CM2604 Machine Learning

Supervised Machine Learning - Part 3

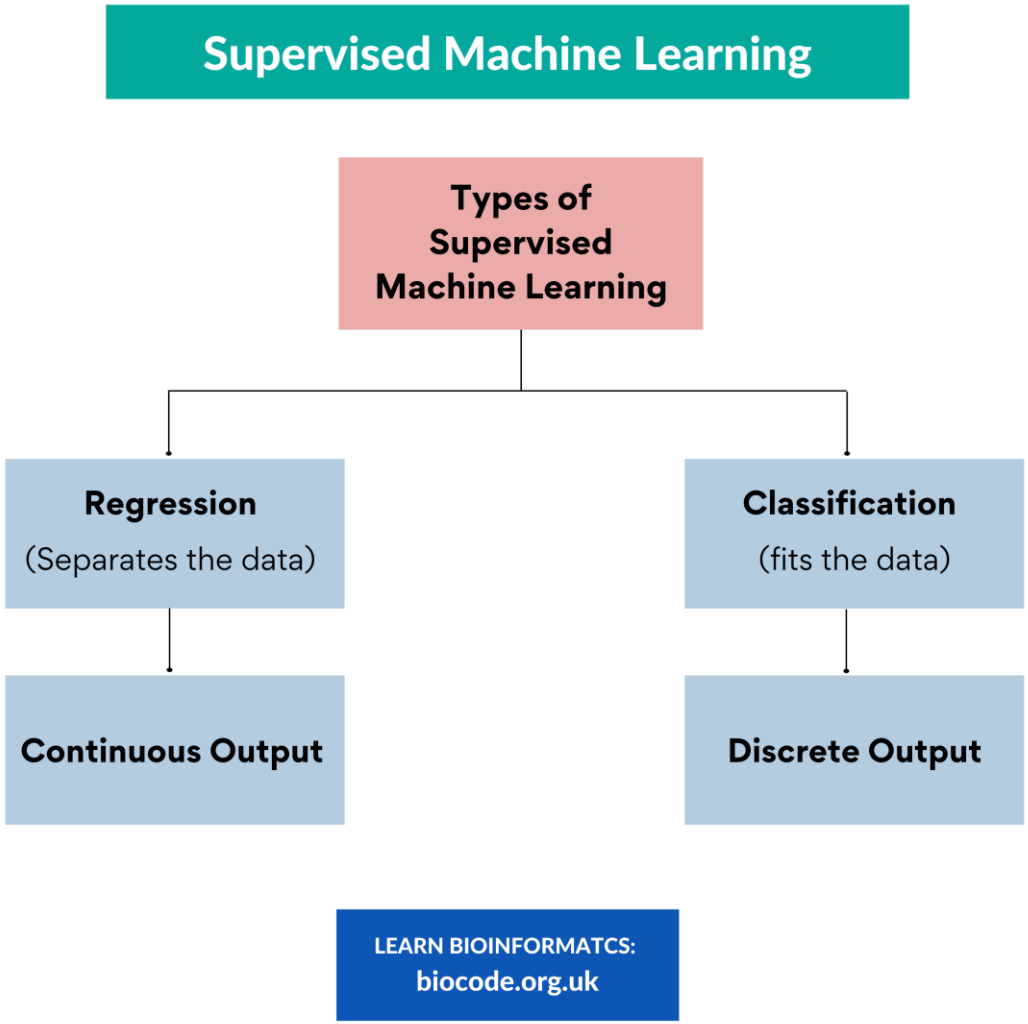
Week 05 | Prasan Yapa

ROBERT GORDON UNIVERSITY ABERDEEN
TEF Gold
INFORMATICS INSTITUTE OF TECHNOLOGY

Overview

- General Framework of Supervised Learning
- Regression in ML
- Linear Regression
- Logistic Regression

General Framework of Supervised Learning



Regression in ML

Regression Analysis in Machine Learning

- Regression analysis is a statistical method to model the relationship between a dependent and independent variables.
- Regression helps to understand how the value of the dependent variable is changing.
- It predicts continuous/real values such as **temperature, age, salary, price**, etc.

Regression Analysis in Machine Learning

- Suppose there is a marketing company A, who does various advertisement every year and get sales on that.
- Now, the company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year.

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Regression Analysis in Machine Learning

- Regression is a supervised learning technique which helps in finding the correlation between variables.
- In Regression, we plot a graph between the variables which best fits the given datapoints.
- Regression shows a line or curve that passes through all the datapoints.
- The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Terminologies

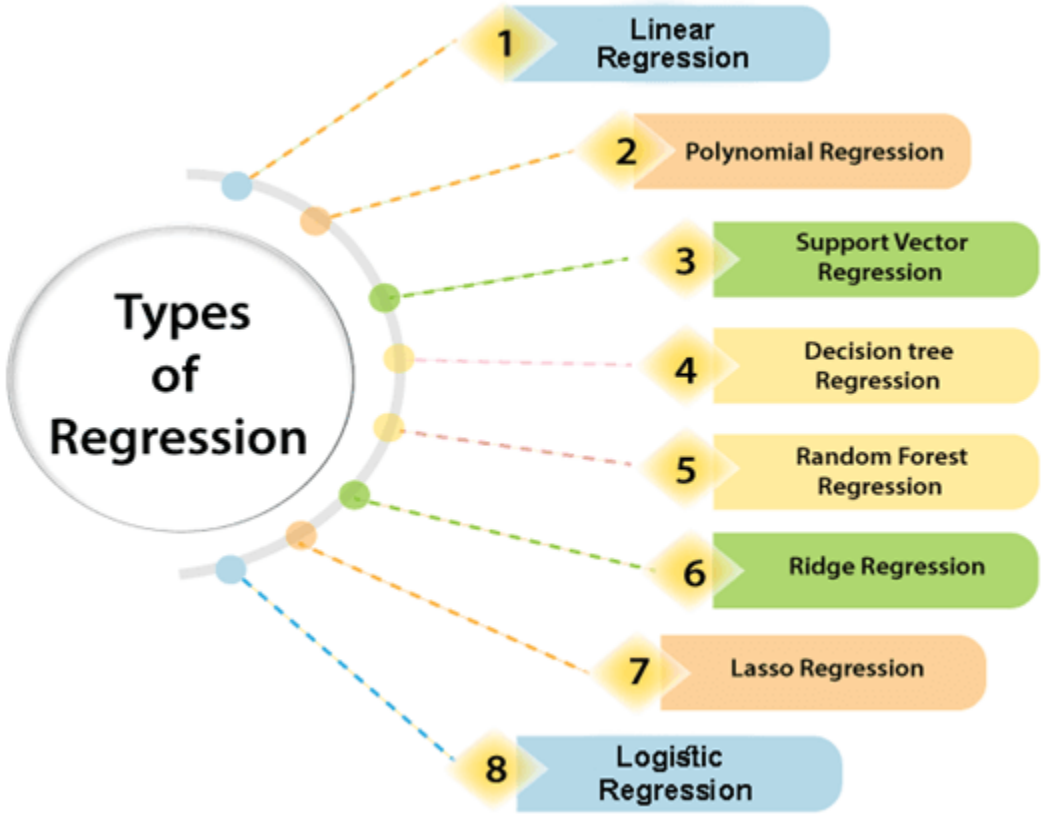
- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.
- **Independent Variable:** The factors which affect the dependent variables, or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our algorithm does not perform well even with training dataset, then such problem is called underfitting.

Why Regression Analysis?

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data.
- It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the most important factor, the least important factor.

Types of Regression

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Ridge Regression
- Lasso Regression



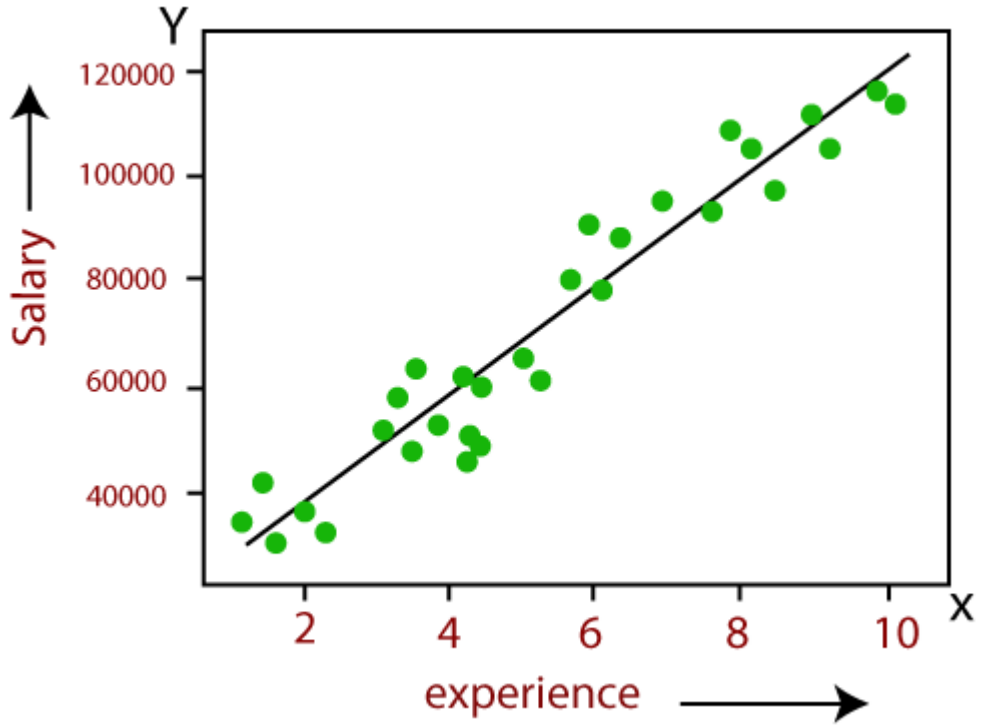
Linear Regression

Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).
- If there is only one input variable (x), then such linear regression is called simple linear regression.
- If there is more than one input variable, then such linear regression is called multiple linear regression.

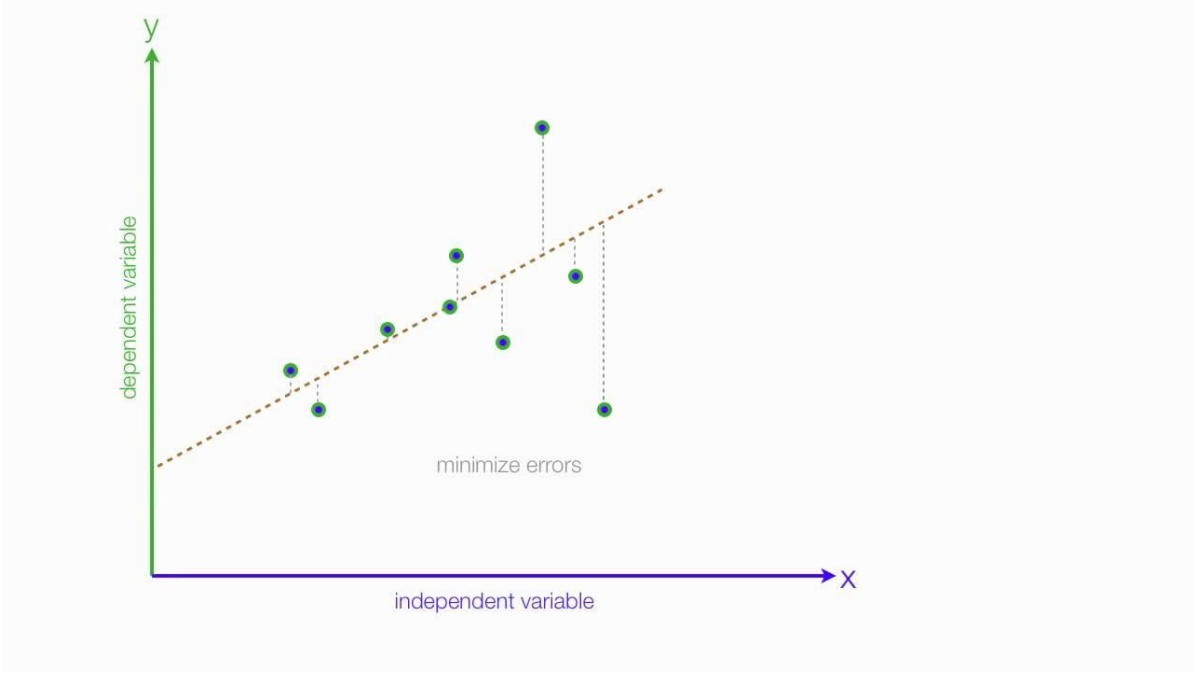
Linear Regression

- $Y = aX + b$
- Y = dependent variables (target variables), X= Independent variables (predictor variables), a and b are the linear coefficients



Linear Regression

- Analyzing trends and sales estimates.
- Salary forecasting.
- Real estate prediction.
- Arriving at ETAs in traffic.



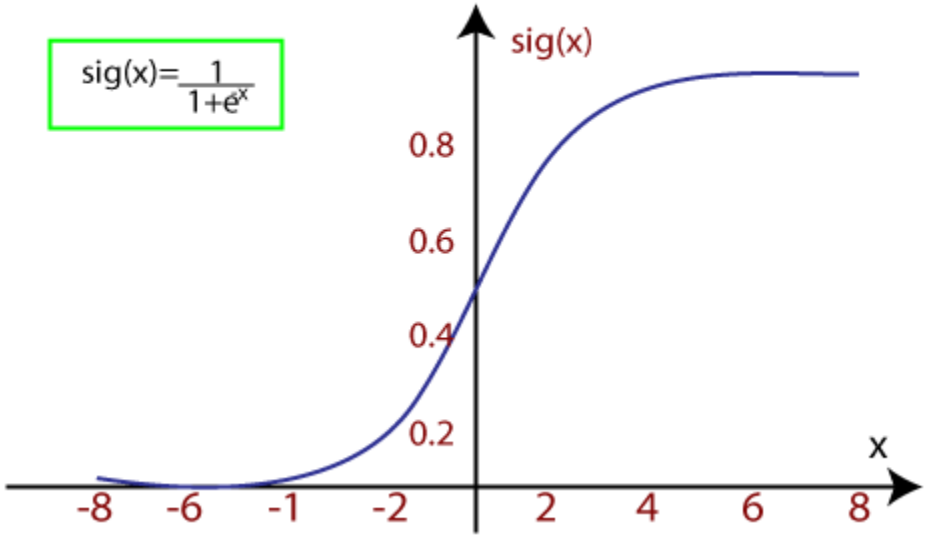
Logistic Regression

Logistic Regression

- Logistic regression is another supervised learning algorithm which is used to solve the classification problems.
- Logistic regression algorithm works with categorical variables.
- It is a predictive analysis algorithm which works on the concept of probability.
- Logistic regression uses **sigmoid function** or logistic function which is a complex cost function.

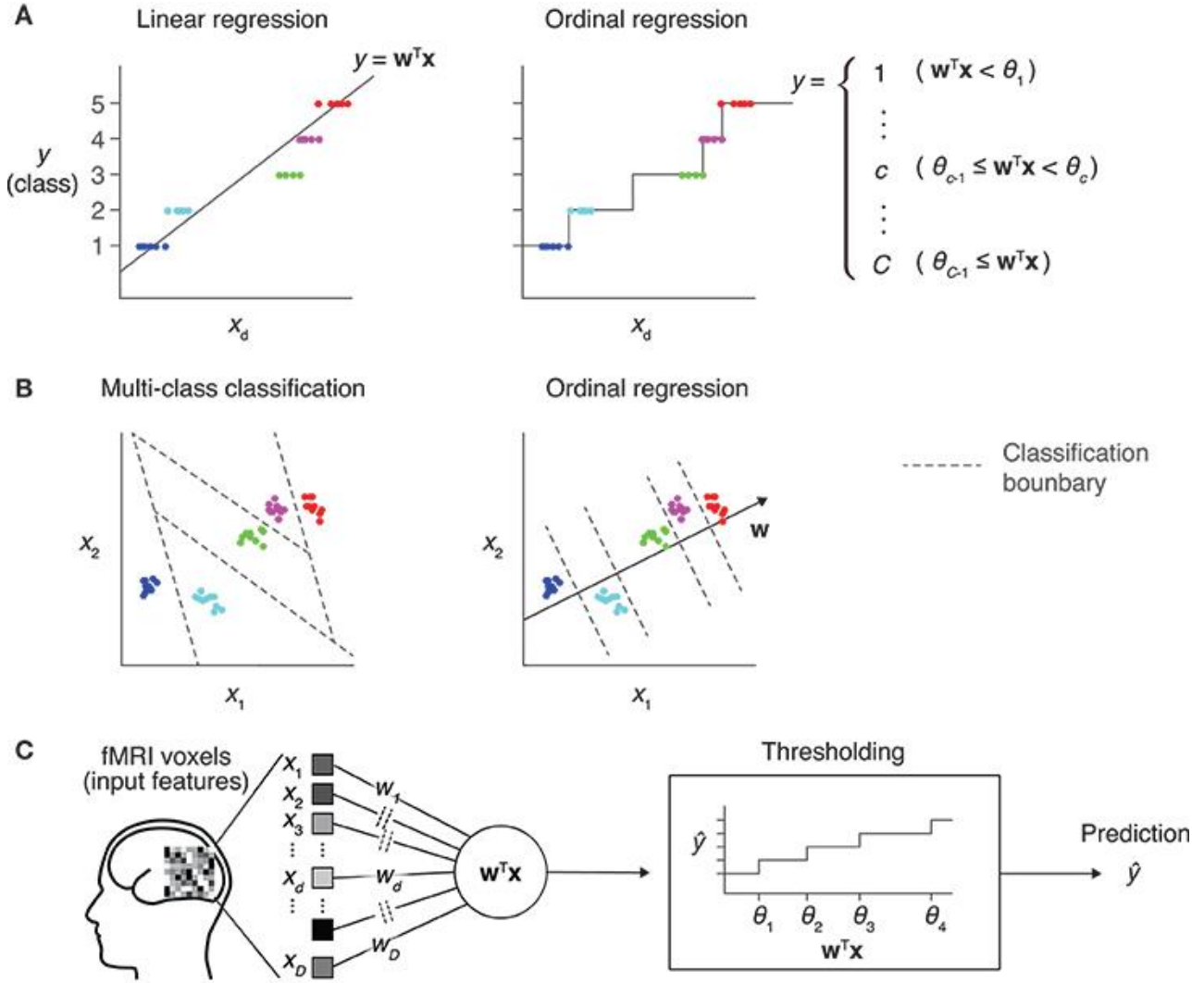
Logistic Regression

- The sigmoid function is used to model the data in logistic regression.
- $f(X) = \frac{1}{1+e^{-x}}$
  - f(x)= Output between the 0 and 1 value.
  - x= input to the function
  - e= base of natural logarithm.
- The S-curve will be as follows.



Logistic Regression

- Types
  - Binary(0/1, pass/fail)
  - Multi(cats, dogs, lions)
  - Ordinal(low, medium, high)



Questions