Cognixia

A Collabera LEARNING SOLUTIONS COMPANY

Data Science with Python

Day 10,11 – Machine Learning Concepts

Today's Agenda

✓ Machine Learning Basics

- What is Machine Learning?
- Supervised & Unsupervised ML Models
- Parametric & Non- Parametric Models
- Covariance and Correlation
- Regression
- Linear Regression
- Linear Regression methods
  - Ordinary Least Square
  - R Squared method
- Why Logistic Regression?
- Logistic Regression Function

✓ Machine Learning Techniques

- ML Model Validation
- Training & Testing
- Under fitting and Overfitting
- Confusion Metrix
- K- Fold Cross Validation
- Regression Evaluation Metrics

Machine Learning Basics

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

Machine Learning Basics

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

Machine Learning Basics

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

Machine Learning Basics

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.

Supervised Machine Learning Models

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

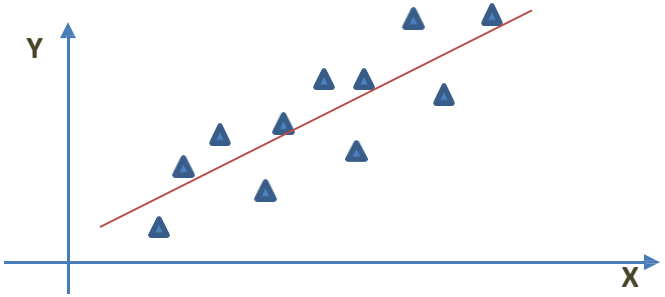
1. **Linear Regression**
   - Simple Linear Regression – Only 1 Independent Variable
   - Multiple linear Regression – More than 1 Independent Variables
2. **Logistic Regression**
   - Simple Logistic Regression– Only 1 Independent Variable
   - Multiple Logistic Regression– More than 1 Independent Variables

Supervised Machine Learning Models

Supervised ML Models (Parametric) - Linear Regression –

"When relationship follows a straight line"

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

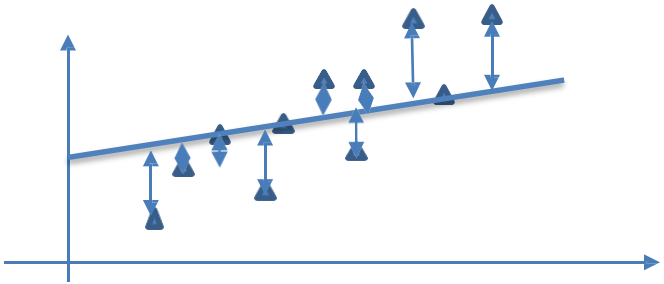


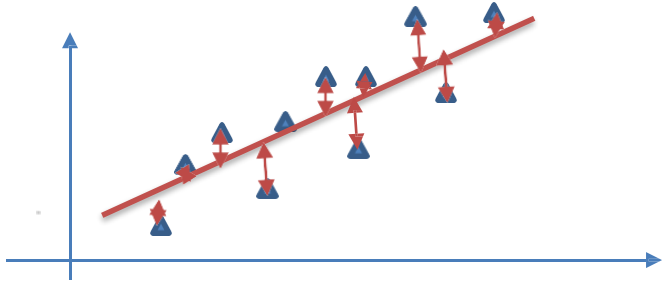
Multiple Linear Regression – More than 1 Independent Variables – $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots\ldots B_nX_n$

Supervised Machine Learning Models

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 allthe points.





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

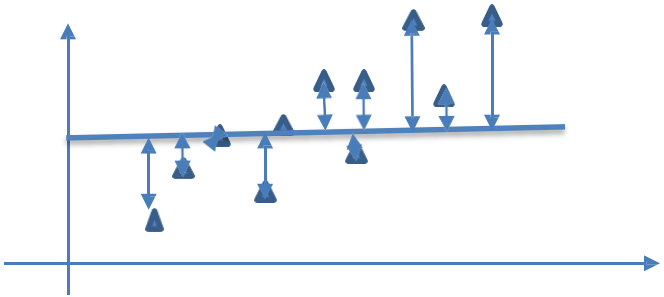
Supervised Machine Learning Models

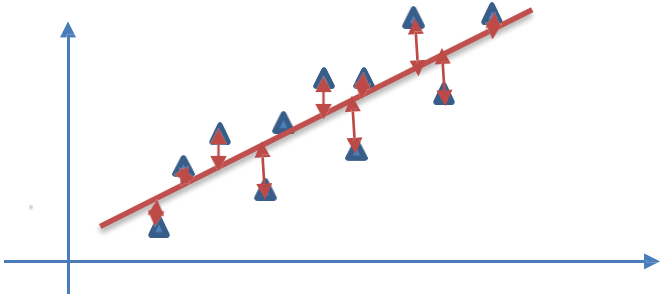
Supervised ML Models (Parametric) - Linear Regression Methods – "how to get best fitting line"

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.

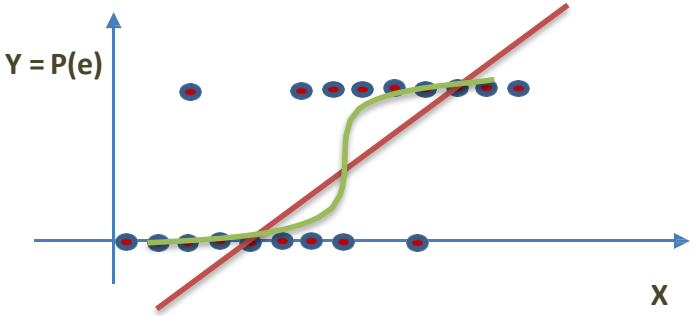




Supervised Machine Learning Models

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

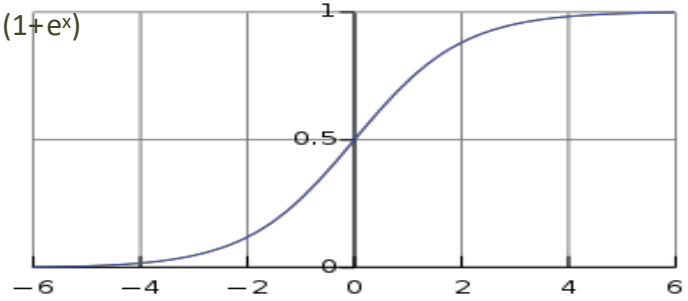
Logistic Regression predicts the probability of an event to occur or not.



Supervised Machine Learning Models

- 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

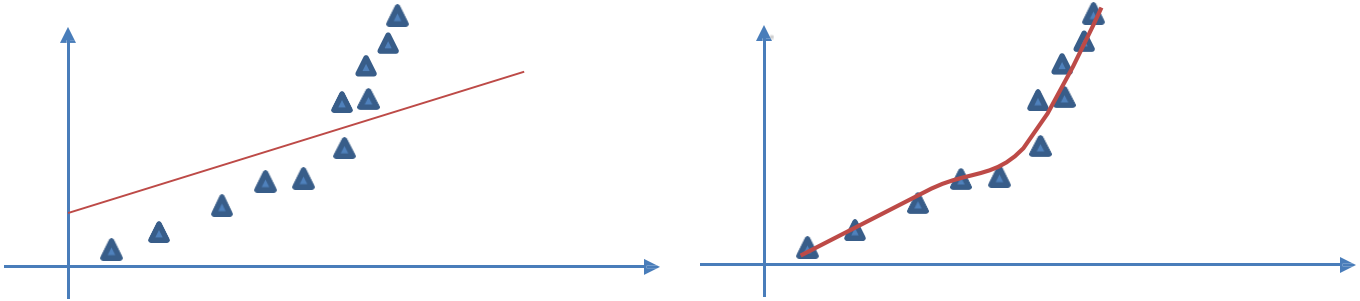


Machine Learning Techniques

Machine Learning – Why Model Validation? - 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.



Machine Learning Techniques

Machine Learning – How to do Model Training & Validation "How to train & validate 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

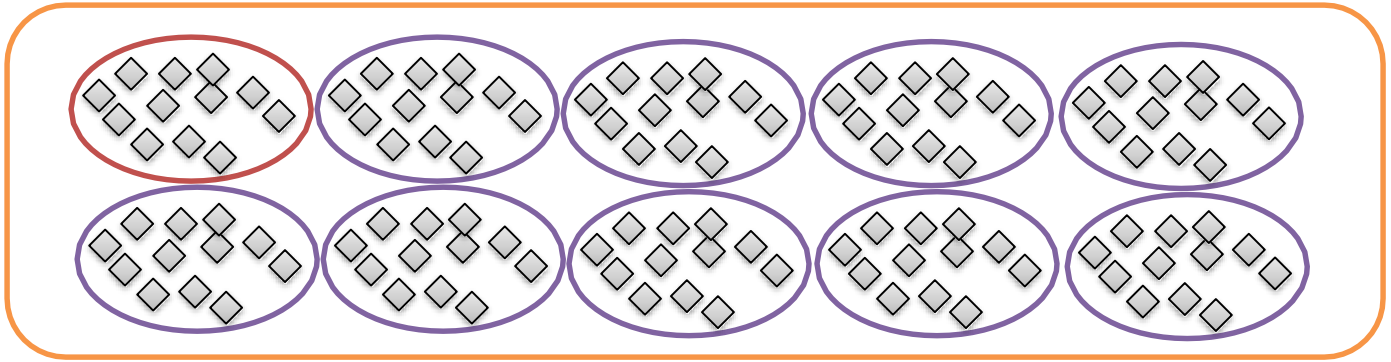
2. Model Evaluation Methods -

1. Confusion Matrix
2. Regression Evaluation metrics

Machine Learning Techniques

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.



Machine Learning Techniques

Machine Learning – Model Evaluation Methods - Confusion Matrix

Confusion matrix can be used in case of classification or Logistics Regression -

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

Machine Learning Techniques

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**

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


Cognixia
A Collabera LEARNING SOLUTIONS COMPANY
THANK YOU
45%