Cognixia

A Collabera LEARNING SOLUTIONS COMPANY

Data Science with Python

Day 13-14 – Machine Learning Models & Case Studies

www.cognixia.com

Today’s Agenda

✓ Non Parametric Machine Learning Models
- Decision Trees - Classification and Regression Tree
- Random Forest

✓ Machine Learning Case Study – Regression using Random Forest

✓ Advanced Machine Learning Models
- Support Vector Machine (SVM)
- Clustering
  - K Means Clustering
  - Hierarchical Clustering

✓ **Machine Learning Case Study – Clustering using K-Means Clustering**

- Time Series Analysis
- ARIMA Time series Models
  - AR
  - MA
  - ARMA
  - ARIMA

✓ **Machine Learning Case Study – Time Series Analysis using ARIMA**

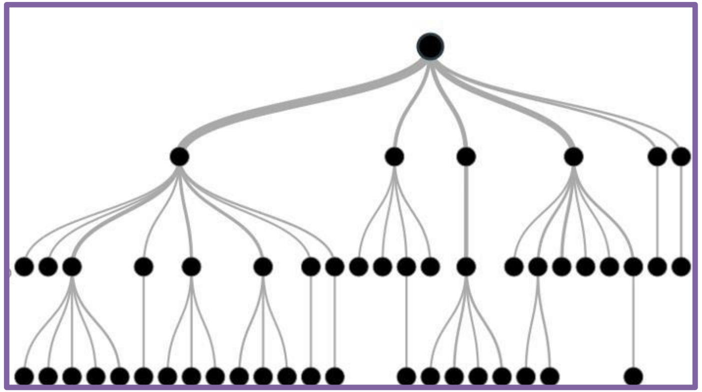
Non Parametric Machine Learning Models

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

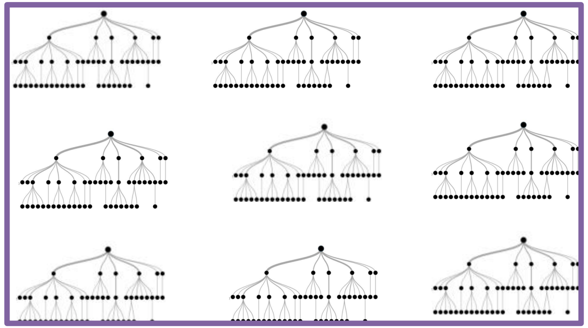


Non Parametric Machine Learning Models

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.

Case Study

- ❑ **Case Study #4 – Random Forest using Regression Tree –** White Wine Quality Rating Prediction

Advanced ML Models

Advanced Machine Learning Models –Support Vector Machine (SVM)

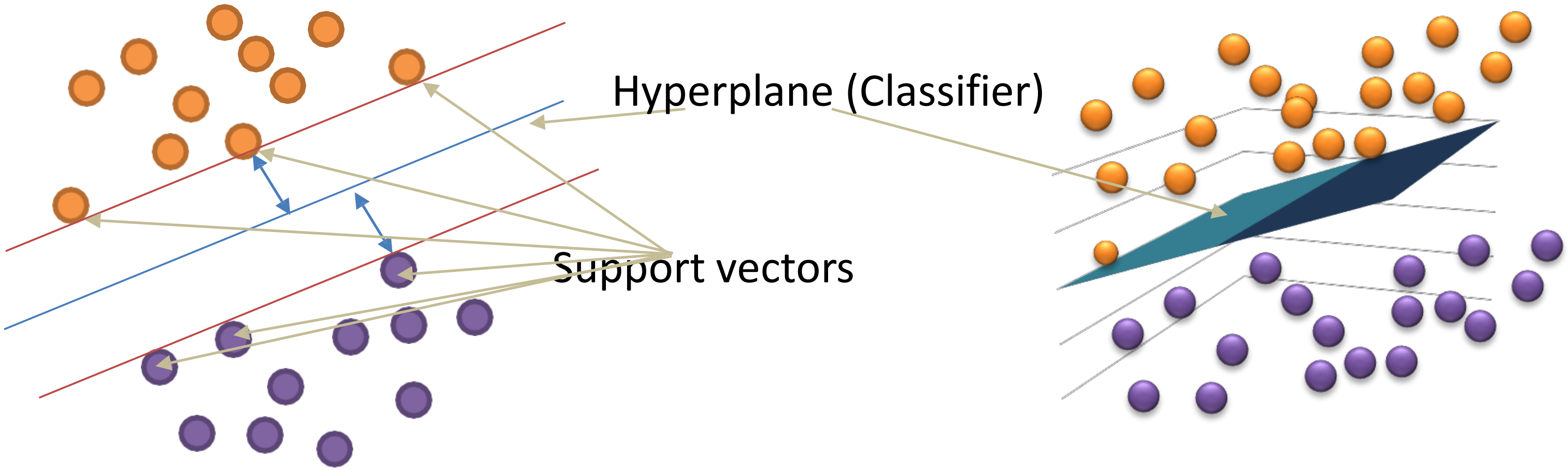
- **Whats is SVM? -** SVM is supervised non-probabilistic binary linear classifier type of algorithm however with the help of kernel trick it can used for non linear classification as well. SVM is used mainly for classification & regression but most popular application is classification.

- **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)

Advanced ML Models

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



Hyperplane (Classifier)

Support vectors

Unsupervised ML Models

Unsupervised ML Models - Clustering - “No Supervision”, When model learns about the data on its own andno 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**

Unsupervised ML Models

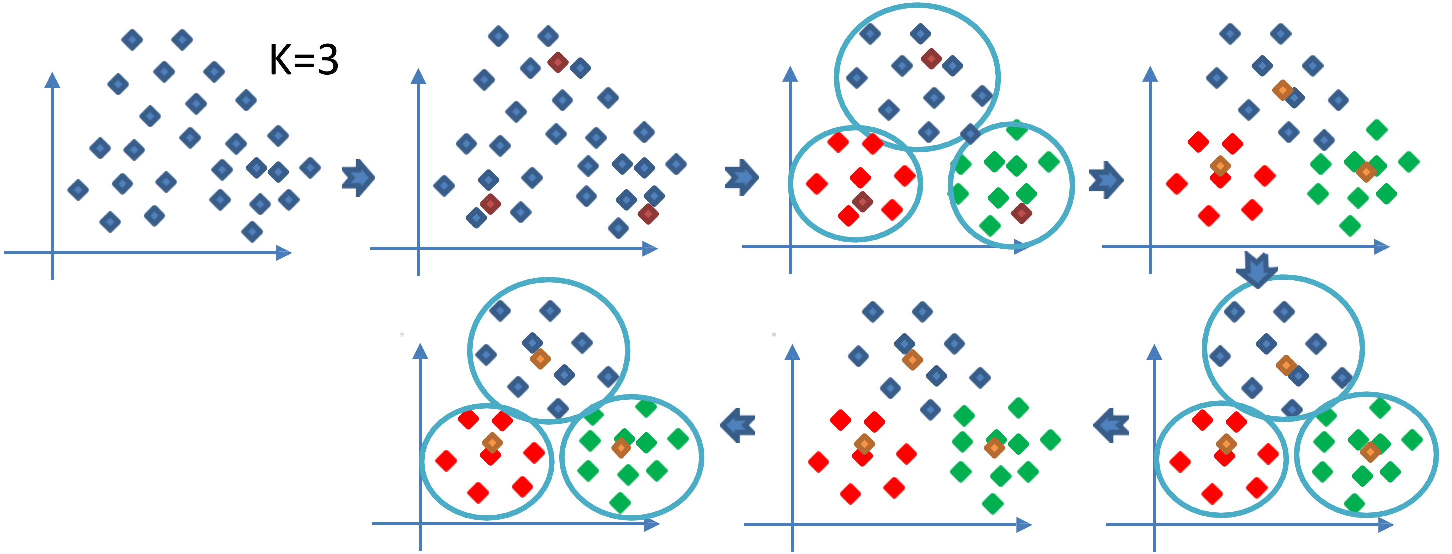
Unsupervised ML Models - Clustering - K Means Clustering

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

Unsupervised ML Models

Unsupervised ML Models - Clustering - K Means Clustering

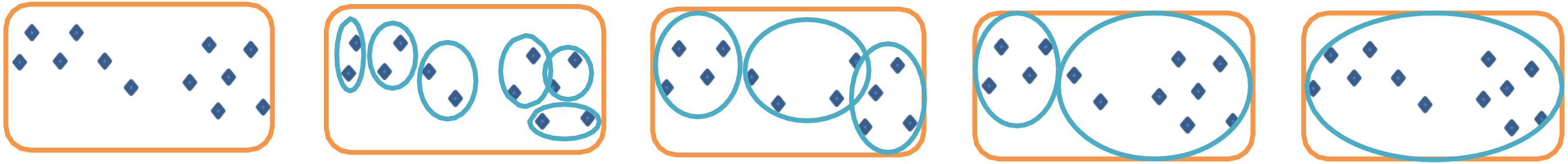


Unsupervised ML Models

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



Case Study

- ❑ **Case Study #5 – Cluster Analysis –** Cab Driver Segmentation Analysis

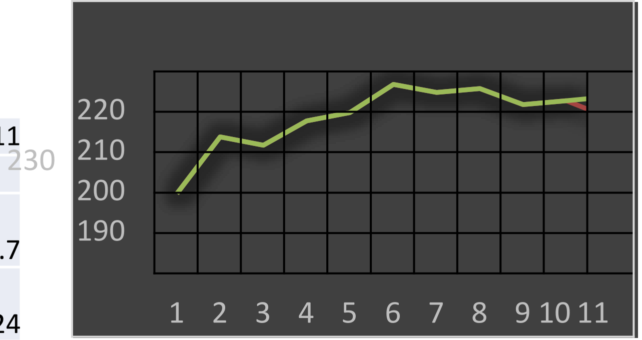
Advanced ML Models

- ## Machine Learning – Time Series Analysis

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

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

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

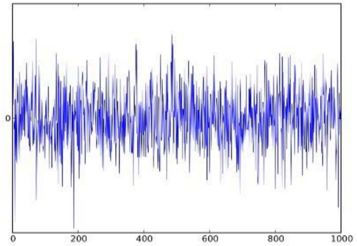


Advanced ML Models

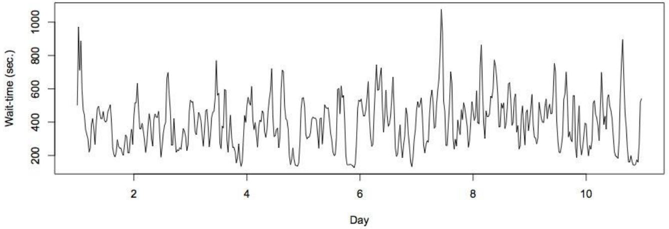
Machine Learning – Time Series Analysis

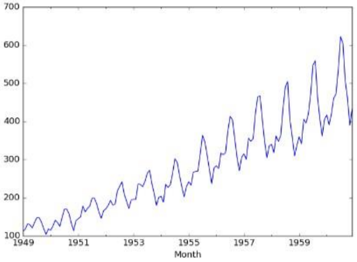
AR, MA, ARMA and ARIMA –

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.





Advanced ML Models

Machine Learning – Time Series Analysis

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.

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

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.

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

Advanced ML Models

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.

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

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

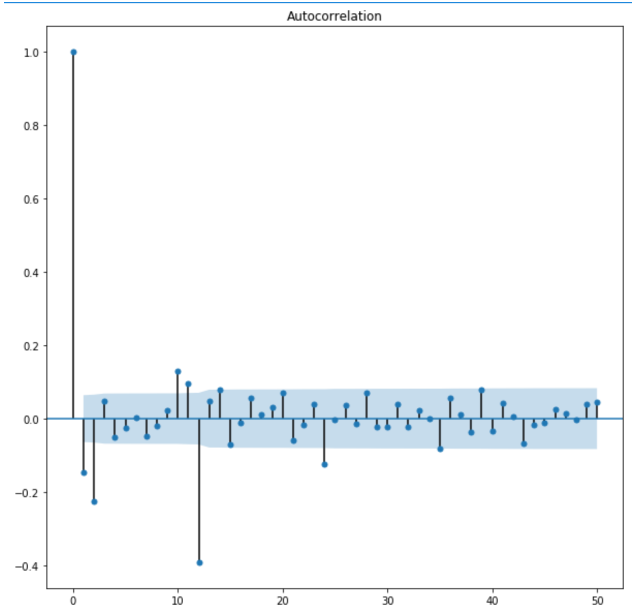
$$\text{Differenced variable: } \Delta y_t = y_t - y_{t-1}$$

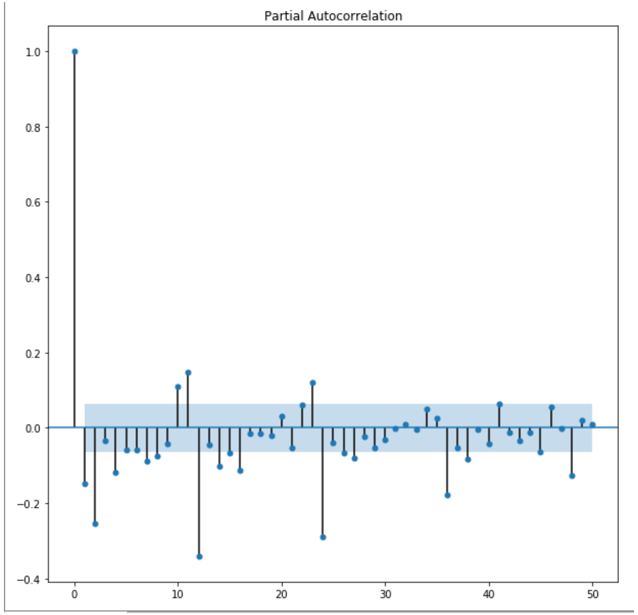
$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \ldots\ldots + \beta_p Y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \ldots. + \phi_q \varepsilon_{t-q}\ .$$

Case Study

- ❑ **Case Study #6 – Time Series Analysis (ARIMA) –** Energy Production Forecasting.

Case Study


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
40
50


Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
10
20
30
40
50

Cognixia

A Collabera LEARNING SOLUTIONS COMPANY

THANK YOU

www.cognixia.com