Today.

1. Introduction to clustering
2. K-means clustering
3. Hierarchical clustering

1. Introduction to Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other.
- It make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

⇒ Two clustering methods:

① K-means clustering — seek to partition the observations into a pre-specified # of clusters.

② Hierarchical clustering — do not know how many clusters we want in advance; in fact, we end up with a tree like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.

2. K-means clustering

- Let $C_1 \dots C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

① $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.

② $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping; no observation belongs to more than one cluster.

i.e. if the ith observation is in the kth cluster, then $i \in C_k$.

- The idea behind the K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within cluster variation for cluster C_k is a measure $WCV(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we try to solve the problem

$$\underset{C_1 \cdots C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} WCV(C_k) \right\} \quad (*)$$

- In words, this formula says that we want to partition the observations into K clusters such that the overall within-cluster variation, summed over all K clusters, is as small as possible.

* How to define within-cluster variation?

- Typically, we use Euclidean distance

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \quad (**)$$

Where $|C_k|$ denotes the # of observations in the kth cluster.

⇒ Combine $(*)$ and $(**)$, the optimization problem that defines K-means clustering

$$\underset{C_1 \cdots C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \quad (***)$$

* K-means algorithm

① Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations. (K-means, hierarchical clustering)

② Iterate until the cluster assignments stop changing:

(2.1) For each of the K-clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

(2.2) Assign each observation to the cluster whose centroid is closest (where "closest" is defined using Euclidean distance)

* properties of the algorithm

① This algorithm is guaranteed to decrease the value of objective (***) at each step why?

② Note: $$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$

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

where $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i \in C_k} x_{ij}$ is the mean for feature j in cluster C_k

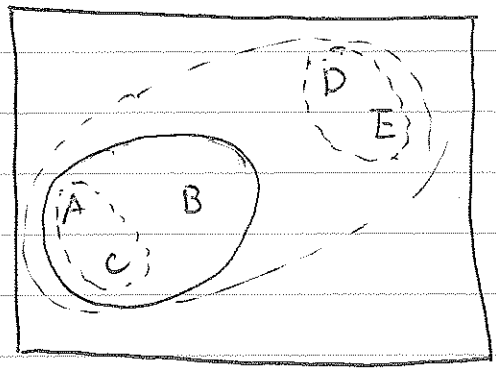
③ However, it is not guaranteed to give the global minimum. (different starting points lead to different local optimum)

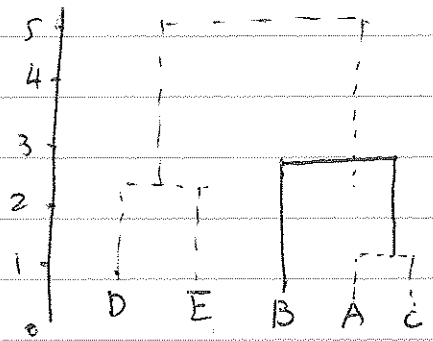
3. Hierarchical clustering

- Instead of pre-specifying K, hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- Here, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

* Hierarchical clustering: the idea

Dendrogram





Algorithm

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat
- Ends when all points in one cluster.

* Type of linkage

Complete = $\max \sum_{i \in C_k, i' \in C_{k'}} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ (maximum inter-cluster dissimilarity)

Single : $\min \sum_{i \in C_k, i' \in C_{k'}} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ (minimum inter-cluster dissimilarity)

Average : $\frac{1}{|C_k| \times |C_{k'}|} \sum_{i \in C_k, i' \in C_{k'}} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$ (average inter-cluster similarity)

Centroid : $\sum_{j=1}^{p} (\bar{x}_{ij} - \bar{x}_{i'j})^2$ (the dissimilarity between the centroid from cluster of i & i')

* Choice of dissimilarity measures
  - So far use Euclidean distance
  - An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.

* Notes : ① Scaling matters : should the observations or features first be standardized in some way? For instance, maybe the variables should be centred to have mean zero and scaled to have standard deviation one.

② what dissimilarity measure should we use? (Hierarchical)

Euclidean distance $h(x, x') = \left[\sum_{j=1}^{p}(x_j - x'_j)^2\right]$ Manhattan distance $h(x, x') = \sum_{j=1}^{p}|x_j - x'_j|$

③ what type of linkage should be used? (Hierarchical)

④ How many clusters to choose? (K-means + hierarchical clustering)

4. Summary

- unsupervised learning is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning.
- It is intrinsically more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective. (like the test set accuracy).
- It is an active field of research, with many recently developed tools such as self-organizing maps, independent components analysis and spectral clustering.