Data Analytics

Mrinal Das

Predictive Analytics

Month	Temp	Time	Speed	Water	Material	Fragrance	Quality Index
March	40	20	300	75	15	10	5
April	39	20	300	70	15	15	7
May	40	18	200	80	15	5	2
June	40	20	250	80	15	5	4
July	40	18	200	75	15	10	?
August	40	20	200	80	15	5	?

How can we predict QI for future ?

Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	?
80	?

Consider this simple case:
Can we predict the QI from Water level ?

We can think of formulating this as a function

QI = f (Water)

Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	?
80	?

QI = f (Water)

How do we find f ?

Lets use a simple function:

QI = a * Water + b

Finding f is equivalent to finding a and b.

Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	?
80	?

QI = a * Water + b

If we know a and b, then

QI-July = a 75 + b
QI-August = a 80 + b

Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	?
80	?

QI = a * Water + b

Using the data we can write:

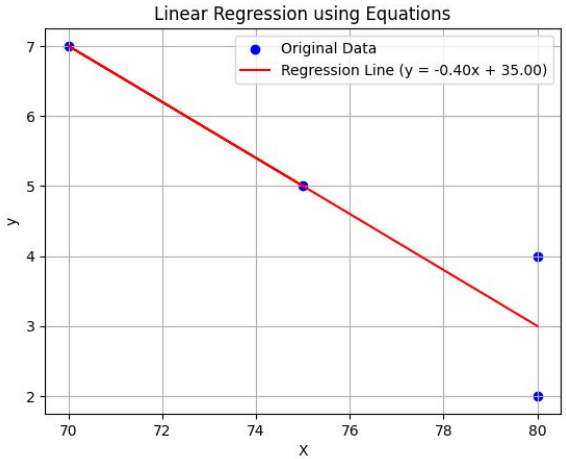
5 = a 75 + b
7 = a 70 + b
2 = a 80 + b
4 = a 80 + b

Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	?
80	?

QI = a * Water + b

Solution we get: a = -0.4, b=35

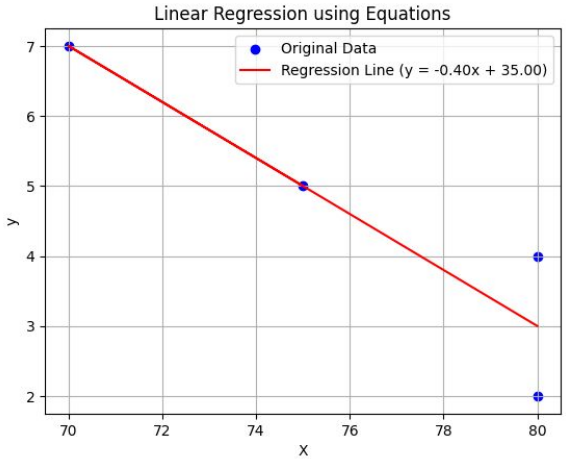


Prediction as Function

Water	Quality Index
75	5
70	7
80	2
80	4
75	5
80	3

QI = a * Water + b

Solution we get: a = -0.4, b=35



Theory Behind

t: target, x: feature, then $t = ax$

Let y is the correct value, then $L = (y - t)^2$ is the loss

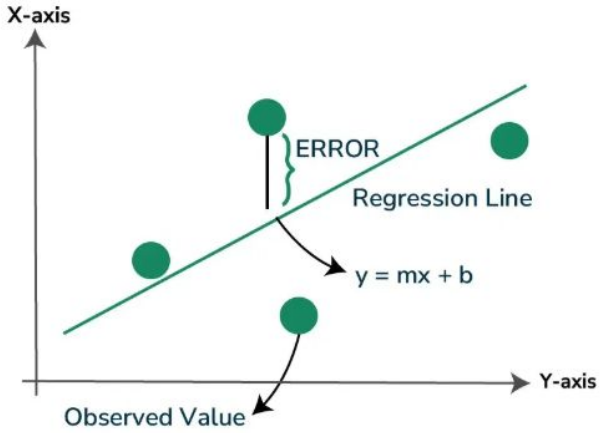
By substituting the value of t, we get $L = (y - ax)^2$

Minimize the Error

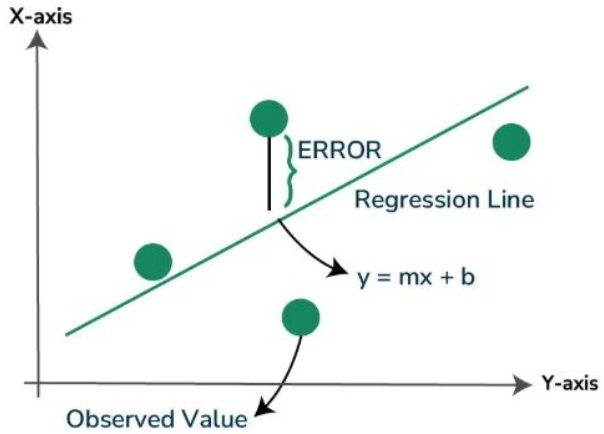
t: target, x: feature, then $t = ax$

Let y is the correct value, then $L = (y - t)^2$ is the loss

By substituting the value of t, we get $L = (y - ax)^2$



Theory Behind

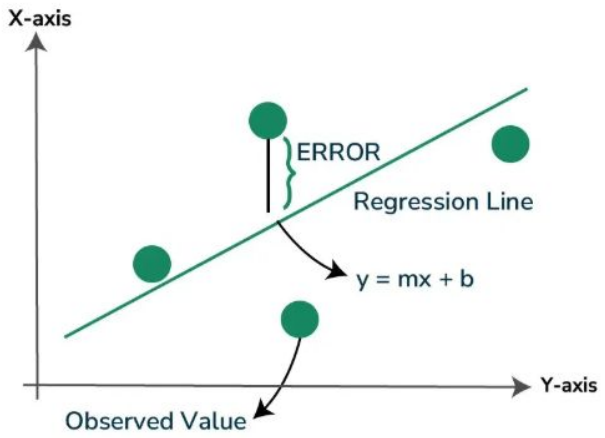


We have to minimize the loss to find a.

We can do that by computing the derivative and equating to 0.

$$\frac{dL}{da} = 2(y - ax)(-x) = 0$$

Theory



We have many samples, not one.

Samples are in pairs (x_i, y_i).

This comes from the data.

Therefore, we write the equation as:

$$t_i = a\ x_i$$

Then the loss is the aggregate over all samples:

$$L = \sum_{i=1}^{n} (y_i - t_i)^2 = \sum_{i=1}^{n} (y_i - a x_i)^2$$

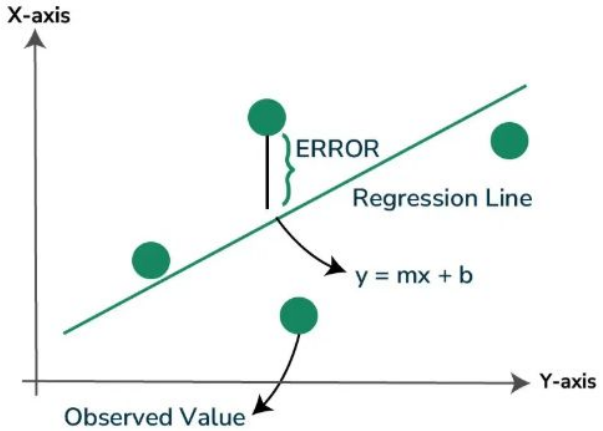
Theory

Thus we have to minimze the loss

$$L = \sum_{i=1}^{n} (y_i - t_i)^2$$

Then again using the value for t_i, we get

$$L = \sum_{i=1}^{n} (y_i - ax_i)^2$$



We can minimize the loss by computing the gradient and equating to 0

$$\frac{dL}{da} = 0$$

Theory

We want

$$\frac{dL}{da} = 0 \qquad L = \sum_{i=1}^{n} (y_i - a x_i)^2$$

$$\frac{dL}{da} = 2 \sum_{i=1}^{n} (y_i - a x_i)(-x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i x_i \qquad \Rightarrow a = \left(\sum_{i=1}^{n} x_i y_i \right) / \left(\sum_{i=1}^{n} x_i^2 \right)$$

Theory

$$\Rightarrow a = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i=1}^{n} x_i^2\right)$$

Single feature

Generally we will have many features, then we can write as follows

Let $X = (x_1, x_2, \ldots, x_n), Y = (y_1, y_2, \ldots, y_n)$

Then, $a = (XX^T)^{-1}(XY^T)$

Revisiting Descriptive Analytics

Often we may not have a target variable to predict.

We want to analyse the situation if everything is going fine or not

Logs are generally used to capture or record events and actions

Logs are useful information to check for analysis

However, generally logs are huge and difficult to manually go through them

Moreover, often several entries are similar and uninformative

Descriptive Analytics for Large Scale

Large Scale Descriptive Analytics

Huge log files

Huge sensor data

Huge inventory management data

.....

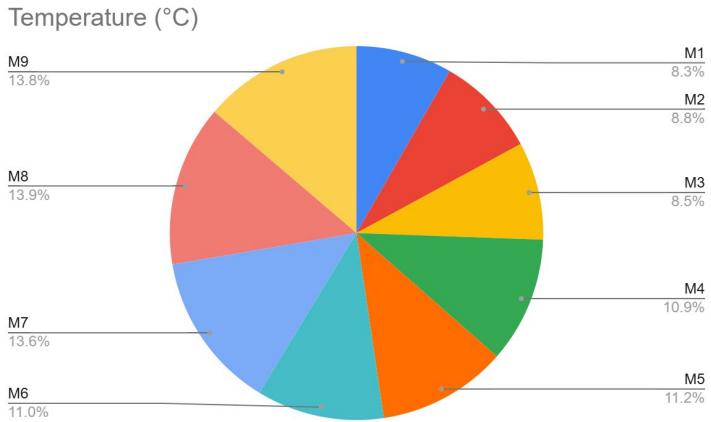
If we are looking for a pattern manually inspecting them is useless

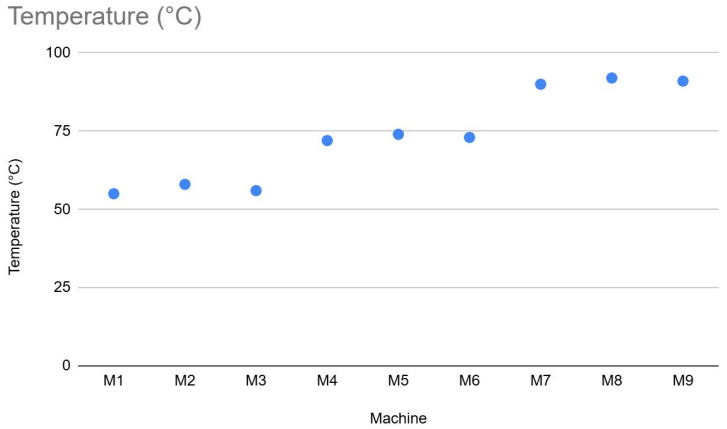
We can summarize the data by grouping and look for patterns

Revisiting Descriptive Analytics

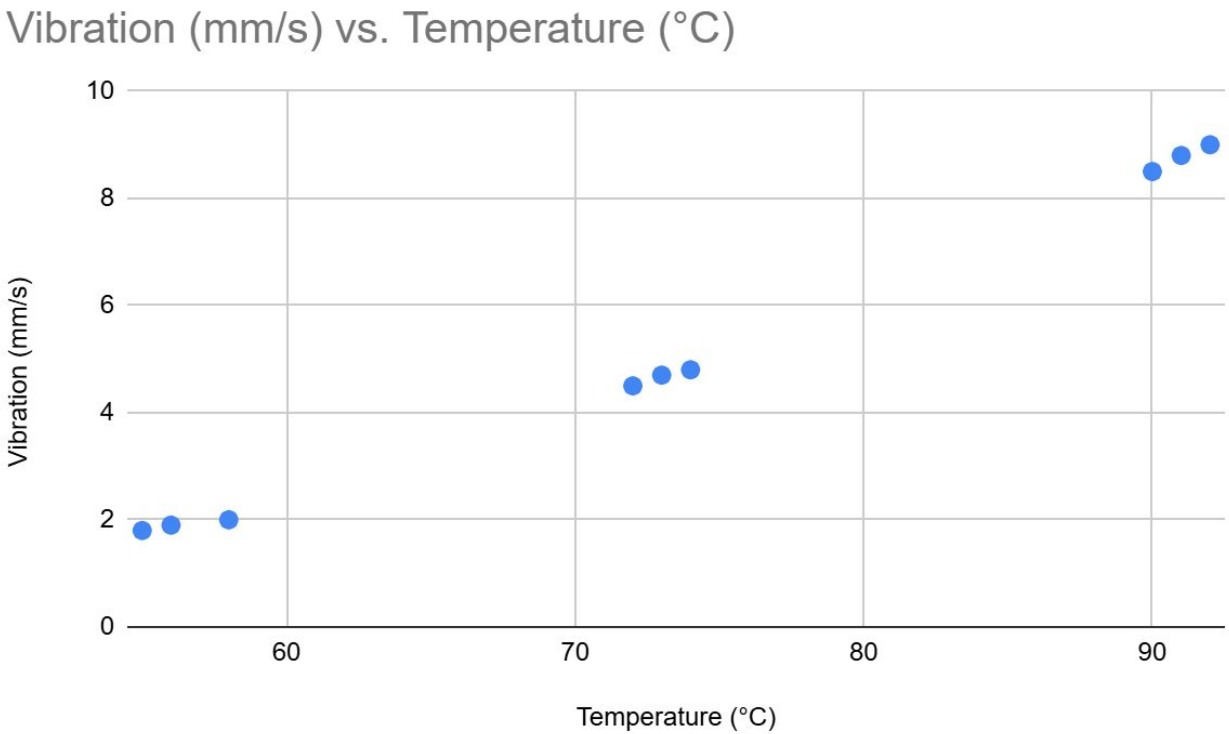
Machine	Temperature (°C)	Vibration (mm/s)	Current (A)	Noise (dB)
M1	55	1.8	18	62
M2	58	2	19	64
M3	56	1.9	18	63
M4	72	4.5	28	78
M5	74	4.8	30	80
M6	73	4.7	29	79
M7	90	8.5	42	96
M8	92	9	44	98
M9	91	8.8	43	97

Analysing by Clustering

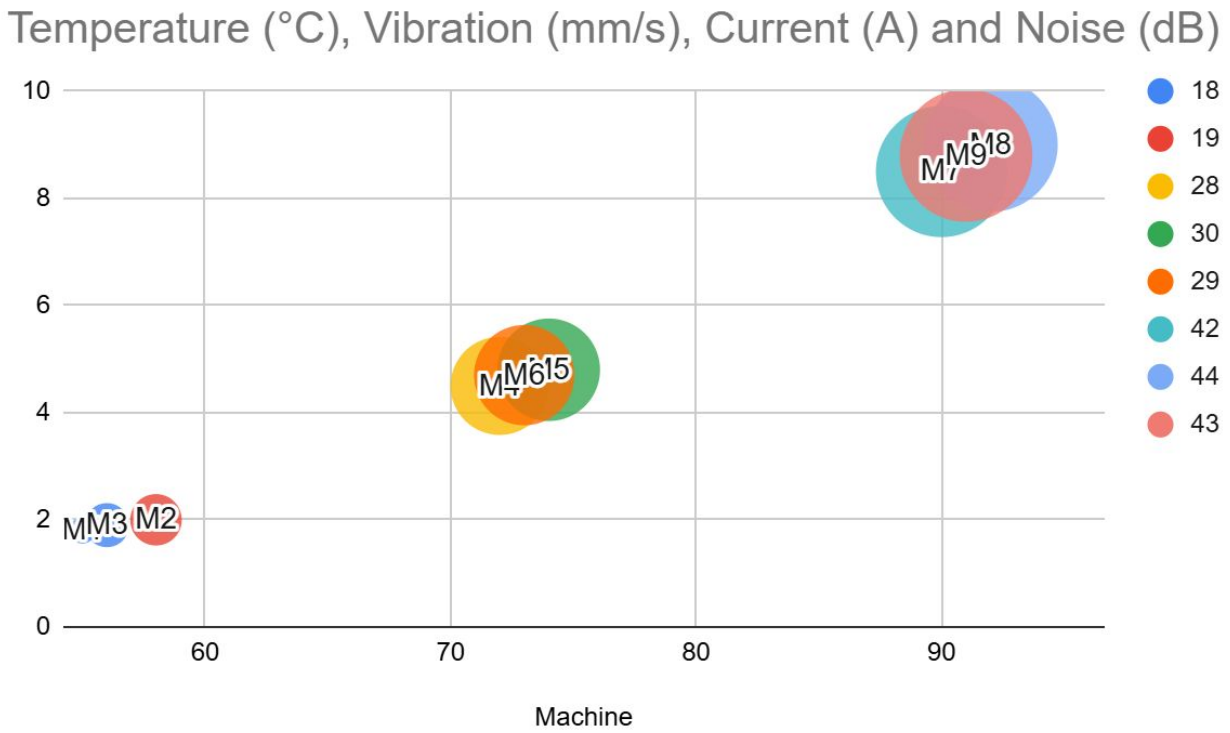




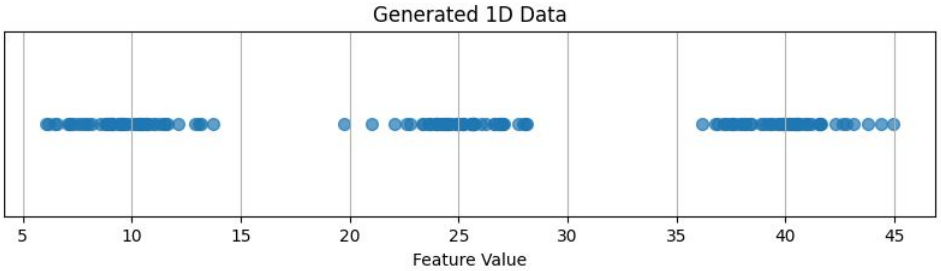
Clustering using 2 Variables



Clustering using all variables

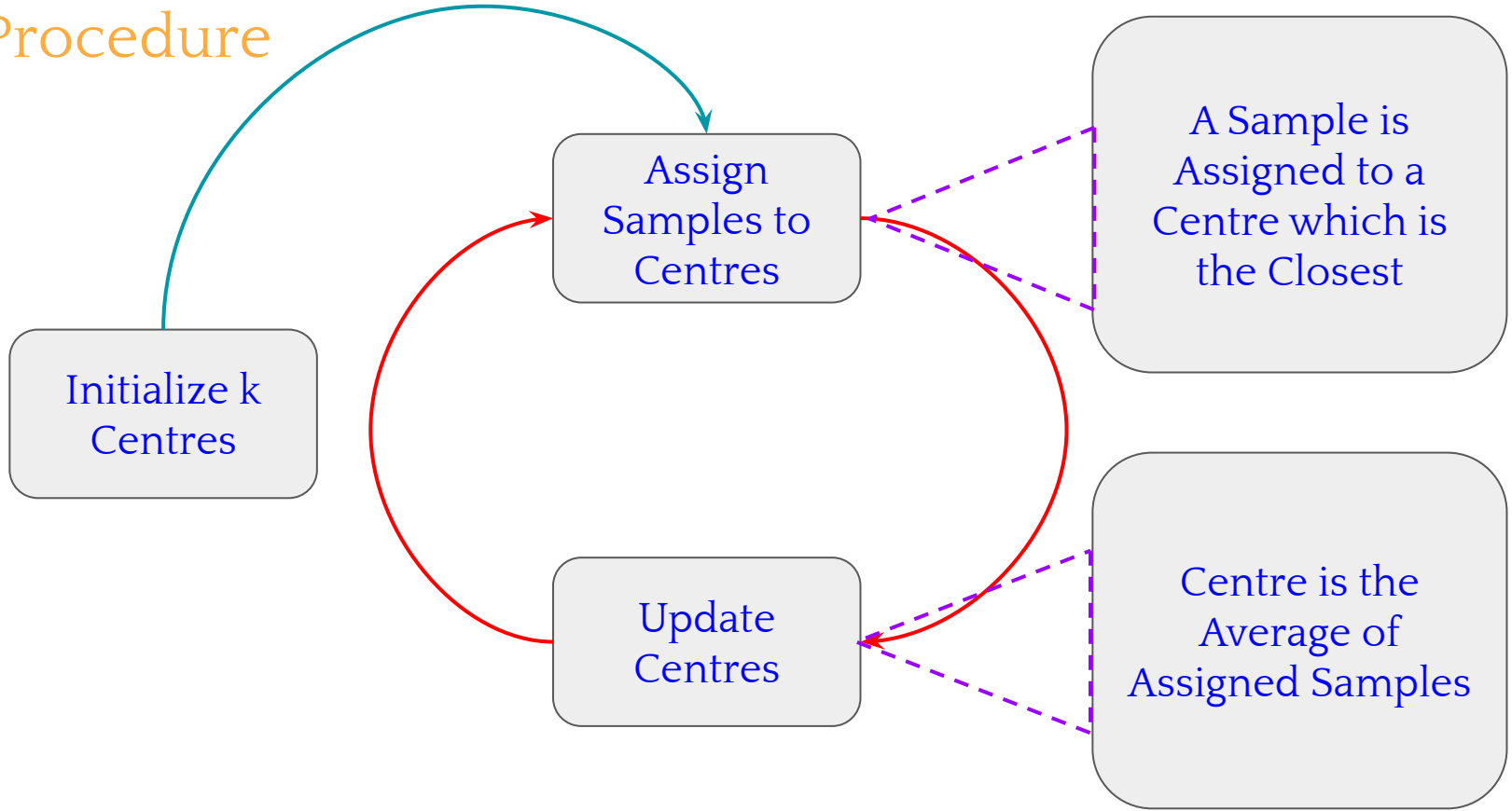


Clustering

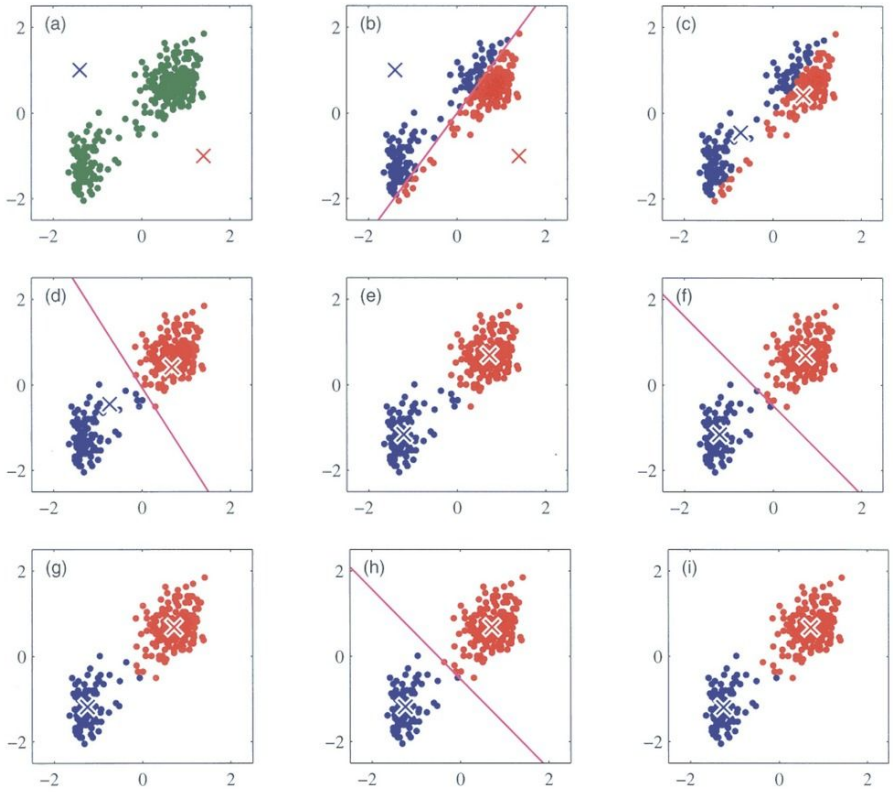

Generated 1D Data
5
10
15
20
25
30
35
40
45
Feature Value


K-Means Clustering Results (k=3)
Cluster 1
Centroid 1
Cluster 2
Cluster 3
5
10
15
20
25
30
35
40
45
Feature Value

Procedure


Initialize k Centres
Assign Samples to Centres
Update Centres
A Sample is Assigned to a Centre which is the Closest
Centre is the Average of Assigned Samples

Clustering Illustration



Quiz-8

1. Write 2 things that you liked/understood today.
2. Write 2 things that you did not like/understand today.

3. Consider the case study you have been developing
    a. Identify two variables and describe how you can develop predictive analytics
    b. Describe a situation when you can use clustering for descriptive analytics