Machine Learning Algorithms

20CP401T

Himanshu K. Gajera
Department of Computer Science & Engineering
Pandit Deendayal Energy University, Gandhinagar

Learning Algorithms

Supervised learning

- Prediction
- Classification (discrete labels), Regression (real values)

Unsupervised learning

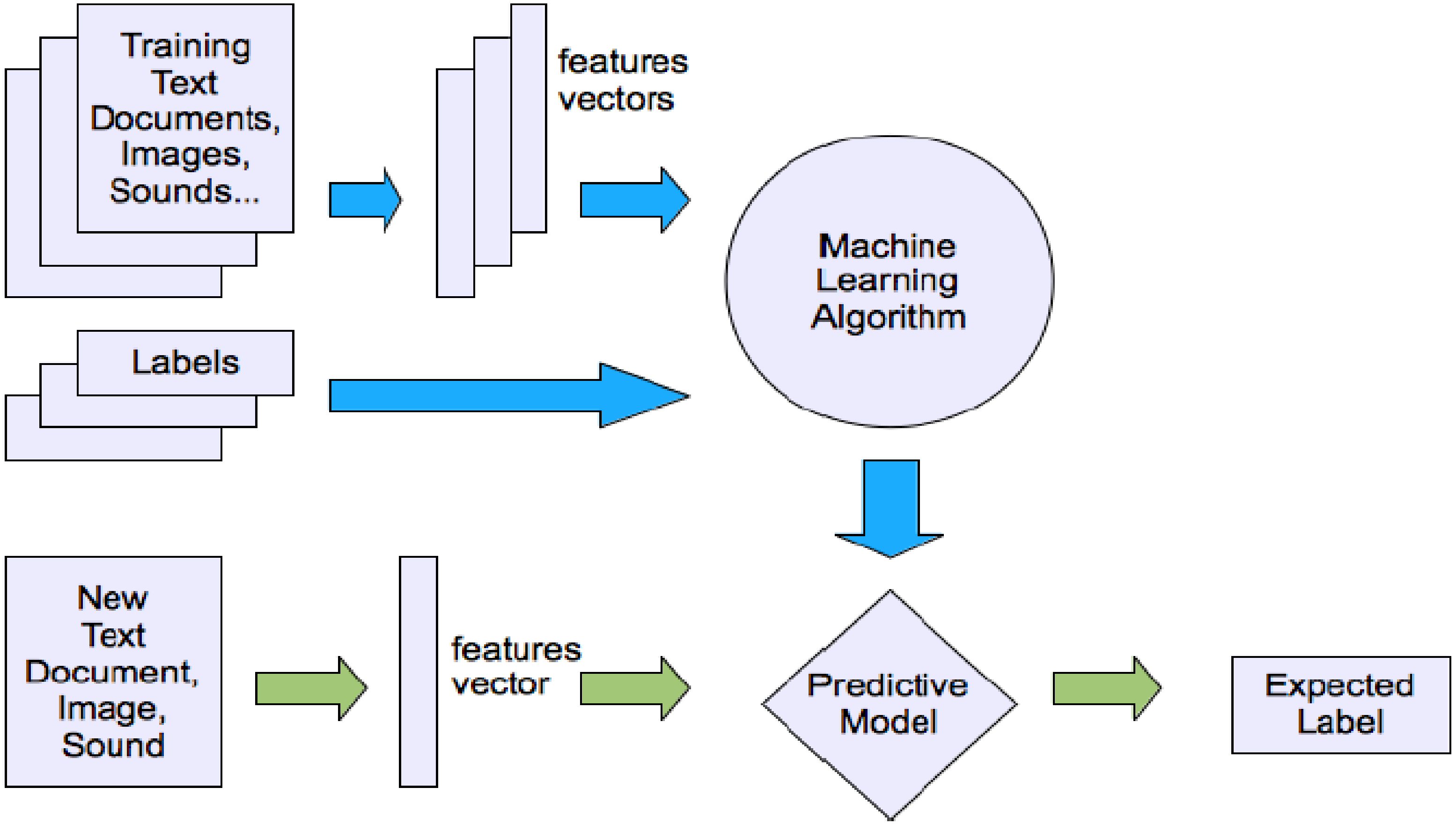
- Clustering
- Probability distribution estimation
- Finding association (in features)
- Dimension reduction (not all)

Semi-supervised learning
Reinforcement learning
Transfer learning
Ensemble learning

Learning Algorithms

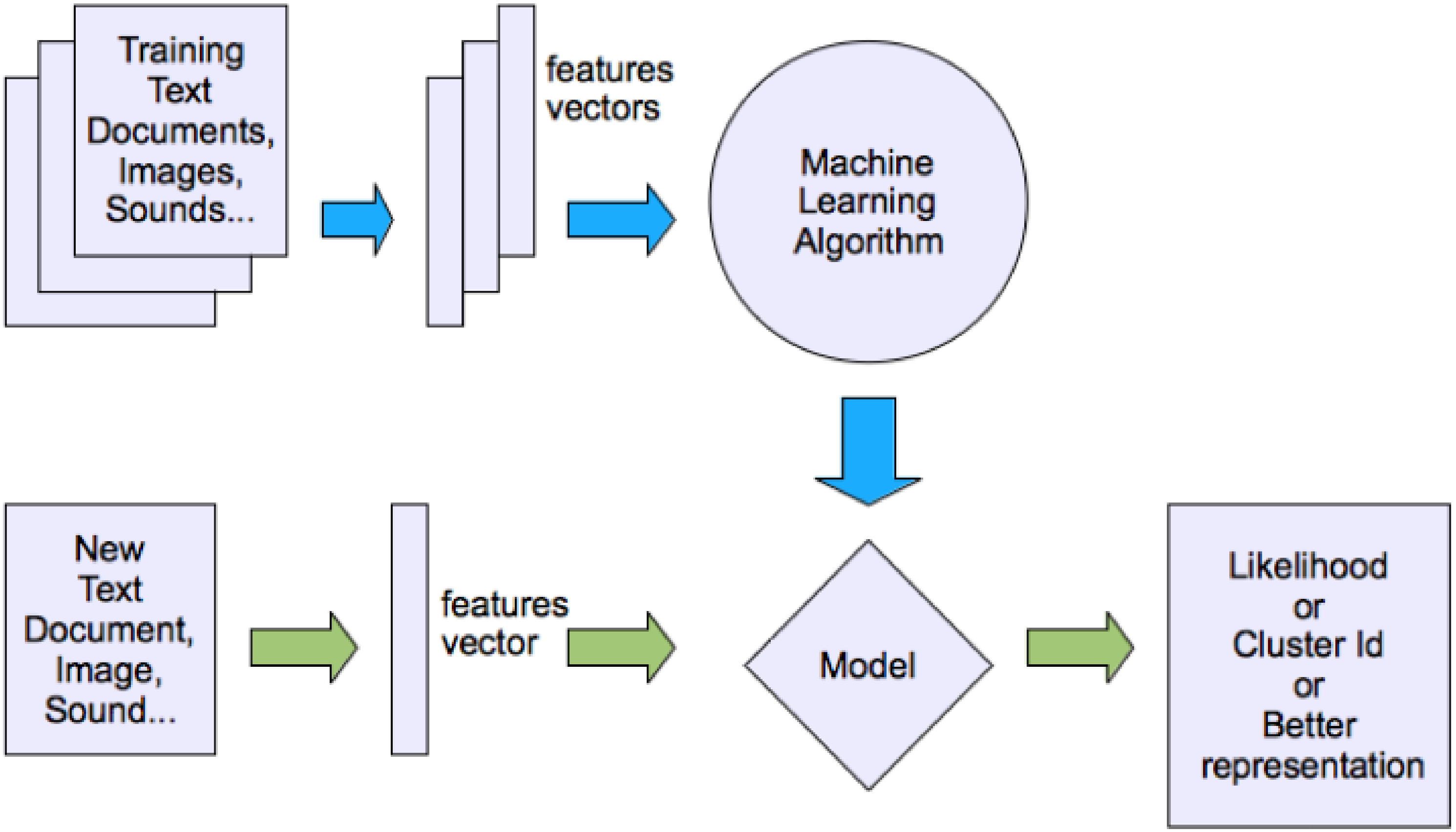
Machine learning structure

Supervised learning



Machine learning structure

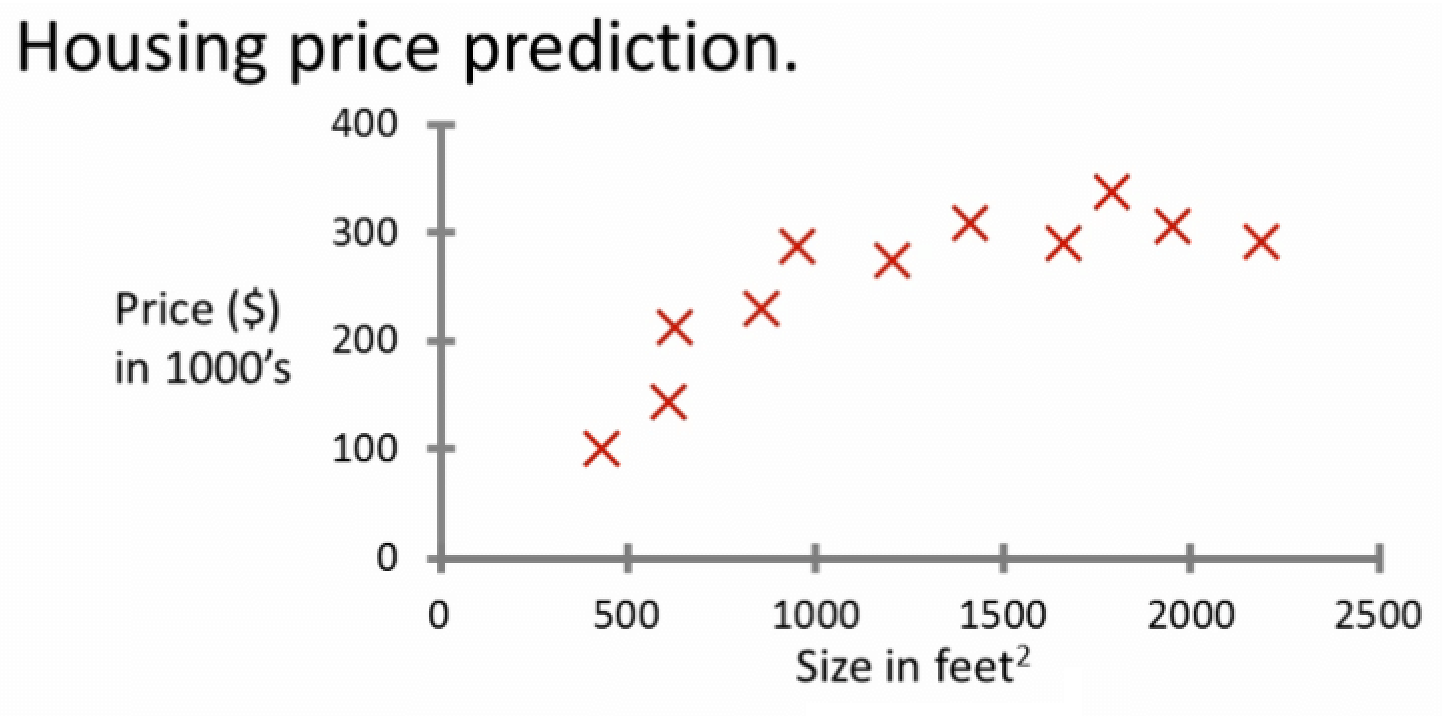
Unsupervised learning



Supervised learning example

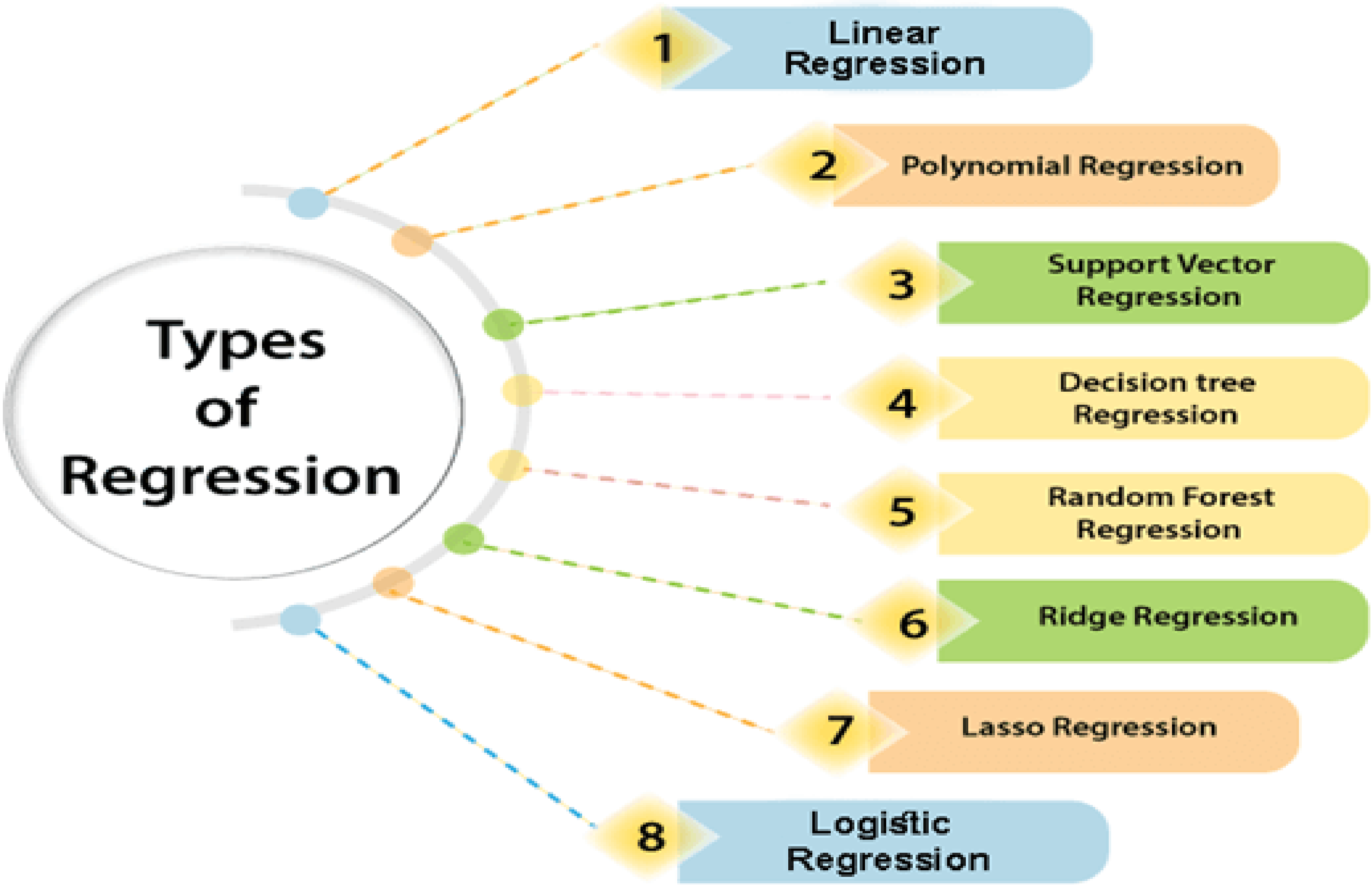
Regression

- If the output of the model is a continuous value.
- It is used to predict a continuous value.



- House price prediction
- Stock market prediction
- Predicting age of a person
- Number of copies a music album will be sold next month

Types of Regression



Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis.

- Shows the relationship between the continuous variables.

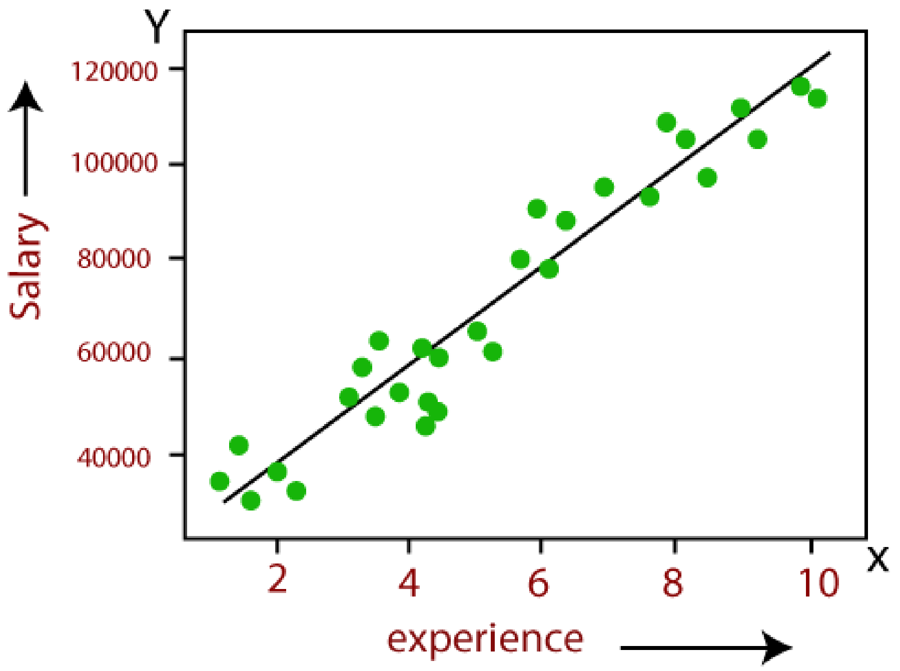
1) **Simple Linear Regression**

- In this case, only a single independent variable and a single dependent variable.

- In simple linear regression, we try to find a relationship between target variable and input variables by fitting a line, known as the regression line.

$$Y = m * X + b$$

Here, Y = dependent variables (target variables), X= Independent variables (predictor variables), m and b are the linear coefficients



Linear Regression

2) **Multiple Linear Regression (MLR).**

- If there is more than one input variable, then such linear regression is called multiple linear regression.
- MLR is used extensively in econometrics and financial inference.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Linear Regression

2) **Multiple Linear Regression (MLR).**

Predict the value of Y for subject 6 from the given dataset that contains values for X1, X2, and Y by using a Multiple Regression Model.

Subject	Y	X1	X2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	**?**	3	2

Linear Regression

2) **Multiple Linear Regression (MLR).**

The Multiple Linear Regression Model with n independent variables is written as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots\ldots\ldots\ldots + b_nX_n + u$$

Where,

Y = The variable needs to be predicted (dependent variable)
X = The variable used to predict Y (independent variable)
a = The intercept
b = The slope
u = The regression residual

Linear Regression

2) **Multiple Linear Regression (MLR).**

Regression of two independent variables can be predicted by using the below formulas such as Intercepts (a), Regression Coefficients (b1, b2)

$$Intercepts\ a = \overline{Y} - b_1(\overline{X}_1) - b_2(\overline{X}_2)$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Linear Regression

2) **Multiple Linear Regression (MLR).**

Regression of two independent variables can be predicted by using the below formulas such as Intercepts (a), Regression Coefficients (b1, b2)

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\bar{Y} = \frac{\sum Y}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\bar{X}_1 = \frac{\sum X_1}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\bar{X}_2 = \frac{\sum X_2}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Linear Regression

2) **Multiple Linear Regression (MLR).**

Step 1 : First, calculate all the values required in the above formulae.

Subject	Y	X_1	X_2	X_1X_2	X_1X_1	X_2X_2	X_1Y	X_2Y
1	-3.7	3	8	24	9	64	-11.1	-29.6
2	3.5	4	5	20	16	25	14	17.5
3	2.5	5	7	35	25	49	12.5	17.5
4	11.5	6	3	18	36	9	69	34.5
5	5.7	2	1	2	4	1	11.4	5.7
SUM	**19.5**	**20**	**24**	**99**	**90**	**148**	**95.8**	**45.6**

Linear Regression

2) **Multiple Linear Regression (MLR).**

Step 2 : Then put these values into the above-mentioned formulae to get the exact predictable values to calculate Regression Coefficients b1 and b2

$$\sum x_1^2 = \sum X_1X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 90 - \frac{20 \times 20}{5} = 10$$

$$\sum x_2^2 = \sum X_2X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 148 - \frac{24 \times 24}{5} = 32.8$$

$$\sum x_1y = \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{N} = 95.8 - \frac{20 \times 19.5}{5} = 17.8$$

$$\sum x_2y = \sum X_2Y - \frac{(\sum X_2)(\sum Y)}{N} = 45.6 - \frac{24 \times 19.5}{5} = -48$$

$$\sum x_1x_2 = \sum X_1X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 99 - \frac{20 \times 24}{5} = 3$$

Linear Regression

2) **Multiple Linear Regression (MLR).**

Step 2 : Then put these values into the above-mentioned formulae to get the exact predictable values to calculate Regression Coefficients b1 and b2

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_1 = \frac{(32.8 \times 17.8) - (3 \times (-48))}{(10 \times 32.8) - (3)^2} = 2.2816$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_1 = \frac{(10 \times (-48)) - (3 \times 17.8)}{(10 \times 32.8) - (3)^2} = -1.672$$

Linear Regression

2) **Multiple Linear Regression (MLR).**

Step 3 : Calculate the value of **Intercept a**

$$a = \overline{Y} - b_1(\overline{X}_1) - b_2(\overline{X}_2) = \frac{19.5}{5} - \frac{2.2816 \times 20}{5} - \frac{(-1.672 \times 24)}{5} = 2.796$$

Step 4 : The final Regression Equation or Model looks as follows:

$$Y = 2.796 + 2.28x_1 - 1.67x_2$$

Therefore, for given x1= 3 and x2 = 2, the value of Y = ? calculated as follows:

$$Y = 2.796 + (2.28 \times 3) - (1.67 \times 2)$$

$$Y = 6.296$$

Locally weighted Linear Regression

- Locally weighted linear regression is a non-parametric algorithm, that is, the model does not learn a fixed set of parameters as is done in ordinary linear regression.

- Rather parameters θ are computed individually for each query point x. While computing θ, a higher "preference" is given to the points in the training set lying in the vicinity of x than the points lying far away from x.

- The modified cost function is: $J(\theta) = \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$

- where, $w^{(i)}$ is a non-negative "weight" associated with training point $x^{(i)}$.

- For $x^{(i)}$ lying closer to the query point x, the value is large, while for $w^{(i)}$ lying far away from x the value of $w^{(i)}$ is small. of w^{(i)}

- For $x^{(i)}$s lying closer to the query point x, the value of $w^{(i)}$ is large, while for $x^{(i)}$s lying far away from x the value of $w^{(i)}$ is small.

Locally weighted Linear Regression

- A typical choice of $w^{(i)}$ is: $w^{(i)} = exp(\frac{-(x^{(i)}-x)^2}{2\tau^2})$
- where τ is called the bandwidth parameter and controls the rate at which $w^{(i)}$ falls with distance from x Clearly.

For example: Consider a query point x= 5.0 and let $x^{(1)}$ and $x^{(2)}$be two points in the training set such that $x^{(1)}$= 4.9 and $x^{(2)}$= 3.0. Using the formula $w^{(i)} = exp(\frac{-(x^{(i)}-x)^2}{2\tau^2})$ with τ= 0.5: $w^{(1)} = exp(\frac{-(4.9-5.0)^2}{2(0.5)^2}) = 0.9802$[Tex]w^{(2)} = exp(\frac{-(3.0 – 5.0)^2}{2(0.5)^2}) = 0.000335 [/Tex]
$So,\ J(\theta) = 0.9802 * (\theta^T x^{(1)} - y^{(1)}) + 0.000335 * (\theta^T x^{(2)} - y^{(2)})$ Thus, the weights fall exponentially as the distance between x and $x^{(i)}$increases and so does the contribution of error in prediction for $x^{(i)}$to the cost. Consequently, while computing θ, we focus more on reducing $(\theta^T x^{(i)} - y^{(i)})^2$for the points lying closer to the query point (having a larger value of $w^{(i)}$).

Locally weighted Linear Regression

- Locally weighted linear regression is a supervised learning algorithm.
- It is a non-parametric algorithm.
- There exists No training phase. All the work is done during the testing phase/while making predictions.
- The dataset must always be available for predictions.
- Locally weighted regression methods are a generalization of k-Nearest Neighbour.
- In Locally weighted regression an explicit local approximation is constructed from the target function for each query instance.
- The local approximation is based on the target function of the form like constant, linear, or quadratic functions localized kernel functions.

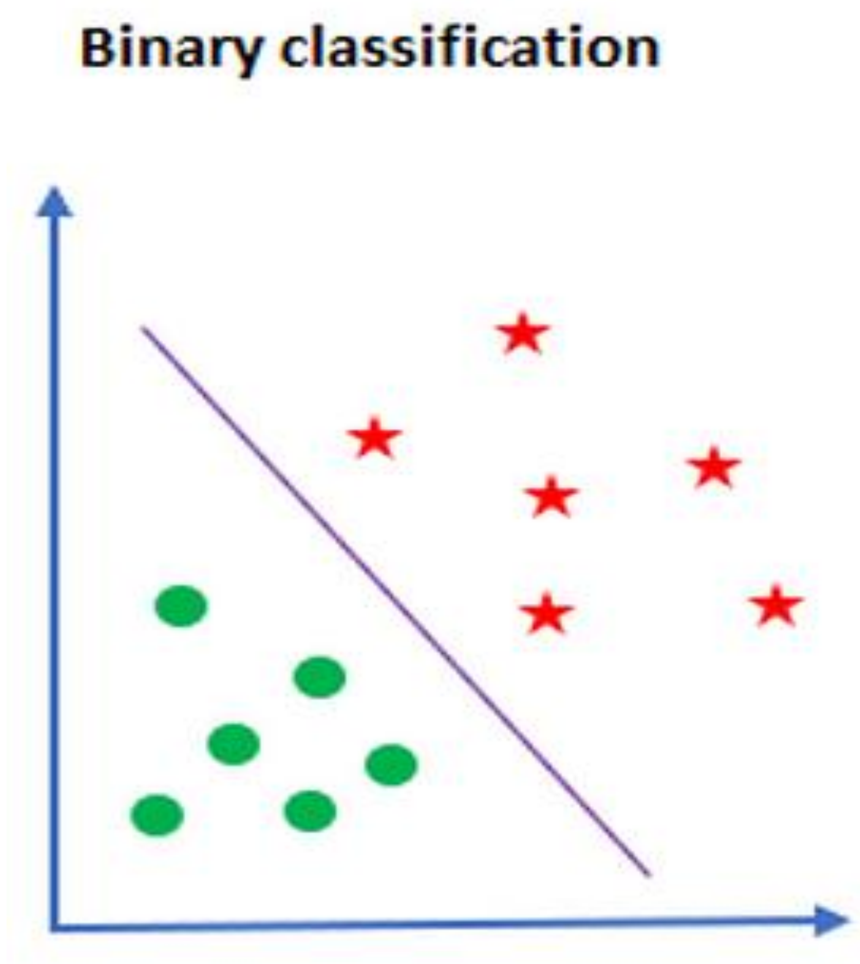
Classification

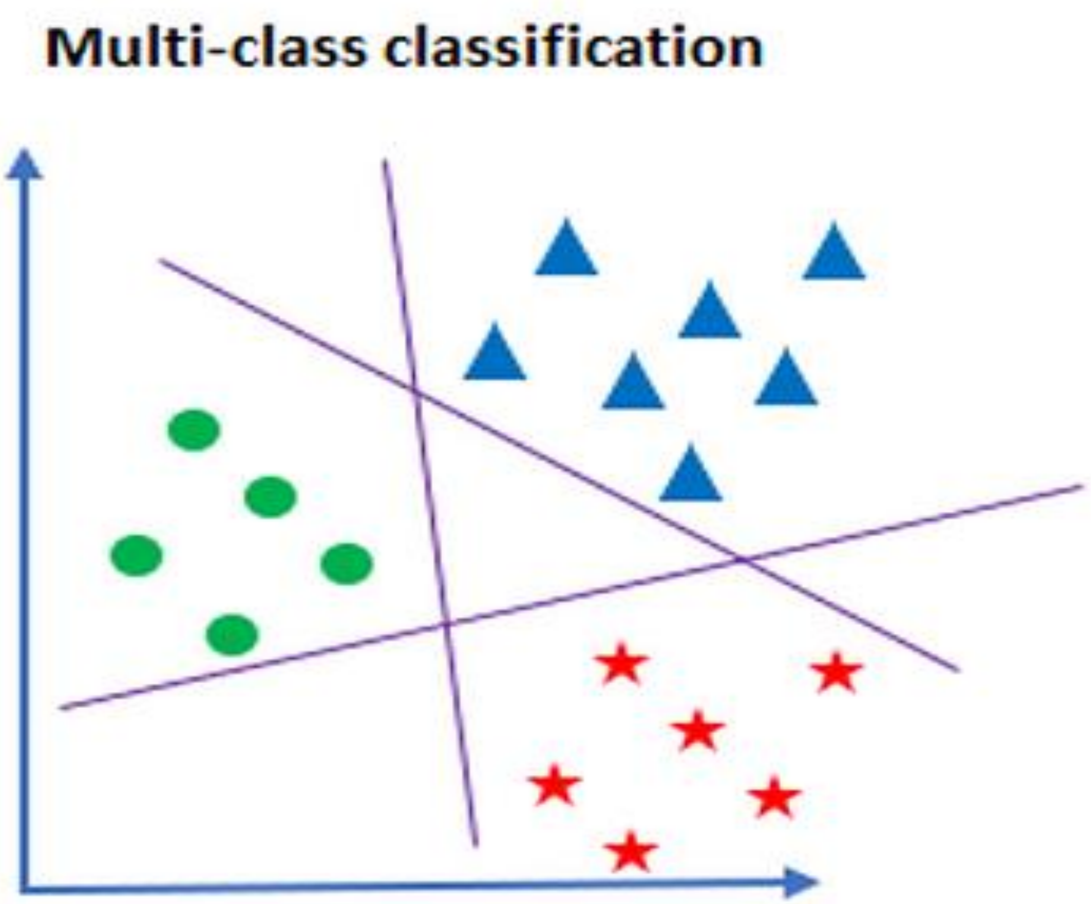
- If the output of the model is categorical.

- In classification we are interested to predict the categorical response value where the data can be separated into specific “classes”.
  - Spam filtering
  - Cat dog classification

- Two types
  - Binary classification: Two class
  - Multiclass classification: More than two class

Classification

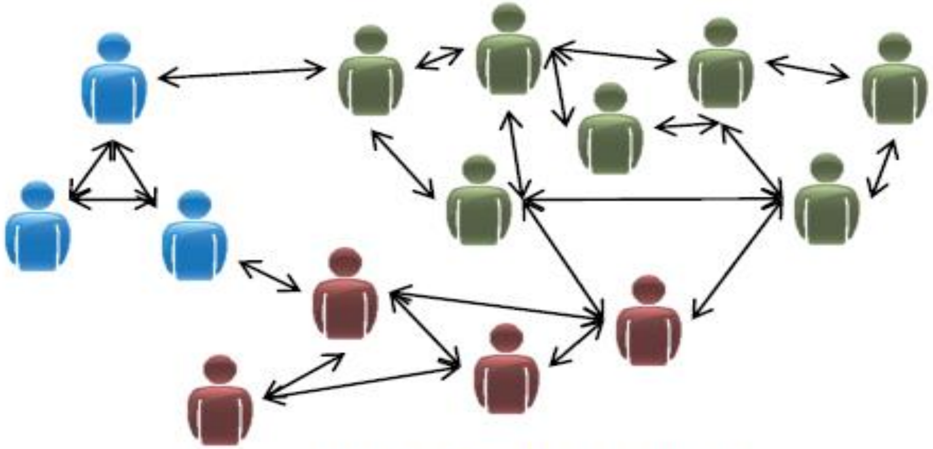




Source: https://wadhwatanya1234.medium.com/multi-class-classification-one-vs-all-one-vs-one-993dd23ae7ca

Unsupervised learning example

Organize computing clusters



Social network analysis

Market segmentation

Astronomical data analysis

Unsupervised learning

Which would you address using an unsupervised learning algorithm?

- Given a database of customer data, automatically discover market segments and group customers into different market segments.
- Given email labeled as spam/not spam, learn a spam filter.
- Given a set of news articles found on the web, group them into set of articles about the same story.
- Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.