CAP 5610: Machine Learning

Lecture 13:
Clustering

Instructor: Dr. Gita Sukthankar
Email: gitars@eecs.ucf.edu

Reference Materials From: GuoJun Qi

Schedule

- Midterm exam next Tuesday
  - No material from the homework
  - No proofs
- Project proposal (ungraded) to be due on Oct 29
  - To be discussed in class next Thursday
- Final homework will be on deep RL
- Remaining topics: more deep learning: reinforcement learning, optimization, graphical models, LSTMs, GANs

Reading

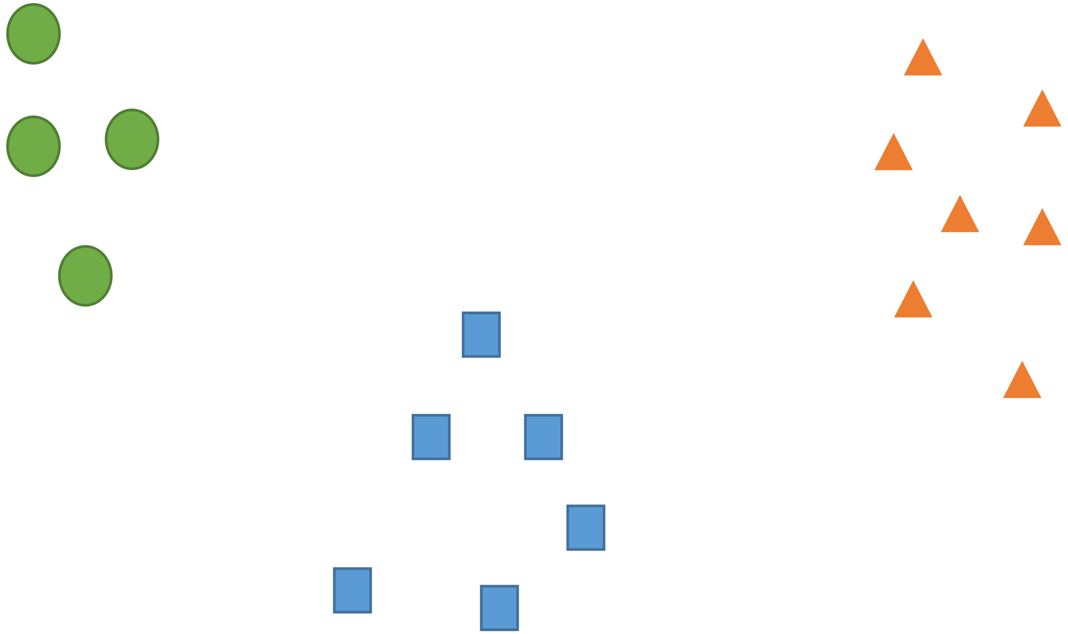
- Bishop Chapter 9
- Marsland Chapter 14
- Murphy Chapter 11 (advanced discussion EM and mixture of Gaussians)

Supervised vs. Unsupervised Learning Algorithms

- Supervised algorithms – training examples are labeled
  - Learning classifiers: KNN, (Naive) Bayes classifiers, SVM
  - Learning low-dimensional subspace: FDA
- Unsupervised algorithms – training examples are unlabeled
  - Learning low-dimensional subspace: PCA, autoencoder
  - Today’s topic: Clustering analysis: grouping a set of objects into (overlapped/disjoint) partitions of similar ones

Clustering Analysis

- Objective: grouping a set of objects into (overlapped/disjoint) partitions of similar ones



How to measure similarity?

- Similar in semantics or similar in appearance
  - It is hard to define a universal similarity measurement
  - E.g., Visually these two images are similar, but semantically they are not.
- Pragmatically, we define similarity in terms of distance between feature vectors.
  - **Euclidean distance**
  - L1 distance
  - Manhattan distance
  - etc.

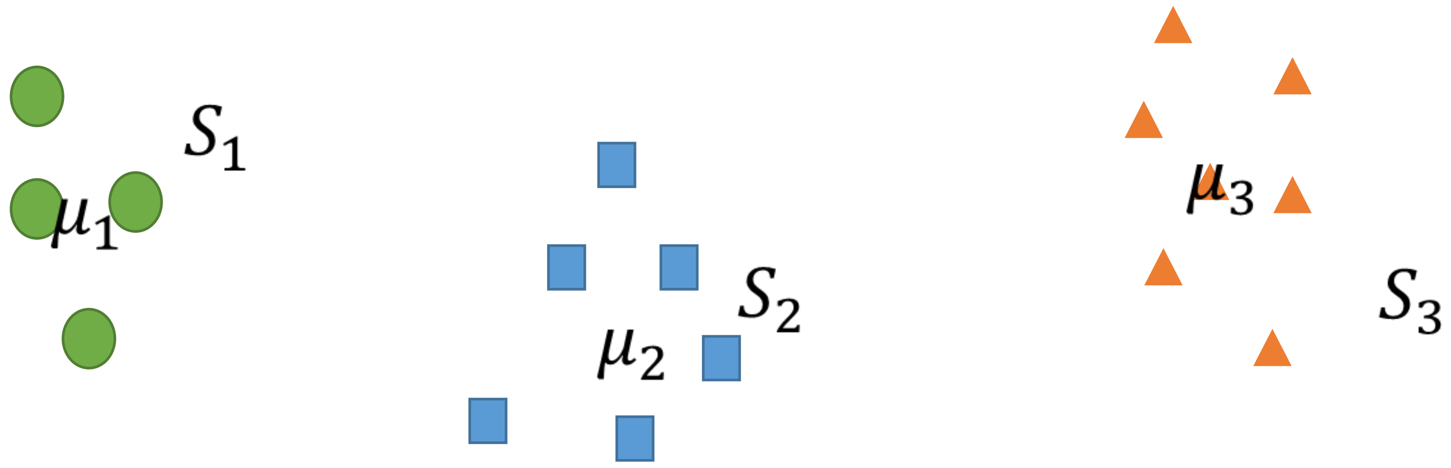
By B. Poczos and A. Singh

K-means clustering

- Given a set of examples $\{X_1, X_2, \ldots, X_n\}$, partition these n examples into k sets $\{S_1, S_2, \ldots, S_k\}$, by minimizing the within-cluster sum of squares:

$$\operatorname{argmin} \sum_{i=1}^{k} \sum_{X_j \in S_i} \left\| X_j - \mu_i \right\|^2$$

Where μ_i is the mean of examples in set S_i.

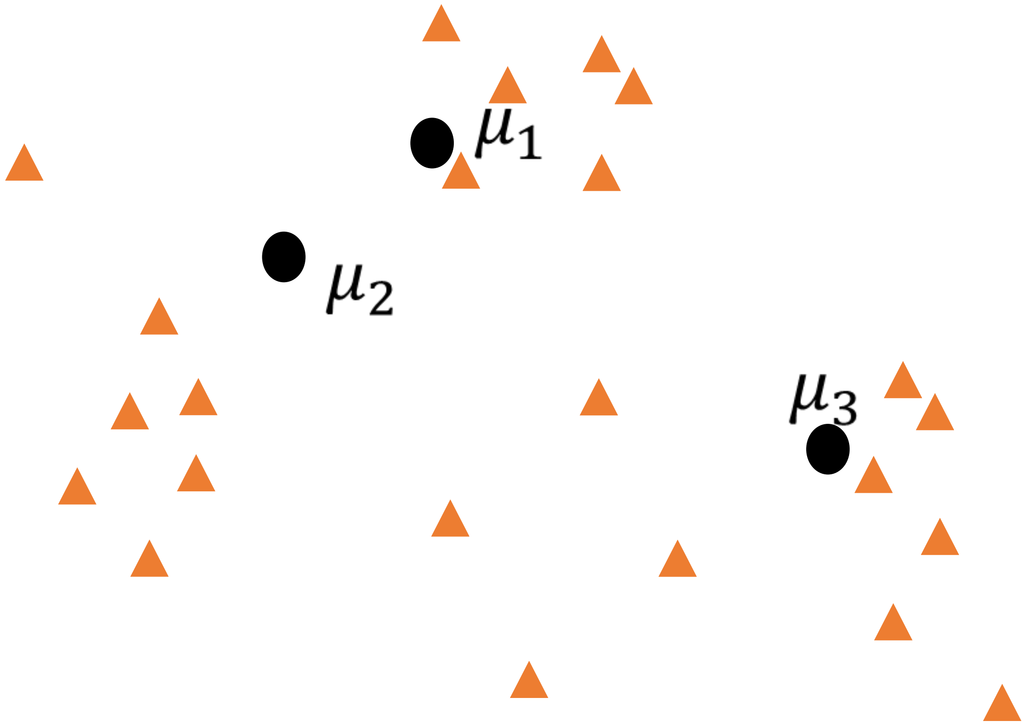


K-means clustering

- The decision variables associated with K-means clustering problem include
  - Assignment of each example to a cluster $\{S_1, S_2, \dots, S_K\}$
  - Mean vectors $\{\mu_i\}$ for each cluster
- Jointly optimization of these two sets of decision variables is NP-hard.
  - Heuristic method: K-means algorithm
  - Alternately updating the two sets of decision variables.

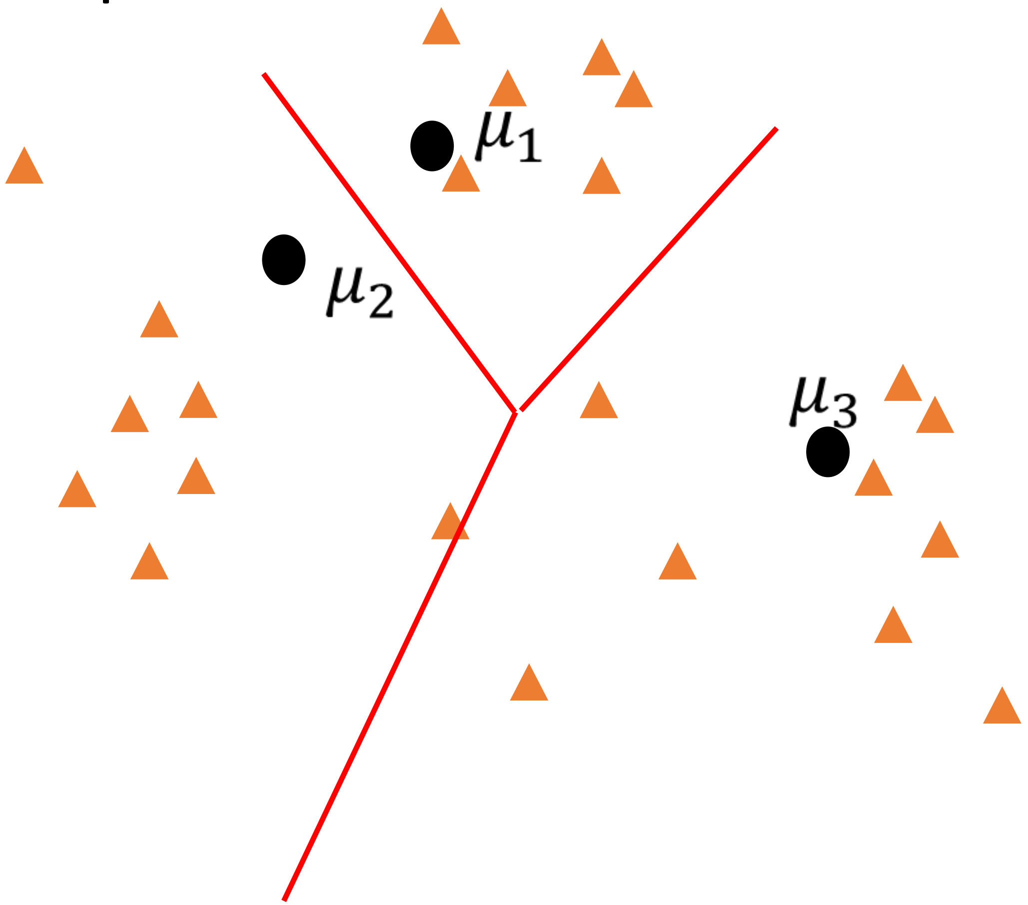
K-means: Step 1

- Randomly guess the cluster means



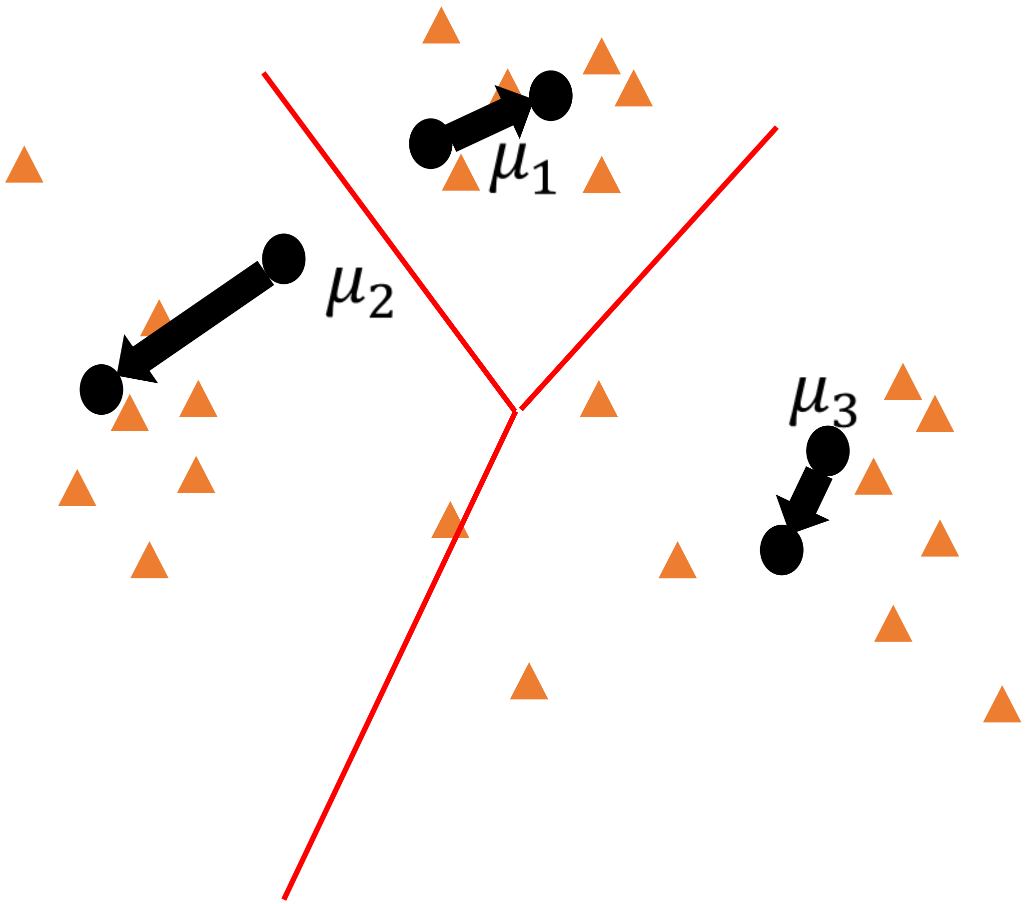
K-means: Step 2

- Assign each example to the nearest cluster mean



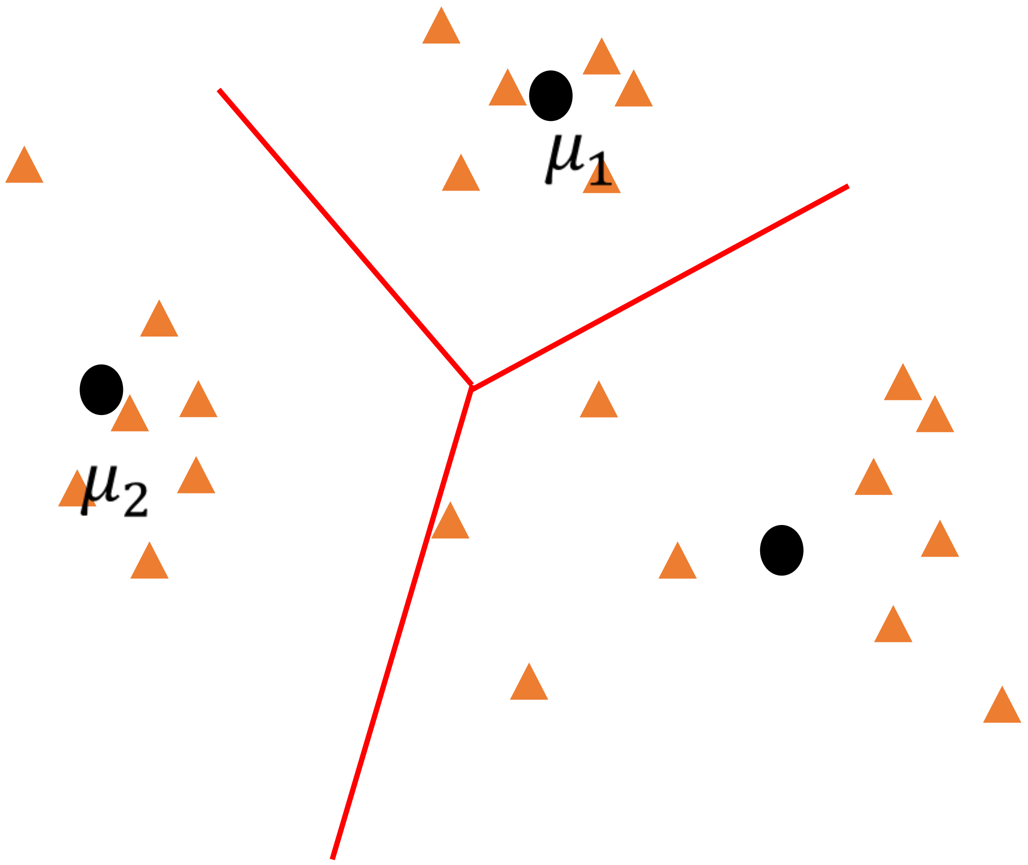
K-means: Step 3

- Update the estimates of cluster means based on current example assignment.



K-means: Step 3

- Reassign each example to the assignment.



K-means: Algorithms

- Step 1: randomly initialize the guess of cluster means
- Repeat
  - Step 2: assign each example to the closest cluster mean
  - Step 3: update the estimate of cluster means based on the current example assignments
- Until convergence (the cluster means and example assignments do not change too much)

Formal treatment of K-means

- Denote by $C(j)$ the assignment of example j to a cluster $C(j)$, then the sum of squared distance between examples and cluster means can be written as

$$\operatorname{argmin} \sum_{i=1}^{k} \sum_{X_j \in S_i} \left\| X_j - \mu_i \right\|^2 = \sum_{j=1}^{n} \left\| X_j - \mu_{C(j)} \right\|^2$$

Optimal solution

- Alternately optimizing μ, C

$$\mathrm{argmin}_{\mu,C} \sum_{j=1}^{n} \left\| X_j - \mu_{C(j)} \right\|^2$$

- Fix μ, the best assignment:

$$\mathrm{argmin}_C \sum_{j=1}^{n} \left\| X_j - \mu_{C(j)} \right\|^2 = \sum_{j=1}^{n} \mathrm{argmin}_{C(j)} \left\| X_j - \mu_{C(j)} \right\|^2$$

  - It is exactly assigning each example to the closest cluster as in K-means.

Optimal solution

- Alternately optimizing μ, C

$$\operatorname{argmin}_{\mu,C} \textstyle\sum_{j=1}^{n} \left\| X_j - \mu_{C(j)} \right\|^2 = \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\| X_j - \mu_i \right\|^2$$

- Fix C, the optimal means:

$$\operatorname{argmin}_{\mu} \sum_{i=1}^{K} \sum_{j:C(j)=i} \left\| X_j - \mu_i \right\|^2 = \sum_{i=1}^{K} \operatorname{argmin}_{\mu_i} \sum_{j:C(j)=i} \left\| X_j - \mu_i \right\|^2$$

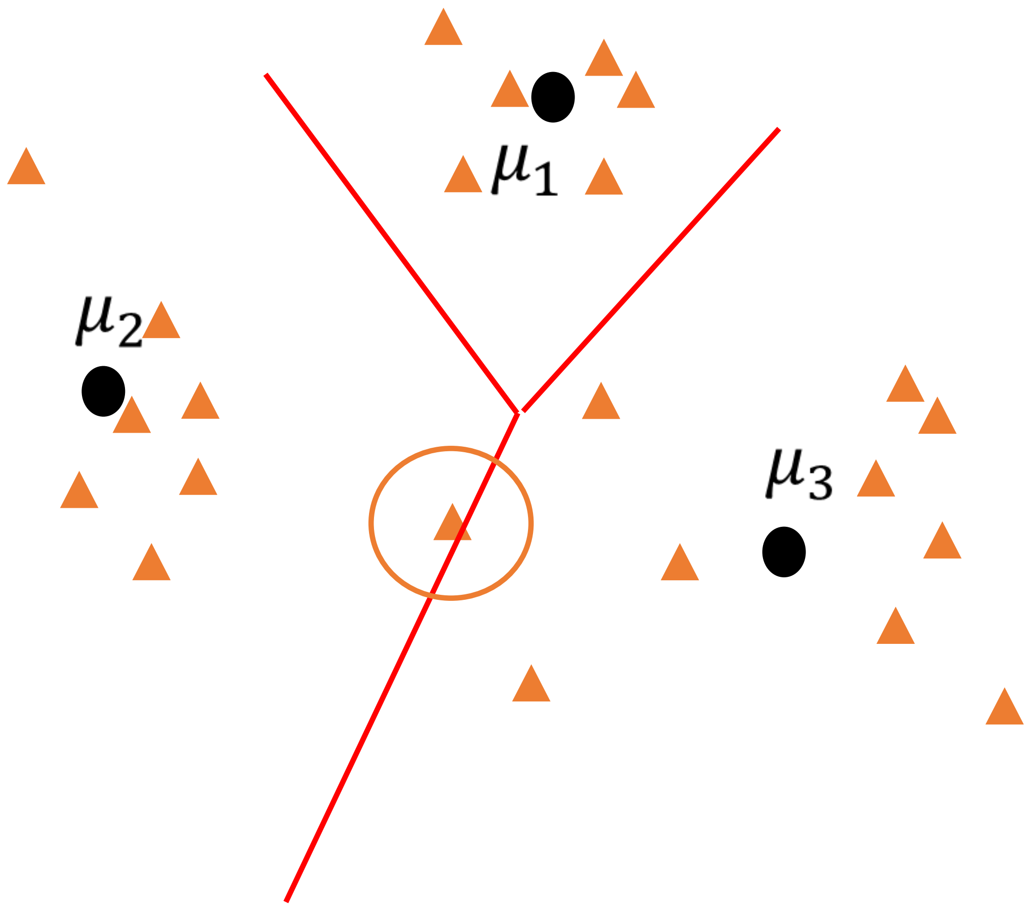
  - The best solution is obtained by setting μ_i to the mean of examples assigned to this cluster.

Expectation-Maximization

- K-means algorithm:
  - Expectation: Fix μ, find the assignment
    - "expectation" means to which cluster we expect to assign an example.
  - Maximization: Fix C, find the best mean for each cluster
    - "maximization" means we maximize the likelihood that all assigned examples belong to this cluster by setting a proper mean.
- We will see a generalization of EM to a probabilistic model.

Problems to be addressed

- K-means makes a hard assignment of an example to a cluster.
  - Ambiguity may exist when we assign an example
  - Soft assignment is preferred.
    - Directly characterizing the probability that an example belongs to a cluster
    - A distribution will be used to model each cluster, e.g., Gaussian
    - The whole distribution for all examples is a mixture of distributions for each cluster, i.e., a mixture distribution model



Convergence

- K-means algorithms can be guaranteed to converge.

Proof: In each step, K-means minimizes the objective function monotonically

$$\sum_{j=1}^{n} \left\| X_j - \mu_{C(j)} \right\|^2$$

This generates a sequence of non-increasing objective values, which is lower bounded by zero; hence convergence occurs.

Computation Complexity

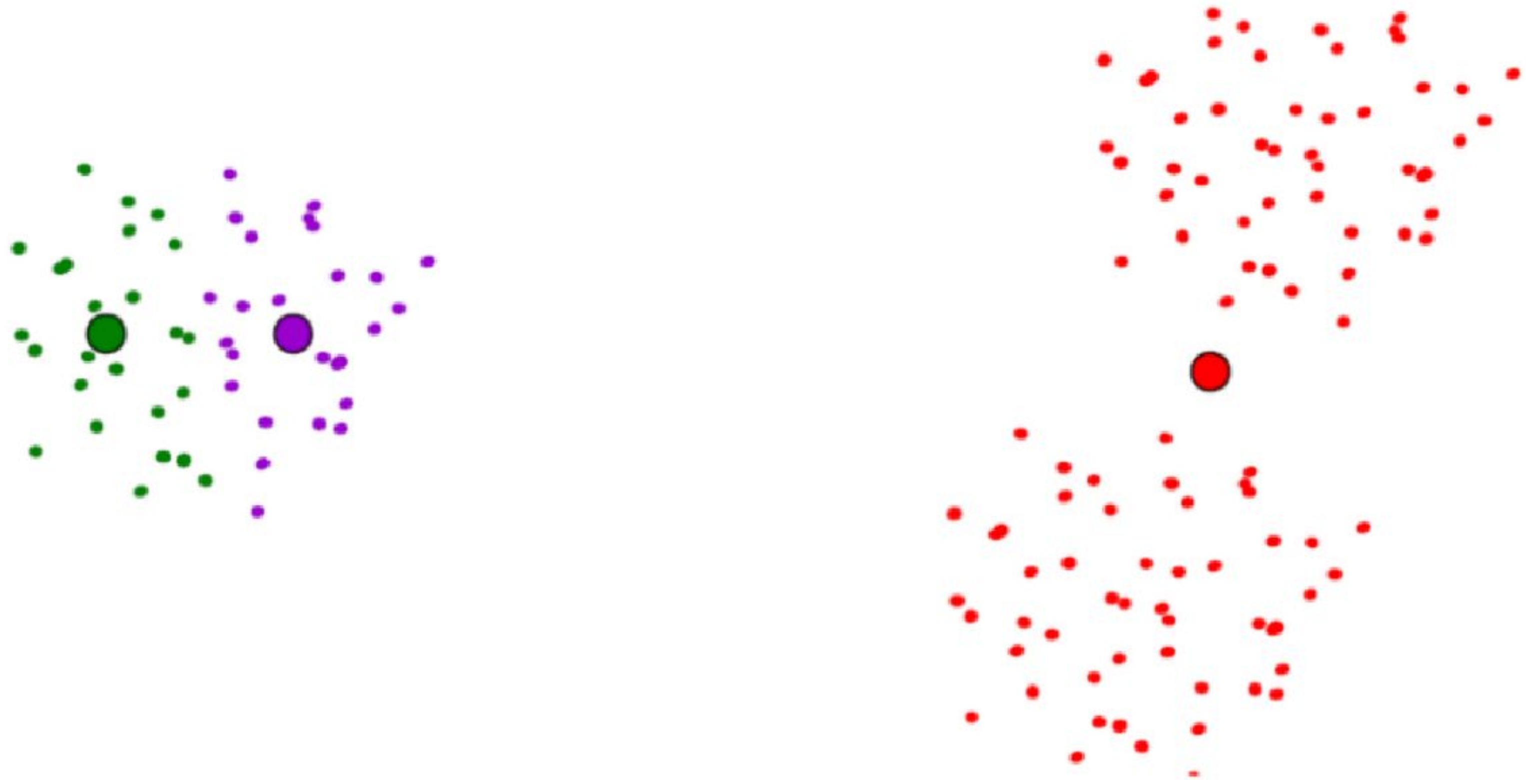
- In each iteration,
  - It costs O(Kn) to compute the distance between each of n examples and K cluster means
  - It costs O(n) to update the cluster means by adding each example to one cluster
- Assume l iterations are done before terminating the algorithm, the computational complexity is O(lkn)

K-means algorithm

- The objective function optimized by k-means is not convex
  - It suffers from many local optima
  - It is sensitive to the initialization of mean vectors (seeds)
    - Bad seeds can result in poor convergence, or converge to bad clustering result.

Bad seeds

Bad seeds



Seed choice

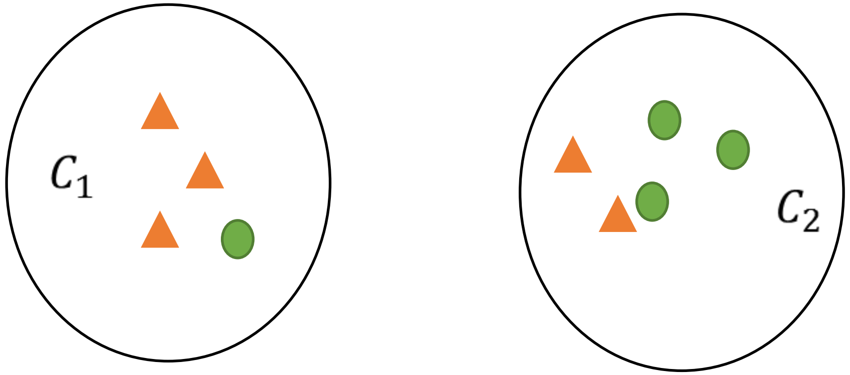
- Choosing good seeds using heuristics
  - Choosing seeds which are least similar to each other
  - Initialize seeds multiple times and choose the results with least objective value (least sum of squares between cluster means and examples)
  - Initialize with the results from another methods

Applications

- Social community discovery – grouping users with the similar profiles
- Image segmentation – grouping pixels with similar values which are spatially close to one another
- Human genetic clustering – clustering similar genetic data to find population structures
- Marketing – grouping customers into market segments based on surveys, sales data, and test panels.
- Many others...

Evaluation metrics

- Internal metrics
  - High intra-cluster similarity and low inter-cluster similarity
- External metrics
  - Assume we have labeled examples
  - Purity – each cluster is assigned to the class with the most frequent label in this cluster, and purity measures the portion of correctly assigned examples.

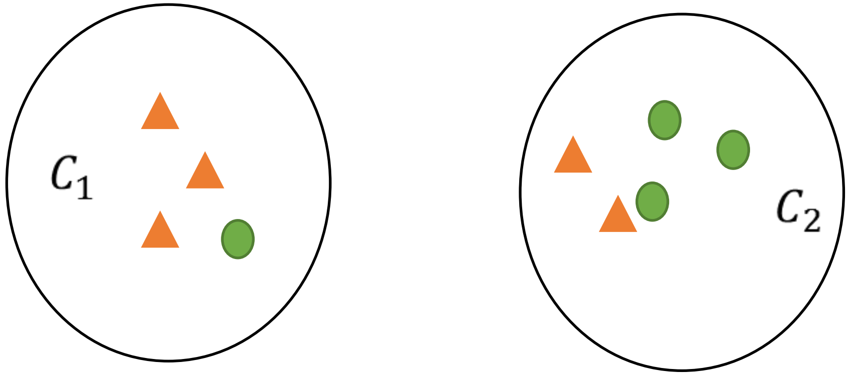


$$\text{purity} = \frac{3+3}{9}$$

Evaluation metrics

- External metrics
  - Rand Index – how good the decision made by clusters in terms of the labels
  - For each pair of examples, define
    - True positive (TP): two examples in the same cluster are assigned the same label
    - True negative (TN): two examples in different clusters are assigned the different labels
    - False positive (FP): two examples in the same cluster are assigned the different labels
    - False negative (FN): two examples in the different clusters are assigned the same label

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$



Summary

- Clustering analysis aims to group similar objects into a set of clusters
- K-means is one of most popular methods
  - Implementing a heuristic EM method to optimize sum of squared distance between cluster means and examples.
  - Guaranteed to converge, but not always converge to global convergence.
  - Sensitive to initialization
- Extension of EM to soft assignment, handling mixture distribution model