CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

Machine Learning

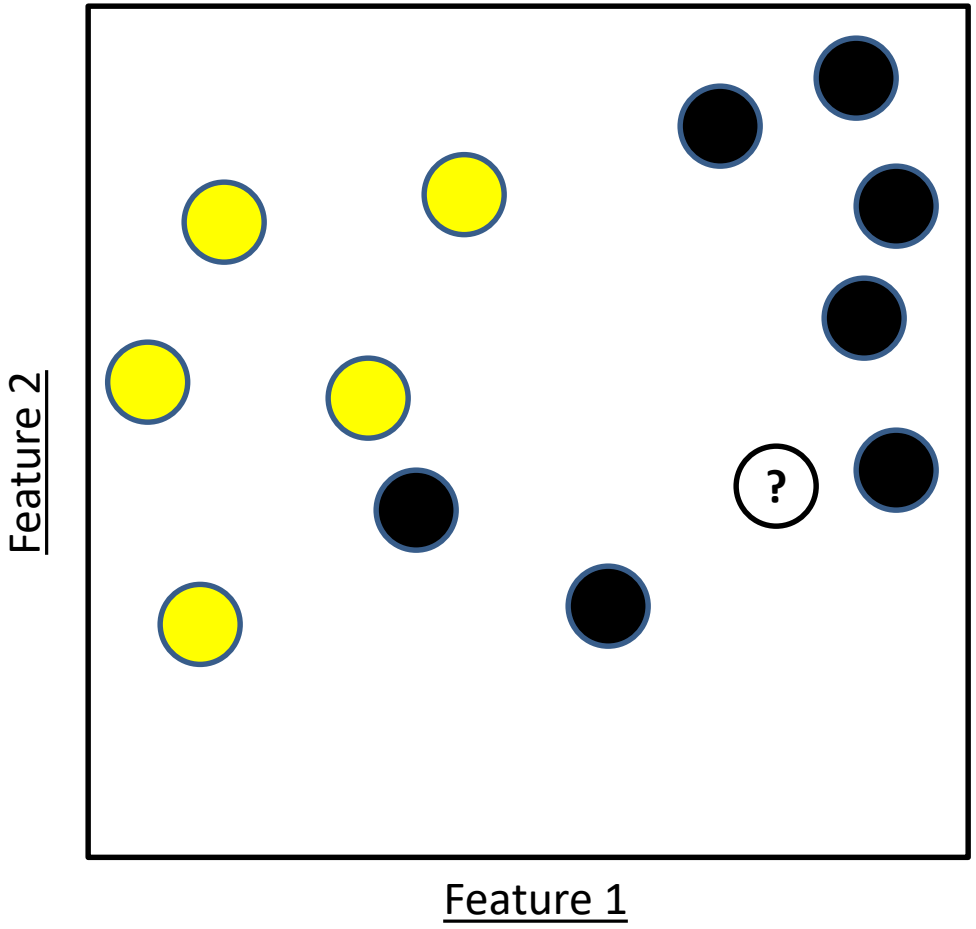
Machine Learning

Lecture: Kmeans Clustering

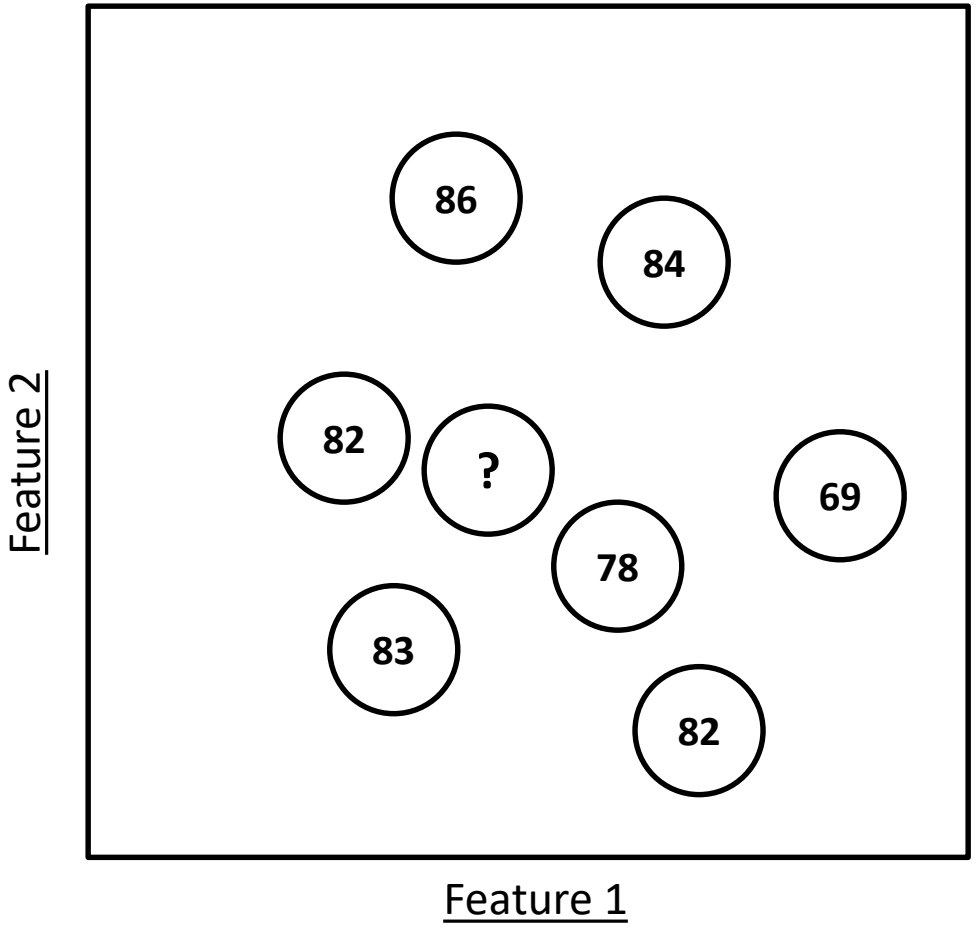
Ted Scully

Nearest Neighbour Example

Feature 1	Feature 2	Colour



Feature 1	Feature 2	Regression Target



Distance Metrics

Euclidean ->

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Mahattan ->

$$d(p, q) = \sum_{i=1}^{n} |q_i - p_i|$$

Minkowski ->

$$d(p,q) = \left(\sum_{i=1}^{n}|p_i - q_i|^a\right)^{\frac{1}{a}}$$

Normalization

- $$newValue = \frac{originalValue - minValue}{maxValue - minValue}$$

Distance Weighted kNN

Regression

Given a query instance xq,

$$f(\mathbf{x}_q) := \frac{\sum_{i=1}^{k} w_i f(\mathbf{x}_i)}{\sum_{i=1}^{k} w_i}$$

Where

$$w_i = \frac{1}{d(\mathbf{x}_q, \mathbf{x}_i)^n}$$

Classification

$$vote(c_j) := \sum_{i=1}^{k} \frac{1}{d(\mathbf{x}_q, \mathbf{x}_i)^n} (c_i, c_j)$$

Basic Measures of Error (Regression)

- The **R^2** coefficient compares the **performance of a model on a test set (sum of squared residuals)** with the performance of an imaginary model that always predicts the **average values from the test set (total sum of squares)**.

- The **R^2** coefficient is calculated as:

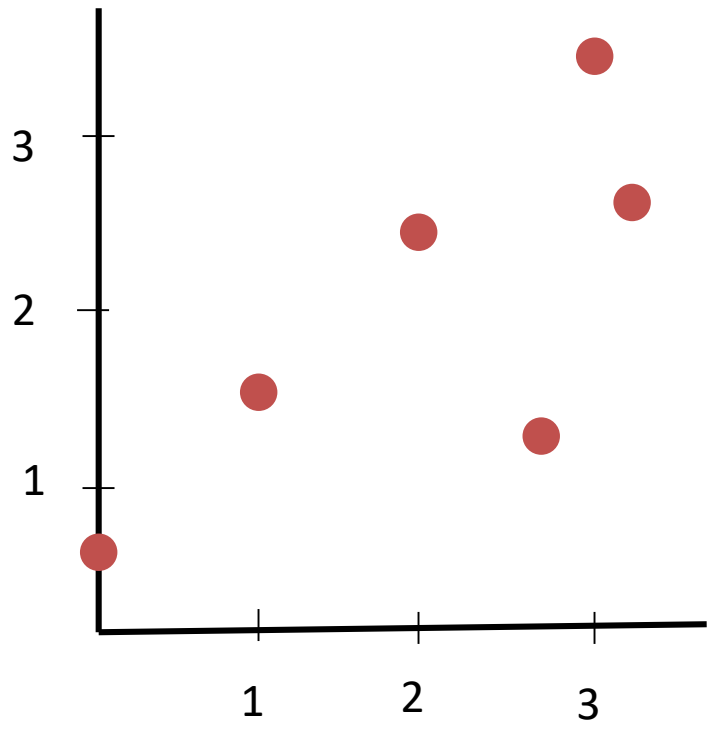
$$R^2 = 1 - \frac{sum\ of\ squared\ residuals}{total\ sum\ of\ squares}$$

- Where

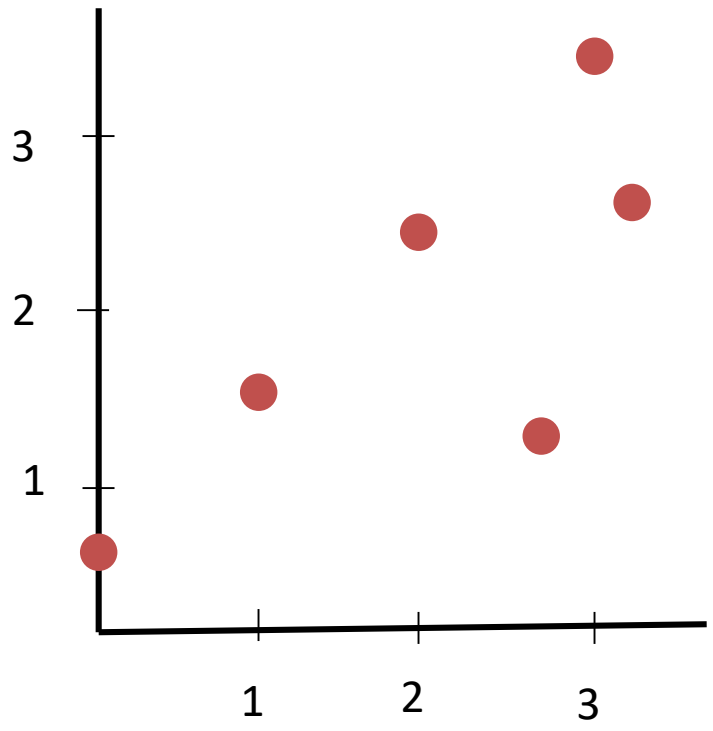
$$sum\ of\ squared\ residuals = \sum_{i=0}^{m} (f(x^i) - y^i)^2$$

$$total\ sum\ of\ squares = \sum_{i=0}^{m} (\bar{y} - y^i)^2$$

$$total\ sum\ of\ squares = \sum_{i=0}^{m} (\bar{y} - y^i)^2$$



$$sum\ of\ squared\ residuals = \sum_{i=0}^{m} (f(x^i) - y^i)^2$$



Basic Measures of Error (Regression)

- The **R^2** coefficient values typically fall in the range [0, 1) and larger values indicate better model performance.

- The **worse the model** produced, the closer the sum of square residuals value will be to the total sum of squares value. Consequently the **smaller the total R^2**.

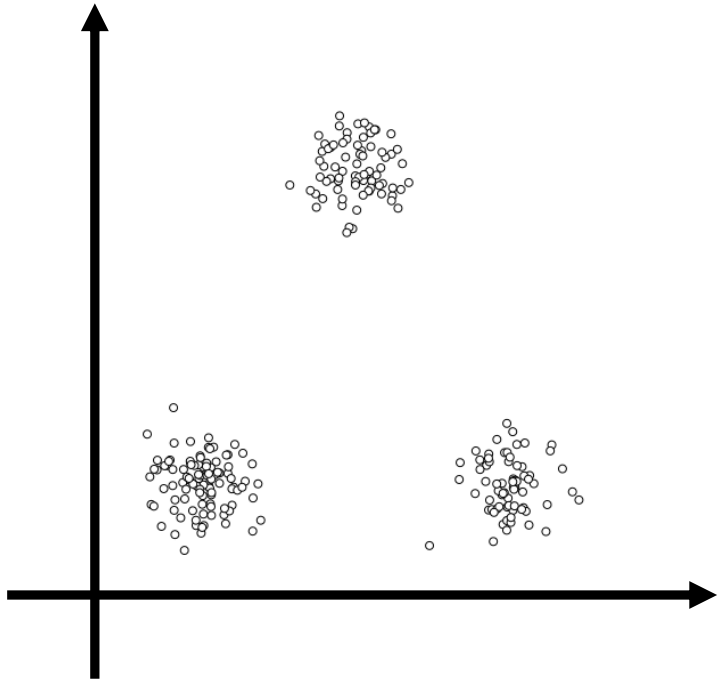
- The **better the model** the smaller the squared residuals (smaller error in the model) and the **larger the R^2** value.

- While it is rare, the model produced could be worse than the total sum of squares. In this case the R^2 would be **negative**. The worse the model the lower the R^2 values. It means that whatever model that you came up with is worse than predicting the mean (not a good sign!).

Unsupervised Learning (Clustering)

- In unsupervised learning the data is **unlabelled**. In other words there is no target class or regression value associated with each data point.

- Therefore, the objective is to discover **patterns or underlying structure** in the data.



- From a visual inspection of the data on the right we can see the data is broadly organized into three separate clusters.
- While relatively easy to visualize and identify the clusters for a feature space consisting of two of three features it becomes much more challenging to identify clusters as the dimensionality of the space increases.

Applications

- Google news uses clustering to group new articles related to their content. In this case articles related to a heatwave in Ireland

Most shops expected to close under new restrictions

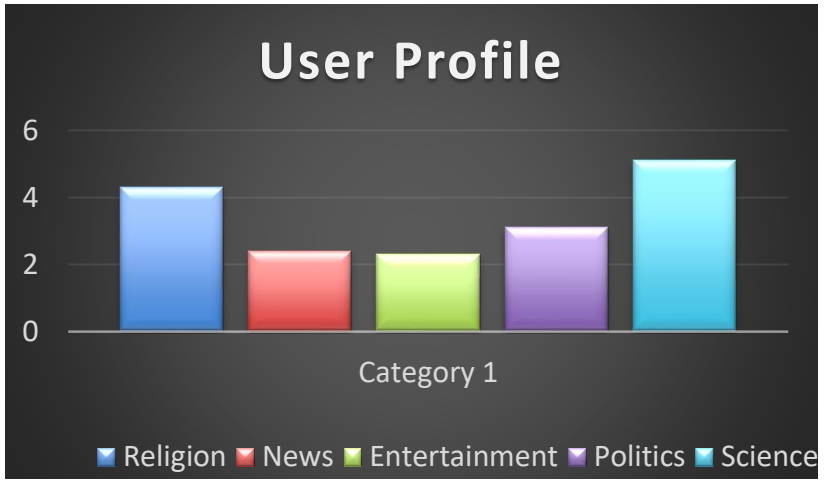
RTE.ie · 3 hours ago

- **Government plans ‘Level 4-plus’ restrictions for up to a month, with schools likely to remain open**
  The Irish Times · 3 hours ago
- **Cabinet to meet today to finalise 'decisive and nationwide' restrictions**
  TheJournal.ie · 20 minutes ago
- **Govt to confirm 'decisive, nationwide' Covid measures**
  RTE.ie · 11 hours ago
- **LIVE Covid-19 Ireland updates as Level Five decision made by Cabinet and list of what is expected to close**
  Irish Mirror · 2 hours ago

View Full coverage

Applications

- If a user is reading an article from one cluster let's say sports then we can use the cluster to which to article belongs in order to recommend further articles for the user.
- We could also profile a user by learning their preferences over a specific set of topics. This in turn can be used to feed into and impact the recommended articles(documents that are recommended to the user).

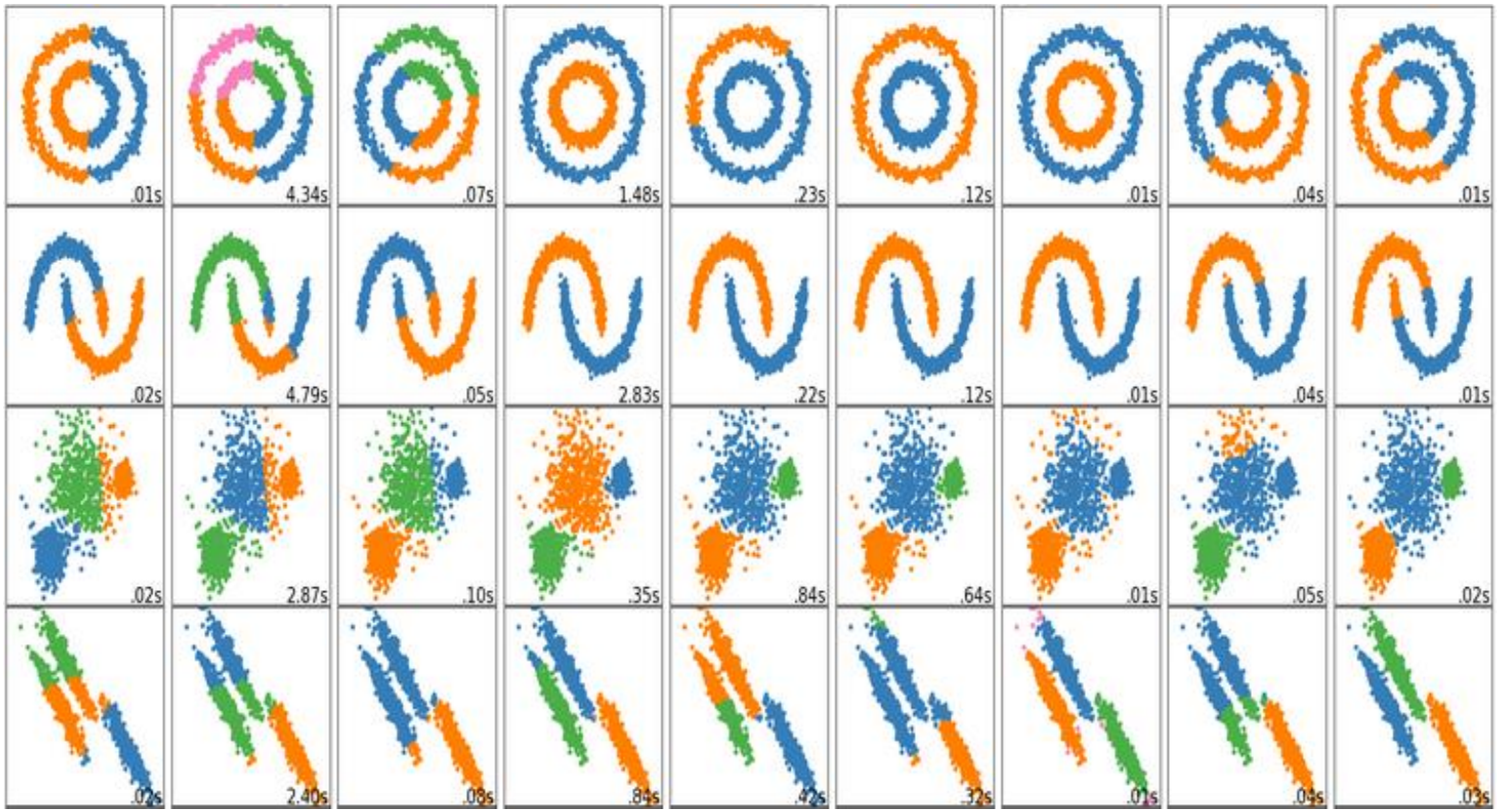


- Used for market-segmentation, that is the identification of sub-groups of consumers based on some type of shared characteristics. That is many may share particular purchasing patterns.

- Many other applications include:
  - Grouping of individuals based on genomic patterns or medical conditions.
  - Clustering neighbourhoods
  - Outlier detection

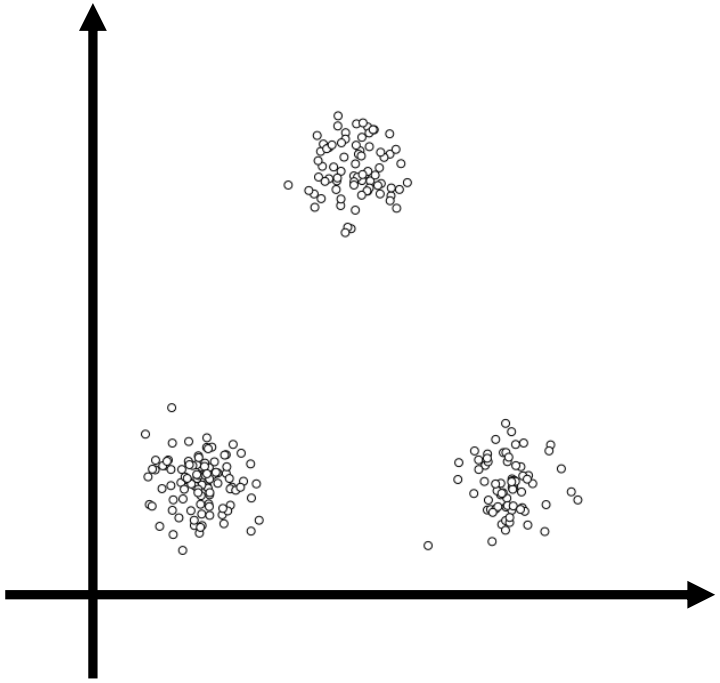
Clustering

- Clustering is a difficult problem and while we commonly view clustering as being spherical, this quite often not the case.
- The following is a image from Scikit-Learn shows some popular ‘toy’ dataset.
- As we can see below different algorithms can return differing solutions.



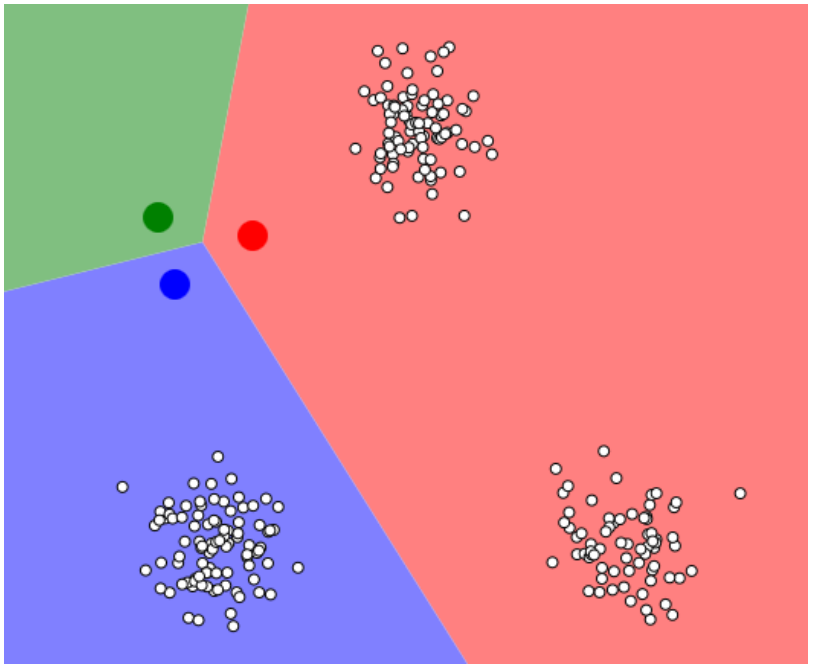
K-Means Clustering

- K- Means clustering is the most widely used unsupervised learning technique.
- It is also a very simple algorithm to understand and implement.
- It is **parameterized**. We must specify the number of centroids in advance.
- It is an iterative algorithm that involves two main steps:
  - **Cluster Assignment**
  - **Move Centroid (Move Cluster Centre)**



K-Means Clustering

- The very first thing we do is we 'randomly' pick **k points** in space, we call each of these points **centroids**.
- In the example below I have "randomly" picked three centroid points in space (green, blue and red circles).
- We then iteratively perform the **cluster assignment** and **move centroid** steps.

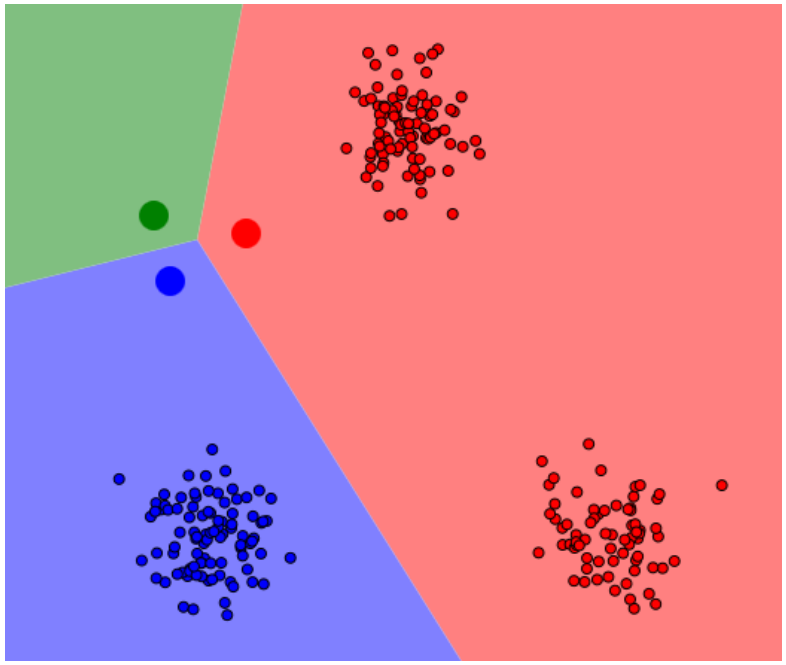


```
for x iterations:
        cluster assignment
        move centroids
```

K-Means Clustering – Cluster Assignment Step

for x iterations:
    cluster assignment
    move centroids

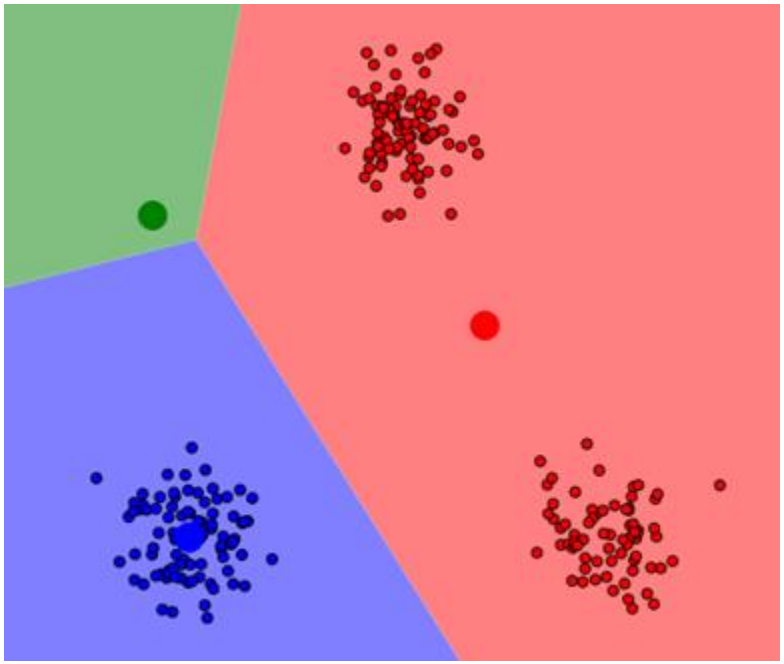
- In the cluster assignment step we assign each training example to the nearest cluster.
- You will notice that all training points are assigned to either the blue or red centroid. The green centroid is further away from all training points.
- We have coloured the training example accordingly.



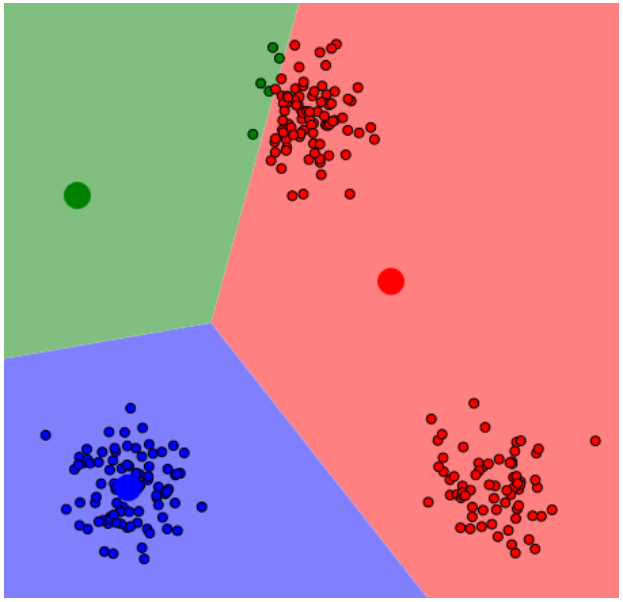
K-Means Clustering – Move Centroid Step

for x iterations:
    cluster assignment
    move centroids

- In the **move centroid** step we assign a new position to each centroid by obtaining the mean or average of all training points assigned to that centroid.
- So you will notice that the blue and red centroid have moved. The red is positioned halfway between two clusters.



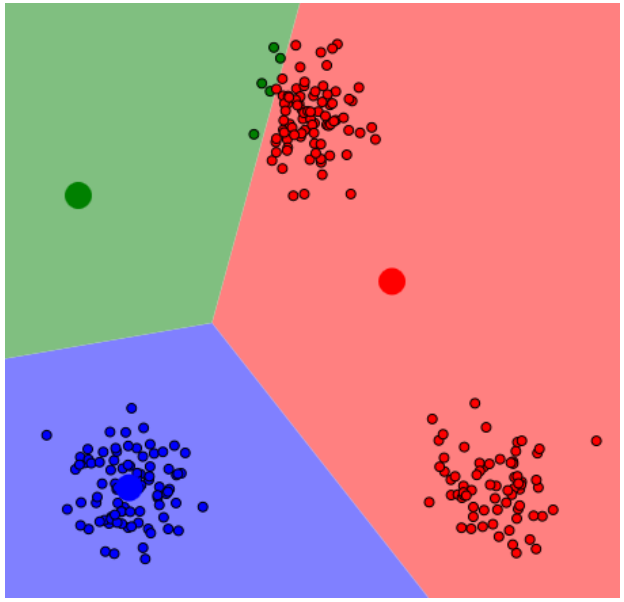
K-Means Clustering



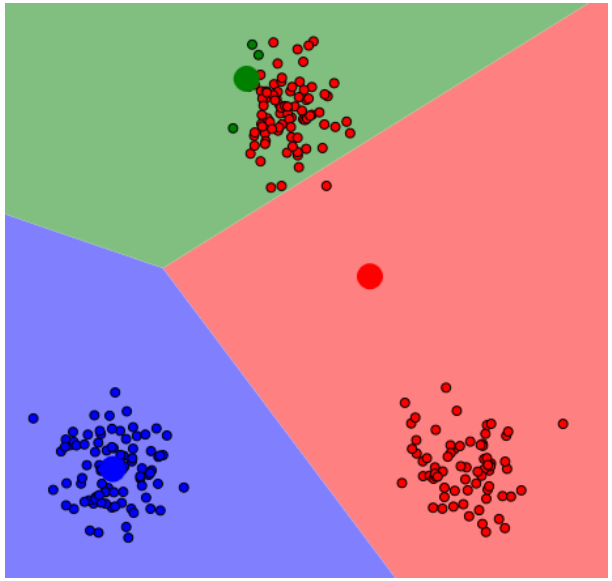
- **Cluster Assignment** step repeated.

- **Move centroid** step repeated.

K-Means Clustering

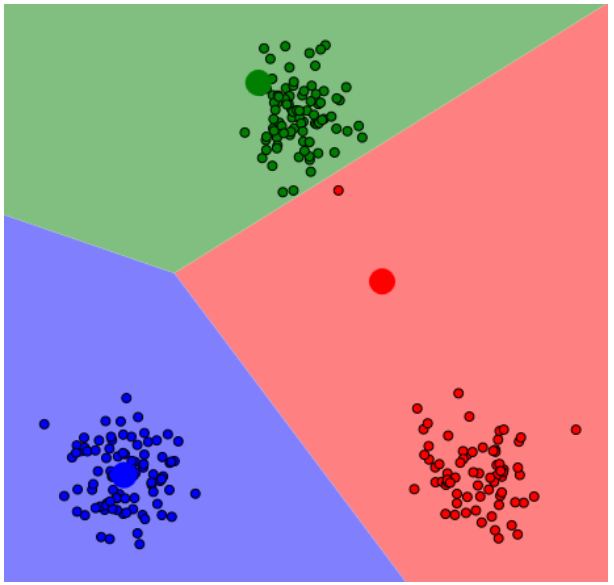


- **Cluster Assignment** step repeated.



- **Move centroid** step repeated.

K-Means Clustering

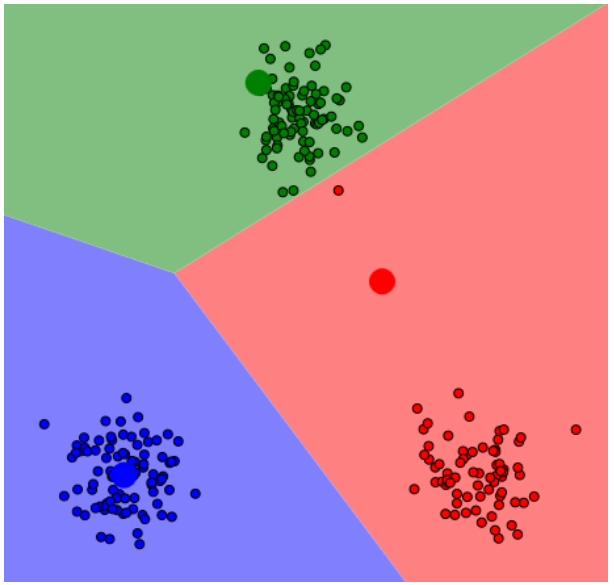


- **Cluster Assignment** step repeated.

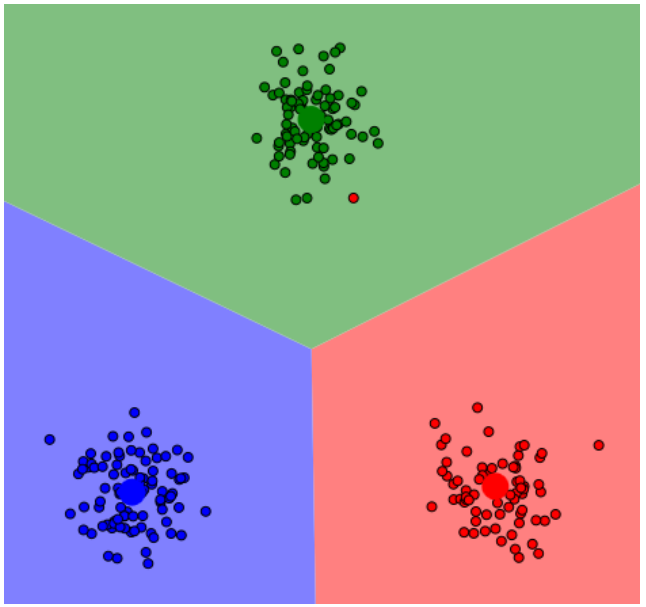
- **Move centroid** step repeated.

K-Means Clustering

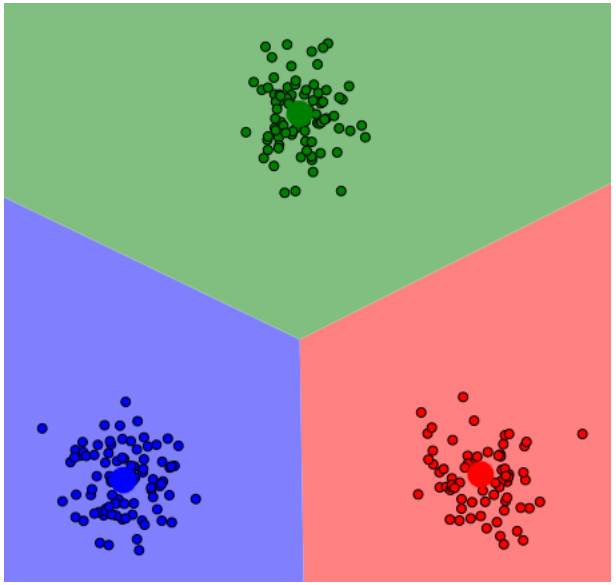
- **Cluster Assignment** step repeated.



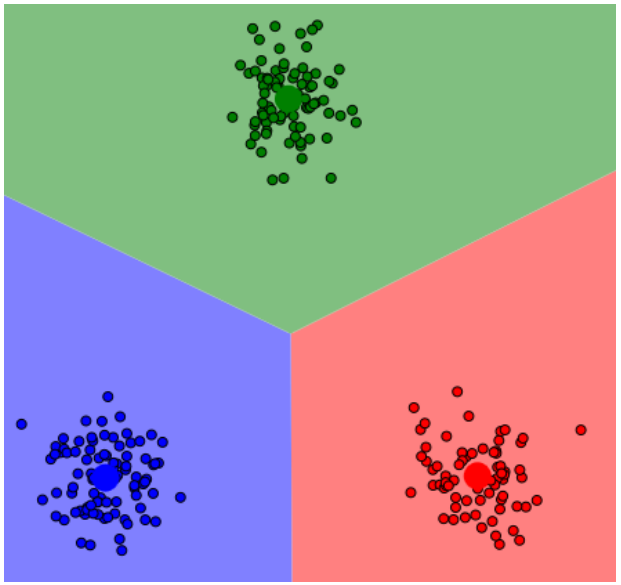
- **Move centroid** step repeated.



K-Means Clustering



- **Cluster Assignment** step repeated.



- **Move centroid** step repeated.
- At this stage the algorithm has converged.

Formal Description of K – Means

- Input
  - K (number of clusters)
  - Unlabelled **training** set $\{x^1, x^2, .. , x^m\}$
    - Where each x^i is a feature vector

Randomly initialize K cluster centroids $u_1, u_2, \ldots u_k$

for x iterations:

for i = 1 to m:

c_i = index (from 1 to K) of cluster centroid closest to x_i

for k = 1 to K:

u_k = mean of points assigned to cluster k

This is the **Cluster assignment** step. It takes each training example and assigns it a centroid index for the nearest centroid. If centroid with index 2 is nearest training example x^5 then it is assigned to that cluster centroid.

$c_i = \min_k || x_i - u_k ||$

Formal Description of K – Means

- Input
  - K (number of clusters)
  - Unlabelled training set $\{x^1, x^2, .. , x^m\}$
    - Where each x^i is a n dimensional vector

Randomly initialize K cluster centroids $u_1, u_2, \ldots u_k$

for x iterations:

    for i = 1 to m:
        c_i = index (from 1 to K) of cluster centroid closest to x_i

    for k = 1 to K:
        u_k = mean of points assigned to cluster k

This is the **move centroid** step. It assigns a new value to each centroid by averaging the value of all training set examples assigned to that centroid

Formal Description of K – Means

- Input
  - K (number of clusters)
  - Unlabelled training set $\{x^1, x^2, .. , x^m\}$
    - Where each x^i is a n dimensional vector

Randomly initialize K cluster centroids $u_1, u_2, \ldots u_k$

for x iterations:

    for i = 1 to m:
        c_i = index (from 1 to K) of cluster centroid closest to x_i

    for k = 1 to K:
        u_k = mean of points assigned to cluster k

Take a scenario where training examples 3, 5, and 12 are currently assigned to cluster centroid 2 (in other words $c_3 = 2$, $c_5 = 2$ and $c_{12} = 2$). Then $u_2 = (x^3 + x^5 + x^{12}) /3$

Distortion Cost Function

- The KMeans algorithm is attempting to minimize a specific cost function (often referred to as the distortion cost function).

- **$D(c_1, c_2, \ldots c_m, u_1, .., u_k) = \frac{1}{m}\sum_{i=1}^{m} | |x^i - U_{c(i)}| |^2$**

- K-Means attempts to find min $D(c_1, c_2, \ldots c_m, u_1, .., u_k)$

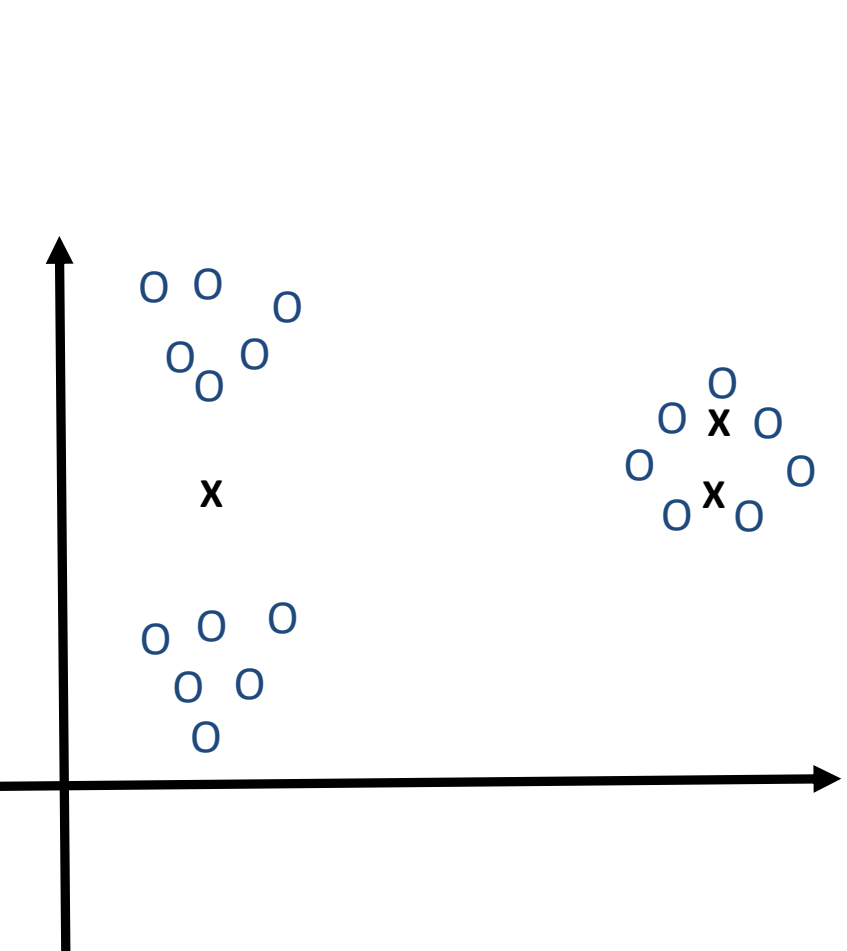
- Remember
  - c_i is the index of the cluster centroid closest to training example i
  - u_n is the cluster centroid n
  - $U_{c(i)}$ is the cluster centroid that training example x_i is connected to.

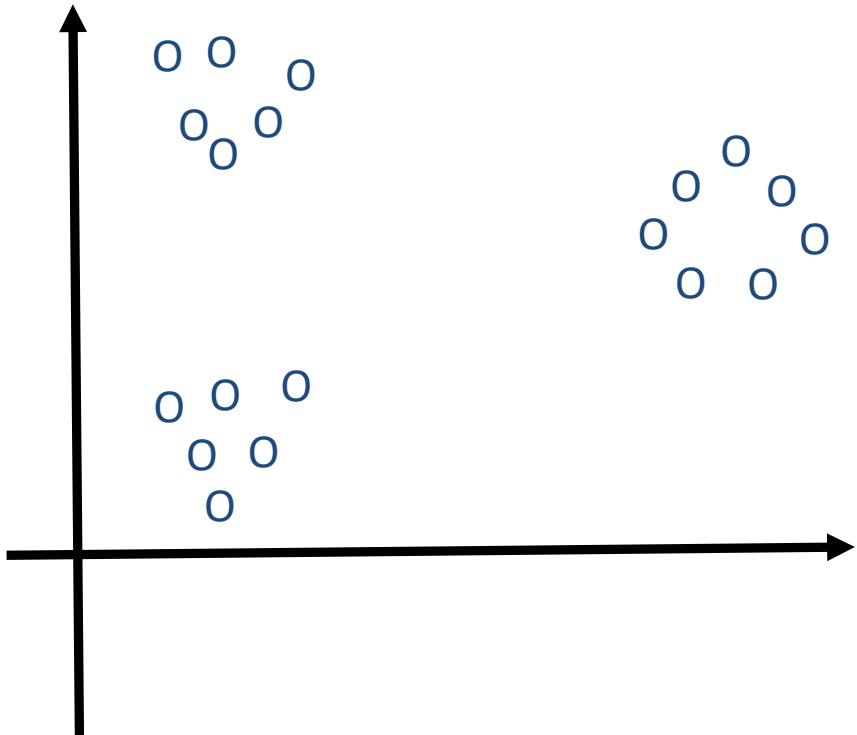
The distortion function D finds the **average** (across all *m* training examples) **squared distance** between each training example and the centroid to which it is connected to.

This function is useful as it can enable us to **monitor the performance** of our k means algorithm to make sure it is working as anticipated.

Local Optima

- KMeans is susceptible to arriving at **local optimums** (a local optima of the distortion function *D*).
- In other words it can converge incorrectly and provides a poor clustering of the training data points.





Local Optima

- KMeans is susceptible to arriving at local optimums (a local optima of the distortion function *D*). In other words it can converge incorrectly and provide a poor clustering of the training data points.

- The typical approach to solving this problem is to **run KMeans many times** and select the version that achieves the minimum distortion function value.

for 1 to 10:
    Randomly initialise Kmeans algorithm
    Run Kmeans
    Compute distortion function D
        (D(c1, c2, … cm, u1, .. , uk))

Select clustering solution that gives minimum distortion cost function value.
min D(c1, c2, … cm, u1, .. , uk)

Random Initialisation? – Not Quite!

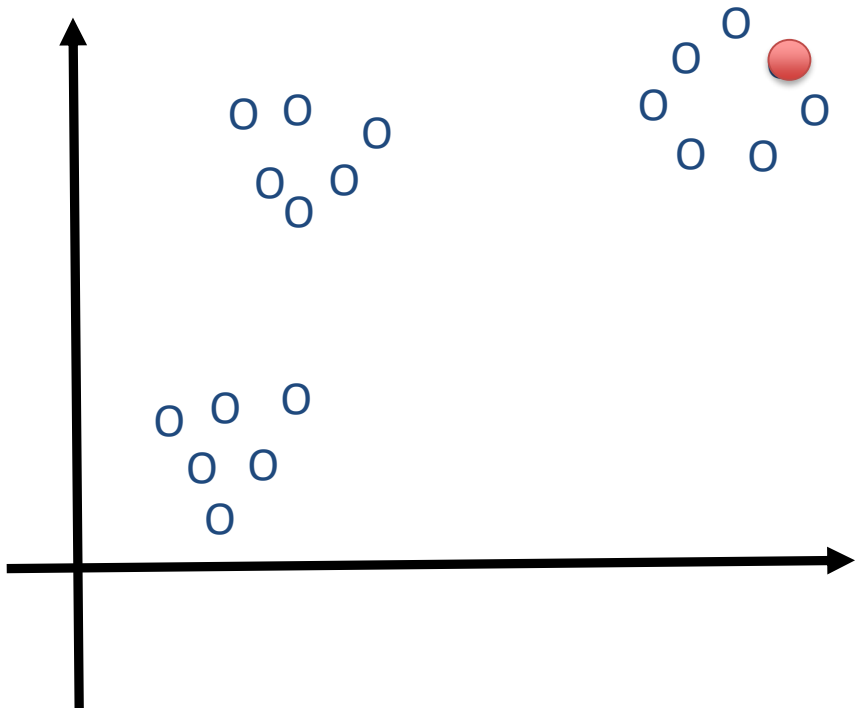
- One of the limitations of the standard k-means algorithm is that it is very sensitive to its initialization.

- The KMeans algorithm specifies that the centroids are randomly initialized.
- However, the random initialization is generally “controlled”.
- If we have specified n clusters then we randomly selected n training points to become the centroids for each cluster.
- This has been shown empirically to outperform versions of KMeans where there is absolute random initialization but is still very sensitive.

- An alternative version of kMeans called kMeans++ implements an initialization strategy that is more robust.

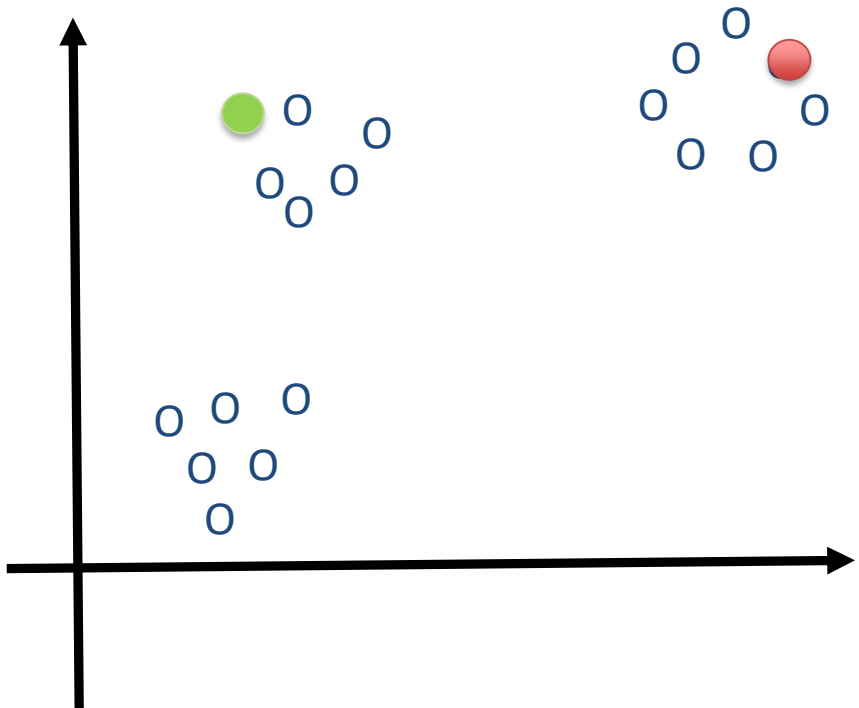
K-Means ++

- kMeans++ is a variant of kMeans that utilizes a different initialization strategy for the centroids.
- With kMeans the **probability** of being chosen as a **centroid** is related to the **minimum distance of a data points from the existing set of centroids**.
  - The closer a data point is to an existing centroid the less likely it will be chosen as a new centroid
  - The further away a data point from it's near centroid the more likely it is to be chosen as a new centroid.



K-Means ++

- kMeans++ is a variant of kMeans that utilizes a different initialization strategy for the centroids.
- With kMeans the **probability** of being chosen as a **centroid** is related to the **minimum distance of a data points from the existing set of centroids**.
  - The closer a data point is to an existing centroid the less likely it will be chosen as a new centroid
  - The further away a data point from it's near centroid the more likely it is to be chosen as a new centroid.



K-Means ++

- Starting with a dataset X of n points (x^1, x^2, … , x^n).

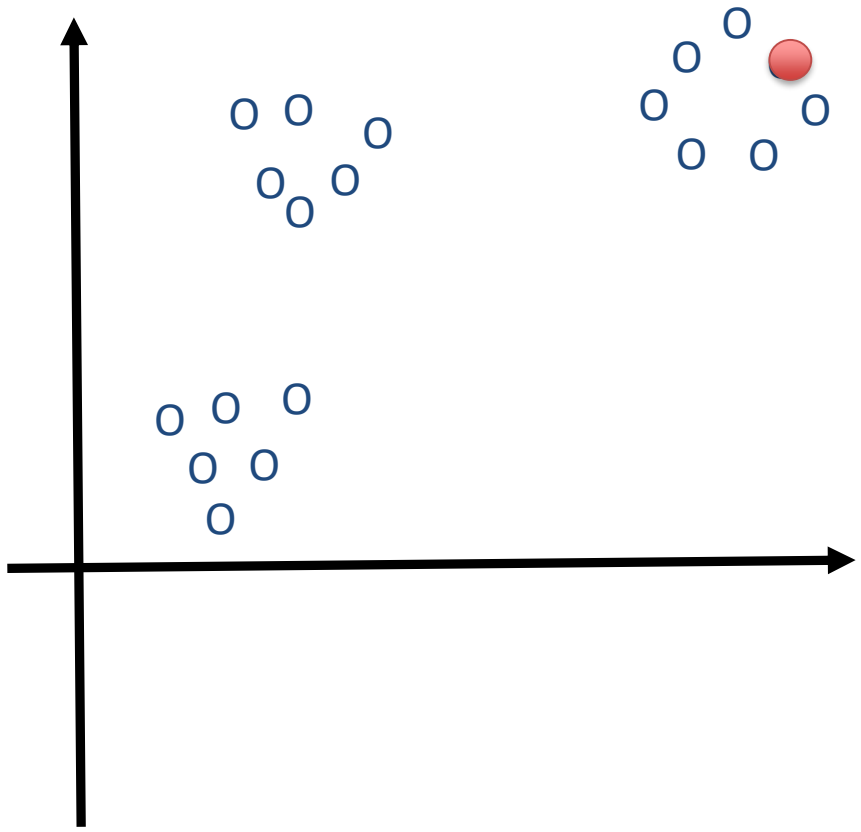
1. Select the first centre C_1 uniformly at random from the set of data points
2. Compute a vector containing the square distances between all points in the dataset and C_1:

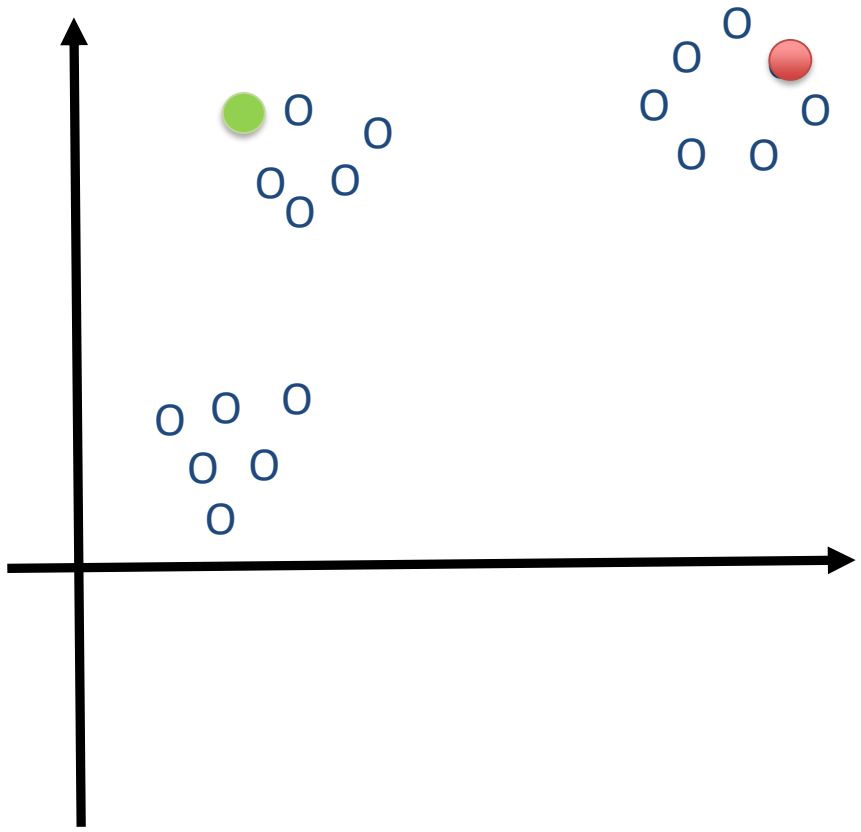
   $$D_i = || X^i - C_1 ||^2$$

3. Choose a second centre $\mathbf{C_2}$ from X randomly based on the probability distribution defined by: $\mathbf{D_i / sum(D)}$

4. Re-compute the distance vector as for each instance x^i we calculate
5. $D_i = \min(||X^i - C_1||^2, ||X^i - C_2||^2)$ and select C_3 as described in step 3.

6. Re-compute the distance vector as Di = $\min(||X^i - C_1||^2, ||X^i - C_2||^2 , \dots ||X^i - C_n||^2$ and select C_i as described in step 3.
7. Stop when K centres are selected.



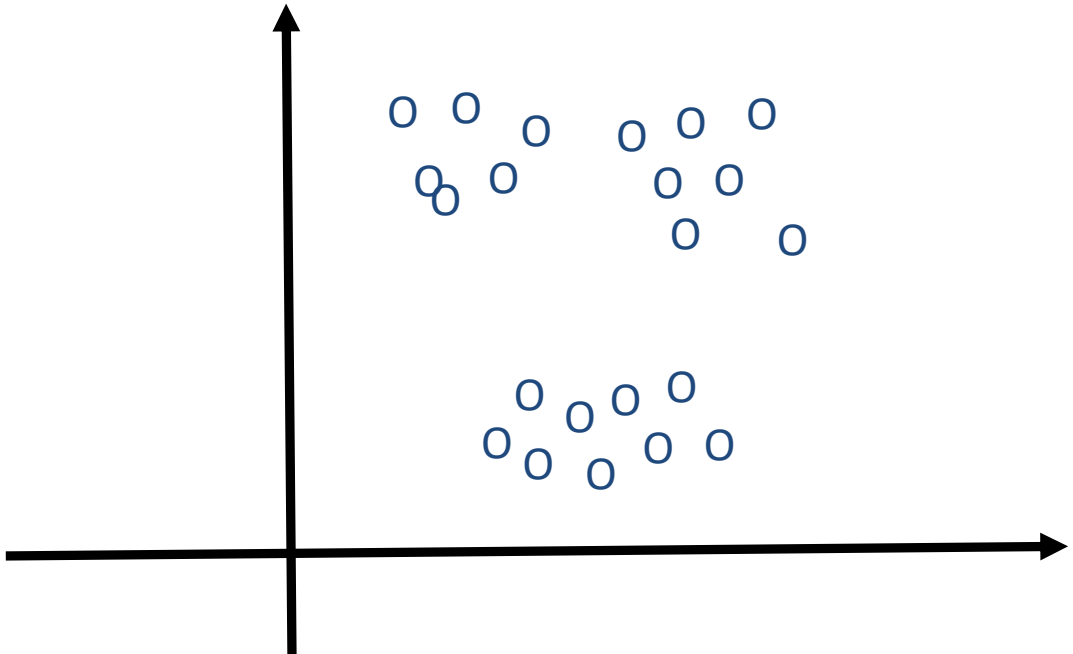


Choosing a Value of K

- Selecting an **appropriate value of K** for the K-Means algorithm is very important and can have a significant impact on the results.

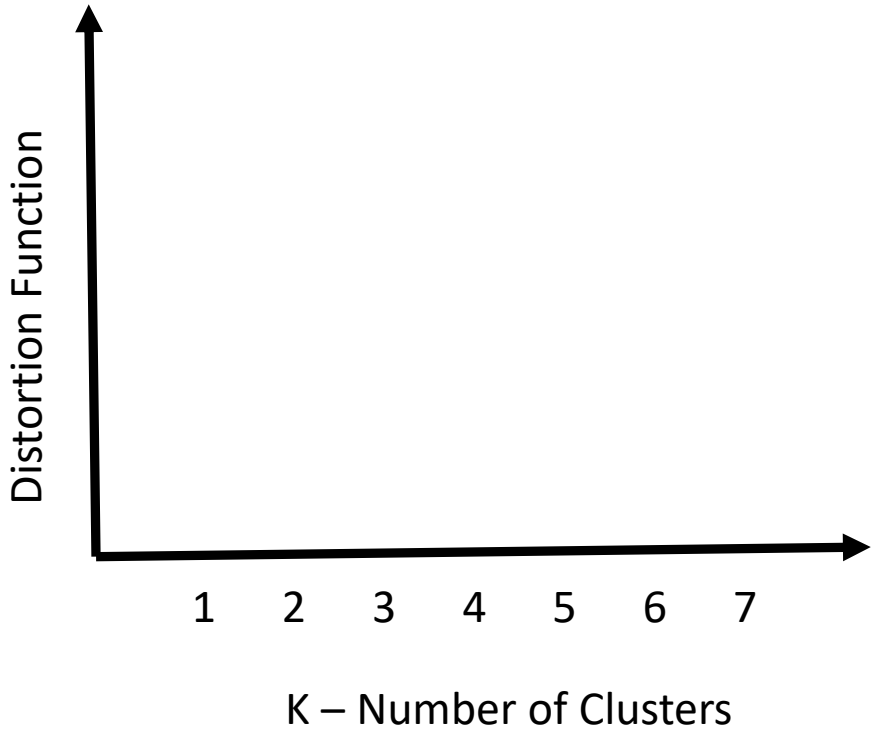
- Unfortunately there is no proven method of obtaining the optimal value for k.

- It can often be very ambiguous how many clusters there are in the data. In the example presented earlier it was very clear there were three clusters.



Choosing a Value of K – Elbow Method

- One method that can be employed is referred to as the elbow method or elbow plot.
- It maps the distortion value achieved by selecting different values of k.
- If you end up with a graph such as the one below that begins to plateaux after K=3, it can give a strong indication that K = 3 is the most appropriate value of K. Note this is still not a very reliable method as the graph may often not take this ideal form.



Choosing a Value of K – Elbow Method

- One method that can be employed is referred to as the elbow method or elbow plot.
- It maps the distortion value achieved by selecting different values of k.
- If you end up with a graph such as the one below that begins to plateaux after K=3, it can give a strong indication that K = 3 is the most appropriate value of K. Note this is still not a very reliable method as the graph may often not take this ideal form.