Finding groups of data with Clustering

Module 6

Clustering

- Clustering is an **unsupervised** machine learning task that automatically divides the data into clusters, or groupings of similar items. It does this without having been told what the groups should look like ahead of time. As we may not even know what we're looking for, clustering is used for knowledge discovery rather than prediction. It provides an insight into the natural groupings found within data.
- Clustering is guided by the principle that records inside a cluster should be very similar to each other, but very different from those outside.

Sample use cases of clustering

- The resulting clusters can then be used for action.
- For instance, you might find clustering methods employed in applications such as: Segmenting customers into groups with similar demographics or buying patterns for targeted marketing campaigns and/ or detailed analysis of purchasing behavior by subgroup
- Detecting anomalous behavior, such as unauthorized intrusions into computer networks, by identifying patterns of use falling outside known clusters
- Simplifying extremely large datasets by grouping a large number of features with similar values into a much smaller number of homogeneous categories

K-Means clustering approach

1. Choose a value of k
2. Select k objects in an arbitrary fashion. Use these as the initial set of k centroids
3. Assign each of the objects to the cluster for which it is nearest to the centroid
4. Recalculate the centroids of the k clusters
5. Repeat steps 3 and 4 until the centroids no longer move

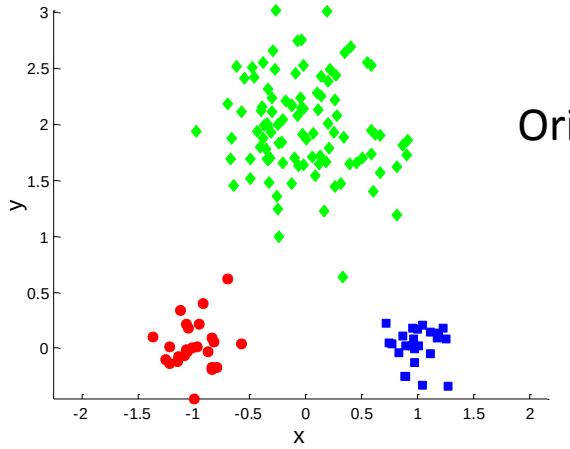
Important design issue in K-Means approach-1/2

- How many clusters (K value) used ?
- How to assign the initial centroid points ?

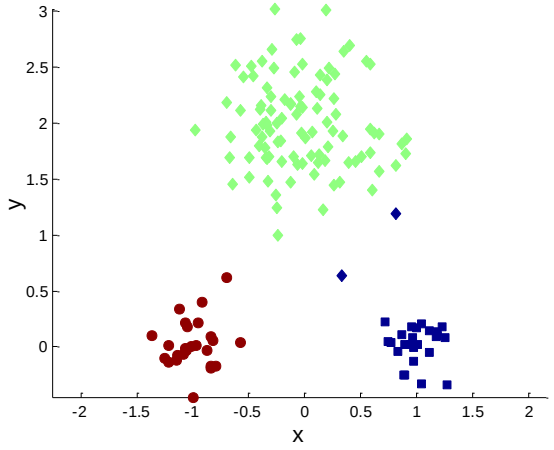
Important design issue in K-Means approach-2/2

- We learned that the algorithm can be sensitive to randomly chosen cluster centers. Indeed, if we had selected a different combination of three starting points in the previous example, we may have found clusters that split the data differently from what we had expected.
- Choosing the number of clusters requires a delicate balance. Setting the k to be very large will improve the homogeneity of the clusters, and at the same time, it risks overfitting the data.
- Ideally, you will have some a priori knowledge (that is, a prior belief) about the true groupings, and you can begin applying k-means using this information.

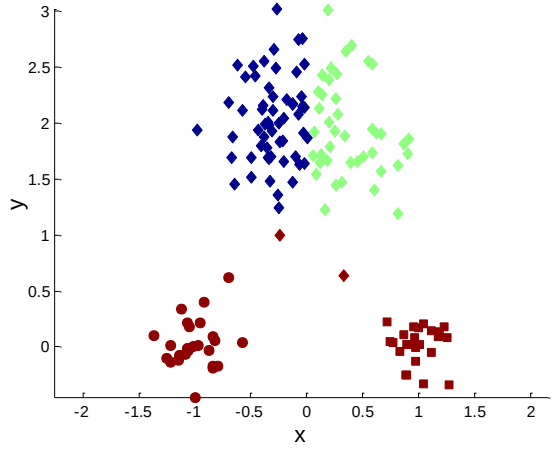
Two different K-means Clusterings



Original Points

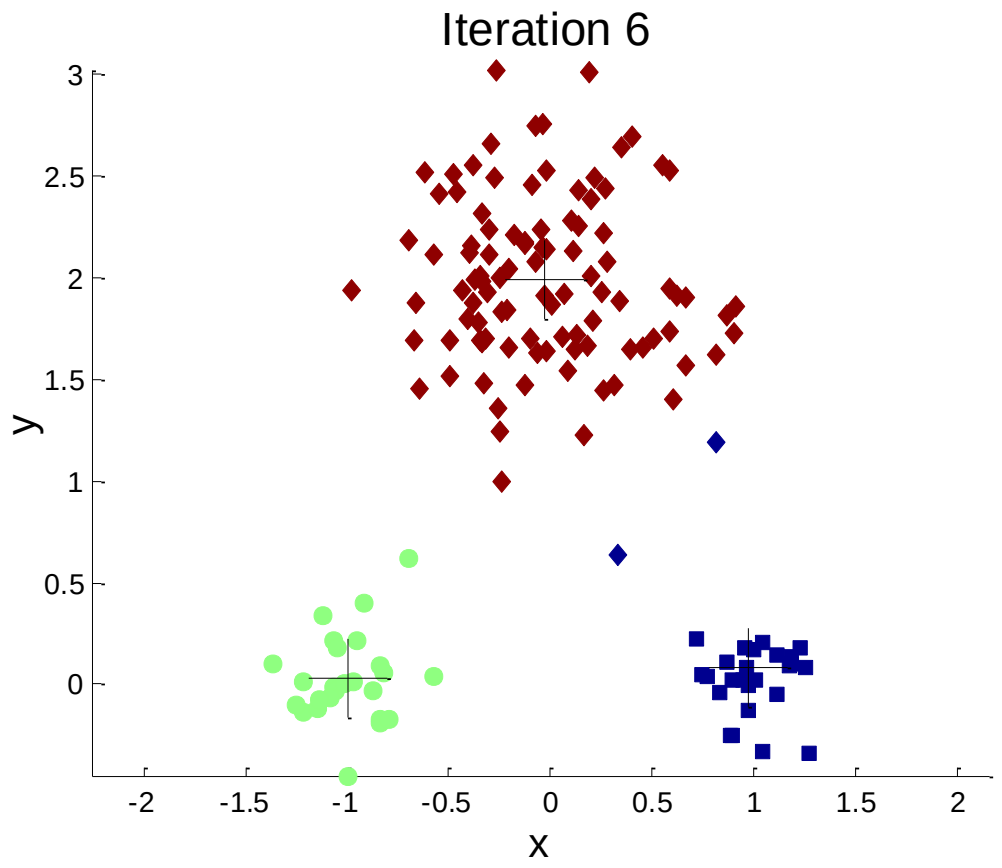


Optimal Clustering

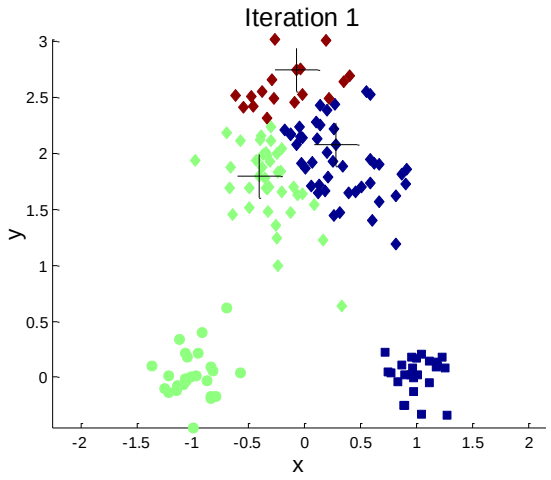


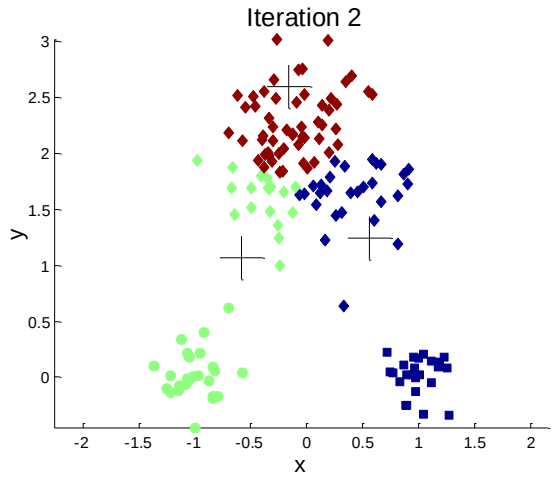
Sub-optimal Clustering

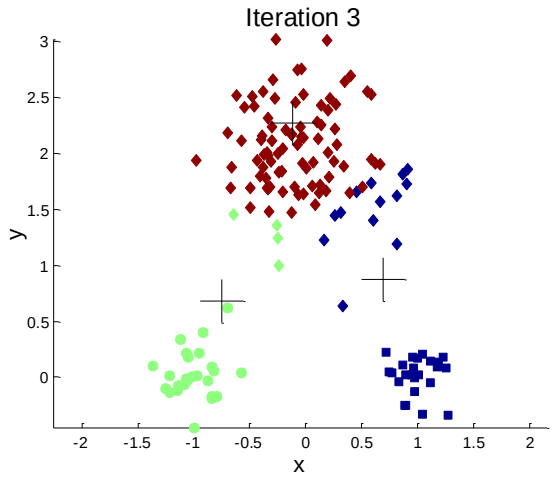
Importance of Choosing Initial Centroids

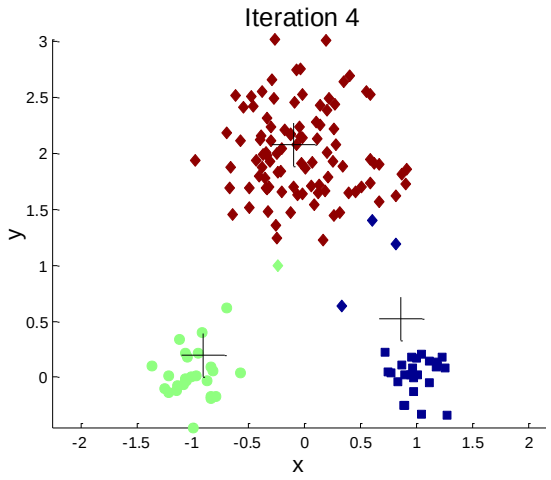


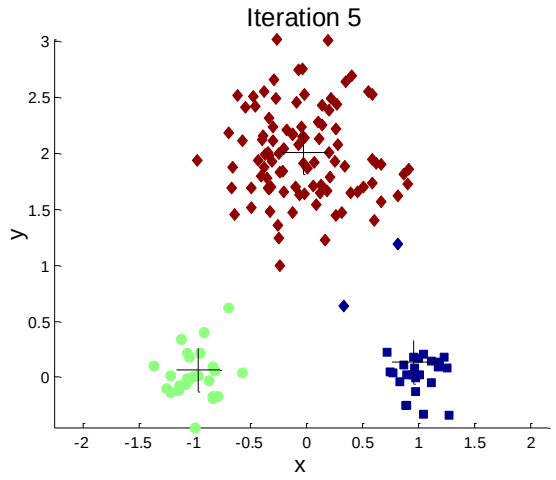
Importance of Choosing Initial Centroids

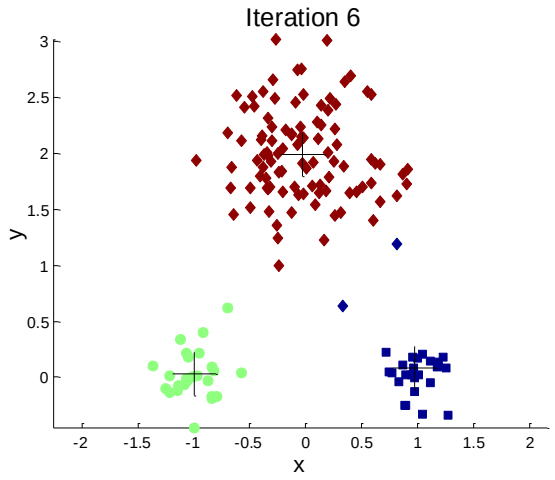










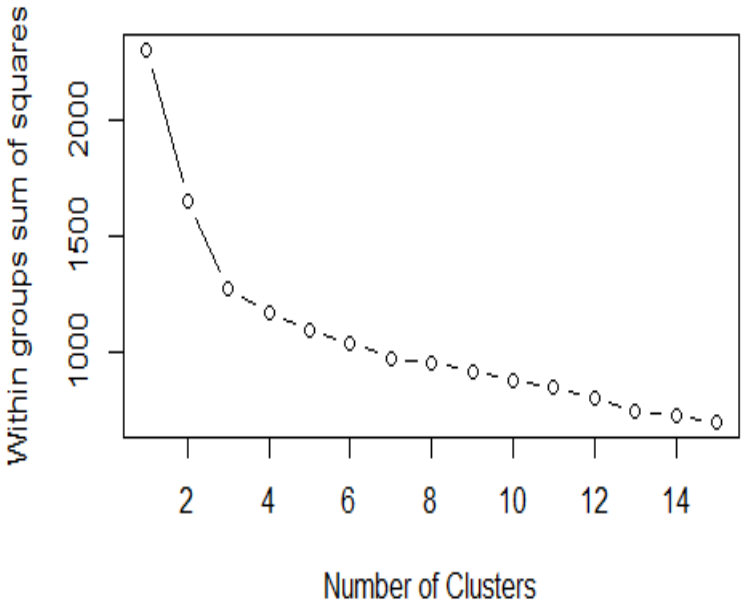


Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - *x* is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K



Exercise

Finding teen market segments using k-means clustering

- Interacting with friends on social networking sites such as Facebook has become a rite of passage for teenagers around the world. Having a relatively large amount of disposable income, these adolescents are a coveted demographic for businesses hoping to sell snacks, beverages, electronics, and hygiene products.
- The many millions of teenage consumers browsing such sites have attracted the attention of marketers struggling to find an edge in an increasingly competitive market. One way to gain this edge is to identify segments of teenagers who share similar tastes, so that clients can avoid targeting advertisements to teens with no interest in the product being sold. For instance, a sports beverage is likely to be a difficult sell to teens with no interest in sports.