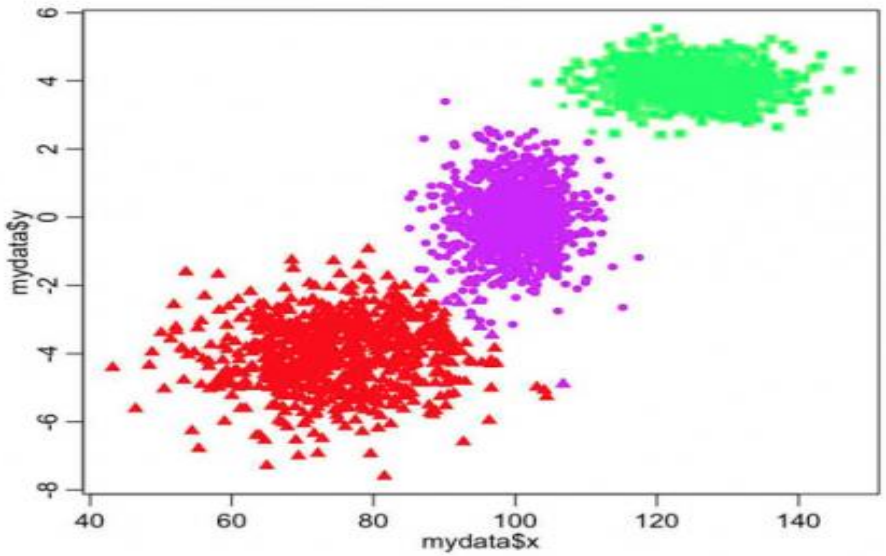
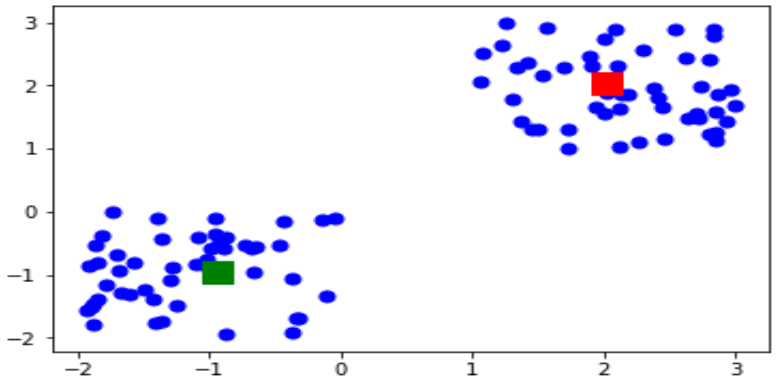
Topic 7.8 Cluster analysis and Machine Learning

- **Cluster analysis** or **clustering**:
  - task of grouping a set of objects
  - objects in same group (a **cluster**) are more similar (in some sense)
  - attribute values of data samples are used to measure similarity



Source: Internet

- **k-means clustering**
  - One of the simplest and most popular **unsupervised** machine learning algorithms
  - Usually, takes k as input and partitions the set into k subsets (clusters), and thus learns to which group an individual sample belongs.
  - 'Distance' is measured with respect to the mean value of the positions of samples in a cluster, called '**center of gravity**' or '**centroid**'.
  - Comparatively lower intra-cluster 'distance' than inter-cluster
  - Number of clusters, k may also be learned using various methods like Elbow method and Silhouette method.



Source: Internet

- For a given k, initially k objects are selected randomly as centroids.
- Two major repeated steps:

  1. Data assignment step
     Each data point is assigned to its nearest centroid.

  2. Centroid update step:
     Centroids are recomputed involving the current data points.

- Termination criteria: no data point changes its cluster; the sum of the distances is minimized; some maximum number of iterations is reached.

- Typically, the squared error criterion is used:

$E = \sum_{i=1:k} \sum_{p \in Ci} |p - m_i|^2$

E – sum of the squared errors of all objects; tried to be minimized
p – point in space representing a given object
m_i is the mean of cluster C_i

➢ Generally NP-hard, but heuristic variants for practical use are there.

➢ Diverse applications:

Segmenting customers by purchase history
Segmenting users by activities on website
Detecting activity types in motion sensors
Separating valid activity groups from bots
etc.