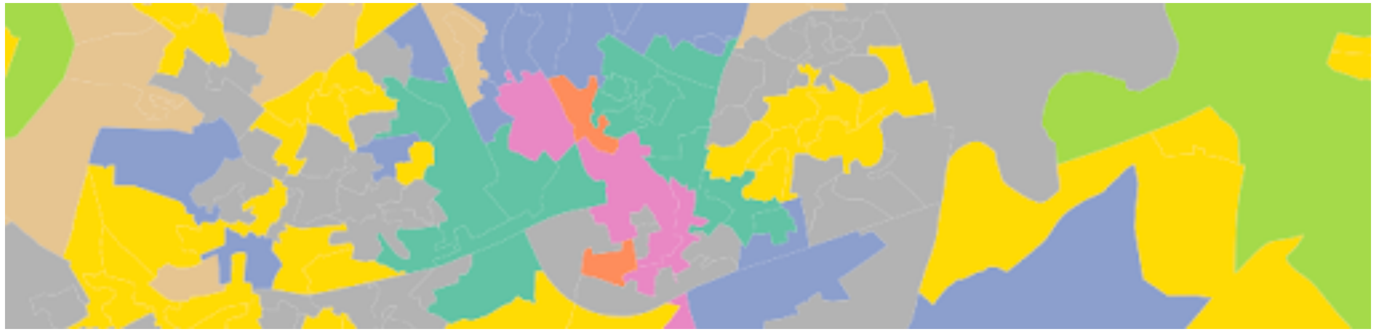
Principles of Spatial Analysis

SHORT LECTURE 02, WEEK 09: K-MEANS CLUSTERING



geodemographics

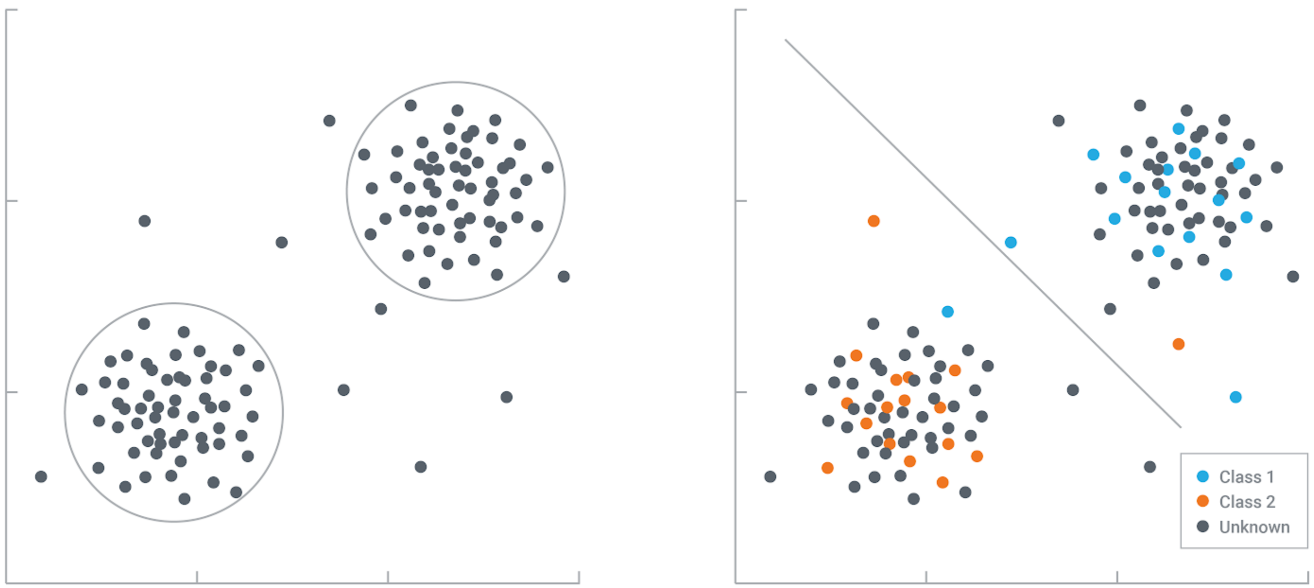
- analysis of people by where they live
- unsupervised machine learning algorithm: k-means clustering

supervised machine learning

- mapping input data to output labels
- logistic regression, naive bayes, support vector machines, artificial neural networks, ensemble methods (such as random forest)

unsupervised machine learning

- clustering, representation learning, density estimation
- k-means clustering, principal component analysis, autoencoders



UNSUPERVISED

SUPERVISED

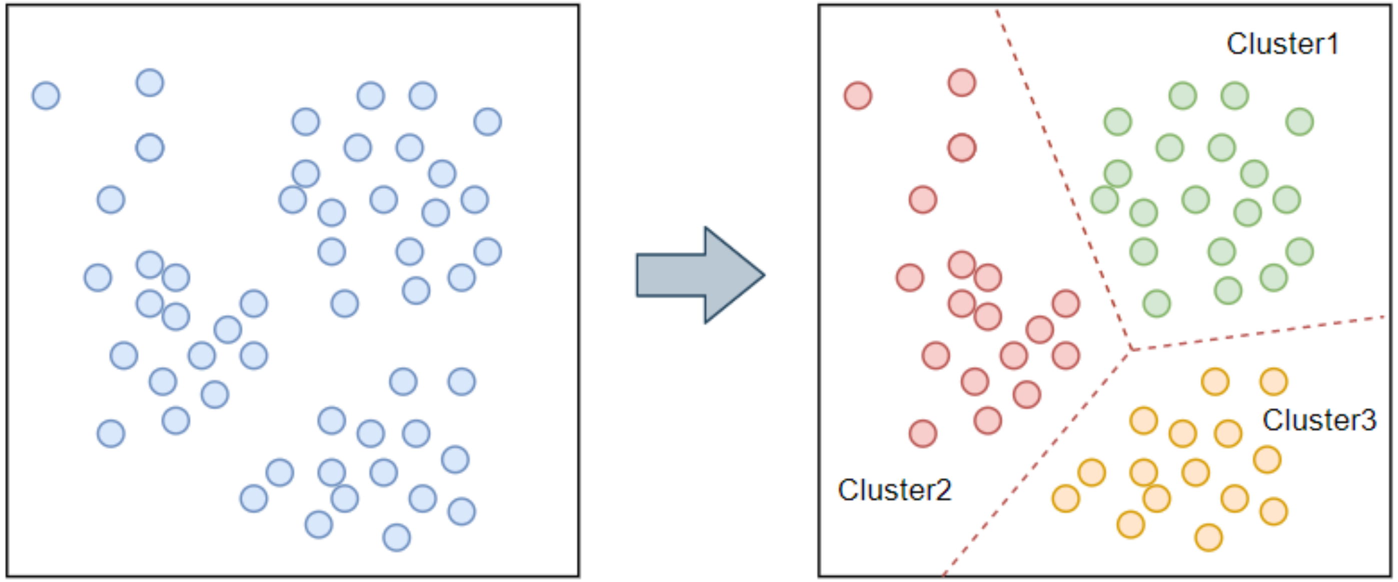
k-means clustering

- assign geographic areas with common underlying attributes to similar classification groups
- Internet User Classification, ONS’ Output Area Classification

k-means clustering

- *k* clusters (pre-defined) of *n* individual observations
- each observation can have any number of attribute data
- choice of data is a balancing act: theory, available data, statistical considerations

Clustering

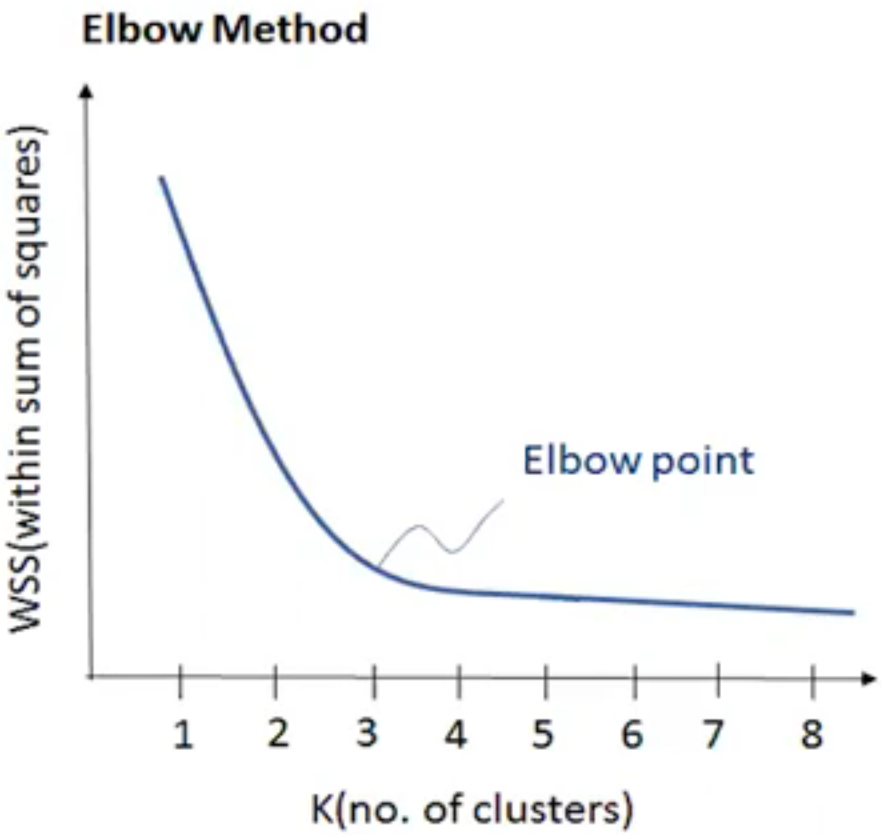


k-means clustering

- minimising the distance between an observations input variables to the means of the respective cluster groups
- maximising the distance between cluster groups
- number of clusters defined *a priori*

number of clusters

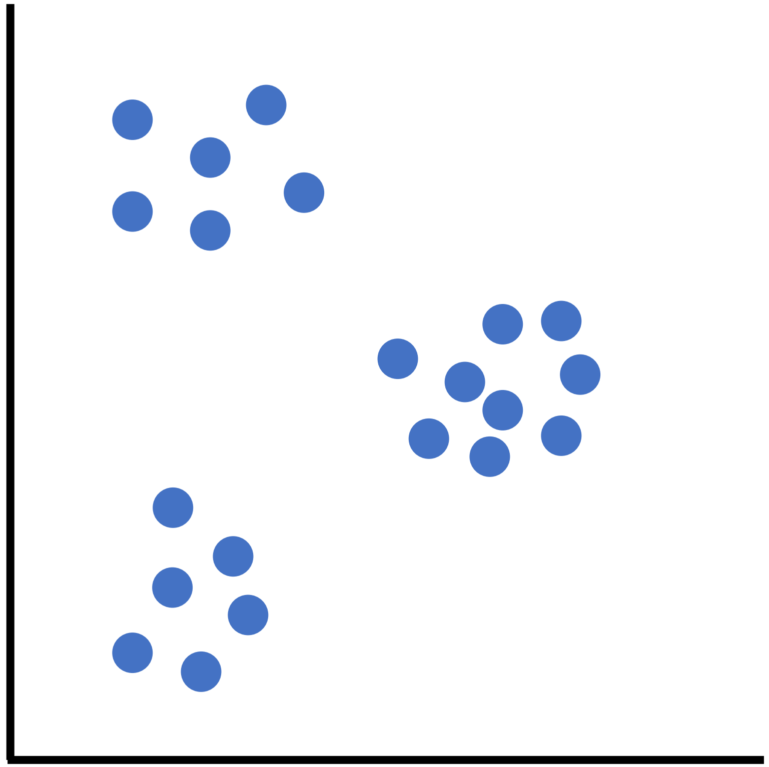
- too few: too much variation within the groups
- too many: overfitting and splitting similar observations
- iterate through the model multiple times
- minimising variation within cluster groups


Elbow Method
WSS(within sum of squares)
Elbow point
1
2
3
4
5
6
7
8
K(no. of clusters)

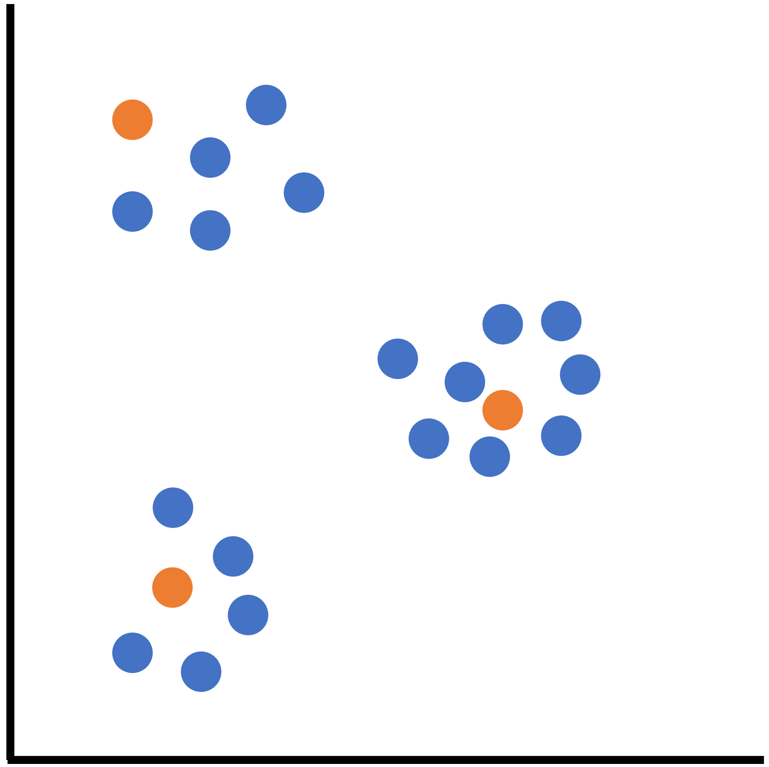
k-means clustering

- Step 1: identify your *k*
- Step 2: randomly identify *k* distinct data points as initial cluster centre
- Step 3: assign each observations to the nearest cluster
- Step 4: calculate the mean of each cluster
- Step 5: repeat with mean value becoming new cluster centre until no change

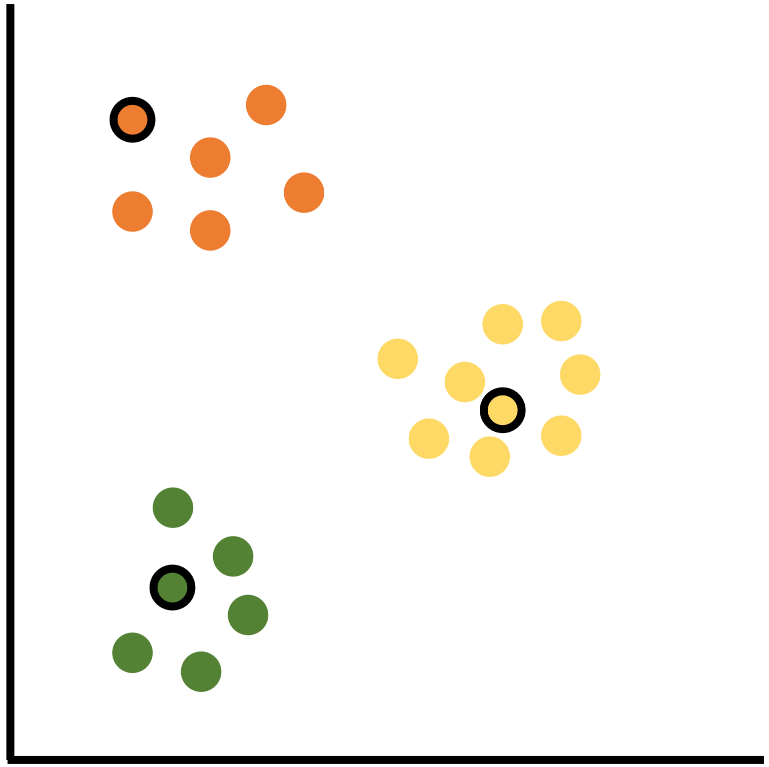
Step 1: identify your *k*



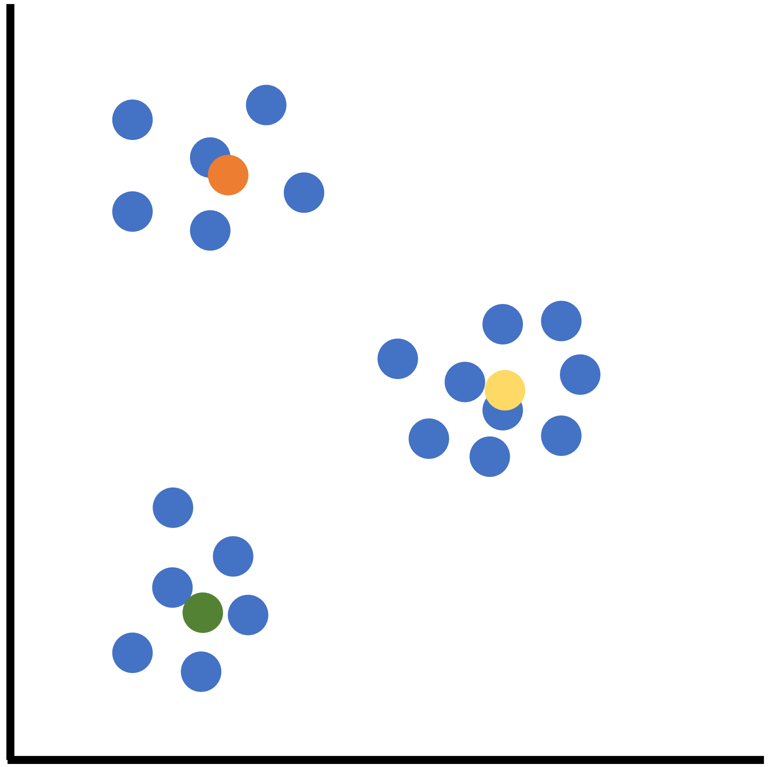
Step 2: randomly identify *k* distinct data points as initial cluster centre



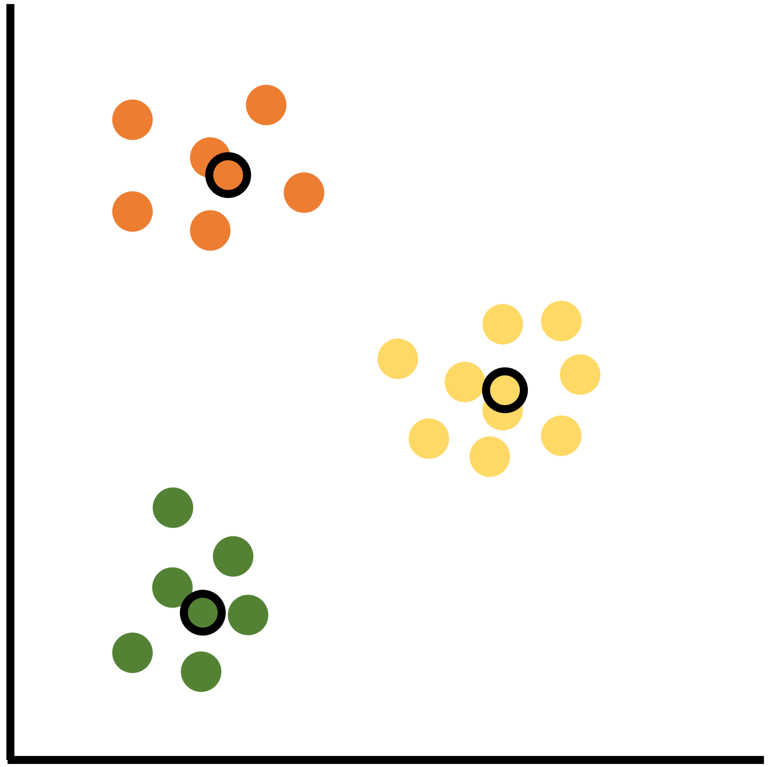
Step 3: assign each observations to the nearest cluster



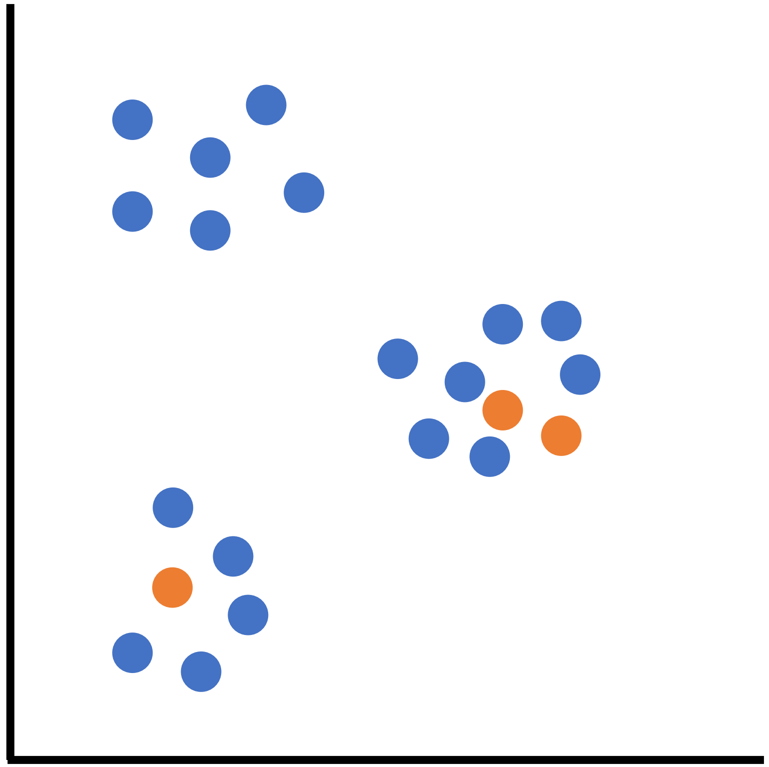
Step 4: calculate the mean of each cluster



Step 5: repeat with mean value becoming new cluster centre until no change



Why run the clustering multiple times?



interpretation of clusters

- look at the means of the input data for each cluster (‘signature’)
- based on the underlying (mean) signature pen portraits can be developed
- not spatial in nature (different to the DBSCAN)
- important to consider collinearity issues

let’s put it into practice