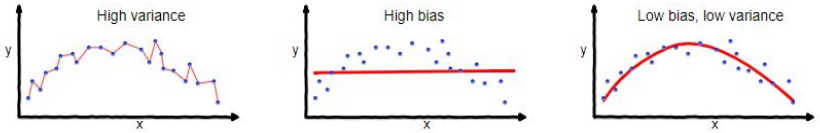
Bias-Variance: Underfitting and Overfitting

- **Bias:** Bias is the difference between the prediction of our model and the correct value which we are trying to predict. Model with high bias gives less attention to the training data and overgeneralize the model which leads to high error on training and test data.

- **Variance:** Variance is the value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the test data. Therefore, such models perform very well on training data but has high error on test data.

- In supervised learning, **underfitting** happens when a model is not able to capture the underlying pattern of the data. These models usually have high bias and low variance whereas, **overfitting** happens when our model captures the noise along with the underlying pattern in data. These models usually have low bias and high variance.

- In supervised learning, **underfitting** happens when a model is not able to capture the underlying pattern of the data. These models usually have high bias and low variance whereas, **overfitting** happens when our model captures the noise along with the underlying pattern in data. These models usually have low bias and high variance.



Bias-Variance Tradeoff

If our model is too simple and has very few parameters, then it may have high bias and low variance. On the other hand, if our model has a large number of parameters, then it's going to have high variance and low bias. So, we need to find the right/good balance between overfitting and underfitting the data.

An optimal balance of bias and variance would neither overfit nor underfit the model.

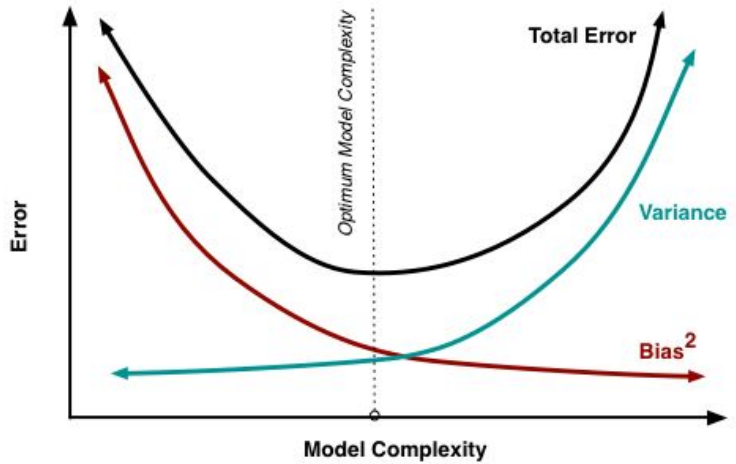


Image Source

Regularization and its types

- Regularization is the process which regularizes or shrinks the coefficients towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- Regularization significantly reduces the variance of the model, without a substantial increase in its bias.
- The two most common types of regularization in regression are:
    - **Lasso Regression:** In this technique, we add $\alpha * \sum|\beta|$ as the shrinkage quantity. It only penalizes high coefficients. It has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter α is sufficiently large. This technique is also called L1 regularization.
    - **Ridge Regression:** In this technique, we modify the residual sum of squares by adding the shrinkage quantity $\alpha * \sum\beta^2$ and use α as the tuning hyperparameter that decides how much we want to penalize the flexibility of our model. This technique is also called L2 regularization.

Cross-Validation and its types

Cross-Validation is a technique in which we train our model using the subset of the dataset and then evaluate using the complementary subset of the dataset.

- It provides some kind of assurance that the model has got most of the pattern from the dataset correct and it is not picking up noise.
- We will be discussing two types of cross validation techniques:
    1. K-Fold Cross-Validation
    2. Leave-One-Out Cross-Validation (LOOCV)

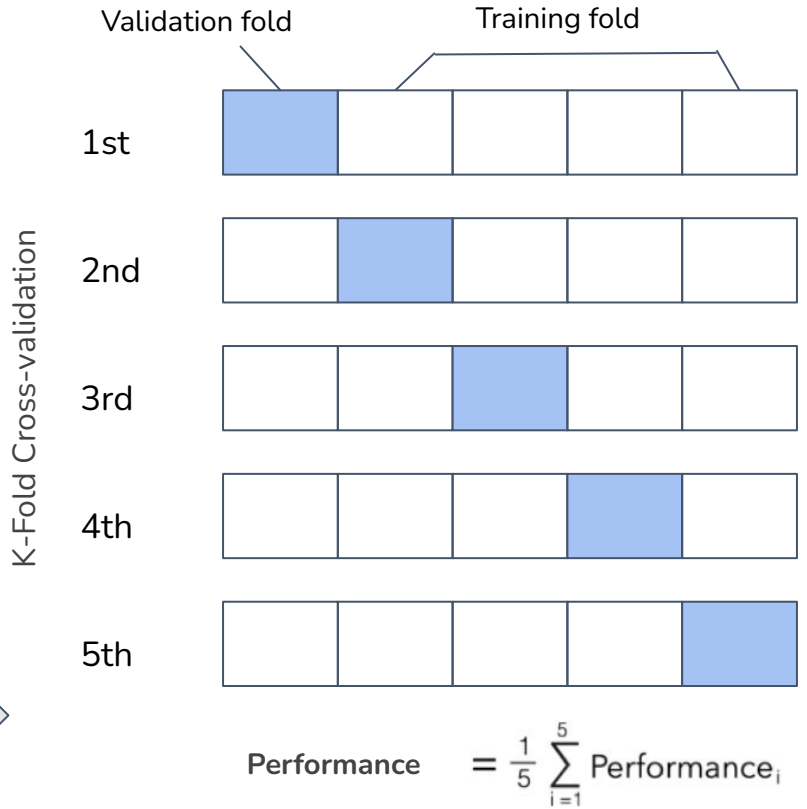
K-Fold Cross-Validation

This algorithm has a single parameter called K that refers to the number of groups that a given dataset is to be split into.

This algorithm has the following procedure:

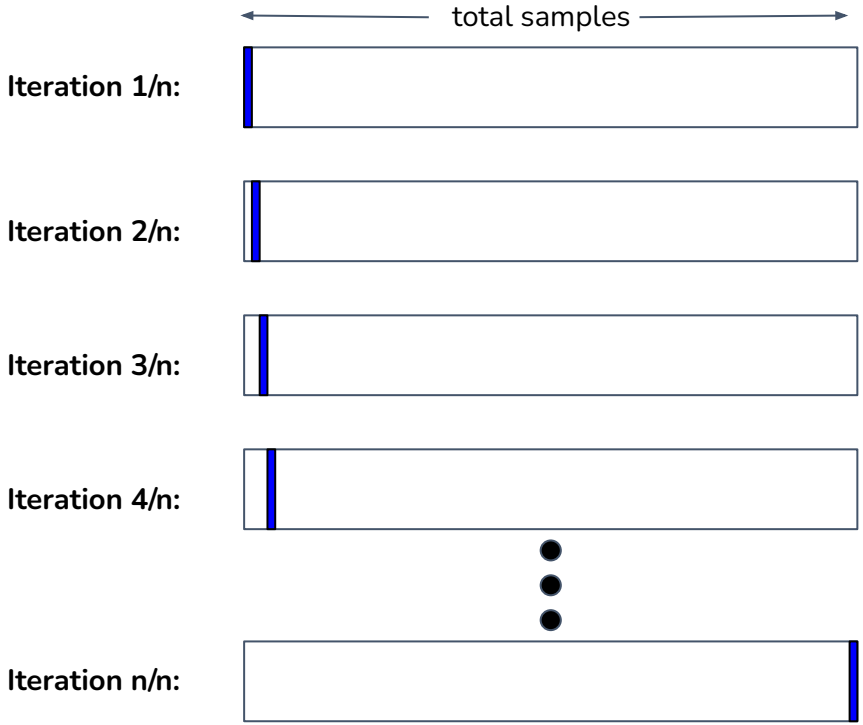
1. Shuffle the dataset randomly.
2. Split the whole dataset into K distinct groups.
3. In each iteration, take one group as a hold out set and the remaining as the training set.
4. Repeat step 3, K times with a different group, as a validation set, in each iteration.
5. Summarize the skill of the model using the average of model evaluation scores of all groups.

Here, K = 5



Leave-One-Out Cross-Validation (LOOCV)

- LOOCV is a special case of K-Fold Cross-Validation where K equals n, n being the number of data points in the dataset.
- This approach leaves 1 data point out of the training data, i.e., if there are n data points in the original dataset, then n-1 data points are used to train the model and 1 data point is used as the validation set.
- This is repeated for all combinations in which the original dataset can be separated this way, and then the error is averaged for all trials, to give overall model performance.
- The number of possible combinations is equal to the number of data points in the original dataset, i.e., n.



Bootstrapping

Bootstrapping (also called Bootstrap sampling) is a resampling method that involves drawing of samples from the data repeatedly with replacement to estimate a population parameter.

It involves the following steps:

1. Choose a number of bootstrap samples to perform
2. Choose a sample size n
3. For each bootstrap sample
    1. Draw a sample with replacement with the chosen size
    2. Calculate the statistic on the sample
4. Calculate the mean of the calculated sample statistics

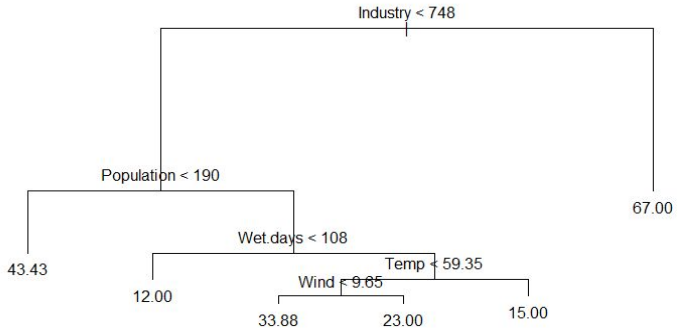
Bootstrap sampling can be used to estimate the parameter of a population, for example, mean, standard error, etc.

Case Study
Regression

Optional Section - Additional Reading

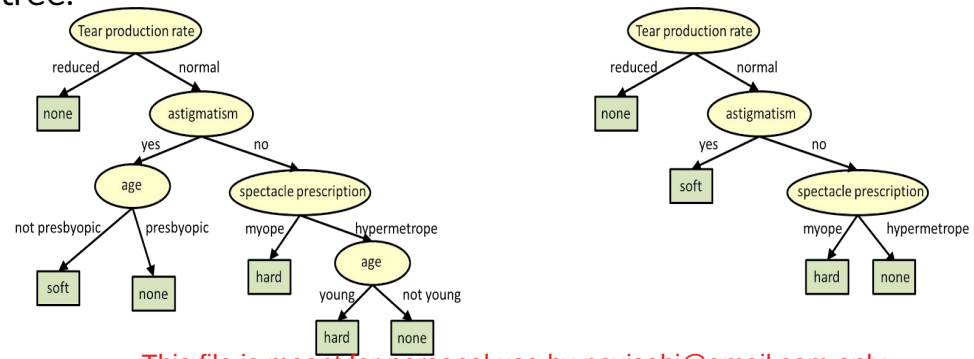
Regression Trees

- A decision tree is one of the most popular and effective supervised learning techniques for Regression and classification problems, that works equally well with both categorical and continuous variables.
- It is a graphical representation of all the possible solutions to a decision that is based on a certain condition.
- In this algorithm, the training sample points are split into two or more sets based on the split condition over input variables.



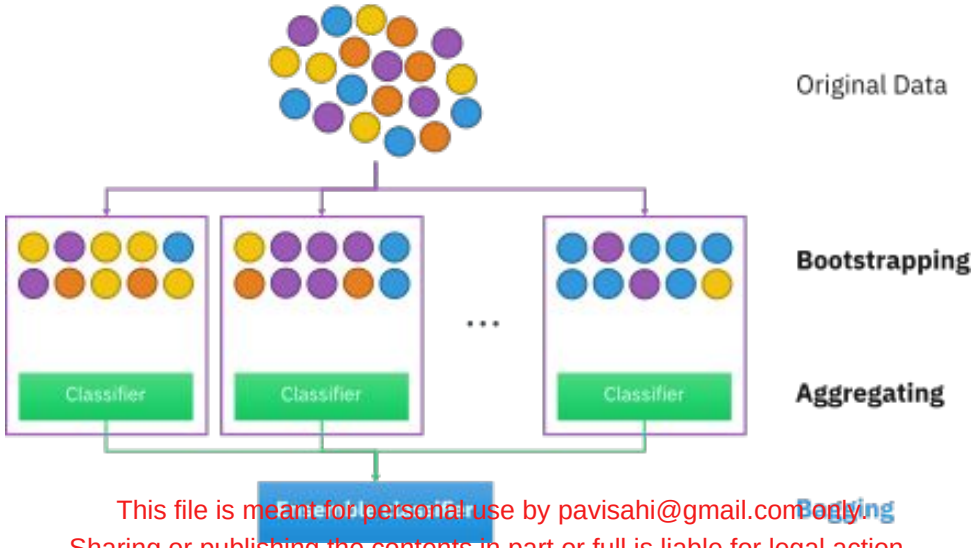
Tree Pruning

- One of the problems with the decision tree is it gets easily overfit with the training sample and becomes too large and complex.
- A complex and large tree poorly generalizes to new sample data whereas a small tree fails to capture the information of the training sample data.
- Pruning may be defined as shortening the branches of the tree. It is the process of reducing the size of the tree by turning some branch node into a leaf node and removing the leaf node under the original branch.
- By removing branches we can reduce the complexity of the tree which helps in reducing the overfitting of the tree.



Bootstrap Aggregation (Bagging)

- Bagging is a technique of merging the outputs of various models to get a final result.
- It reduces the chances of overfitting by training each model only with a randomly chosen subset of the training data. Training can be done in parallel.
- It essentially trains a large number of “strong” learners in parallel (each model is an overfit for that subset of the data).
- Then it combines (averaging or voting) these learners together to "smooth out" predictions.



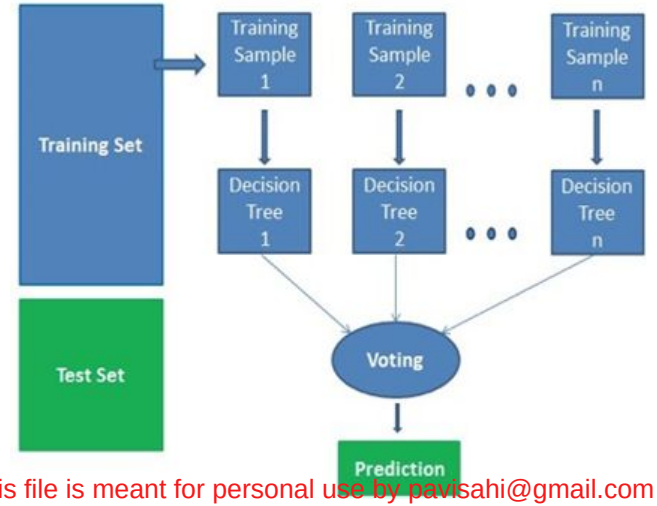
Random Forest

- Random Forest is a supervised machine learning algorithm which can be used for both classification and regression.
- It generates small decision trees using random subsamples of the dataset where the collection of the generated decision tree is defined as forest. Every individual tree is created using an attribute selection indicator such as entropy, information gain, etc.
- In classification, problem voting is done by each tree and the most voted class is considered the final result whereas in case of regression the average method is used to get the final outcome.
- Random Forest is used in various domains such as classification of images, feature selection and recommendation engines.

Random Forest: Steps involved

The following steps are involved in this algorithm:

1. Selection of a random subsample of a given dataset.
2. Using attribute selection indicators create a decision tree for each subsample and record the prediction outcome from each model.
3. Applying the voting/averaging method over predicted outcomes of individual models.
4. Considering the final results as the average value or most voted value.



Great Learning

Happy Learning !