Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 2

Statistical learning methods

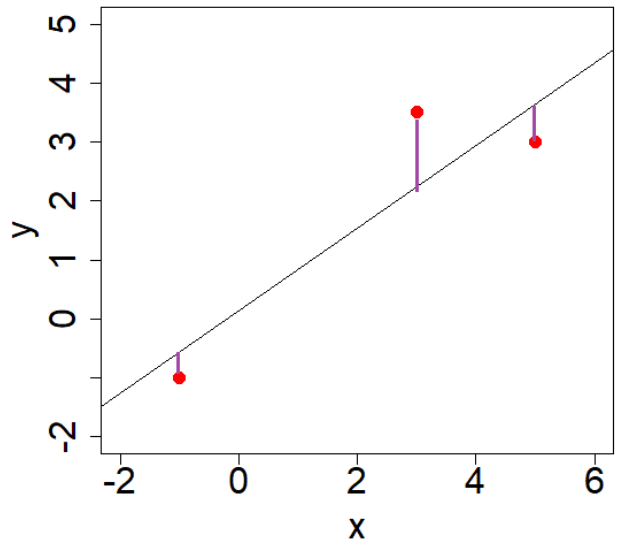
Supervised learning

- In data science, many applications involve making predictions about the outcome y based on a number of predictors x
- Often we assume models of the form

$$y = f(x)$$

where $f(x)$ is a function that maps the predictor(s) to the outcome.

Supervised learning



- For example, in simple linear regression with only one predictor, we assume a model of the form

$$y \approx f(x) = \beta_0 + \beta_1 x.$$

- We have also learnt how to estimate the unknown parameters β_0 and β_1 via the method of *least squares*.

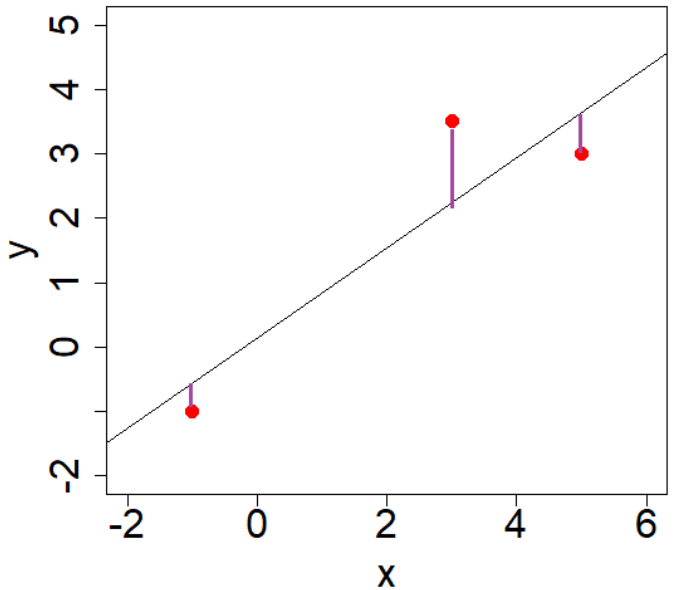
Supervised learning

- Simple linear regression is an example of *supervised learning* since we are training the model

$$y \approx f(x) = \beta_0 + \beta_1 x$$

based on both the response data y as well as the predictor variable x.

Supervised learning



i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1	-1	-1	$\beta_0 + (-1)\beta_1$	$-1 - (\beta_0 + (-1)\beta_1) = -1 - \beta_0 + \beta_1$
2	3	3.5	$\beta_0 + (3)\beta_1$	$3.5 - (\beta_0 + (3)\beta_1) = 3.5 - \beta_0 - 3\beta_1$
3	5	3	$\beta_0 + (5)\beta_1$	$3 - (\beta_0 + (5)\beta_1) = 3 - \beta_0 - 5\beta_1$

Supervised learning

Source: The Straits Times

- It is call 'supervised learning' because we have data on both the outcome y and the predictor x.
- Therefore, the data can 'teach' us, given a certain predictor value for x, what is the corresponding outcome y.

Supervised learning

Source: The Straits Times

- In data science, *supervised learning* usually entails training the model $y = f(x)$ based on data for both x and y.
- For example in simple linear regression, training the model

$$y \approx f(x) = \beta_0 + \beta_1 x$$

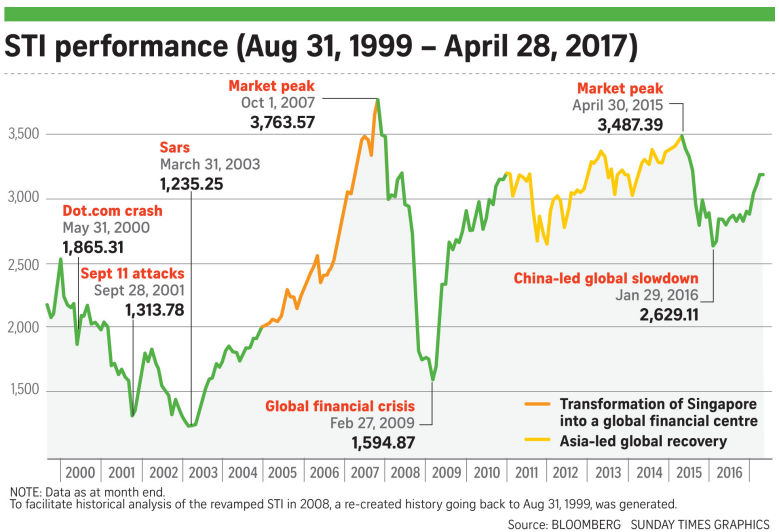
involves estimating the unknown parameters β_0 and β_1 via the method of *least squares*.

Examples of supervised learning

Source: The New Paper

- For self-driving cars, it is vital to identify pedestrians based on images.
- Training such recognition algorithms usually involves millions of image data (x) as well as labelling each image with whether there are pedestrians in it or not (y).
- With the trained model, the car can then predict whether there are pedestrians based on new image data.

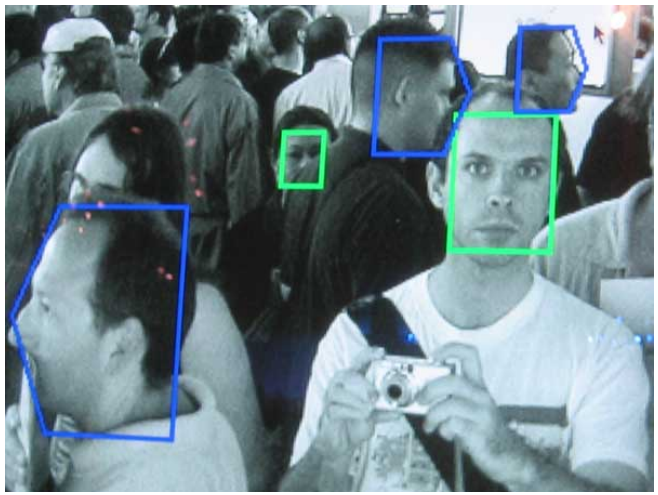
Examples of supervised learning



Source: Bloomberg, Sunday Times Graphics

- Automated 'robo-advisors' in financial technology industry can provide financial management advice with minimal human intervention
- Training a model for stock investment usually requires data on the predictors (x) such as geo-political risks, as well as the outcome stock price (y).

Examples of supervised learning



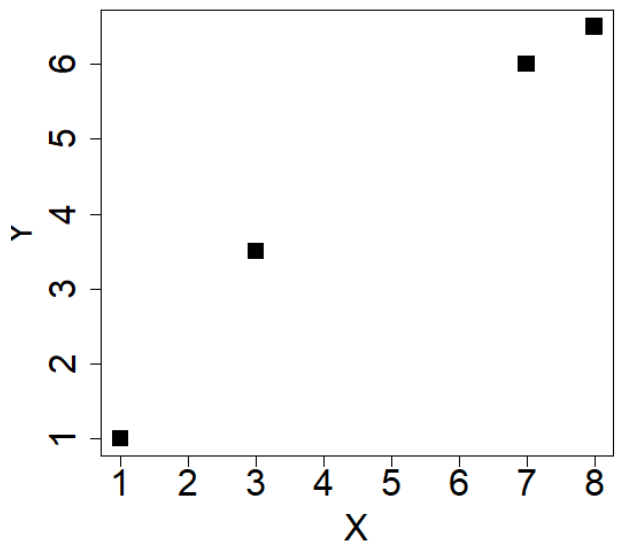
Source: Computerworld

- Automated surveillance programs need to recognize faces based on images to detect potential intruders.
- Here, we need to train a facial recognition model based on millions of image data (x) as well as labelling each image with whether it shows human face or not (y).
- With the trained model, the surveillance program can then predict whether there are potential intruders based on image data.

Unsupervised learning

- Suppose we only have data on the predictors x, but not the outcome y.
- *Unsupervised learning* is the task of inferring hidden structure based on data without the outcome y.
- We call such data without y 'unlabeled' data.

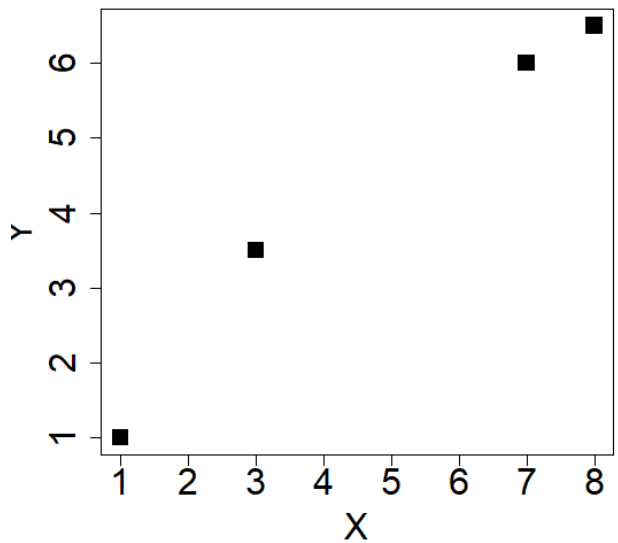
Unsupervised learning



- For example, in simple linear regression, we assume that we have data on both x and y.

i	x_i	y_i
1	1	1
2	3	3.5
3	7	6
3	8	6.5

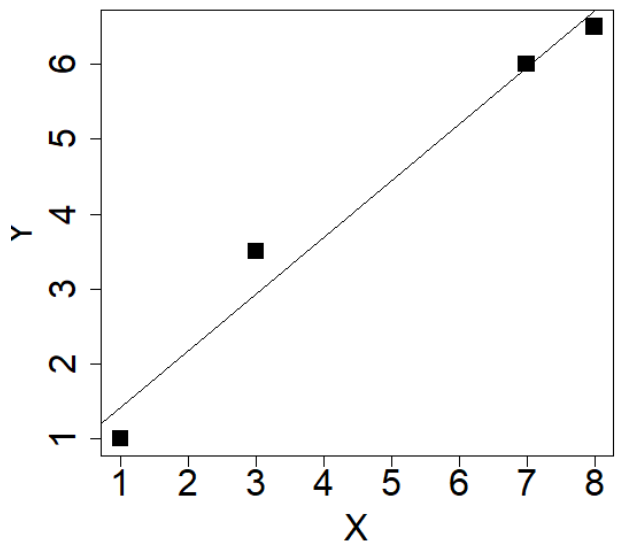
Unsupervised learning



- Let us assume that x refers to the number of Instagram followers and y refers to the total retail spending (in S$1,000) in a month.

i	x_i	y_i
1	1	1
2	3	3.5
3	7	6
3	8	6.5

Unsupervised learning

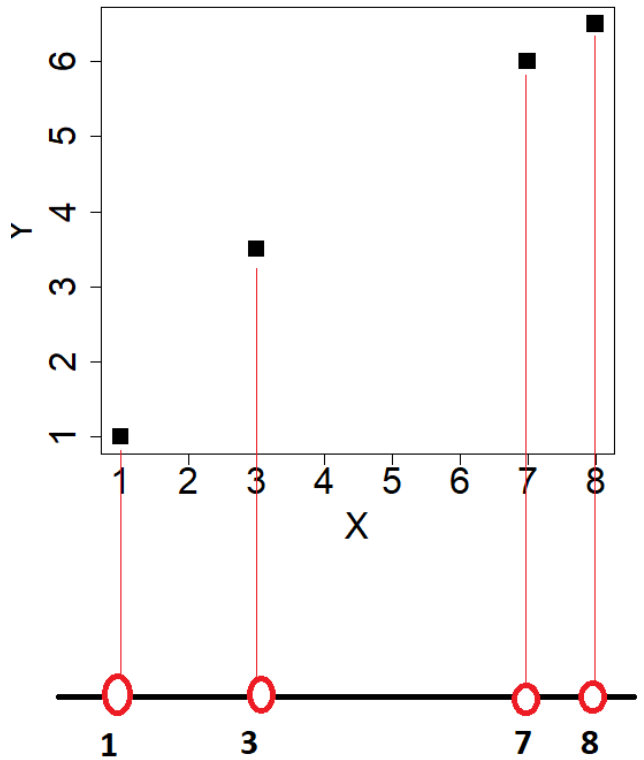


- Using the method of *least squares*, we can estimate the unknown parameters in the model

$$y \approx f(x) = \beta_0 + \beta_1 x.$$

- We can then perform prediction for total retail spending for a person with a given number of Instagram followers, using the R function `predict`.

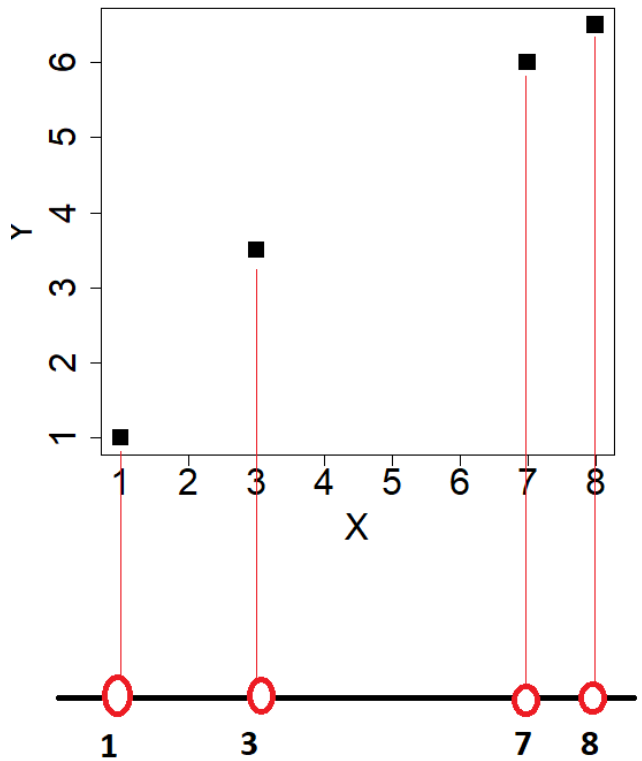
Unsupervised learning



- Now suppose we only have data on x, the number of Instagram followers

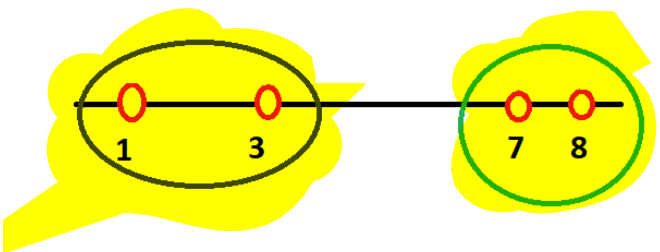
i	x_i
1	1
2	3
3	7
3	8

Unsupervised learning



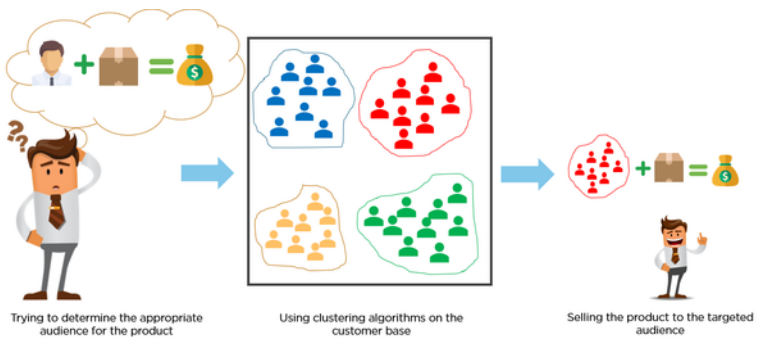
- Then we cannot train our predictive model based on simple linear regression
- Is there something else we can do with the available data?

Unsupervised learning



- We can still explore and find structures within our data on x, the number of Instagram followers
- For example, we can cluster the customers into two groups, one group with low number of followers and another group with high number of followers
- These data analytic findings may provide insights for developing marketing strategies for different group of customers.

Example: unsupervised learning



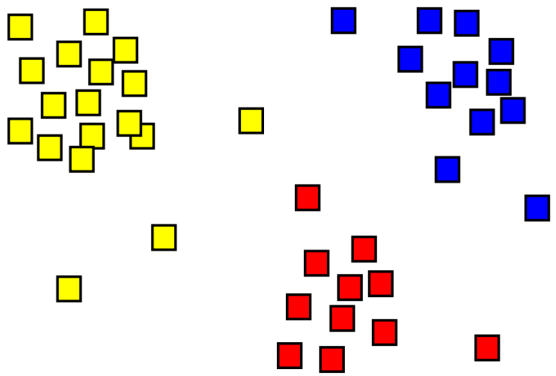
- For example, you are trying to introduce a new product to the market, so you do not know how much a customer is willing to spend on the product.
- However, you can still perform clustering analysis on the customers based on attributes such as gender, age and social media activity.
- This exploratory data analysis will allow you to gain a better understanding of the potential customers for your new product.

Examples of supervised and unsupervised learning algorithms

Supervised	**Unsupervised**
Linear regression	k-means
Decision trees	Association rules
k-nearest neighbor	Hierarchical clustering
Linear discriminant analysis	Deep belief nets
Naive Bayes	Self-organizing map

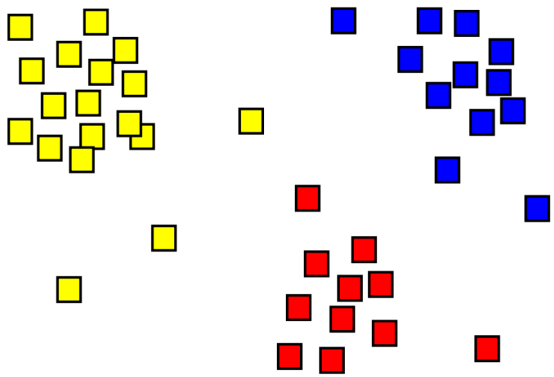
Clustering Methods

Clustering



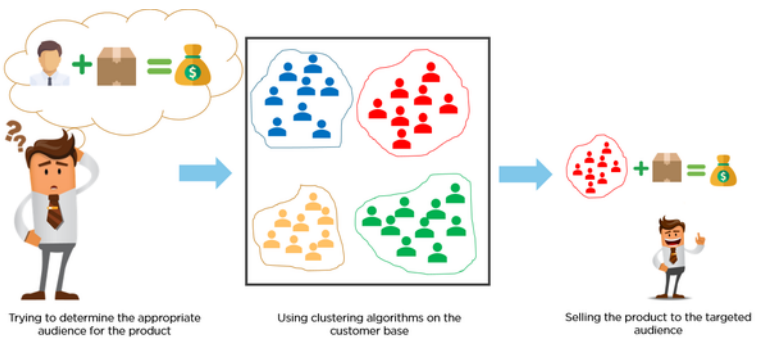
- *Clustering* refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- When we cluster the observations of a data set, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other.

Clustering



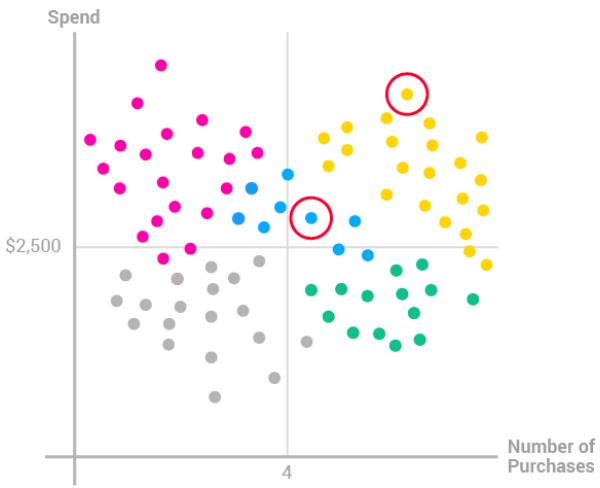
- Clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes.
- Clustering is an example of *unsupervised learning*.
- *Unsupervised learning* tries to find hidden structure in the unlabeled data.

Clustering



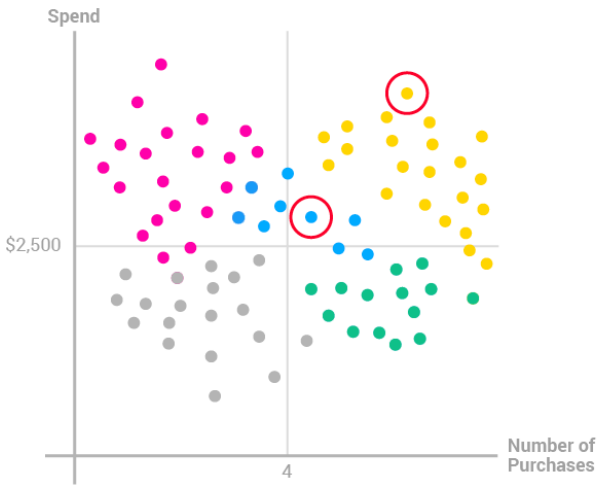
- In marketing, clustering methods from data science can discover distinct groups of customers
- This will allow for more targeted marketing strategies
- There is no group label for each customer, just his or her amount spent and number of purchases

Clustering



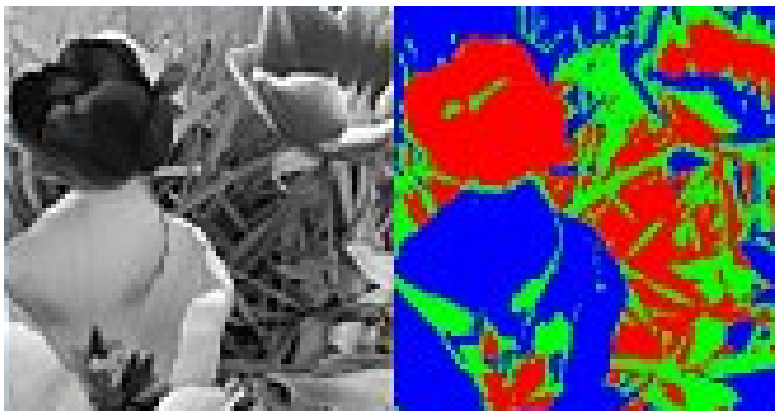
- Aim is to find certain natural groups of customers based on unlabeled data

Clustering



- Clustering is an example of *unsupervised learning*.
- There is no group label for each customer, just his or her amount spent and number of purchases.
- Clustering tries to find hidden structure in the unlabeled data.

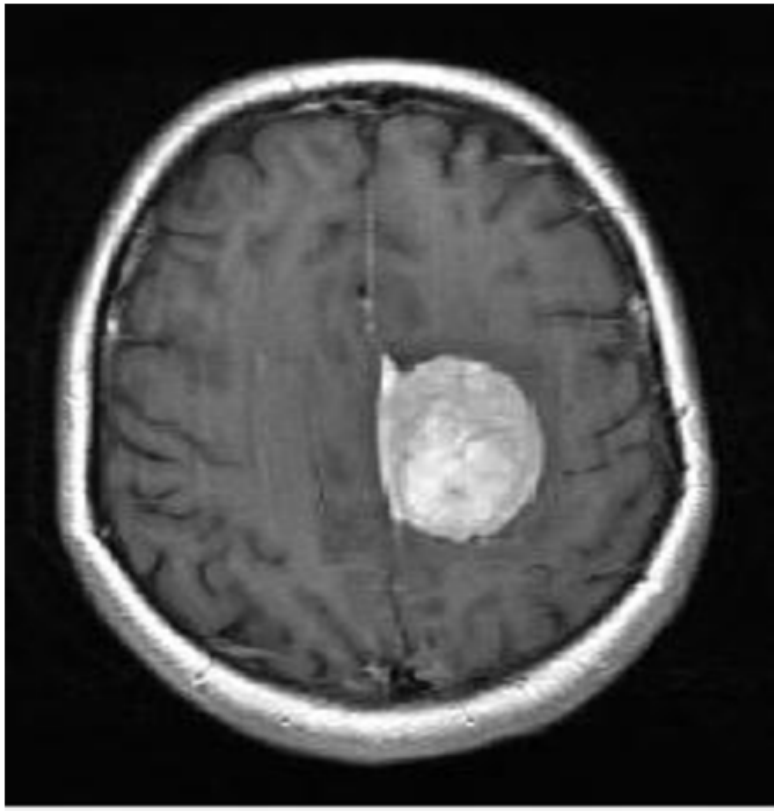
Clustering



Source: Wikipedia

- Clustering can be used in image processing to segment objects
- The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame.

Clustering



- Clustering can aid medical diagnostics
- e.g. identification of clusters with darker pixels

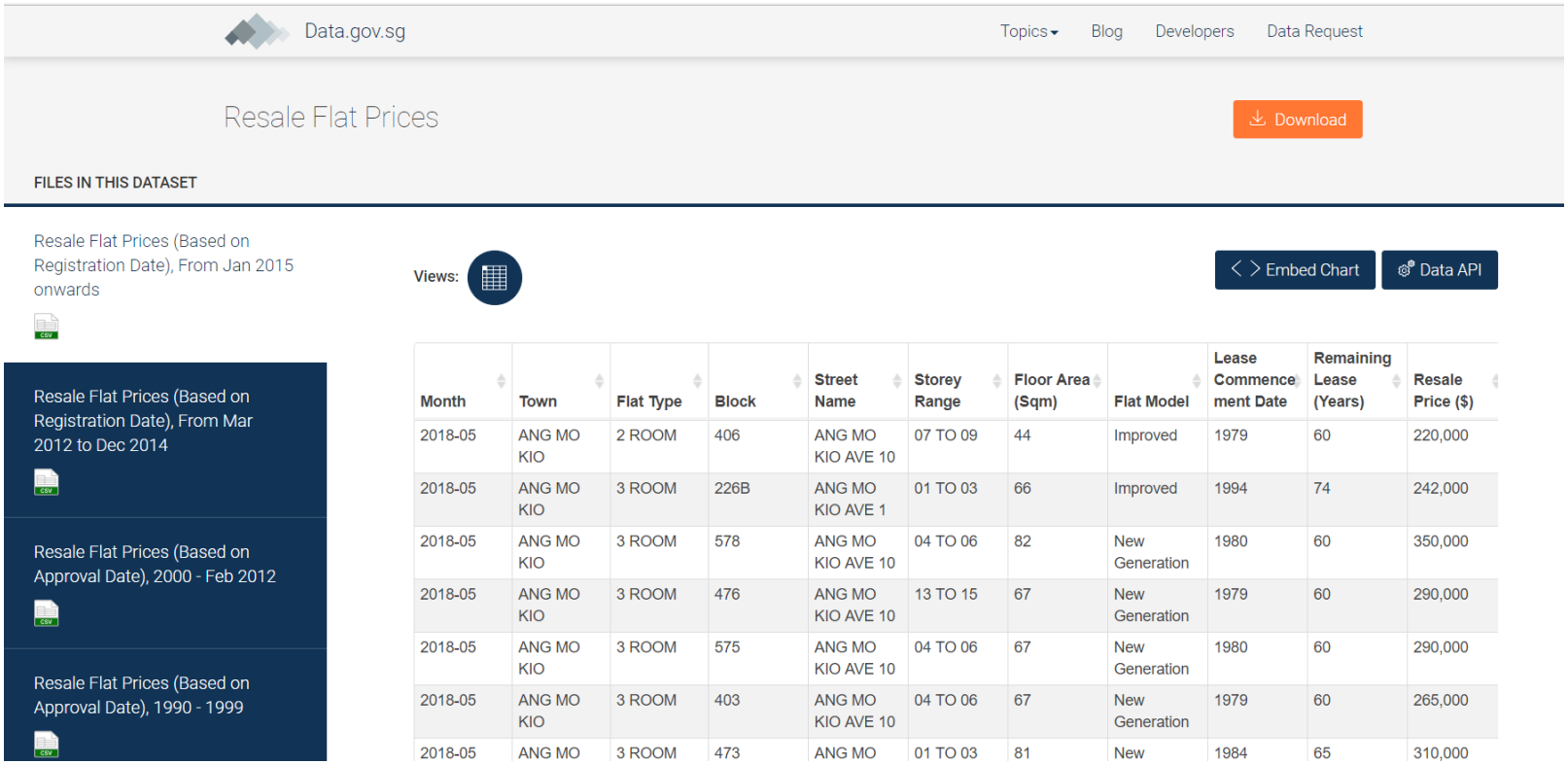
- Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters.
- These clusters could be used to target individuals for specific preventive measures or clinical trial participation.
- Clustering, in general, is useful in biology for the classification of plants and animals as well as in the field of human genetics.

Example closer to home...

Source: The Straits Times

- Data on resale HDB prices based on registration date is publicly available from https://data.gov.sg/dataset/resale-flat-prices.
- We have extracted a subset of all the resale records for three-room flats from March 2012 to December 2014.
- For the purposes of this analysis, I have also added artificially generated variables.
- Available as the data set `HDBresale_cluster.csv` on the course website.

Example closer to home...



Data.gov.sg

Topics Blog Developers Data Request

Resale Flat Prices

Download

FILES IN THIS DATASET

Resale Flat Prices (Based on Registration Date), From Jan 2015 onwards

Resale Flat Prices (Based on Registration Date), From Mar 2012 to Dec 2014

Resale Flat Prices (Based on Approval Date), 2000 - Feb 2012

Resale Flat Prices (Based on Approval Date), 1990 - 1999

Views:

Embed Chart Data API

Month	Town	Flat Type	Block	Street Name	Storey Range	Floor Area (Sqm)	Flat Model	Lease Commence ment Date	Remaining Lease (Years)	Resale Price ($)
2018-05	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	07 TO 09	44	Improved	1979	60	220,000
2018-05	ANG MO KIO	3 ROOM	226B	ANG MO KIO AVE 1	01 TO 03	66	Improved	1994	74	242,000
2018-05	ANG MO KIO	3 ROOM	578	ANG MO KIO AVE 10	04 TO 06	82	New Generation	1980	60	350,000
2018-05	ANG MO KIO	3 ROOM	476	ANG MO KIO AVE 10	13 TO 15	67	New Generation	1979	60	290,000
2018-05	ANG MO KIO	3 ROOM	575	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1980	60	290,000
2018-05	ANG MO KIO	3 ROOM	403	ANG MO KIO AVE 10	04 TO 06	67	New Generation	1979	60	265,000
2018-05	ANG MO	3 ROOM	473	ANG MO	01 TO 03	81	New	1984	65	310,000

Source: data.gov.sg

Example closer at home...

- Let us take a look at the data set in `HDBresale_cluster.csv`:

```
> resale=read.csv("HDBresale_cluster.csv")
> head(resale)
     X flat_type floor_area_sqm amenities
1  580         3          ROOM           74         32
2  581         3          ROOM           74         37
3  582         3          ROOM           73         34
4  583         3          ROOM           59         38
5  584         3          ROOM           68         39
6  586         3          ROOM           68         36
```

Example closer to home...

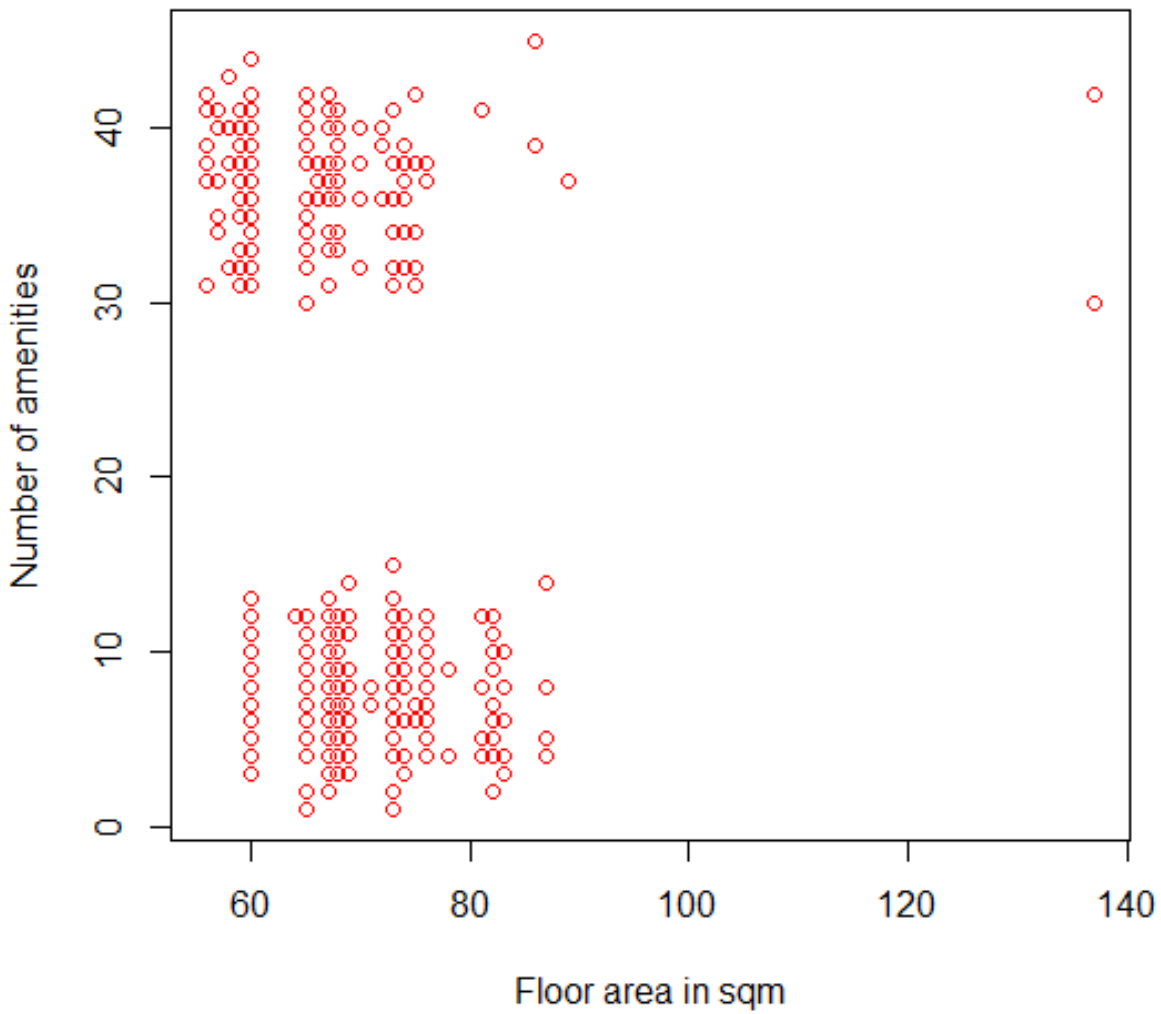
- Let us take a look at the two features in this dataset:

(i) Floor area in square meters

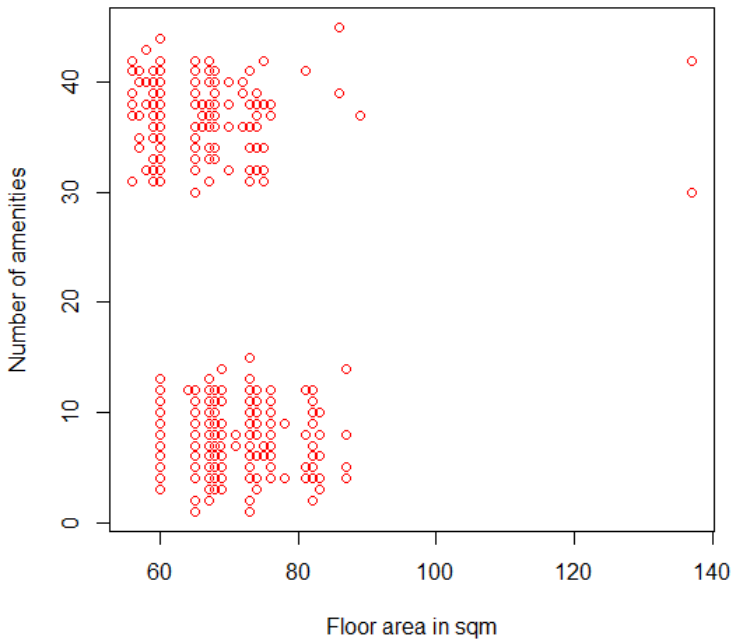
(ii) The number of amenities in the vicinity of the HDB unit

```
plot(x=resale$floor_area_sqm, y=resale$amenities,
xlab="Floor area in sqm", ylab="Number of
    amenities", col="red")
```

Example closer to home...

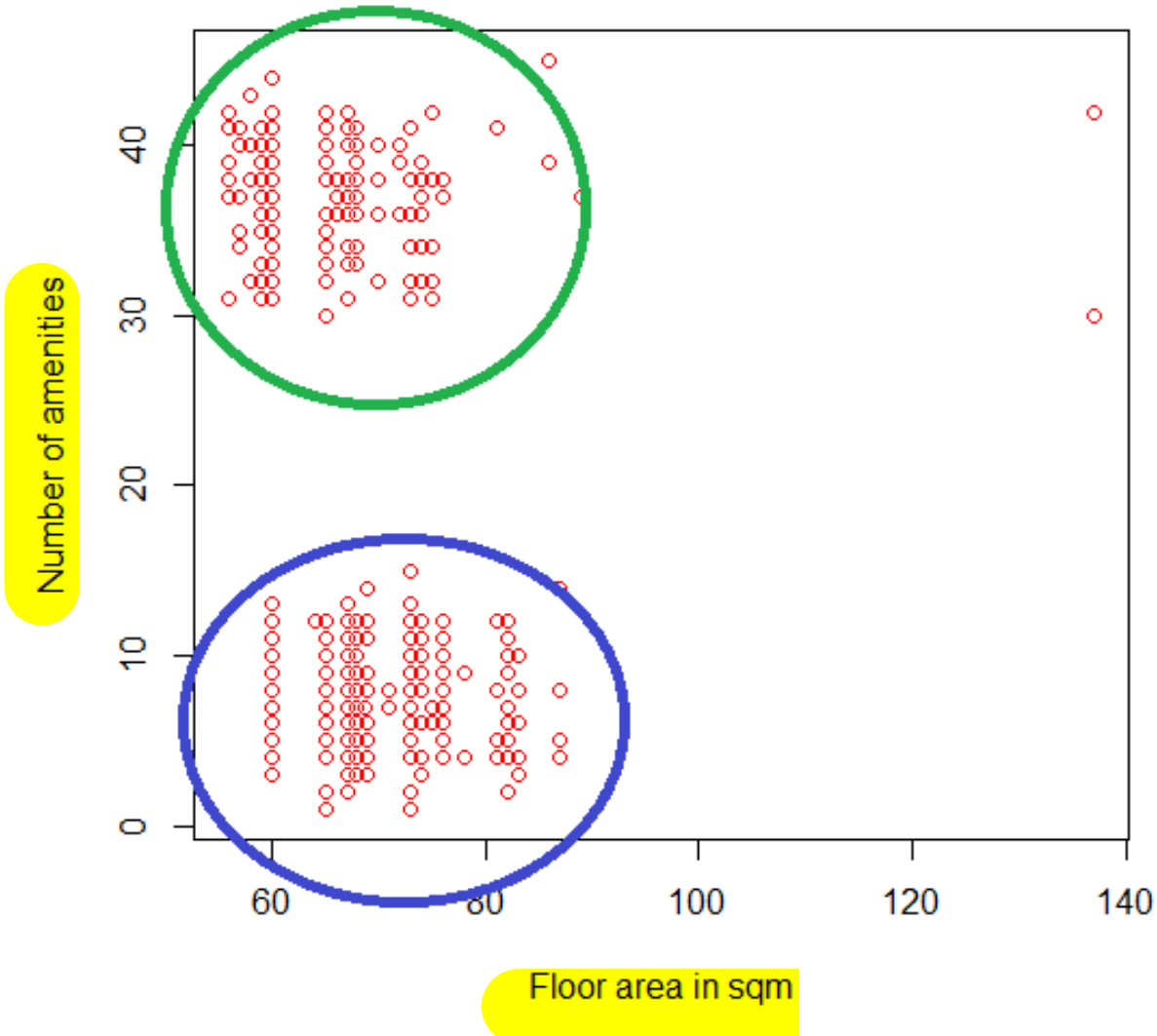


Example closer to home...



- In unsupervised clustering, you start with this data and then proceed to divide it into subsets.
- These subsets are called clusters and are comprised of data points that are most similar to one another.
- It appears that there are two clusters, one in the higher number of amenities group and the other in the lower number of amenities group.

Example closer to home...

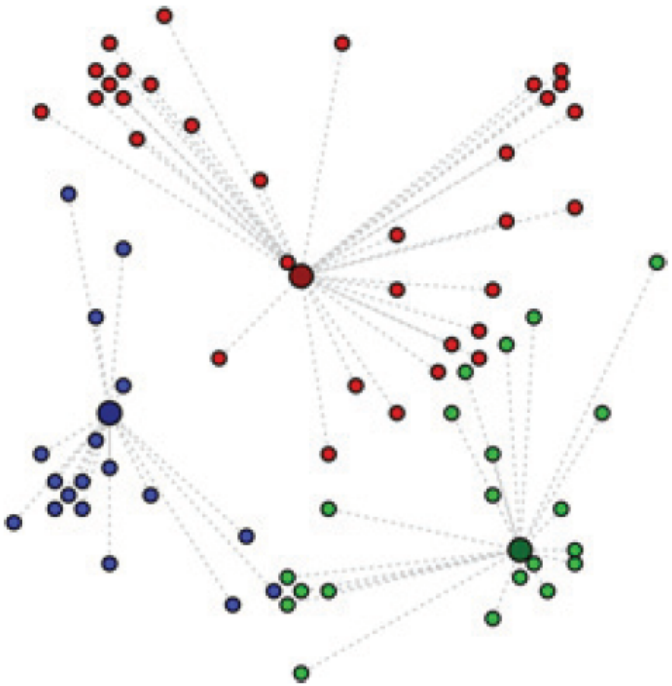


- Before running the clustering algorithm, you don't have an exact idea as to the nature of the subsets (clusters).
- However, by visual inspection, it seems that there are two distinct clusters in our HDB resale data set.
- We need an algorithm that provides us with a principled and automatic way for clustering.

- There are many clustering methods in data science that provide valuable insights from our data.
- Examples include K-means clustering, mean-shift clustering, density-based spatial clustering of applications with noise (DBSCAN), expectation-maximization (EM) clustering using Gaussian mixture models (GMM) and agglomerative hierarchical clustering.
- In this course we will focus on the popular k-means clustering algorithm.

K-means clustering

k-means clustering

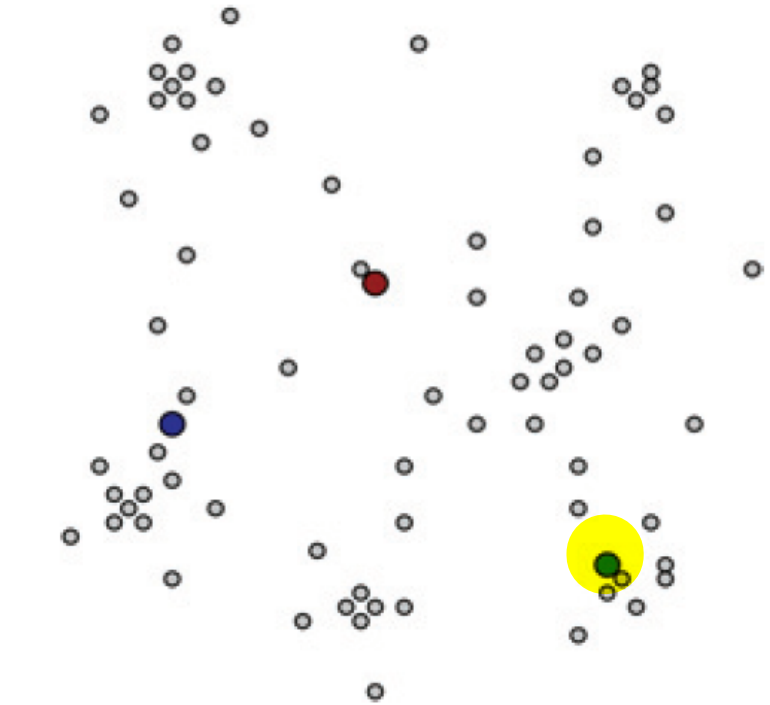


Possible k-means clusters for k=3

- Given a collection of objects each with n measurable attributes, *k-means* is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups.
- The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.
- We will illustrate with an example shortly.

- In this example, we want to find k clusters from a collection of M objects with n attributes (or features).
- We will consider the case with two features. For example, in the HDB resale example we have floor area and resale price, so that $n = 2$.
- Each HDB resale record therefore corresponds to a point (x_i, y_i) in two-dimensional space. For example, x is the floor area and y is the resale price, for the i^{th} HDB resale record.

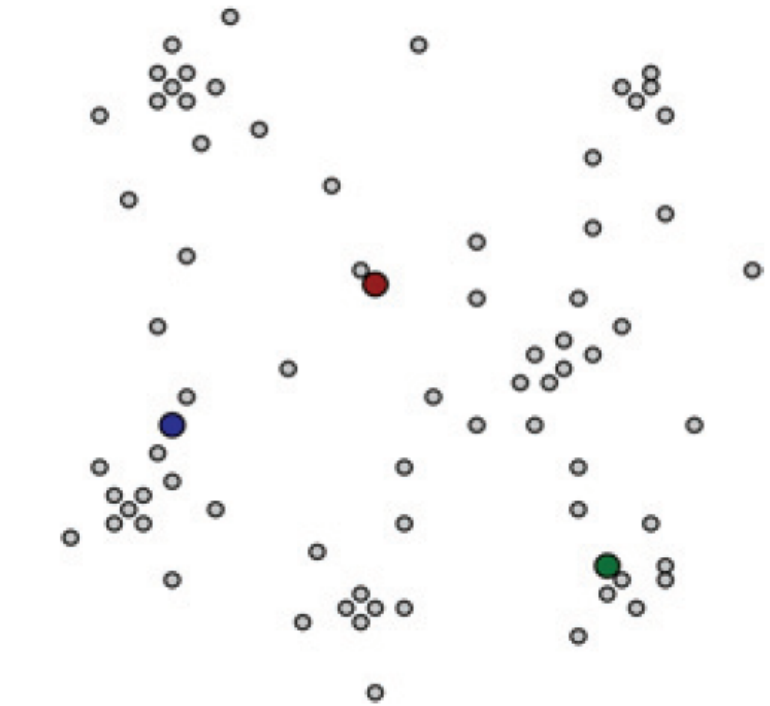
k-means clustering: 1^{st} step



Initial starting points for the centroids

- Choose the value of k and the k initial guesses for the centroids.
- In this example, $k = 3$, and the initial centroids are indicated by the points shaded in red, green, and blue.

k-means clustering: 2^{nd} step



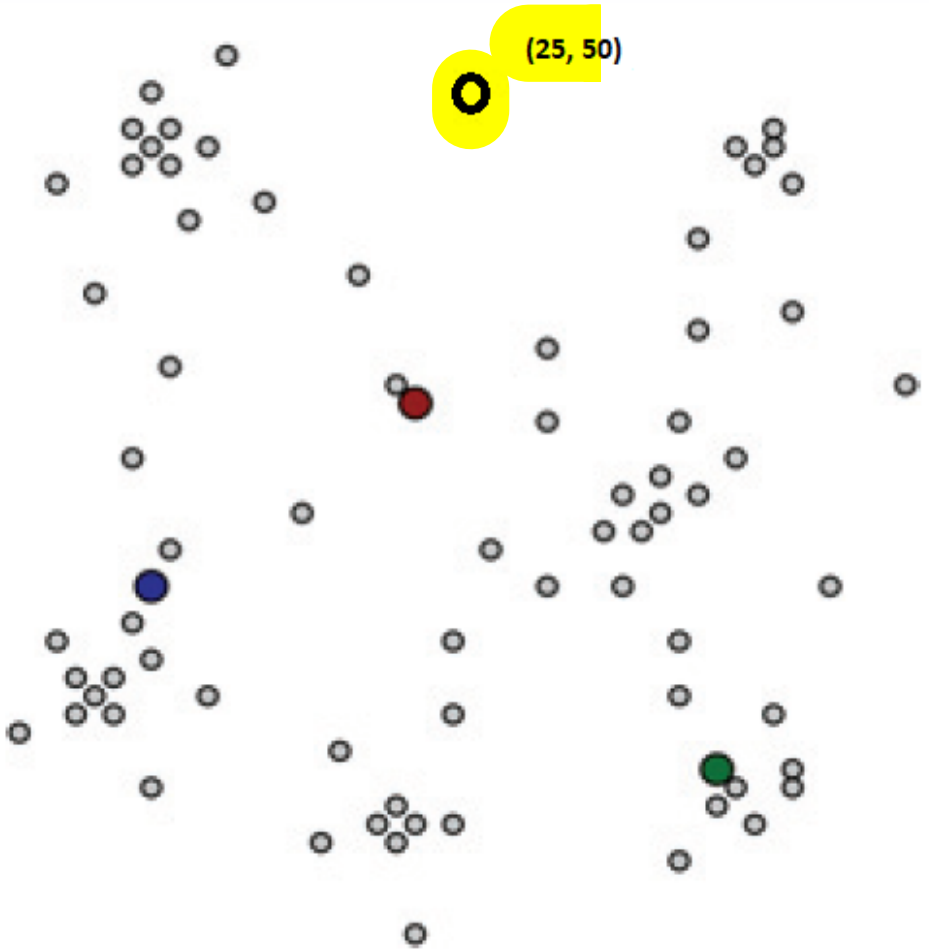
Initial starting points for the centroids

- Compute the distance from each data point (x_i, y_i) to each centroid.
- The distance d between any two points (x_1, y_1) and (x_2, y_2) is given by

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

- We will illustrate this calculation for a chosen point to help you understand this procedure.

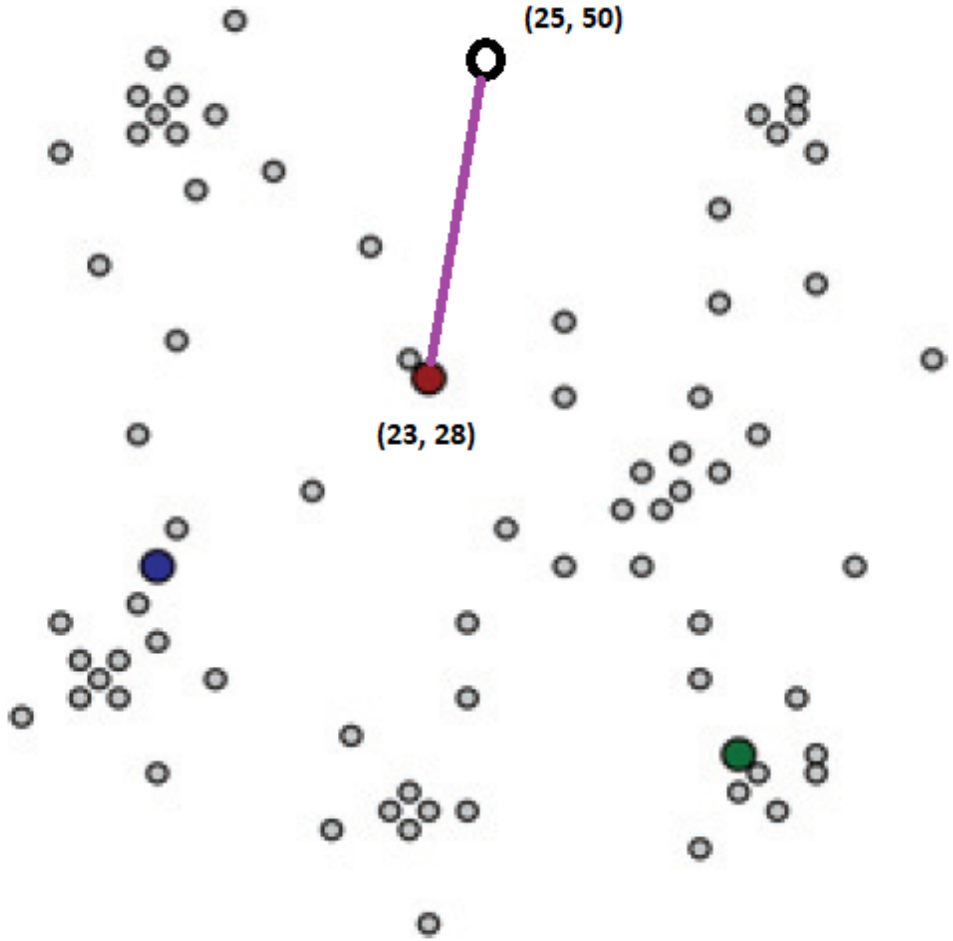
k-means clustering: 2^{nd} step



Initial starting points for the centroids

- Suppose the bold circled point has coordinates $(25, 50)$.

k-means clustering: 2^{nd} step



Initial starting points for the centroids

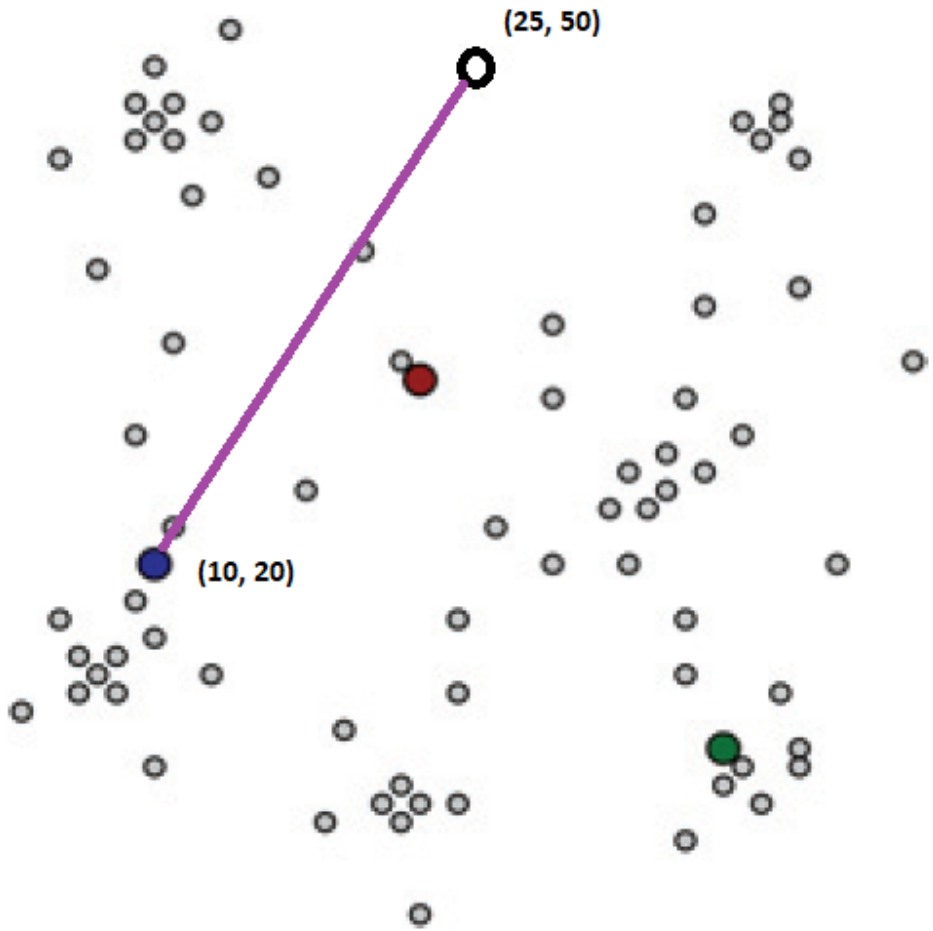
- The distance from this point to the first centroid (in red) is

$$\sqrt{(25-23)^2+(50-28)^2}$$

$$=\sqrt{488}.$$

- This is also the length of the purple line.

k-means clustering: 2^{nd} step



Initial starting points for the centroids

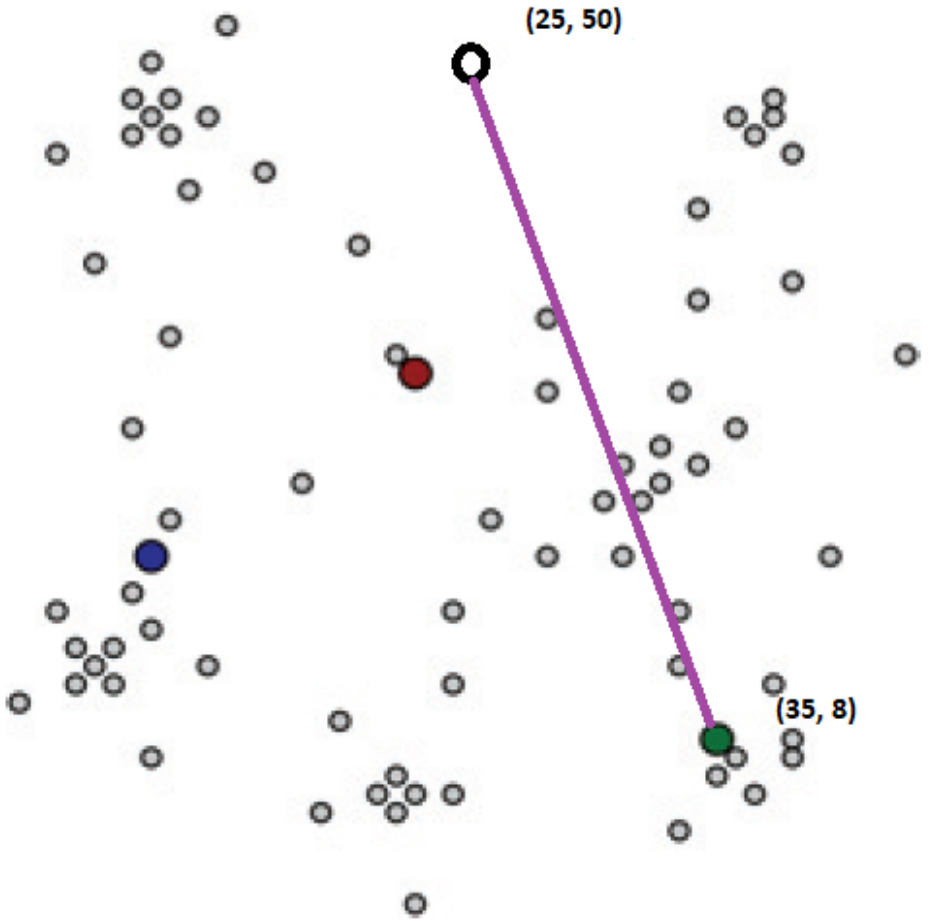
- Next, the distance from this point to the second centroid (in blue) is

$$\sqrt{(25-10)^2+(50-20)^2}$$

$$=\sqrt{1125}.$$

- This is also the length of the purple line.

k-means clustering: 2^{nd} step



Initial starting points for the centroids

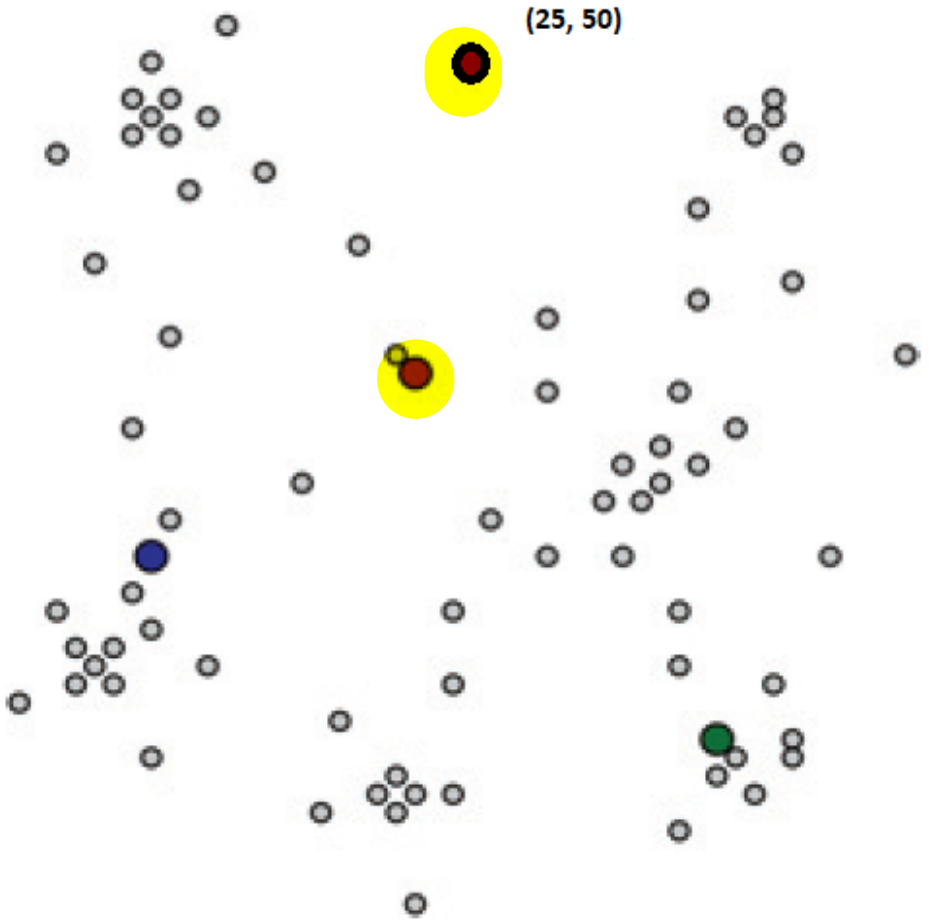
- Finally, the distance from this point to the third centroid (in green) is

$$\sqrt{(25-35)^2+(50-8)^2}$$

$$=\sqrt{1864}.$$

- This is also the length of the purple line.

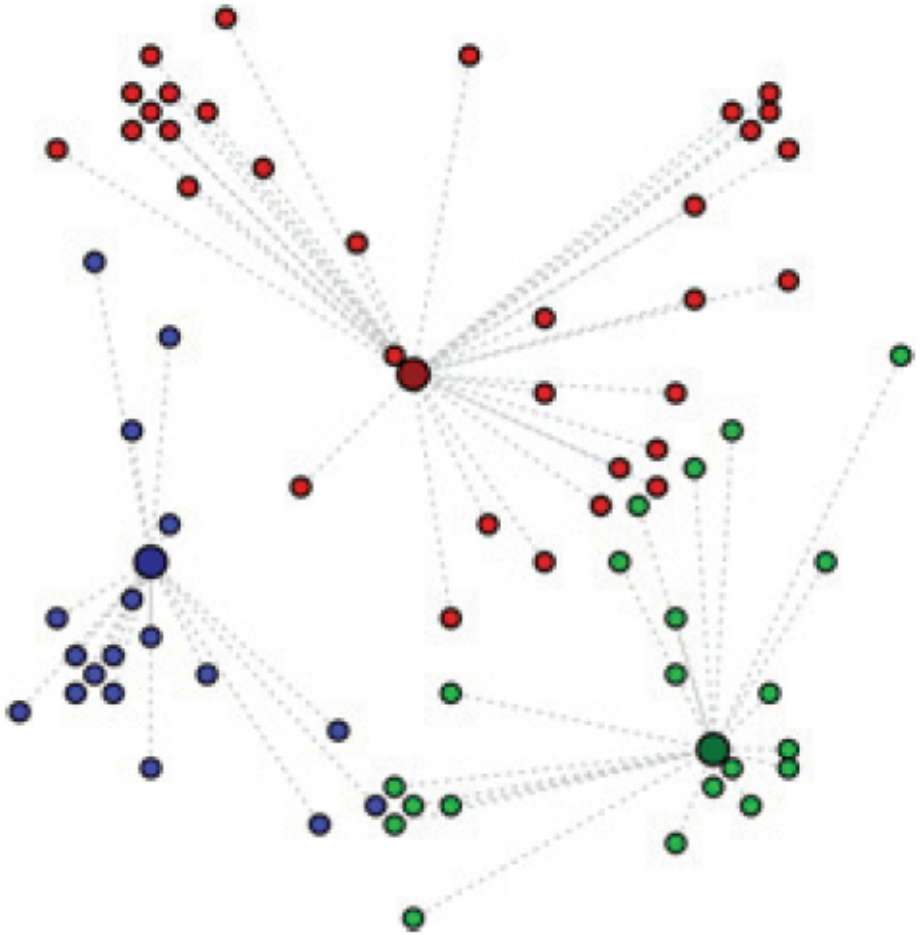
k-means clustering: 2^{nd} step



Initial starting points for the centroids

- Since the distance from this point to the first centroid (in red) is the shortest at $\sqrt{488}$, we classify this point as belonging to the red group.

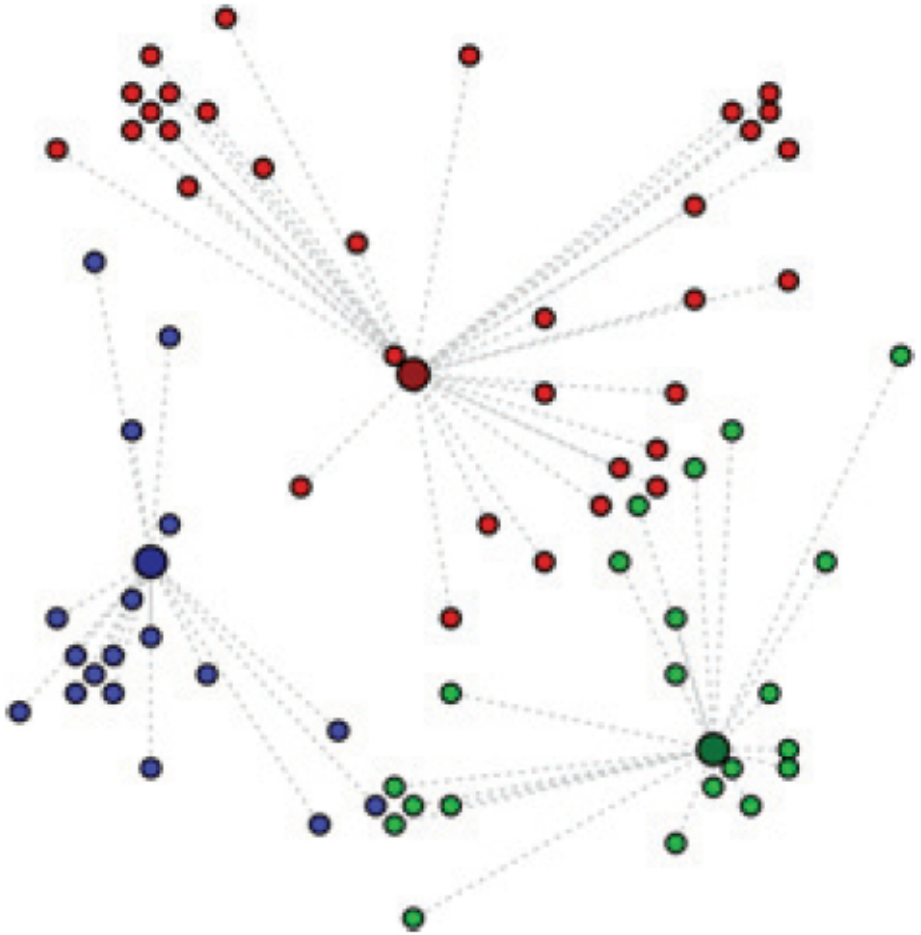
k-means clustering: 2^{nd} step



Points are assigned to the closest centroid

- We repeat this process for all the other points.
- Assign the points to their closest centroid.

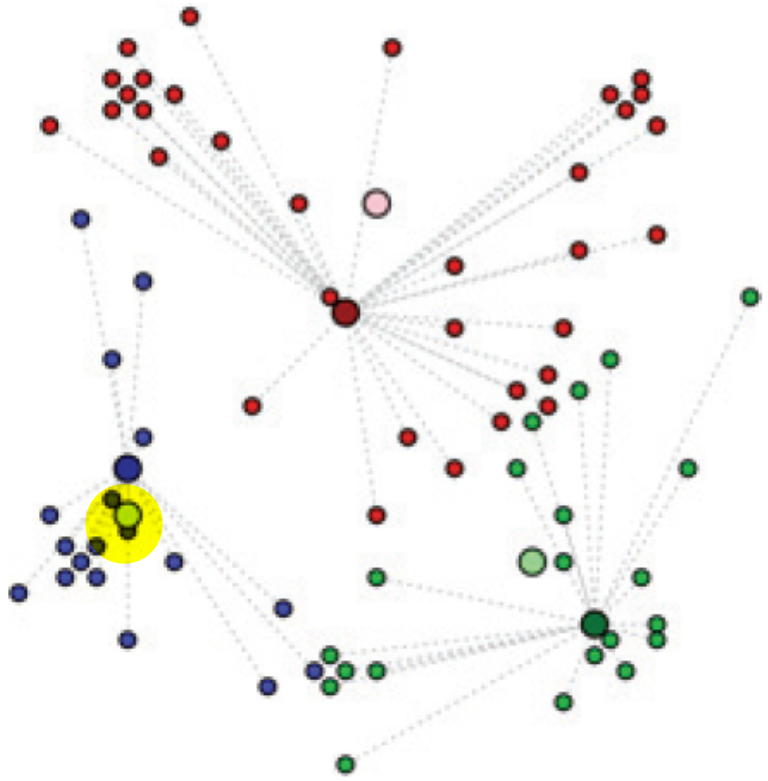
k-means clustering: 3^{rd} step



Points are assigned to the closest centroid

- We now have three clusters of points (blue, red, green).
- Compute the centroids in each cluster.

k-means clustering: 3^{rd} step



Compute the mean of each cluster

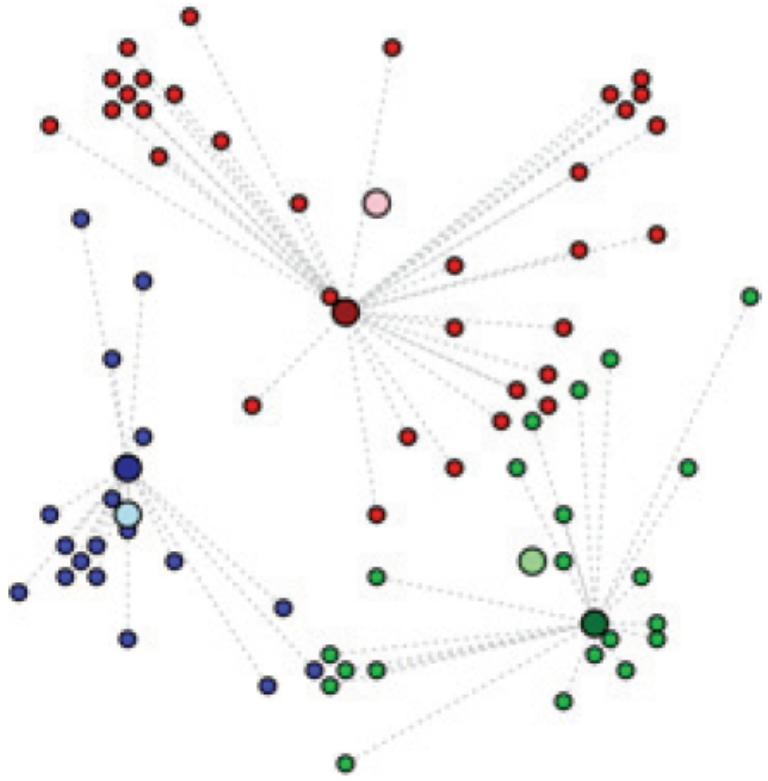
- For example, there are 18 points belonging to the blue cluster.
- The blue centroid is calculated as

$$(x_{blue}, y_{blue}) = \left(\frac{1}{18}\sum_{i=1}^{18} x_i, \frac{1}{18}\sum_{i=1}^{18} y_i\right)$$

based on the coordinates of the 18 points in the blue cluster.

- The blue cluster centroid is indicated by the pale blue point.

k-means clustering



Compute the mean of each cluster

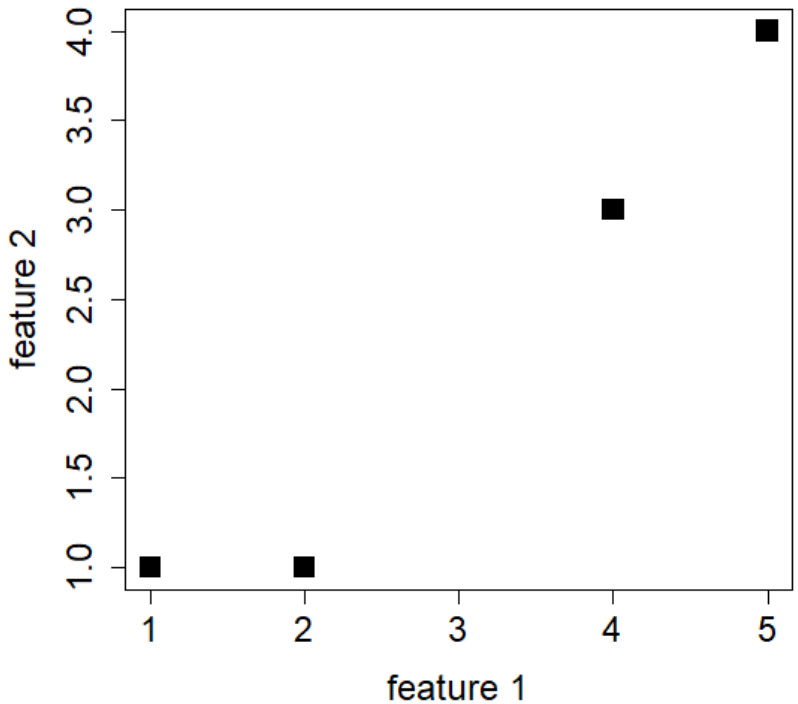
- Now we have three new centroids indicated by the pale blue, pale red and pale green dots.
- Repeat steps one and two until the algorithm converges to an answer:

(i) Assign each point to the closest centroid computed in Step 3.

(ii) Compute the centroid of newly defined clusters.

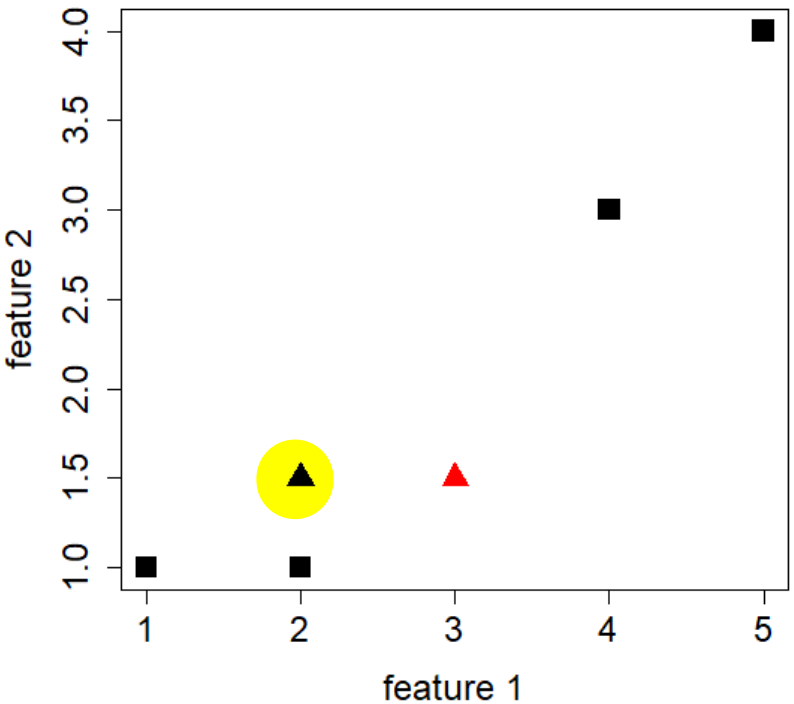
(iii) Repeat until the algorithm converges to an answer (when the centroids are more or less stable).

k-means clustering: a simple complete illustration of the algorithm**Exam



- Suppose we have four data points:
- (1 ,1)
- (2, 1)
- (4, 3)
- (5, 4)

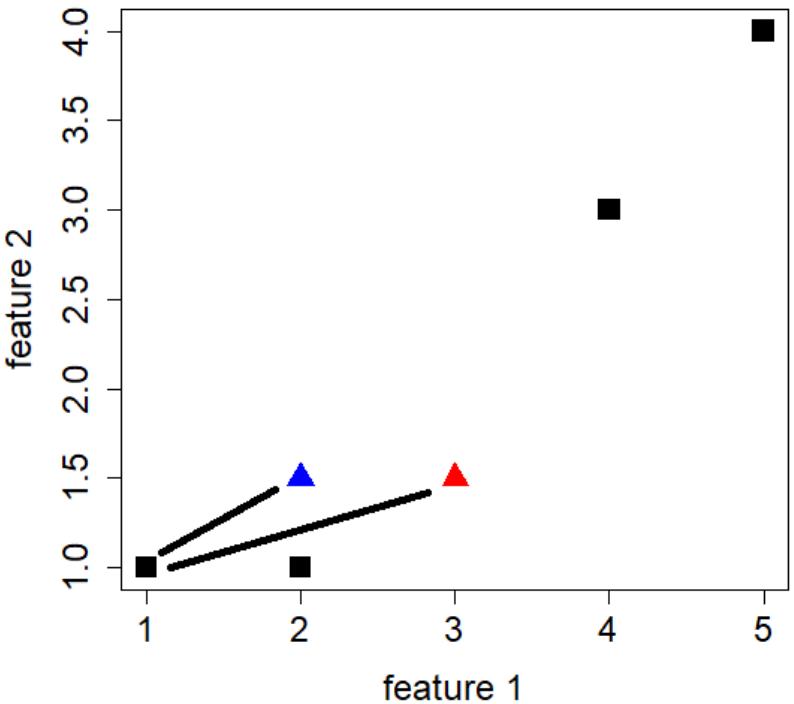
k-means clustering: 1^{st} iteration



May use data points as centroids

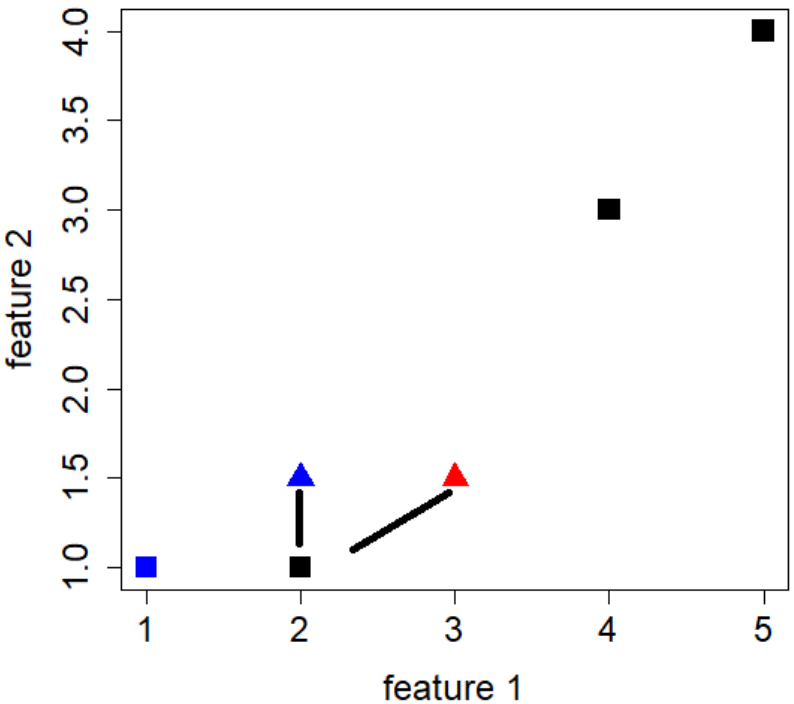
- We want $k = 2$ clusters, and randomly selected two starting centroids:
- Blue centroid: (2, 1.5)
- Red centroid: (3, 1.5)

k-means clustering: 1^{st} iteration



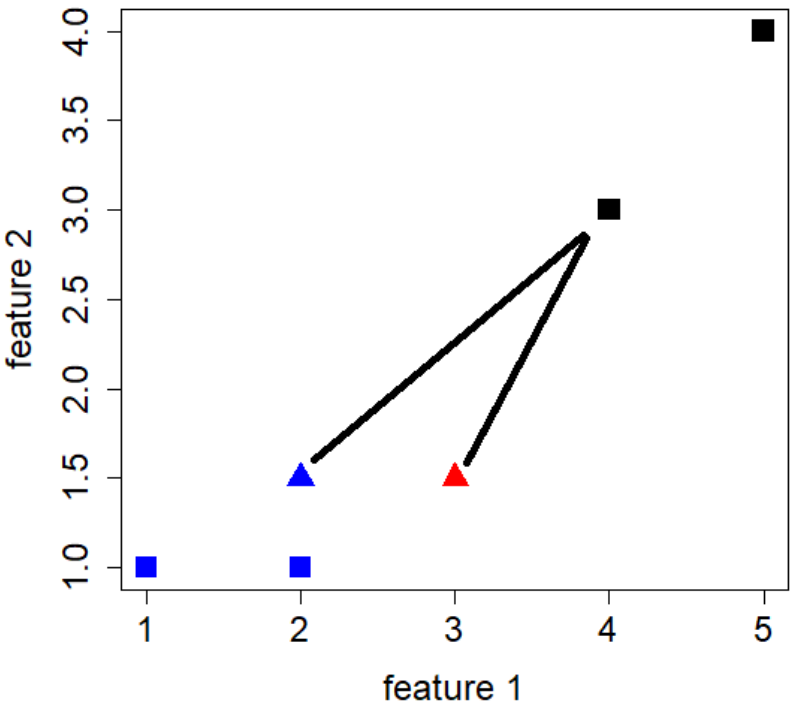
- Distance from first point to blue centroid: $\sqrt{(1-2)^2+(1-1.5)^2} = \sqrt{1.25}$
- Distance from first point to red centroid: $\sqrt{(1-3)^2+(1-1.5)^2} = \sqrt{4.25}$
- So the first point belongs to blue group...

k-means clustering: 1^{st} iteration



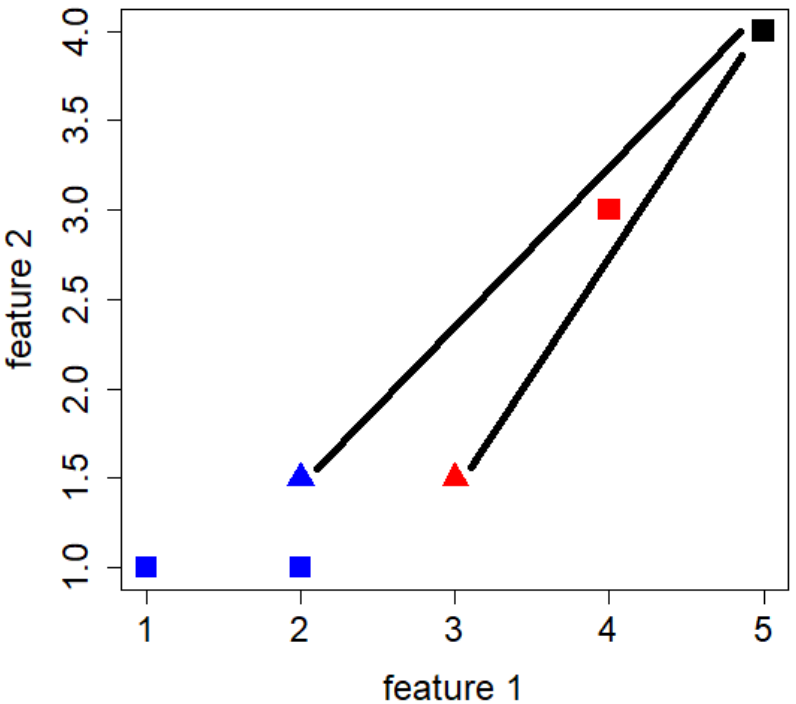
- Distance from second point to blue centroid: $\sqrt{(2-2)^2 + (1-1.5)^2} = \sqrt{0.25}$
- Distance from second point to red centroid: $\sqrt{(2-3)^2 + (1-1.5)^2} = \sqrt{1.25}$
- So the second point belongs to blue group...

k-means clustering: 1^{st} iteration



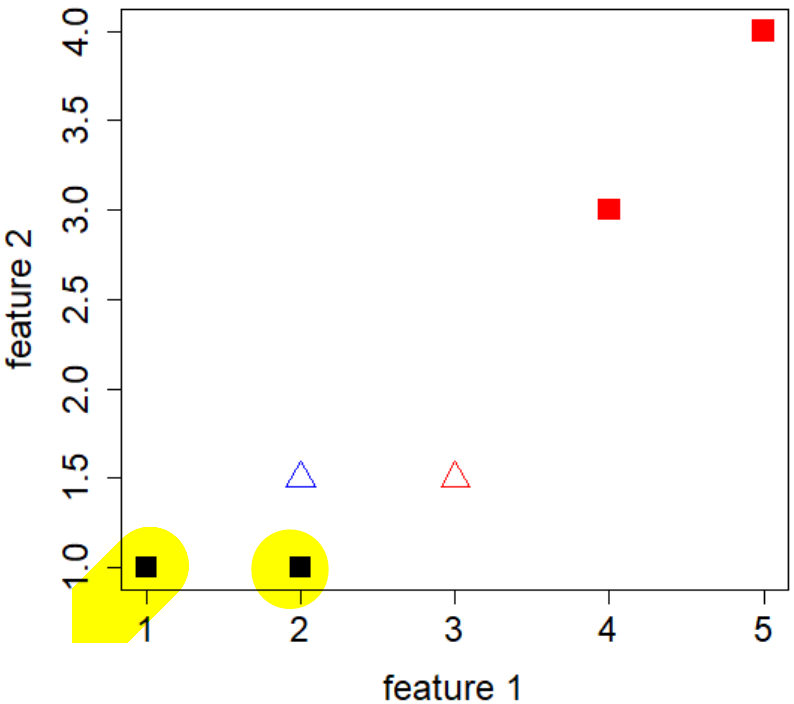
- Distance from third point to blue centroid: $\sqrt{(4-2)^2+(3-1.5)^2} = \sqrt{6.25}$
- Distance from third point to red centroid: $\sqrt{(4-3)^2+(3-1.5)^2} = \sqrt{3.25}$
- So the third point belongs to red group...

k-means clustering: 1^{st} iteration



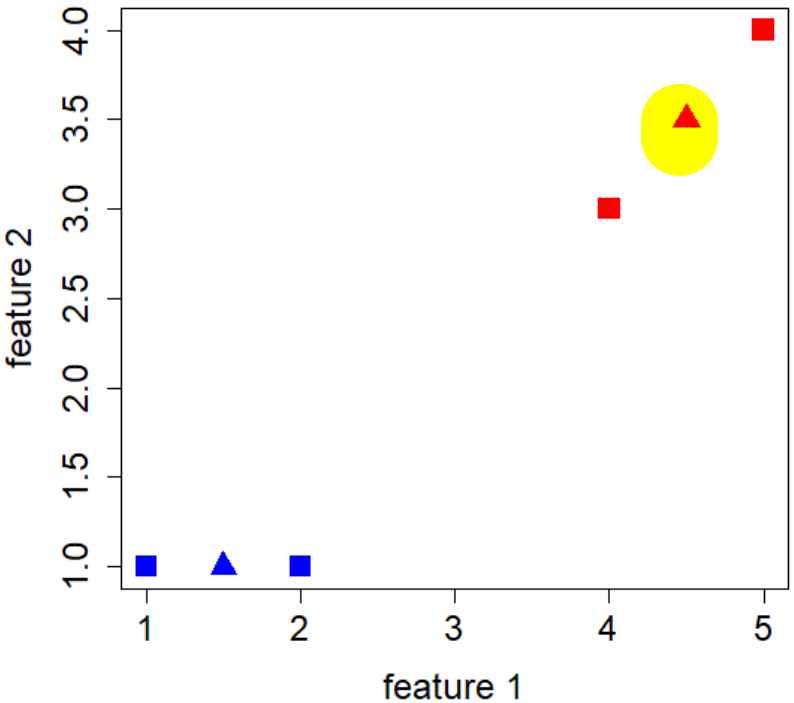
- Distance from fourth point to blue centroid: $\sqrt{(5-2)^2+(4-1.5)^2} = \sqrt{15.25}$
- Distance from fourth point to red centroid: $\sqrt{(5-3)^2+(4-1.5)^2} = \sqrt{10.25}$
- So the fourth point belongs to red group...

k-means clustering: 1^{st} iteration



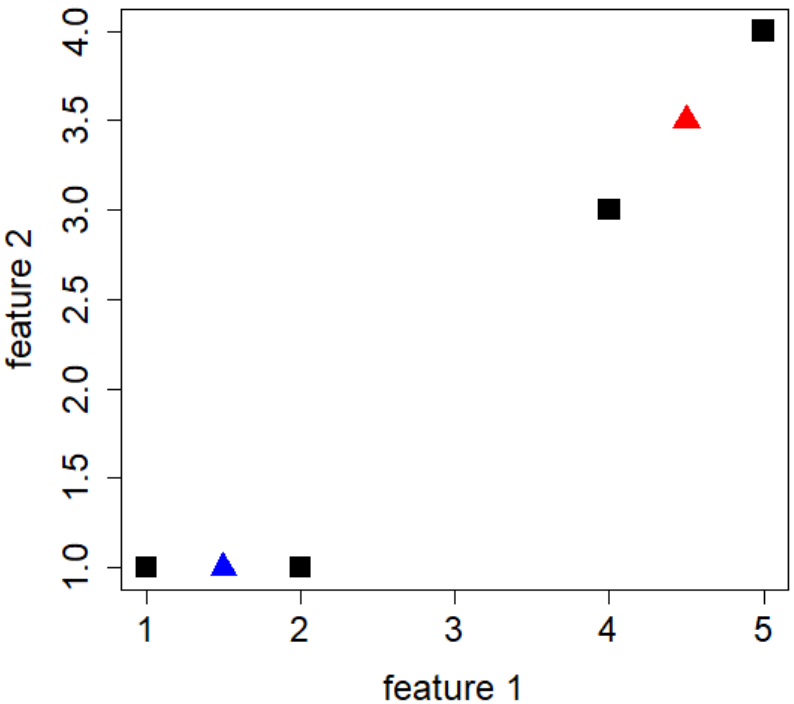
- We have completed the 1^{st} iteration of the $k - means$ algorithm
- Now, we calculate the new centroids in the two clusters.

k-means clustering: 1^{st} iteration



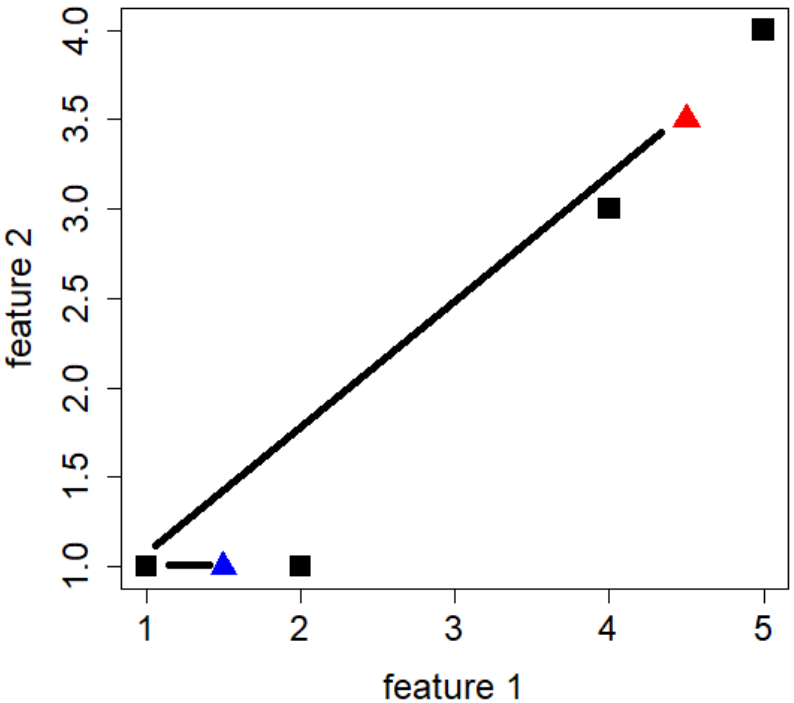
- New blue centroid: $(\frac{1+1}{2}, \frac{2+1}{2}) = (1, 1.5)$
- New red centroid: $(\frac{4+5}{2}, \frac{3+4}{2}) = (4.5, 3.5)$

k-means clustering: start of 2^{nd} iteration



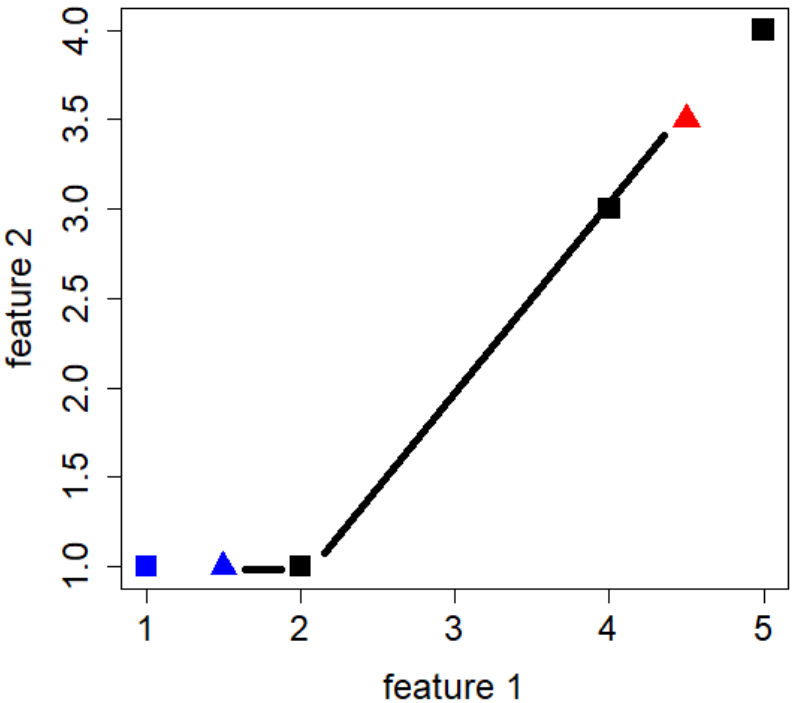
- We move on to the next iteration of the $k - means$ algorithm...

k-means clustering: 2^{nd} iteration



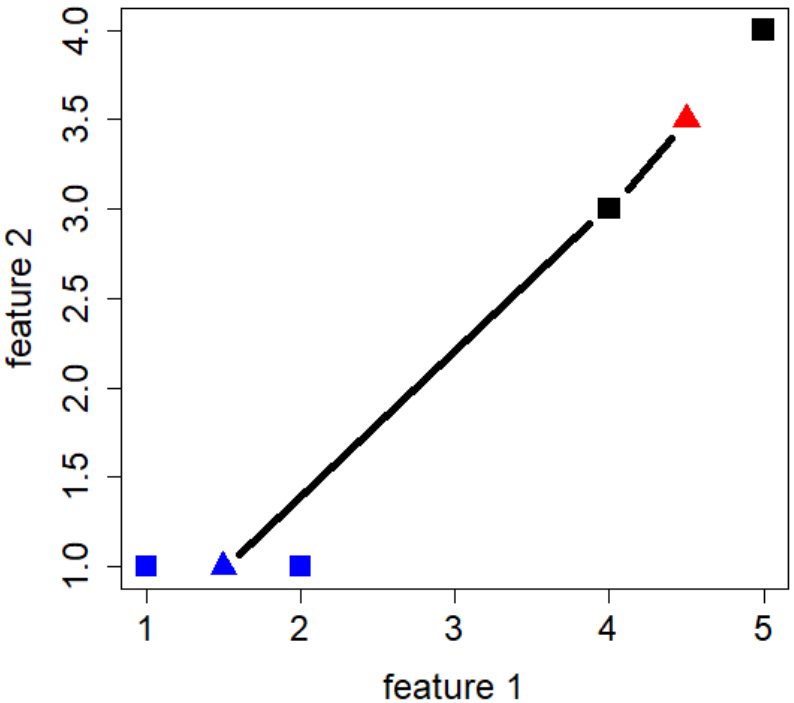
- First point belongs to blue group...

k-means clustering: 2^{nd} iteration



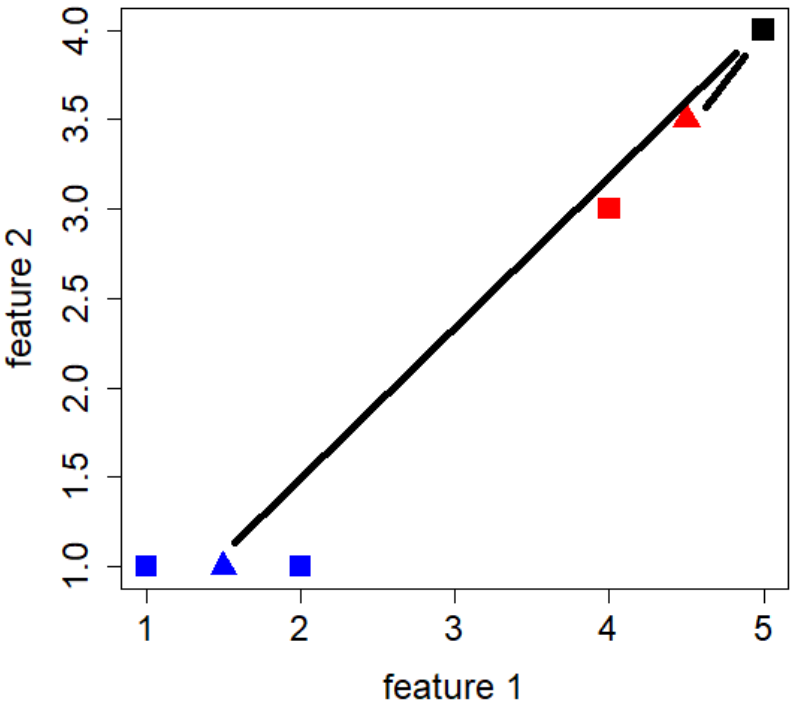
- Second point belongs to blue group...

k-means clustering: 2^{nd} iteration



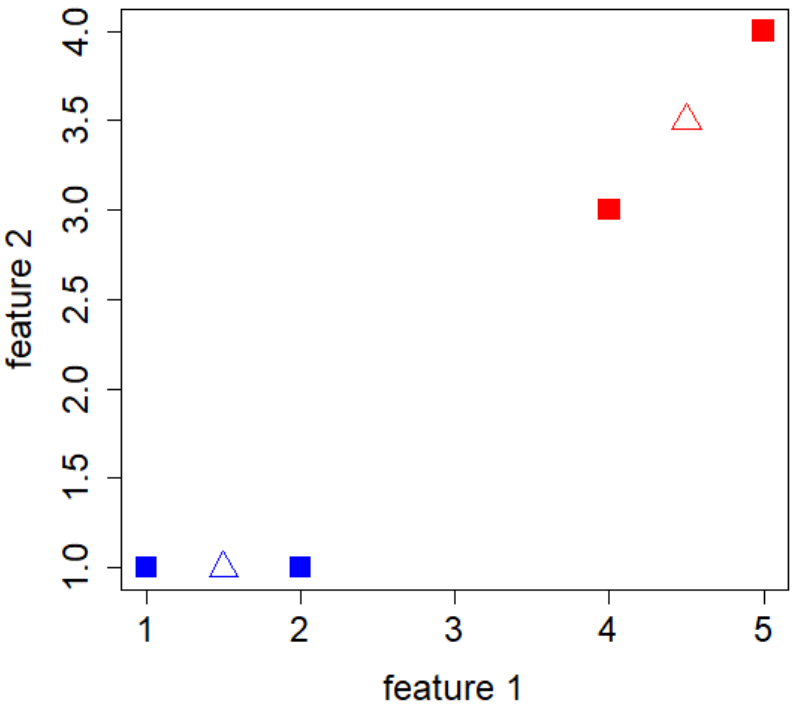
- Third point belongs to red group...

k-means clustering: 2^{nd} iteration



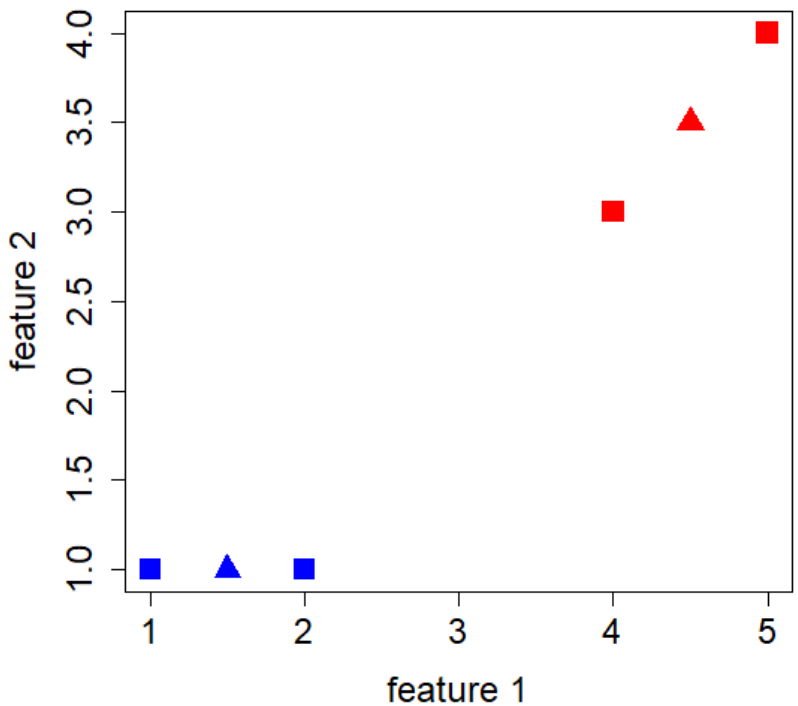
- Fourth point belongs to red group...

k-means clustering: 2^{nd} iteration



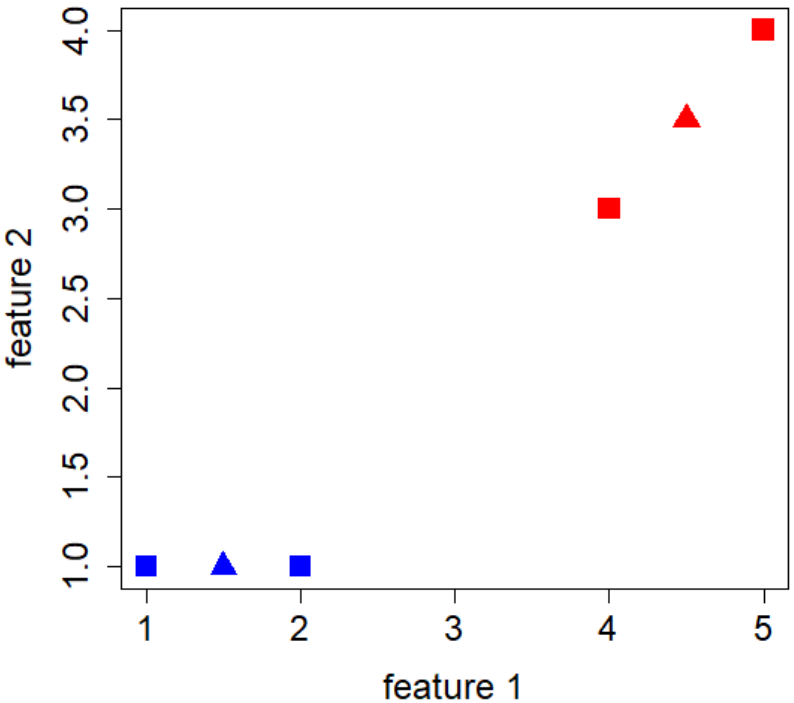
- We have completed the 2^{st} iteration of the $k - means$ algorithm
- Now, we calculate the new centroids in the two clusters.

k-means clustering: 2^{nd} iteration



- New blue centroid: $(\frac{1+1}{2}, \frac{2+1}{2}) = (1, 1.5)$
- New red centroid: $(\frac{4+5}{2}, \frac{3+4}{2}) = (4.5, 3.5)$

k-means clustering: 2^{nd} iteration



- We notice that the computed centroids do not change in their coordinates from the start of the 2^{nd} iteration.
- The algorithm has converged.

k-means clustering: yet another example

- For yet another example of the k-means clustering algorithm in action with two clusters, please watch the Stanford University lecture below, from approximately 2:30 to 6:00 minutes
- Video on k-means clustering algorithm

- The k-means clustering algorithm can be generalized to cluster objects with more than two features.

Example closer at home...

Source: The Straits Times

- Going back to our HDB resale example, let's apply $k - means$ clustering to our data set.

Example closer at home...

- Let us take a look at the data set in `HDBresale_cluster.csv`:

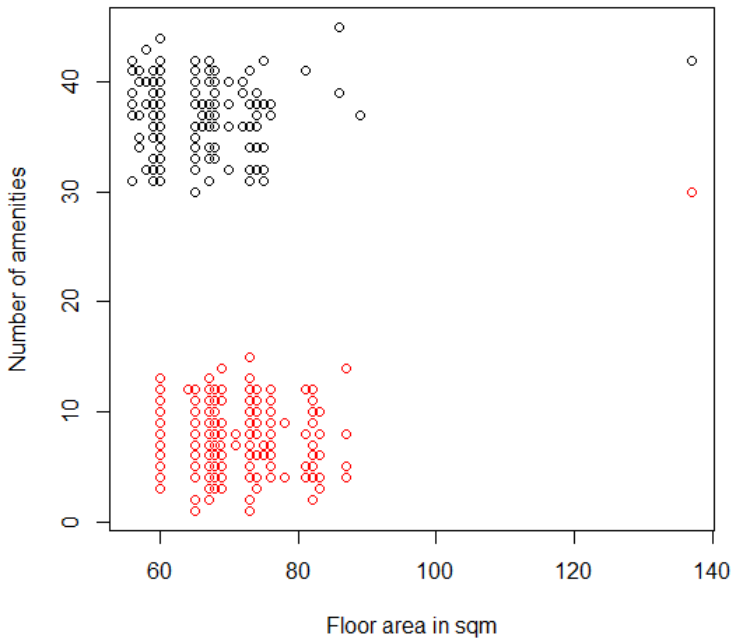
```
> resale=read.csv("HDBresale_cluster.csv")
> head(resale)
     X flat_type floor_area_sqm amenities
1  580         3          ROOM           74         32
2  581         3          ROOM           74         37
3  582         3          ROOM           73         34
4  583         3          ROOM           59         38
5  584         3          ROOM           68         39
6  586         3          ROOM           68         36
```

Example closer at home...

- We can use the `kmeans()` function in R to perform $k - means$ clustering:

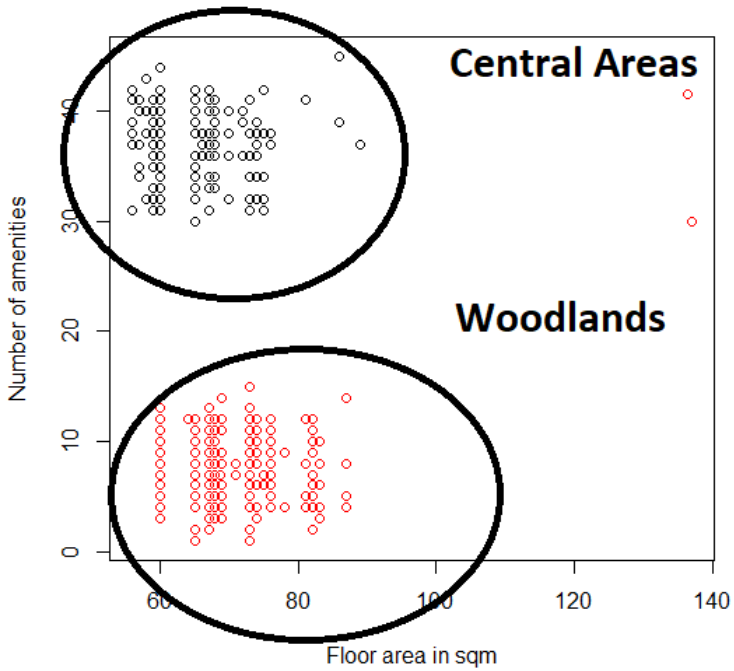
```
1 kout <- kmeans(resale[,c("floor_area_sqm","
      amenities")],
2         centers=2)
3
4 plot(resale$floor_area_sqm,
5       resale$amenities,
6       col=kout$cluster,
7       xlab="Floor area in sqm",
8       ylab="Number of amenities")
```

Example closer at home...



- We see that the results from $k - means$ clustering agree with our initial visual inspection.

Example closer at home...



- In fact, the HDB resale records are sampled from either central area or Woodlands in Singapore.
- We see that $k - means$ clustering has helped us uncover this hidden structure, based on the original unlabeled data.