ACTL3142 Week 10 - Unsupervised Learning

Tadhg Xu-Glassop

2025T2

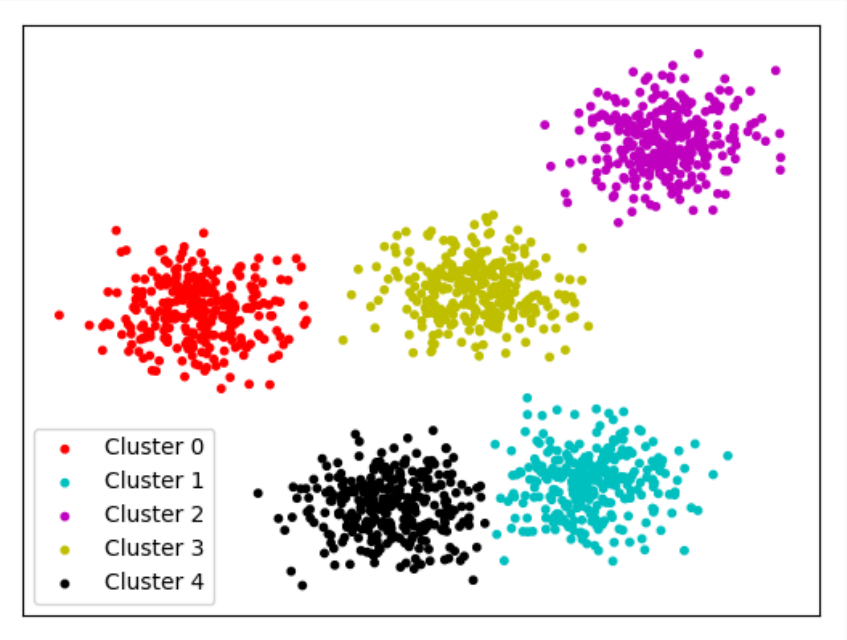
Unsupervised Learning

Unsupervised learning differentiates itself from supervised learning by not having a target variable Y; instead of finding relationships between X and Y, we just try to find meaningful relationships and *simplify* X.

Clustering

Clustering

Clustering aims to group the observations into "clusters" - geometric groups where observations belonging to the same class have some meaningful similarities.



K-means Clustering

K-means clustering

For some predetermined number of clusters K, sort the data into K clusters.
The resulting clusters should minimise the Within-Cluster Variation,

$$W(C_k) = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \overline{x}_{kj})^2.$$

- Found by repeatedly running the K-means algorithm, where we randomly assign centroids and recompute the centroid until convergence (see website).
- K can be chosen by considering many reasons, both practical and statistical.
    - Select K where improvement in WCV plateaus;
    - Want to specifically segment data into K groups.

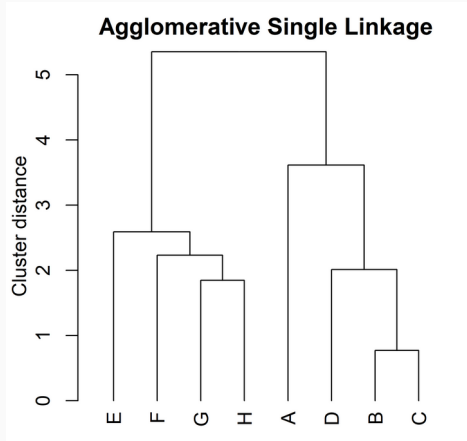
Hierarchical Clustering

Heirarchical Clustering Method (see website for demo!)

For some distance metrics;

1. Start with all observations as their own clusters.
2. Find the pairwise distance between all clusters, and join the two clusters with the least distance.
3. Repeat 2 until everything is clustered together.

Useful when you don't have a value of K in mind and want to see the implications of a different number of clusters. The above can be visualised with a **dendogram**.



Notes on Heirarchical Clustering

There are two choices to be made before:

Choice of distance metric

- We could pick any distance metric to calculate the dissimilarity between observations, and there is no best one!

Linkage (how to find distance between clusters)

- How do we measure the distance between two groups of observations? We could;
  - **Complete** - take the largest difference;
  - **Single** - take the smallest difference;
  - **Average** - take the mean of all distances;
  - **Centroid** - take the distance between centroids.

Ideally, a good clustering should be robust to these choices.

Dimension Reduction

Suppose p is large. Then training models could be very slow !

If only there was a way to reduce p, perhaps projecting our predictor space onto a lesser dimensional subspace could be beneficial...

PCA

Principal Component Analysis

Let $X_1, X_2, \ldots, X_p$ be the columns of our data matrix. Then, the ith principal component for $1 \leq i \leq p$ is

$$Z_i = \phi_{i1}X_1 + \phi_{i2}X_2 + \cdots + \phi_{ip}X_p, \qquad \sum_{j=1}^{p} \phi_{ij}^2 = 1.$$

Each Z_i aims to capture the maximum possible variance.

Geometrically, begin with an empty basis $B = \{\}$.

1. What vector can we add to B that minimises the distance between $\mathrm{span}(B)$ and our observations ? We find it, then add it to B.
2. We repeat this p times. the ith vector we added is the ith principal component.

Notes on PCA

After we find our PC's, we can perform dimension reduction by projecting our data onto $B_k = \text{span}\{Z_1, Z_2, \ldots Z_k\}$.

The choice of k can be found by choosing when additional PCA's don't explain much more variance, or by assessing the performance of an ML model using B_k with cross-validation for different values of k.