



THE UNIVERSITY OF
MELBOURNE

Workshop 6

COMP20008

Elements of Data Processing
Zijie Xu





Agenda

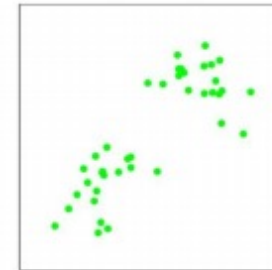
- Unsupervised machine learning
 - Clustering
- Clustering algorithms
 - *k*-means clustering
 - Agglomerative hierarchical clustering
- Principal Component Analysis (PCA)

Unsupervised learning

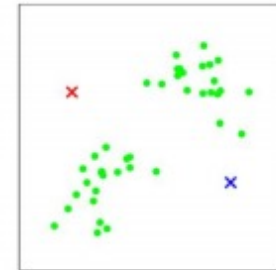
- In supervised learning
 - Labelled data
 - Learns a function that predicts the correct output for unseen input
 - e.g. Classification, regression
- In unsupervised learning
 - Unlabelled data
 - Learns to discover patterns, structures, or relationships within the data without any specific guidance on what to look for
 - e.g. Clustering

k-Means Clustering

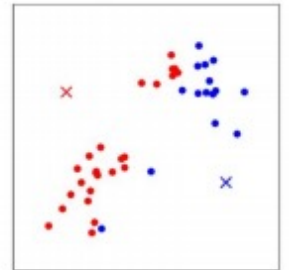
- Assumption: Similar data points belong to same cluster
- Idea: Alternate between
 - Assigning data points to centroids
 - Adjusting centroids based on current assignment
- Data points are categorised by their closet centroid



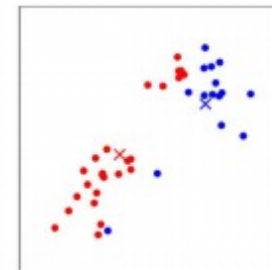
(a)



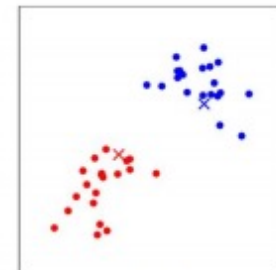
(b)



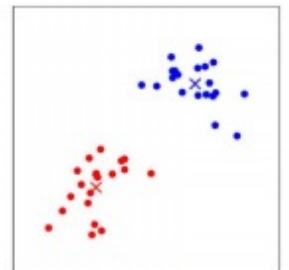
(c)



(d)



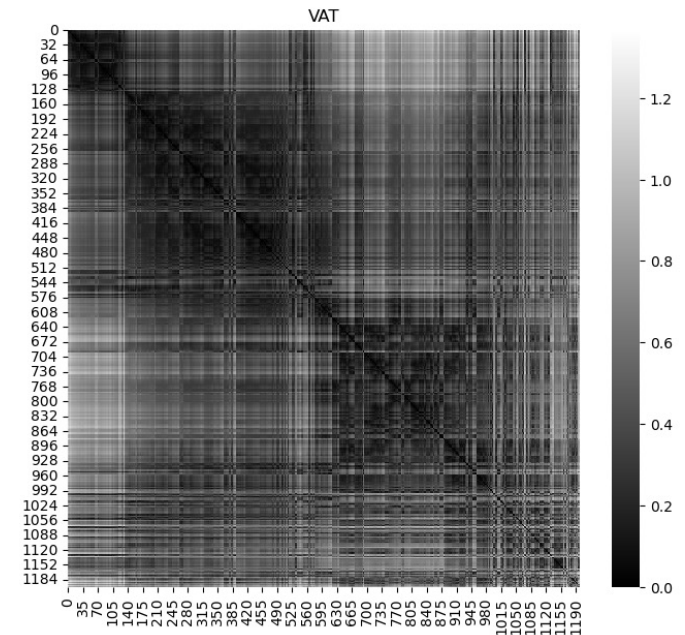
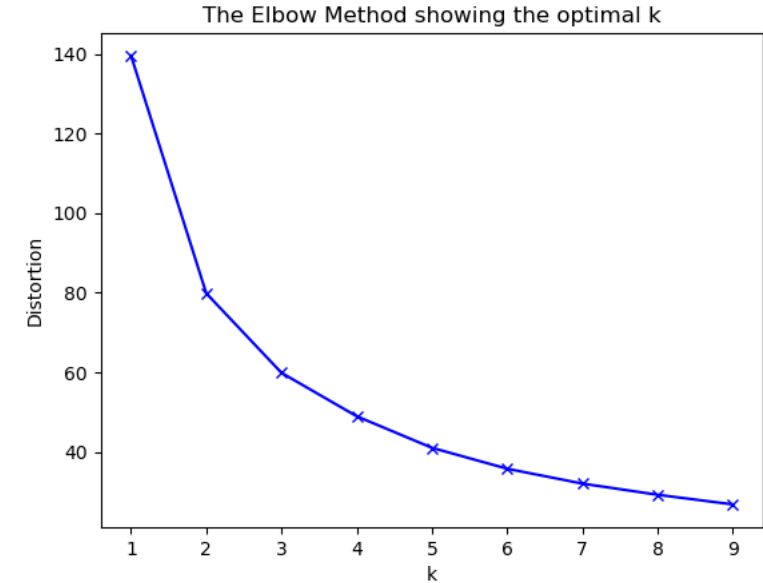
(e)



(f)

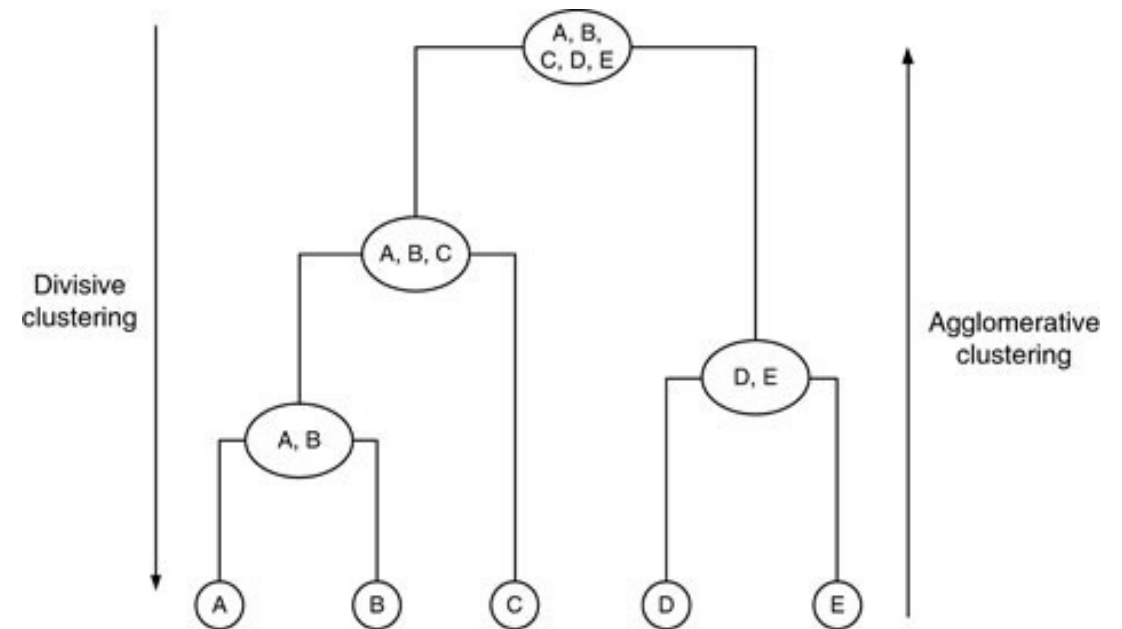
k-Means Clustering

- Need to decide on:
 - Number of clusters k
 - Distance metric (may need normalisation)
- Some methods to help on deciding k
 - Elbow method: Sum of Squared Errors (SSE) for different k
 - VAT: Heatmap on a reordered dissimilarity matrix



Agglomerative hierarchical clustering

- Assumption: There is a hierarchical structure within the data
- Divisive: Top-down
- Agglomerative: Bottom-up
- Idea of agglomerative HC:
 - Merge two closest cluster
 - Update dissimilarity matrix
- Data are put into a dendrogram



Agglomerative hierarchical clustering

- Need to decide on
 - Distance metric between clusters
 - Examples:
 - Single linkage: two most similar parts (closest points)
 - Complete linkage: two least similar parts (furthest points)
 - Average linkage: centre of clusters

Principal Component Analysis (PCA)

- Too many features can cause problems (curse of dimensionality)
- PCA provides one way to perform dimensionality reduction
- Idea: transforms the original variables into principal components (PC)
 - PCs are linear weighted combinations of the original variables
 - Each PC captures as much variance in data as possible
 - Each PC are linearly uncorrelated with each other



Principal Component Analysis (PCA)

- Once we performed PCA:
 - The first few PC capture most of the variance in the dataset
 - We can discard some higher order PC as they don't tell us much about the data, thereby reducing the dimension of the data
- E.g. Reducing 10-D data into 2D/3D allows for visualisation



THE UNIVERSITY OF
MELBOURNE

Thank you

More Resources: Canvas