Unsupervised Learning – Part 3: The k-Means Algorithm

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Discuss the basics of data clustering

Objective

Illustrate the k-Means Algorithm

Finding the clusters/groupings of the samples

A few basic questions to answer

- How to represent the clusters?

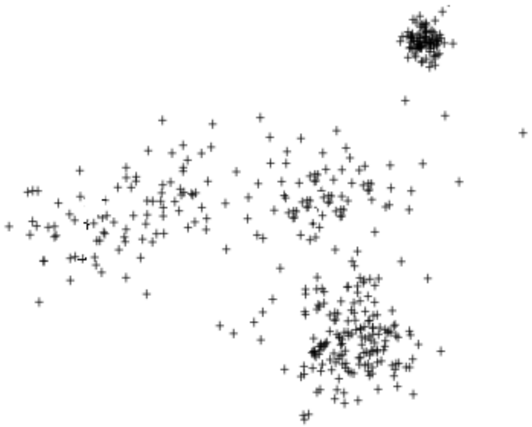
  ➔ We will use the centroid to represent a cluster.

- Which cluster a sample should be assigned to (e.g., membership)?

  ➔ We will use the similarity to the centroid to determine the membership.

- What similarity measure to use?

  - E.g., Euclidean distance



More on Similarity Measures

- If we use Euclidean distance as the measure:
  - It is invariant to translations & rotations of the feature space.
  - But not to more general transformations.

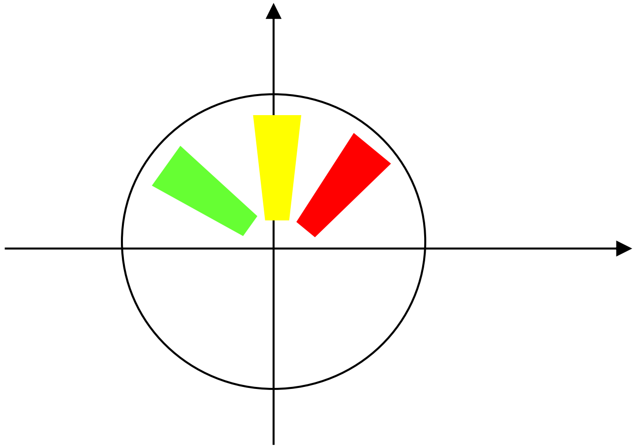
- E.g., if one feature is scaled.

More on Similarity Measures (cont'd)

- Other types of similarity measures

- E.g., cosine similarity
  - For clustering colors in the hue-saturation space

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|}$$



- E.g., distance on a graph, like shortest path.

Clustering as Optimization

The sum-of-squared-error criterion/cost

- Let D_i be the subset of samples from class i.
- Let n_i be the number of samples in D_i, and $\mathbf{m}_i$ the mean of those samples

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{x}$$

- The sum of squared error is:

$$J_e = \sum_{i=1}^{C} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

➔ Well-separated, compact data "clouds" tend to give small errors when the clusters coincide with the clouds.

Clustering as Optimization (cont'd)

$$J_e = \sum_{i=1}^{C} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

- An optimization problem to solve for finding a "good" clustering: to find the partition of the data that minimizes J_e

- If the membership of a sample is determined by the distance to the means $\mathbf{m}_i$
  - Then the task is to find the optimal set of $\{m_i\}$
  - The problem is NP-hard.

k-Means Clustering

- Input: Given n data samples

- Goal: Partition them into k clusters/sets D_i, with respective center/mean vectors $\mu_1, \mu_2, \ldots, \mu_k$, so as to minimize

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in D_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2$$

- Comparing with the mixture models:
  - Here we do “hard” assignment of the membership to a sample (simply based on its distance to the cluster center).

The Basic k-Means Algorithm

```
Given: n samples, a number k.
Begin
    initialize μ1, μ2, …, μk(randomly
    selected)
          do classify n samples according to
                       nearest μi
              recompute μi
          until no change in μi
    return μ1, μ2, …, μk
  End
```

Illustrating the Algorithm