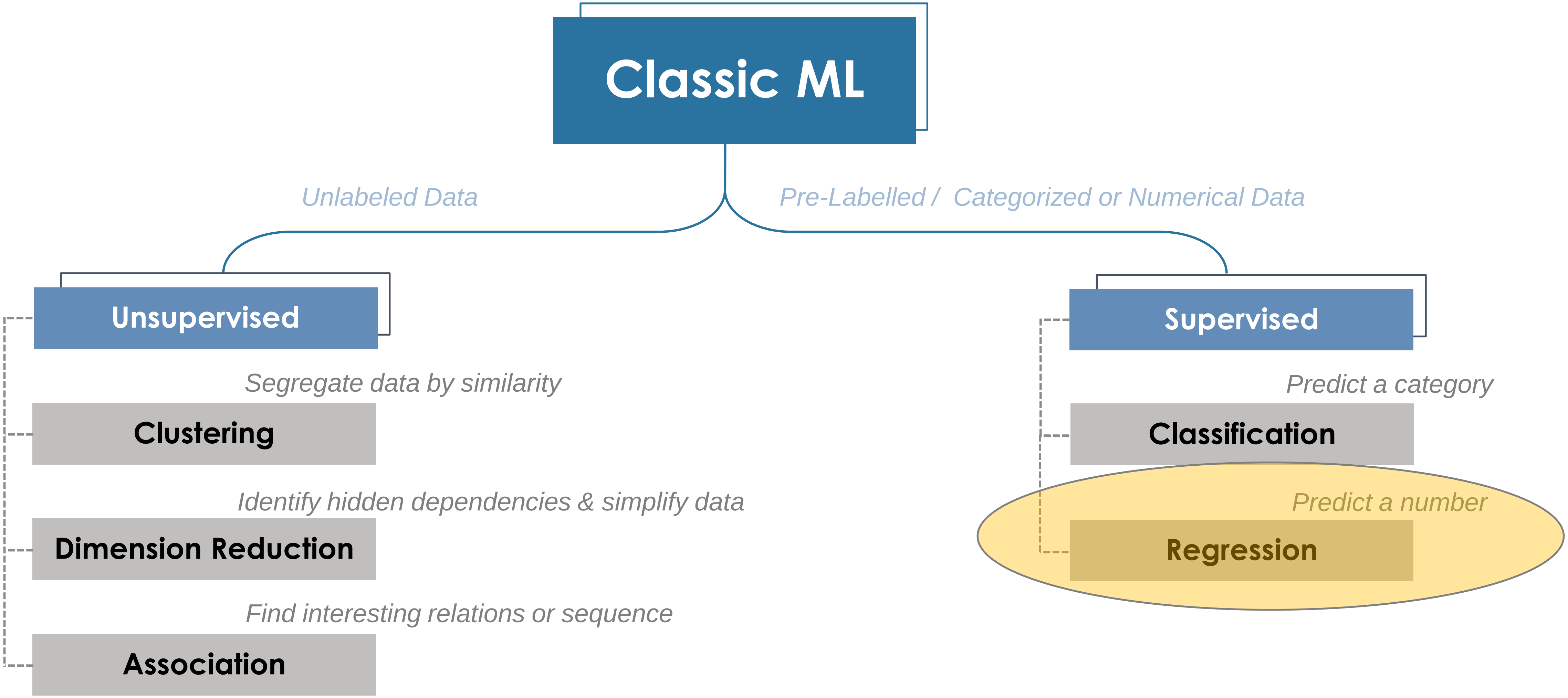
Machine Learning - Part V

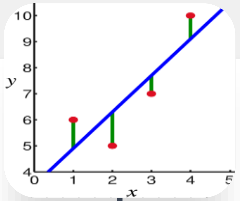
Curated by Arockia Liborious

Linear Regression

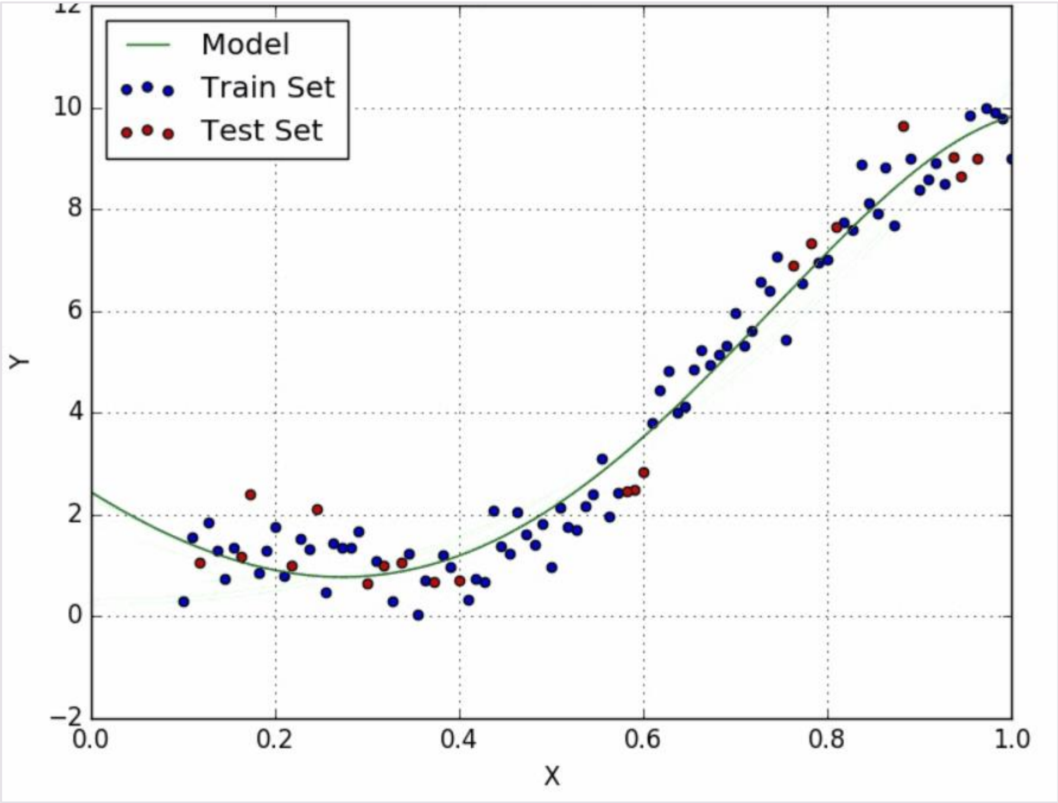
Classic Machine Learning Methods



Supervised Learning - Regression



How it works?



What is it?

- Predicting a specific point on a numerical axis-based relationship between a dependent (target) and independent variables (predictor).
- Like separating shirts by color or size.
- There are two types: Simple and Multiple

Realtime use cases:

- Sales prediction
- Stock price forecasts
- Accident prediction over time

Popular algorithms: Linear and Polynomial regression

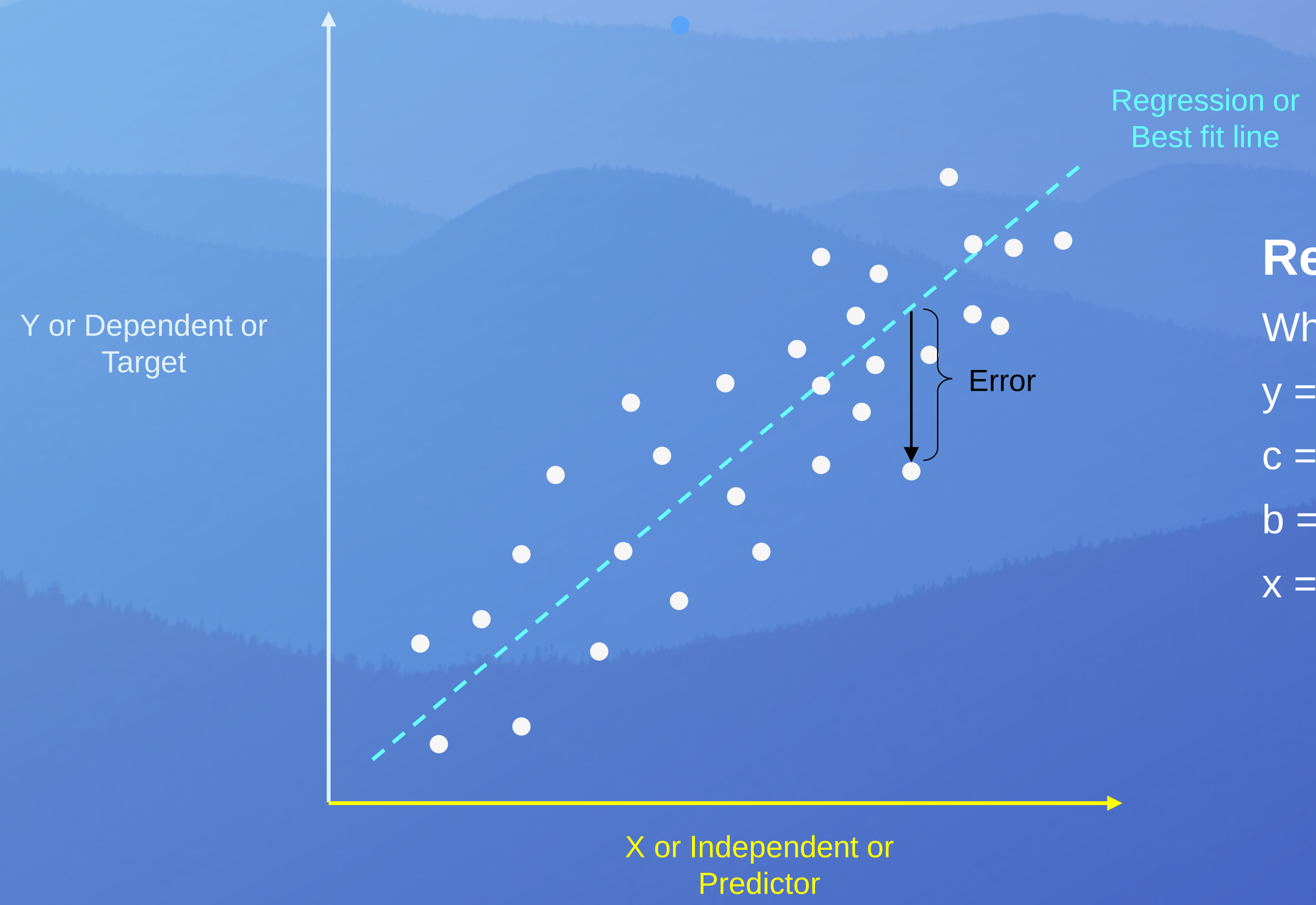
Business Solutions Leveraging Linear Regression

- Business Growth Drivers
- Category Landscape Assessment
- Marketing Mix Modelling

Linear Regression

- Find a line that best fits the data
- The best fit line is the one for which total prediction error (all data points) are as small as possible
- Error is the distance between the point to the regression line

Linear Regression



Regression Equation: y = c + b*x

Where,

y = dependent variable

c = constant

b = regression coefficient

x = independent variable

Additonal Resources:-

- https://www.statisticssolutions.com/what-is-linear-regression/
- https://machinelearningmastery.com/linear-regression-for-machine-learning/

Key Assumptions of Linear Regression

- Normality
- Multicollinearity
- Homoscedasticity
- Autocorrelation

Normality: It is assumed that the error terms, are normally distributed. Check if data is normal distributed to avoid bias

Multicollinearity: Variables must be independent of each other. Use correlation matrix to check this.

Homoscedasticity: It is assumed that the residual terms have the same (but unknown) variance, σ^2. Check for homogeneity of data to reduce variance of output

Autocorrelation: No autocorrelation of residuals. Autocorrelated dependent also results in autocorrelated error

For more details:-
http://r-statistics.co/Assumptions-of-Linear-Regression.html

GitHub Sample Code: https://github.com/itsual/Linear-Regression

Python Coding Steps for Linear Regression:-

1. Import the relevant libraries (Numpy, Pandas, Sklearn, Seaborn, Matplotlib.pyplot)
2. Define some helper functions
3. Load the data
4. Clean the data
5. Feature Engineering
6. Analyze the dataset / EDA
7. Divide the dataset into training and test dataset
8. Train several models and analyzing their performance
9. Select a model and evaluate using test dataset
10. Improve the model by finding the best hyper-parameters and features
11. Analyze the residuals

✓ Tips:
   - Regplot can be used for simple linear regression
   - To analyze Multiple Linear regression outputs use below code
     - print(fit().summary())

Metrics for model evaluation

01 R Squared Value (R^2)

Ranges from 0 to 1. “1" = predictor perfectly accounts for all the variation in Y. "0“ = that predictor X accounts for no variation in Y

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

02 Regression sum of squares (SSR)

Tells how far estimated regression line is from the horizontal no relationship line (Avg. actual output)

$$\text{Error} = \sum_{i=1}^{n}(\text{Predicted\_output} - \text{average\_of\_actual\_output})\text{\^{}2}$$

03 Sum of Squared error (SSE)

Tells how much the target value varies around the regression line (predicted value)

$$\text{Error} = \sum_{i=1}^{n}(Actual\_output - predicted\_output)\text{\^{}2}$$

04 Root Mean Square Error (RMSE)

Some kind of normalized distance b/w the vector of predicted values and the vector of observed values

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

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values
$y_1, y_2, \ldots, y_n$ are observed values
n is the number of observations

05 Correlation co-efficient (r)

Related to R^2 & it ranges from -1 to 1

r = (+/-) sqrt(r^2)

06 Null-Hypothesis and P-value

Null hypothesis is the initial claim.
Low P-value: Reject null hypothesis.
High P-value: Fail to Reject

- Null hypothesis is the initial claim that researcher specify using previous research or knowledge.
- Low P-value: Rejects null hypothesis indicating that the predictor value is related to the response
- High P-value: Changes in predictor are not associated with change in target

Reference: Towards Data Science

To know more… Follow me

Website

arockialiborious.com/

LinkedIn

linkedin.com/in/arockialiborious/

Email

arockialiborious@gmail.com