ACTL3142

Week 10: Unsupervised Learning

Tetian Madfouni

t.madfouni@unsw.edu.au

Announcements

MyExperience (I think you have a tiny bit of time left)

Congrats on finishing the assignment, next time don’t rely on an extension ☺

Recap: Unsupervised Learning

What is it?

- Instead of (X, y) we now just have X and want to discover interesting relationships or things about X (subgroups normally)

The next few slides cover the 3 types of unsupervised learning covered in the lecture, who can name the first?

Recap: K-Means Clustering

How does it work?

- Effectively say there are K clusters (groups) in the data, and try to find which observations should belong to each $C_1, \dots, C_k$ in order to minimise the “within-cluster variation”. Most common choice of function for within-cluster variation is

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} W(C_k). \qquad W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad W(C_k) = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

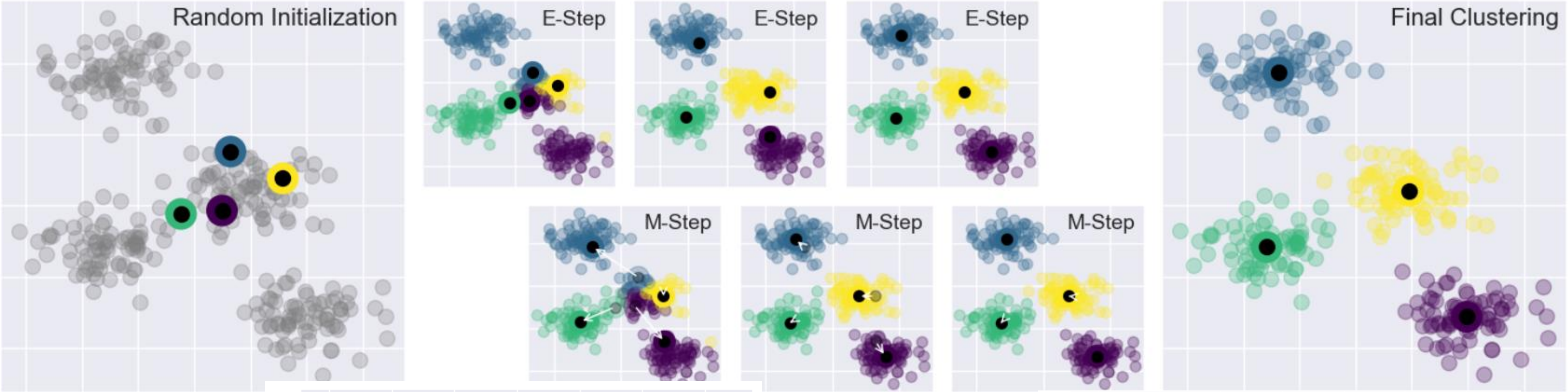
How does the algorithm work?

- Initialise K centroids (middle of each cluster). Assign each observation to the closest cluster and then move the centroids to be the centre of all observations assigned to them.

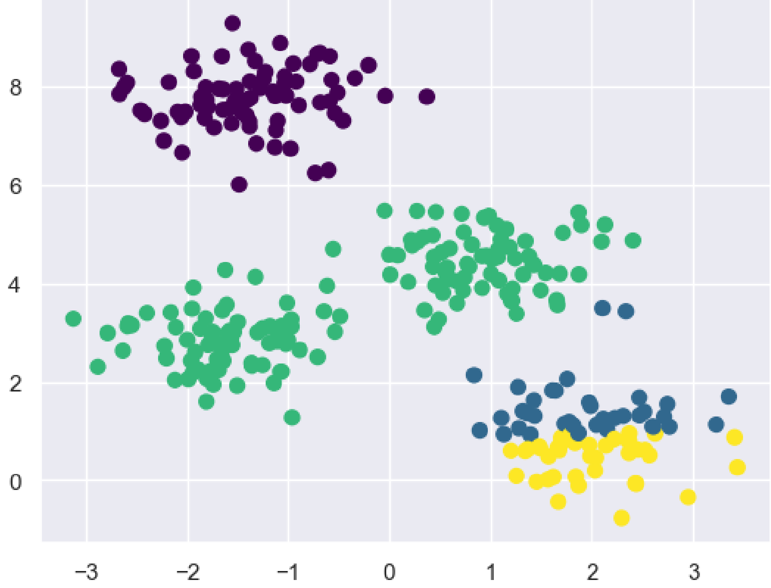
What are a couple of its biggest problems?

- Must define K in advance (there are other algorithms like DBSCAN which avoid this) and it is **highly** sensitive to what we initialise the centroids as.

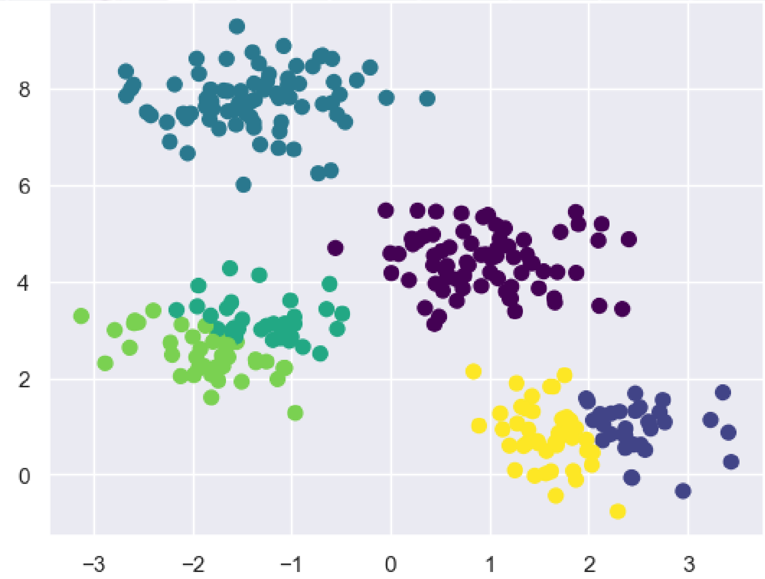
Recap: K-Means Clustering ctd.



Wrong initialisations



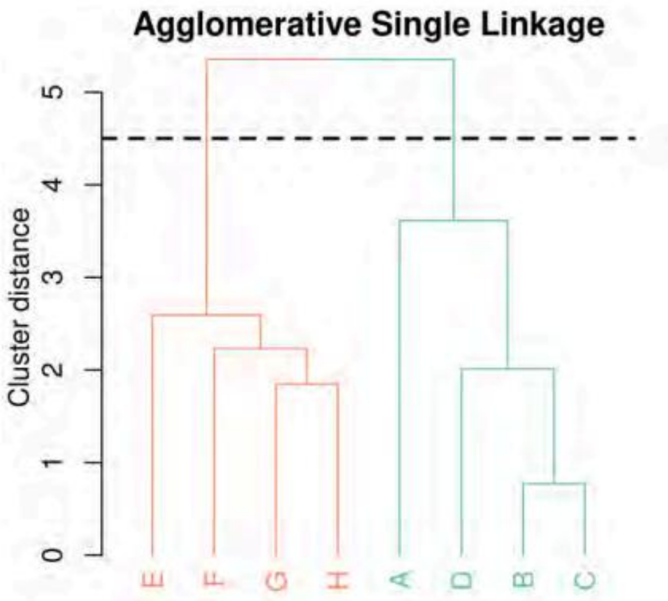
Wrong value of K

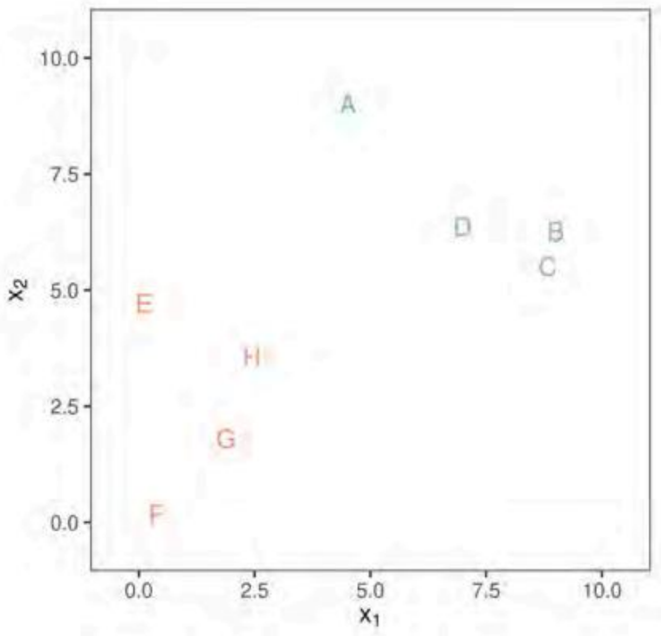


Recap: Hierarchical Clustering

How does it work?

- There’s technically also top-down hierarchical clustering, but we focus on bottom up. Start with each observation as its own cluster, then compute all pairwise dissimilarity scores between the clusters, and merge the two clusters that have the least dissimilarity.
- There’s various dissimilarity scores you can use (Euclidean distance, Manhattan Distance, etc.)
- To actually compare the groups, there’s 4 methods, they’re explained pretty clearly on the slides





Recap: Clustering

In general, there’s a decent number of decisions to be made for both of these

- K-means needs the initialisation and number of clusters
- Hierarchical needs a dissimilarity measure, type of linkage (how we compare groups), and a cut-off point

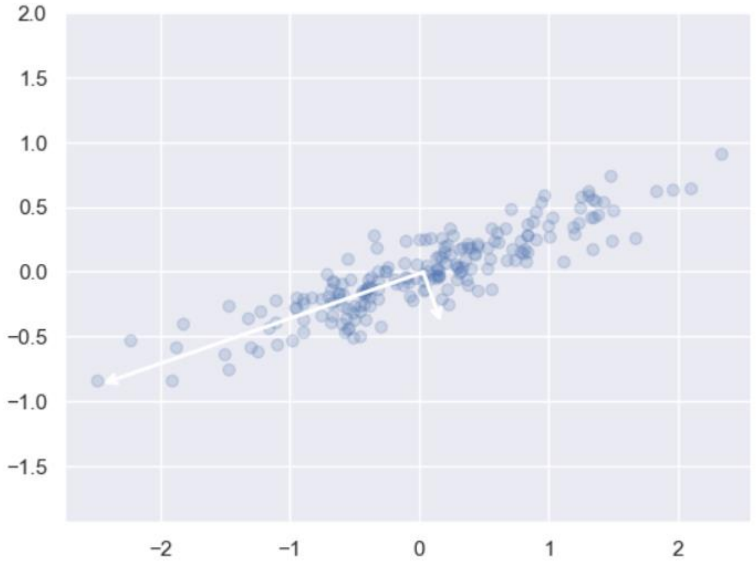
Always use many methods and validate the clusters they create against each other. If specific clusters come up heaps, they’re probably reasonable

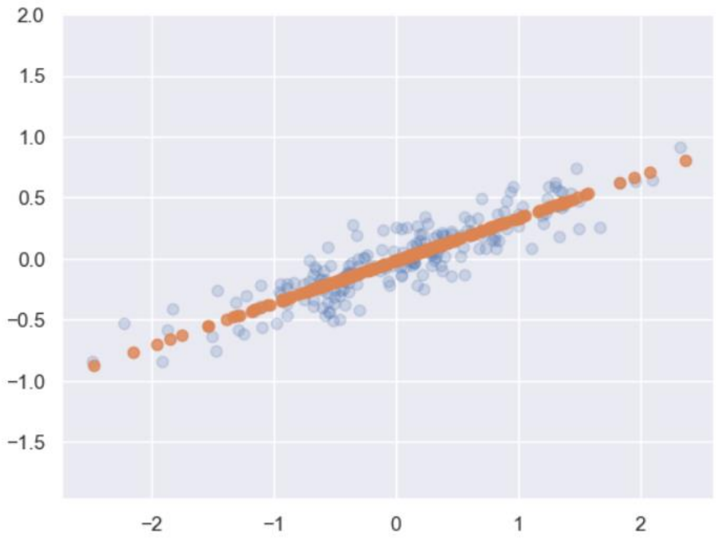
Also google clustering methods, there’s many more (much more advanced) methods for clustering.

Recap: Dimension Reduction

What's the method that was introduced in the lecture for dimension reduction and how does it work?

- Principal Components Analysis. Sequentially finds principal components (linear combination of predictors) that capture the maximum variance in the data set. Note that each successive PC must be fully uncorrelated from all the ones before it (orthogonal direction). We keep adding PCs until the number of PCs matches the number of original dimensions.
- This effectively allows us to capture a lot of information in a data set, while reducing its dimensionality and complexity.

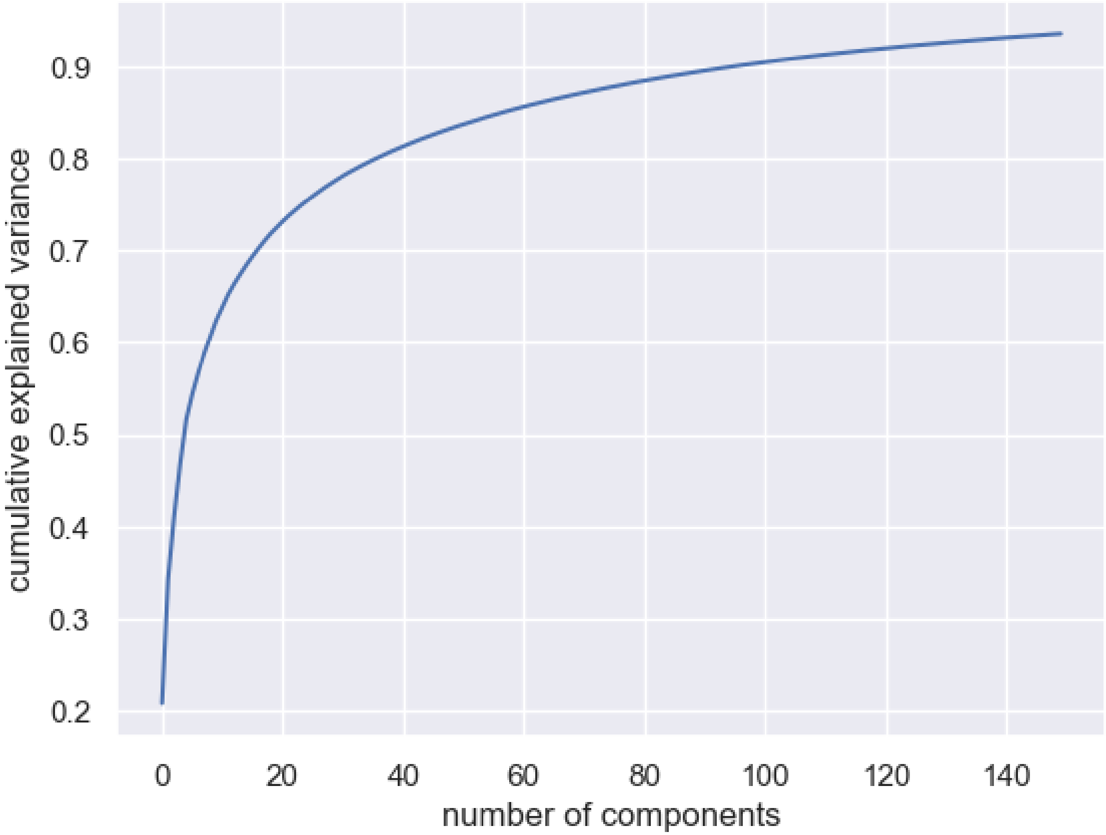




Recap: Dimension Reduction ctd

To select how many principal components we should use, we often consider a cumulative explained variance graph

By discarding the least important directions of variance, we can effectively also use PCA as a noise filtering tool


cumulative explained variance
number of components