K-means Clustering

Rebecca C. Steorts, Duke University

STA 325, Chapter 10 ISL

Agenda

- Clustering
- K-means clustering

Clustering

What is clustering?

Clustering

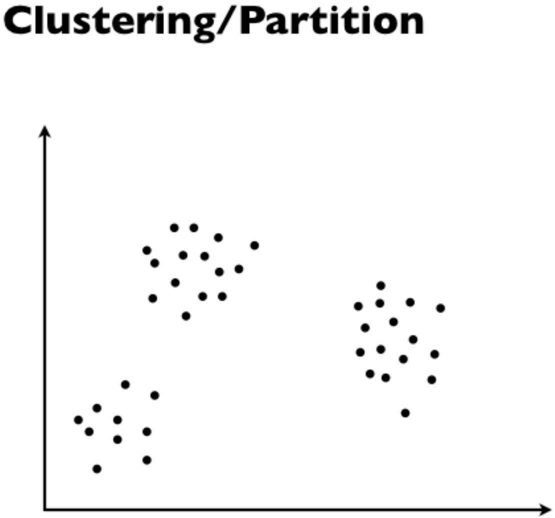


Figure 1: default

Clustering

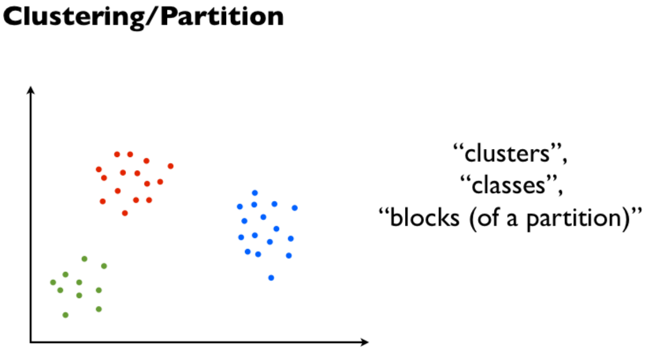


Figure 2: default

Clustering

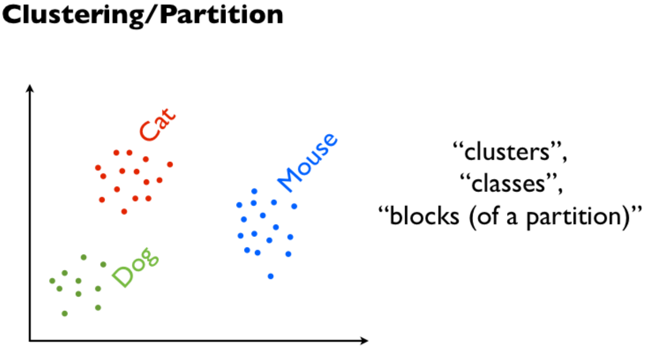


Figure 3: default

K-means clustering algorithm

- K-means clustering: simple approach for partitioning a dataset into K distinct, non-overlapping clusters.
    1. To perform K-means clustering: specify the desired number of clusters K.
    2. Then the K-means algorithm will assign each observation to exactly one of the K clusters.
- Figure 4: results obtained from performing K-means clustering on a simulated example, using $K = 2, 3, 4$.

K-means clustering algorithm

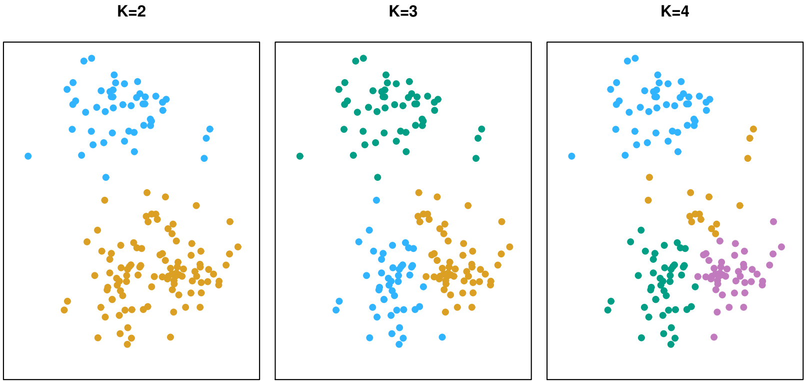


Figure 4: 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K. The cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.

K-means clustering algorithm

Given the number of clusters k and data vectors $\vec{x}_1, \vec{x}_2, \ldots \vec{x}_n$,

1. Randomly assign vectors to clusters
2. Until nothing changes
   a Find the mean of each cluster, given the current assignments
   b Assign each point to the cluster with the nearest mean

There are many small variants of this.

- For instance, the R function `kmeans()` randomly chooses k vectors as the initial cluster centers
- Instead of randomly assigning all the vectors to clusters at the start.

K-means clustering algorithm

- The mean of $x_1, x_2, \ldots x_n$, is

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- But it is also true that $\overline{x} = \arg\min_m \sum_{i=1}^{n} (x_i - m)^2$
- Property extends to vectors:

$$\frac{1}{n}\sum_{i=1}^{n} \vec{x}_i = \arg\min_{\vec{m}} \sum_{i=1}^{n} \|\vec{x}_i - \vec{m}\|^2.$$

K-means clustering

Define $C_1, C_2, \ldots C_k$ denoting sets containing the indices of the observation of each cluster. That is, each $\vec{x}_i$ is in one and only one C_j.

This means that these sets satisfy two properties:

1. $$C_1 \cup C_2 \cup \cdots C_K = \{1, \ldots, n\}.$$

   This means that each observations belongs to at least one of the K clusters.

2. $$C_k \cap C_{k'}$$

   This means the clusters are non-overlapping and so no observation belongs to more than one cluster.

K-means clustering

Recall $C_1, C_2, \ldots C_k$ denotes sets containing the indices of the observation of each cluster. That is, each $\vec{x}_i$ is in one and only one C_j.

- For each cluster we have a center, $\vec{m}_j$, and a sum of squares,

$$Q_j \equiv \sum_{i:\vec{x}_i \in C_j} \|\vec{x}_i - \vec{m}_j\|^2 = \sum_{i,i' \in C_j} \sum_{k=1}^{p} (x_{ik} - x_{i'k})^2$$

K-means clustering

Recall $C_1, C_2, \ldots C_k$ denotes sets containing the indices of the observation of each cluster. That is, each $\vec{x}_i$ is in one and only one C_j.

- For each cluster we have a center, $\vec{m}_j$, and a sum of squares,

$$Q_j \equiv \sum_{i:\vec{x}_i \in C_j} \|\vec{x}_i - \vec{m}_j\|^2 = \sum_{i,i' \in C_j} \sum_{k=1}^{p} (x_{ik} - x_{i'k})^2$$

- Define $V_j = Q_j / n_j$, n_j is the number of points in cluster j.
  - This is the **within-cluster variance**.

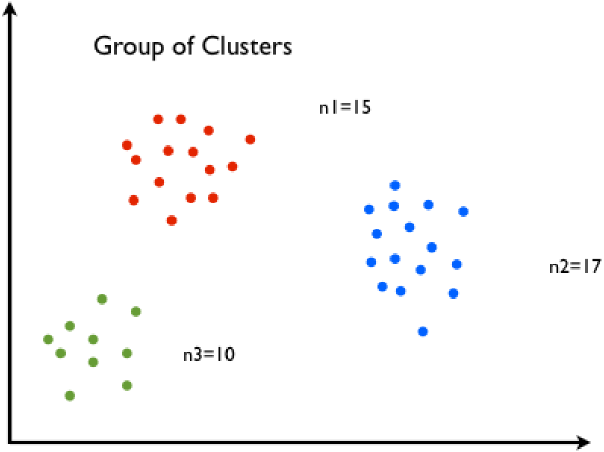
K-means clustering

- We have an over-all sum of squares for the whole clustering

$$Q \equiv \sum_{j=1}^{k} Q_j = \sum_{j=1}^{k} n_j V_j$$

Write a_i for the cluster to which vector i is assigned.

K-means clustering



K-means clustering

- Substitute in the definition of Q_j into that of $Q \implies$

$$Q = \sum_{i=1}^{n} \|\vec{x}_i - \vec{m}_{a_i}\|^2, i.e.$$

  the sum of squared distances from points to their cluster centers.
- K-means tries to reduce Q.
  - Step 2a: adjust $\vec{m}_j$ to minimize Q_j, given the current cluster assignments.
  - Step 2b: adjust a_i to minimize Q, given the current means.
  - At every stage Q either decreases or stays the same.
  - Q is the **objective function** for k-means, what it "wants" to minimize.

K-means as a search algorithm

```
\emph{$K$-means is a {\bf local search} algorithm: it makes
  solution that improve the objective.  This sort of search
  stuck in {\bf local minima}, where the no improvement is
  small changes, but the objective function is still not op
```

K-means as a search algorithm

- K-means: different starting positions correspond to different initial guesses about the cluster centers.
- Changing those initial guesses will change the output of the algorithm.
- Typically randomized, either as k random data points, or by randomly assigning points to clusters and then computing the means.
- Different runs of k-means generally give different clusters.
- Can make use of this: if some points end up clustered together in many different runs, that's a good sign that they really do belong together.