Data Science with R

Day 9 – Introduction to Machine Learning

TACT 360° Training

LEARN. ACHIEVE. STANDOUT

Today's Agenda

✓ Machine Learning

- What is Machine Learning?
- Prediction and Inference
- Supervised & Unsupervised ML Models
- Parametric & Non- Parametric Models
- Supervised ML Models
- Covariance and Correlation
- Regression
    - Linear Regression
    - Polynomial Regression

✓ Machine Learning

- Linear Regression methods
    - Ordinary Least Square
    - R Squared method
- Why Logistic Regression?
- Logistic Regression Function

Machine Learning P1

What is Machine Learning?

Machine learning is making machines(computers) to learn or understand the pattern in data without being explicitly programmed. Machine learning makes use of mathematical and statistical algorithms in order to make the prediction about result or to take some decision about the data.

$$\text{ML Model} = \text{Algorithm(Data)} \ \& \ \text{Data} = X + Y$$

X-> Set of independent variables or features

Y-> Output variable or responder

The objective of ML is to estimate target function **(f)** that best maps input variables (X) to an output variable (Y).

$$Y = f(X) + e$$

Here “e” is the irreducible error because no matter how good we get at estimating the target function (f), we cannot reduce this error

Hemant Rathore

Machine Learning P1

Why should we estimate f(x) ?

There are mainly 2 objectives behind estimating "f(x)" –

- ✓ **Prediction –** When output variable Y is not easily available, so if we have any such relationship defined we can predict Y using X. Ex. Predicting the risk of side effects of any drug on a patient using Patient's blood characteristics.

  $\hat{Y} = \hat{f}(x)$ | $\hat{f}$ -> estimate for f ; $\hat{Y}$ is prediction for Y.

  $Y - \hat{Y}$ - Reducible Error that we want to minimize.

- ✓ **Inference –** When we want to understand the impact on output variable Y when input variables X are changed, here we are not interested in prediction and knowing the relationship is the goal. Ex. Which particular drug will be more effective to cure the disease.

Hemant Rathore

Machine Learning P1

Type of Machine Learning Algorithms

Based on the learning method -

- **Supervised Machine learning –** "Supervise the model", When we tech the model first using some training data to make it learning the data and patterns.
- **Unsupervised Machine Learning –** "No Supervision", When model learns about the data on its own and no training data is provided, in other words there is no response variable Y available that we want to predict or analyze.

Based on the type/nature of target function -

- **Parametric Machine learning Algorithms** – Simplifying the target function to a known form
- **Non-Parametric Machine Learning Algorithms –** No assumptions about the target function form

Hemant Rathore

Machine Learning P1

Supervised ML Models

Covariance – "How 2 variables change together"

The covariance between two jointly distributed random variables X and Y is defined as the expected product of their deviations from their individual expected values.

$$COV(X,Y) = \sum [X-\mu_x]*[Y-\mu_y] / (n-1)$$

Positive Covariance indicates Positive relationship, Negative Covariance indicates Negative relationship

https://en.wikipedia.org/wiki/Covariance

Hemant Rathore

Machine Learning P1

Supervised ML Models

Correlation – "The measurement of relationship"

The covariance doesn't tell anything about the strength of the relation so we need Correlation.

$$CORR(X,Y) = COV(X,Y) / [\sigma_x * \sigma_y] => \sum [X-\mu_x]*[Y-\mu_y] / [(n-1)*\sigma_x*\sigma_y]$$

Sign shows type of relationship, while value represents the strength of relationship, higher the value of CORR stronger the relationship.

CORR will always be between -1 to +1. -1 = Strong negative, +1 Strong Positive, 0 neutral relationship.

Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Regression

Finding relationship between 1 dependent variable and 1 or more independent variables and then finding some equation which can closely define this relationship.

Dependent variable is also known as Responder and independent variable is know as Predictor. Regression is used when dependent variable is continuous and relationship follows some defined pattern like a straight line, polynomial curve or logistic curve.

Type of Regression

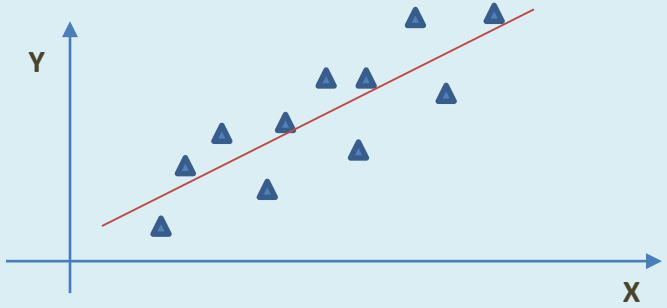
1. **Linear Regression**
   - Simple Linear Regression – Only 1 Independent Variable
   - Multiple linear Regression – More than 1 Independent Variables
2. **Logistic Regression**
   - Simple Logistic Regression– Only 1 Independent Variable
   - Multiple Logistic Regression– More than 1 Independent Variables

Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Linear Regression – "When relationship follows a straight line"

Simple Linear Regression – Only 1 Independent Variable – $Y = B_0 + B_1X$ { $y = mx + c$ }



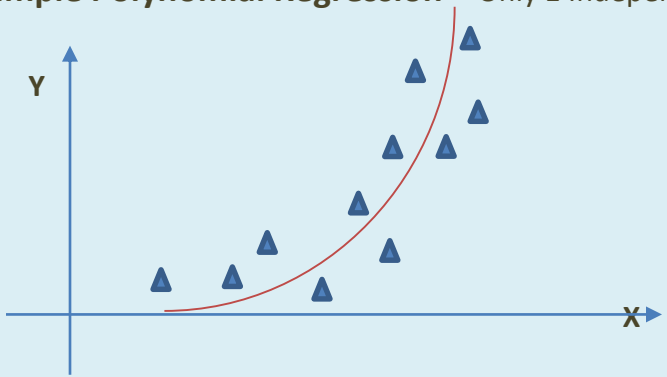
Multiple Linear Regression – More than 1 Independent Variables – $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots B_nX_n$

Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) – Polynomial Regression – "When relationship follows Non Linear Curve"

Simple Polynomial Regression – Only 1 Independent Variable – $Y = B_0 + B_1X + B_2X^2$



Multiple Polynomial Regression – More than 1 Independent Variables –

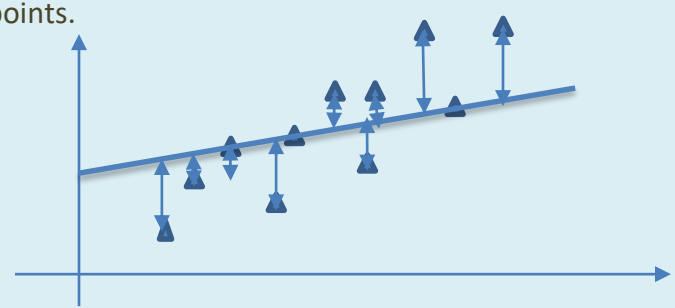
$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X^2_1 + B_4X^2_2 + B_5X_1X_2$$

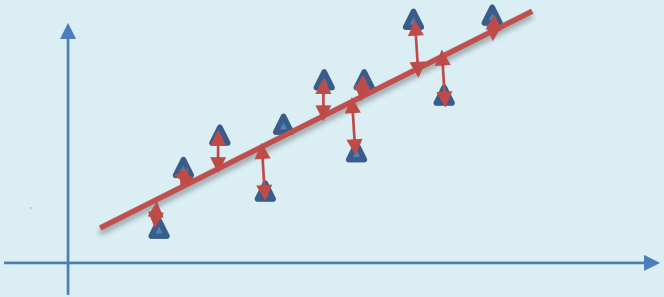
Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Linear Regression Methods – "how to get best fitting line"

1. Ordinary Least Square method – For which line the Sum of Square Residual (SSR or RSS) distance is minimum for all the points.





$$SSR \text{ or } RSS = \sum [\, y(i) - y(model) \,]^2$$

Hemant Rathore

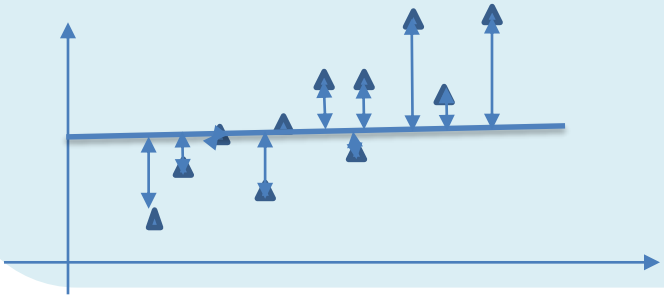
Machine Learning P1

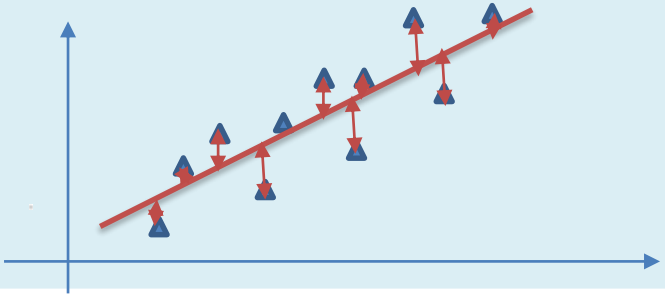
Supervised ML Models (Parametric) - Linear Regression Methods – "how to get best fitting line"

2. R Squared method – Sum of Square Residual (SSR or RSS) for average line is know as Total Sum of Square (SS tot), in this method we calculate R squared for all the possible regression line using following formula

R squared = 1 – RSS/SS tot

Ultimately the line with highest value of R squared is selected. R squared lies between 0 and 1.



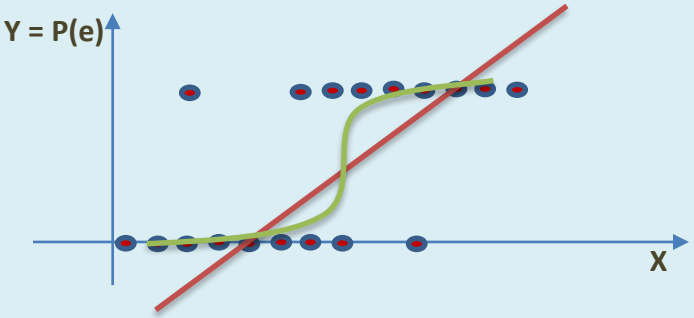


Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Why Logistic Regression – "When there can be only 2 outcomes like TRUE or FALSE"

Logistic Regression predicts the probability of an event to occur or not.



Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Logistic Regression Function

Logit Function –

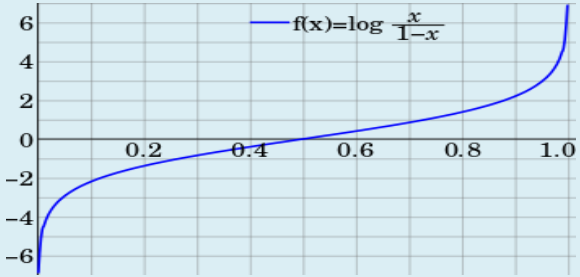
Probability – P ; Odds - P / 1-P

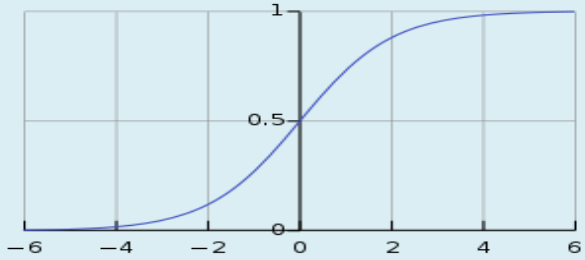
Logit(P) = ln (P / 1-P)

Sigmoid Function – Inverse of Logit function

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$

Logistic Function is one type of sigmoid function





https://en.wikipedia.org

Hemant Rathore

Machine Learning P1

Supervised ML Models (Parametric) - Logistic Regression Function

In case of logistic regression our dependent variable is the Probability P and P is dependent on independent Variable X (or X1, X2…Xn in case of multiple regression)

Only 1 independent variable x with no weightage and constant : $p = e^x / (1+ e^x)$

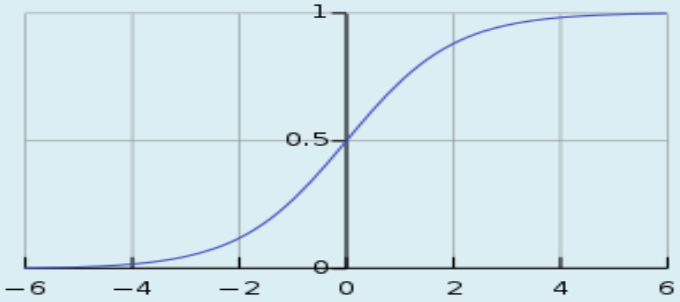
2 Independent variables with weightages : $b0 + b1x1 + b2x2$

$p = e^{(b0 + b1x1 + b2x2)} / (1+ e^{(b0 + b1x1 + b2x2)})$

$$\ln(p/1-p) = b_0 + b_1x_1 + b_2x_2$$

This is the Logistic Regression formula.

Simple : $\ln(p/1-p) = b_0 + b_1x$



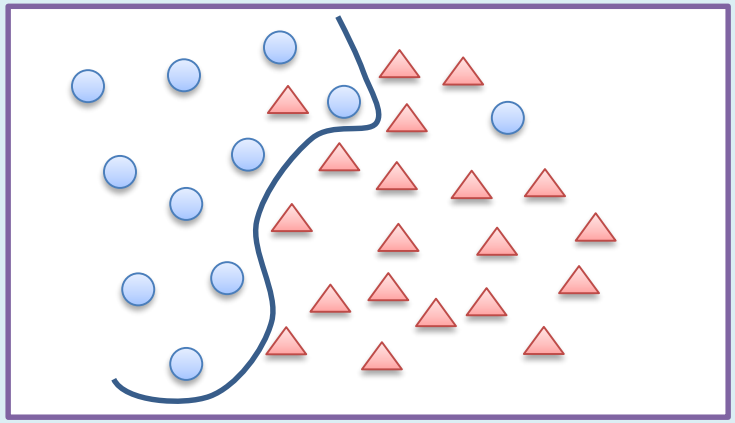
https://en.wikipedia.org/wiki/Logit

Hemant Rathore

Machine Learning P1

Supervised ML Models (Non - Parametric) - Classification

Classifying objects into different classes, drawing Boundaries between groups. Classification is used when dependent variable is categorical.



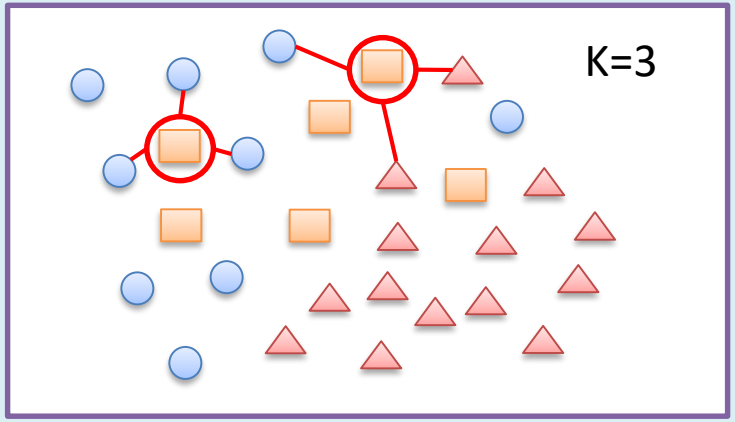
https://en.wikipedia.org/wiki/Logit

Hemant Rathore

Machine Learning P1

Supervised ML Models (Non - Parametric) - Classification Models – K Nearest Neighbor (KNN) - KNN works on majority vote concept.

1. Select some value for K
2. Select 1 unknown object ‘o’ to be classified
3. Search for K nearest neighbors for the unknown object ‘0’
4. Check for the most occurring group type for among neighbors
5. Type of ‘o’ = most occurring group type
6. Repeat this for all the objects of unknown class/type



https://en.wikipedia.org/wiki/Logit

Hemant Rathore

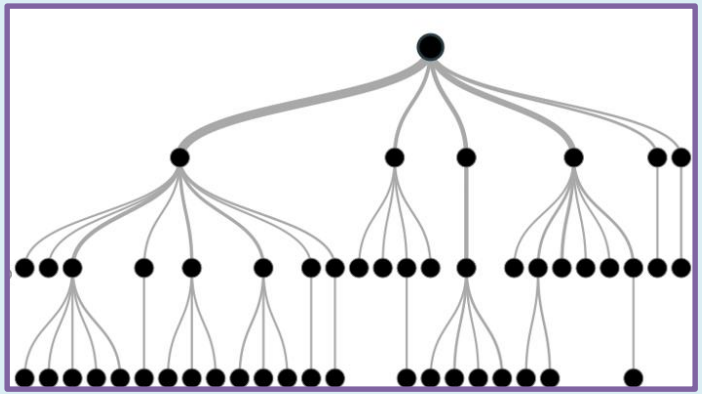
Machine Learning P1

Supervised ML Models (Non - Parametric) - Classification Models – Decision Tree

Decision trees are made by splitting the dataset into nodes, each non leaf node is assigned some conditions and all the data points under that node satisfy the node condition. Leaf nodes give the final result. Decision trees are used when dependent variable is either categorical or continuous but relationship doesn't follow any defined pattern.

Decision trees types (Classification And Regression Tree (CART))

1. **Classification Decision Tree –**
   Categorical Dependent variable
2. **Regression Decision tree**
   Continuous Dependent variable



https://en.wikipedia.org/wiki/Logit

Hemant Rathore

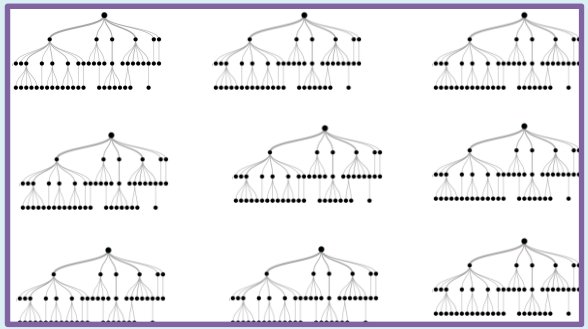
Machine Learning P1

Supervised ML Models (Non - Parametric) - Classification Models – Random Forest

"Collection of Decision Trees"

Random Forest is a type of Ensemble learning (using multiple algorithms or iterations collectively).Decision trees are sometimes not very useful for very large dataset so comes Random Forest.

1. Select some number K – number of data points in each tree and N – number of trees in forest
2. Create a Decision tree using these K data points
3. Repeat step 2 until all the N trees are created
4. For any new object or data pointed to be predicted, predict the value using each tree in the forest
5. Calculate the Average/most frequent of predictions from all the trees, this is the final prediction for given data point.



https://en.wikipedia.org/wiki/Logit

Hemant Rathore

"Qs & As"

Data Science with R

Day 10,11 – Machine Learning - Advanced Models

TACT 360° Training

LEARN. ACHIEVE. STANDOUT

Today's Agenda

✓ ML – Advanced Models

- Unsupervised ML Models
  - Clustering
    - K Means Clustering
    - Hierarchical Clustering
- ML Model Evaluation
  - Under fitting and Overfitting
  - Confusion Metrix
  - K- Fold Cross Validation
  - Regression Evaluation Metrics

✓ ML – Advanced Models

- Time Series Analysis
  - Moving Average
  - Exponential Smoothing
  - ARIMA Time series Models
- Support Vector Machine (SVM)

Machine Learning P2

1. Unsupervised ML Models - Clustering - "No Supervision", When model learns about the data on its own and no training data is provided.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups (clusters).

We can say that Clustering is unsupervised classification where we don't know the classification already.

Type of Clustering Algorithms

1. **Centroid Based Clustering – K Means Clustering**
2. **Connectivity Based Clustering – Hierarchical Clustering**

Hemant Rathore

Machine Learning P2

Unsupervised ML Models - Clustering - K Means Clustering

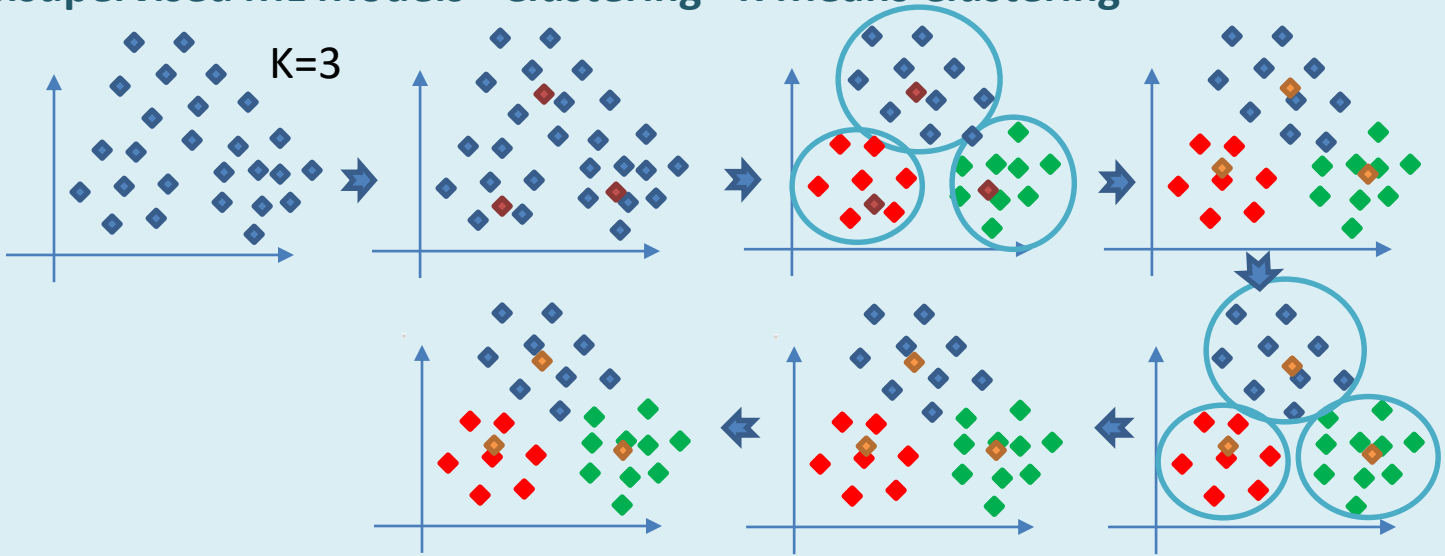
K Means Clustering is centroid based clustering, it makes clusters by finding and grouping the elements near the centroid into same cluster.

1. Randomly select some number K – number of centroids, 1 centroid for each cluster
2. Calculate the distance of all the data points from each centroid
3. Assign each data point to a centroid based on the distance to Centroid (the closest centroid)
4. Once all the points are assigned, calculate the mean for each cluster, and this mean will become the new centroid of the cluster
5. Repeat steps 2,3 until centroids stop moving

Hemant Rathore

Machine Learning P2

Unsupervised ML Models - Clustering - K Means Clustering



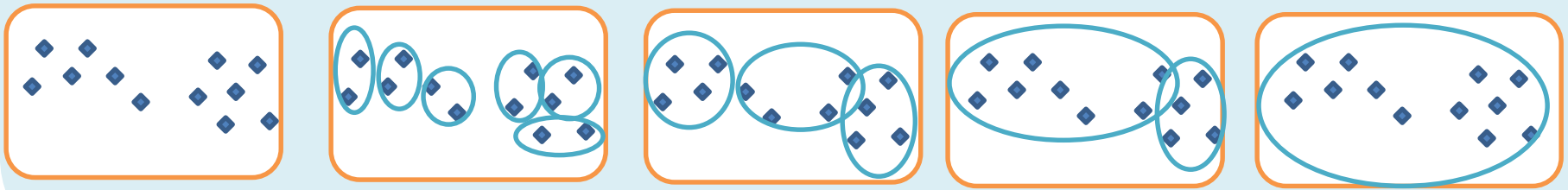
Hemant Rathore

Machine Learning P2

Unsupervised ML Models - Clustering – Hierarchical Clustering

Hierarchical Clustering is a connectivity based based clustering, it makes clusters by merging / dividing the clusters, there are 2 type of Hierarchical Clustering –

1. **Agglomerative – Bottom Up -** Firstly all the points are considered as separate clusters and then nearby clusters are merged together to farm a bigger cluster, this process is repeated until we get a single cluster.
2. **Divisive – Top Down –** Firstly all the data points are considered as part of single one cluster then this single cluster is divided into sub clusters based on proximity until we get separate clusters for each data point



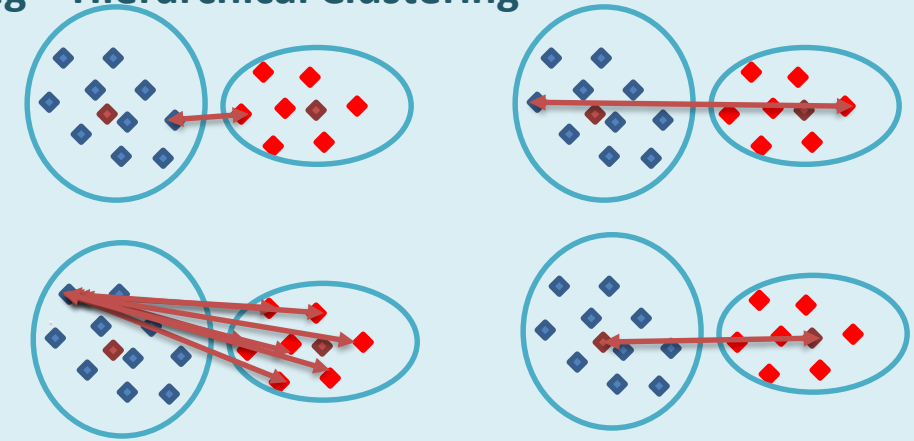
Hemant Rathore

Machine Learning P2

Unsupervised ML Models - Clustering – Hierarchical Clustering

Calculating distance between 2 clusters

1. Single Linkage – Minimum Distance
2. Complete Linkage – Maximum Distance
3. Average Distance – Mean Distance
4. Centroid Linkage – Distance between means



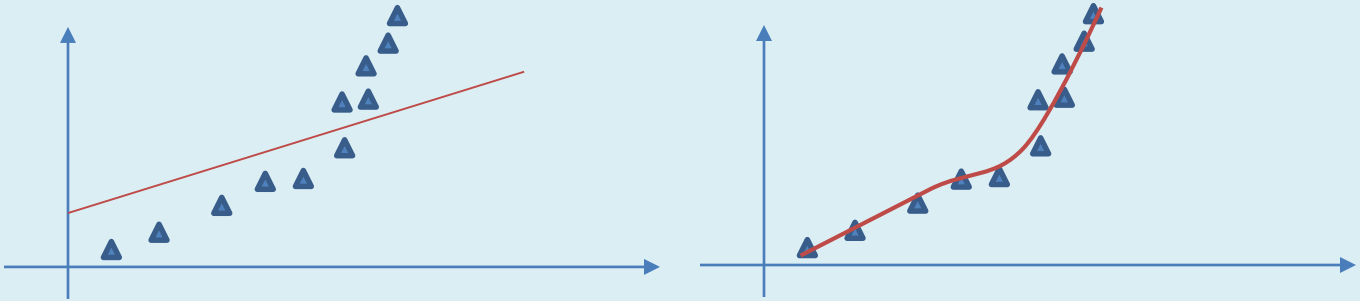
Hemant Rathore

Machine Learning P2

Machine Learning – Why Model Evaluation? - Under fitting and Overfitting

Under fitting – When our model is over generalized and there is the possibility of finding better fitting model this is known as Under fitting, it happens when we don't consider all the possible models/parameters thoroughly or select the model/parameters randomly.

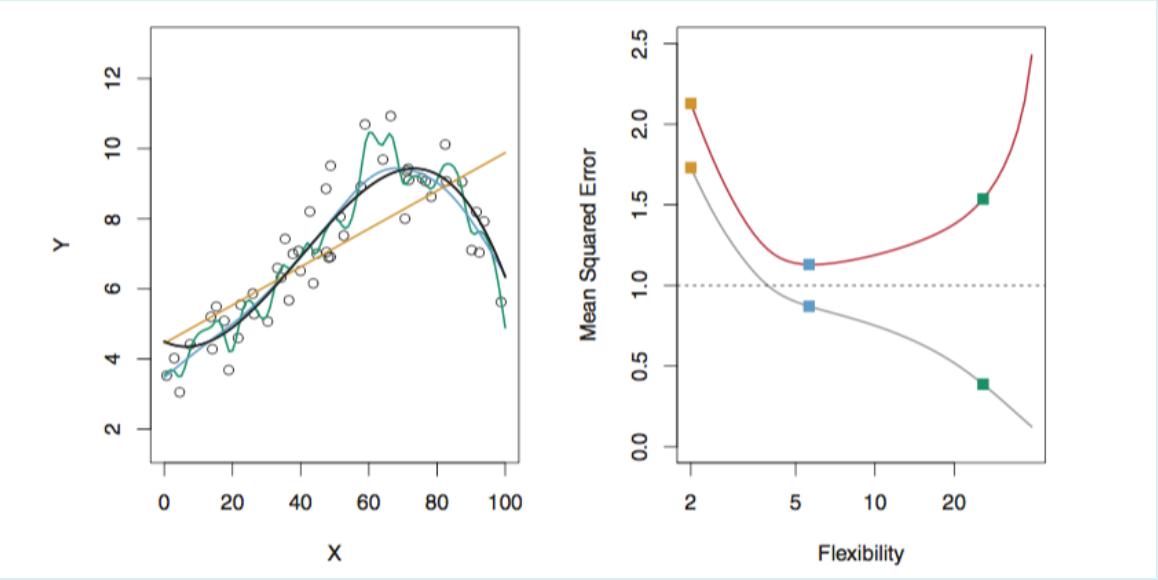
Over fitting – When our model is less generalized and very specific to training data and due to which it gives great result for training data but poor result for test or out of sample data.



Hemant Rathore

Machine Learning P2

Machine Learning – Why Model Evaluation? - Under fitting and Overfitting



Book ref - An Introduction to Statistical Learning

Hemant Rathore

Machine Learning P2

Machine Learning – How to do Model Training & Evaluation "How to train & evaluate the accuracy of the model"

1. Training ,Testing and Cross Validation of ML Model -

1. Train and Test on Same Dataset
2. Train and Test data Split
    1. No overlapping between training and Test data
    2. Overlapped Training and Test data – K fold Cross validation

2. Model Evaluation Methods -

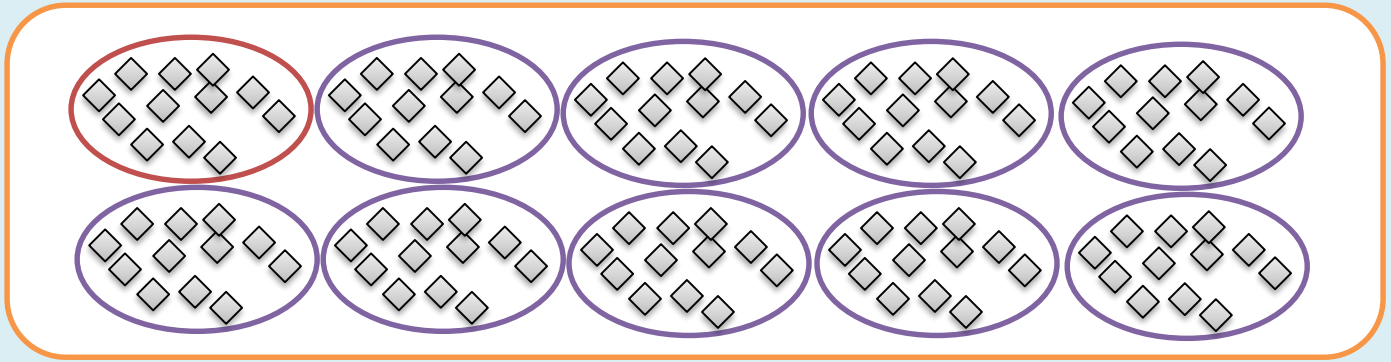
1. Confusion Matrix
2. Regression Evaluation metrics

Hemant Rathore

Machine Learning P2

Machine Learning – Training and Testing of Model – K Fold Cross Validation

We split the total data set into multiple mutually excusive training subset, this number is represented by 'K', now the model is trained on K-1 training sets and tested on the remaining kth data set, this process is repeat until the model is trained and tested for all the K datasets.



Hemant Rathore

Machine Learning P2

Machine Learning – Model Evaluation Methods - Confusion Matrix

Confusion matrix can be used in case of classification or Logistics regression where output is either 100% correct or 100% incorrect

		Actual Result	
		Positive	Negative
Model Outcome	Positive	True Positive	False Positive (Type I)
	Negative	False Negative (Type II)	True Negative

Accuracy % = (True positive + True Negative) / Total

Hemant Rathore

Machine Learning P2

Machine Learning – Model Evaluation Methods – Regression Evaluation Metrics

In case of regression where we have continuous dependent variable confusion matrix will not help so we can use Regression Evaluation Metrics

1. **Mean Absolute Error –**

$$MAE = 1/n \sum [\, y(i) - y(model) \,]$$

2. **Mean Squared Error**

$$MSE = RSS/n = 1/n \sum [\, y(i) - y(model) \,]^2$$

3. **Root Mean squared error**

$$RMSE = \sqrt{MSE} = \sqrt{\{1/n \sum [\, y(i) - y(model) \,]^2\}}$$

Hemant Rathore

Machine Learning P2

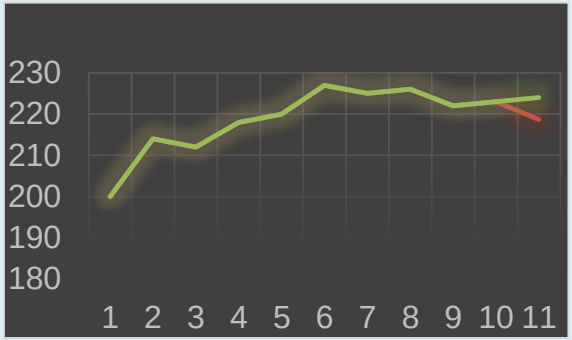
Machine Learning – Time Series Analysis

When we analyze some metric over the period of time so we can say independent variable here is time.

1. Simple Moving Average –The forecast for the value of Y at time t+1 that is made at time t equals the simple average of the most recent m observations

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \dots + Y_{t-m+1}}{m}$$

Obs	1	2	3	4	5	6	7	8	9	10	11
Value	200	214	212	218	220	227	225	226	222	223	
Simple Avg	200	214	212	218	220	227	225	226	222	223	218.7
Moving Avg	200	214	212	218	220	227	225	226	222	223	224



Hemant Rathore

Machine Learning P2

Machine Learning – Time Series Analysis

2. Exponential smoothing –SMA treats all the last n observation equally and completely ignores all the previous observations, Exponential Smoothing helps this by applying exponentially decreasing weightages from latest to older observations, the constant which drive this weightages is known as Smoothing Constant (**α**)

$$\hat{Y}_{t+1} = \alpha[Y_t + (1-\alpha)Y_{t-1} + (1-\alpha)^2 Y_{t-2} + (1-\alpha)^3 Y_{t-3} + \ldots]$$

Smoothing Constant (α) will always be between 0 and 1, lets take α=0.9 then

Y(t+1) = 0.9* y(t) + 0.9*(1-0.9)* y(t-1) + 0.9*(1-0.9)*(1-0.9)* y(t-2) + 0.9*(1-0.9)*(1-0.9)*(1-0.9)* y(t-3) ....

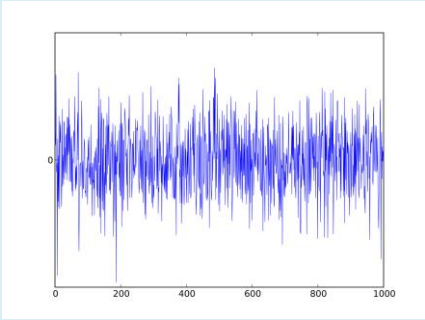
Y(t+1) = 0.9* y(t) + 0.09* y(t-1) + 0.009*y(t-2) + 0.0009* y(t-3) ....

Machine Learning P2

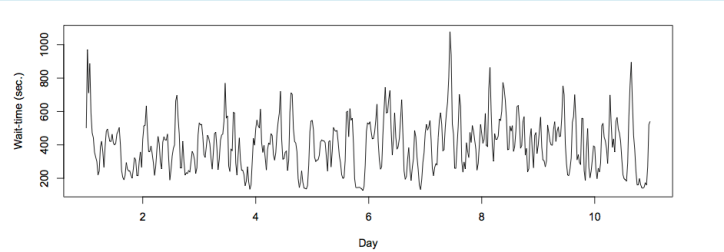
Machine Learning – Time Series Analysis

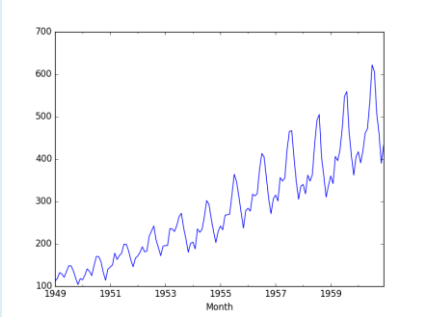
3. AR, MA, ARMA and ARIMA –

White Noise – White Noise can be understood as the signal or TS with Mean = 0 and and some finite variation/standard deviation :



Stationary TS – An stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time.





Machine Learning P2

Machine Learning – Time Series Analysis

3. AR, MA, ARMA and ARIMA –

AR(p) – Auto-Regressive, when next value is dependent only on last p values (p lags), Ex. Supply of umbrella in raining season is dependent on the supply in last season.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_p Y_{t-p} + \varepsilon_t$$

MA(q) – Moving Average, when next value is dependent only on last q errors (q lags), Ex. Demand of Umbrella in raining season is dependent on shortfall in demand last year due to less rainfall, there is still good stock available in market.

$$Y_t = \beta_0 + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots. + \phi_q \varepsilon_{t-q}$$

ARMA(p,q) – Auto-Regressive Moving Average, when next value is dependent on last p values and last q errors, Ex. Supply of Umbrella is dependent on last season's supply and last season's shortfall in supply due to high demand.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots. + \phi_q \varepsilon_{t-q} .$$

Machine Learning P2

Machine Learning – Time Series Analysis

ARIMA (Auto-Regressive Integrated Moving Average) –

ARMA model cannot be applied for Non stationary data. We use Differencing to make the data Stationary.

Differenced variable: $\Delta y_t = y_t - y_{t-1}$

The variable y is integrated of order one denoted by *I* (1).

ARIMA(p,d,q) – Auto-Regressive Integrated Moving Average, when next value is dependent on last p values and last q errors and last d differences. Ex. Stock price for some stock is dependent on last day closing value, last day's rise/fall, trend present in TS

Differenced variable: $\Delta y_t = y_t - y_{t-1}$

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots + \beta_p Y_{t-p} + \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \Phi_3 \varepsilon_{t-3} + \ldots + \Phi_q \varepsilon_{t-q} .$$

Machine Learning P2

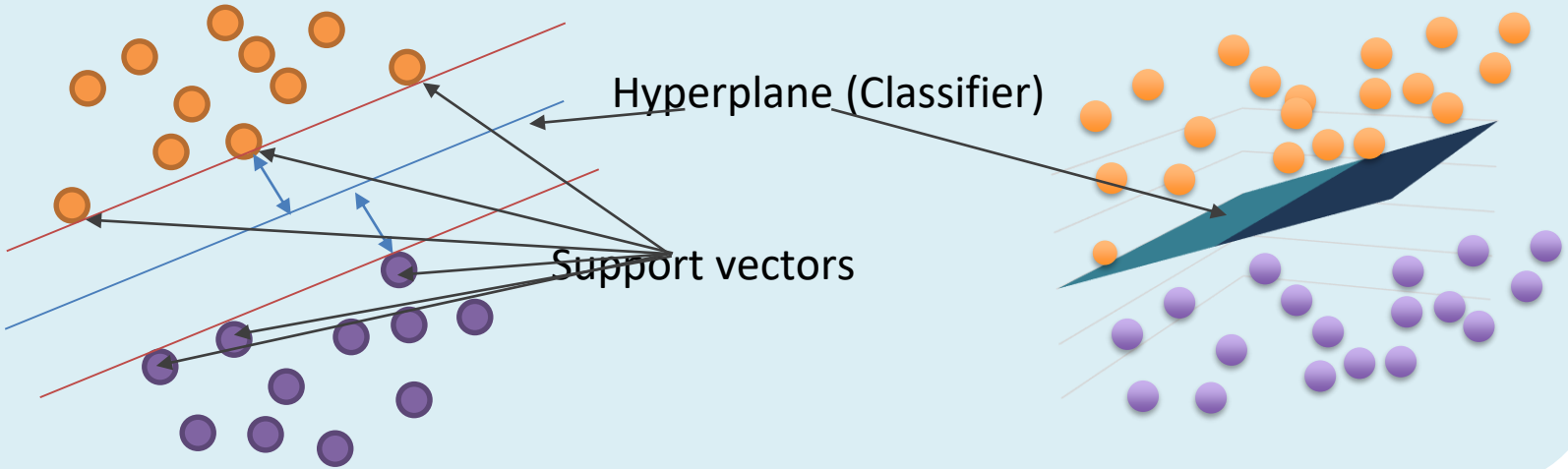
Advanced Machine Learning Models –Support Vector Machine (SVM)

- **Whats is SVM? -** SVM is supervised non-probabilistic binary linear classifier type of algorithm however with the help of kernel trick it can used for non linear classification as well. SVM is used mainly for classification & regression but most popular application is classification.
- **Why SVM?** SVM classification helps in handling the extreme cases by providing a very good margin of classification so it is more accurate as compared to other algorithms.
- **When to use SVM?** SVM works great for smaller data set (~<=1k)

Hemant Rathore

Machine Learning P2

Advanced Machine Learning Models –Support Vector Machine (SVM) – How does it work?



Hemant Rathore

"Qs & As"