CSE 575: Statistical Machine Learning (Spring 2021)
Instructor: Nupur Thakur

Unsupervised Learning

Table of contents

What is unsupervised learning?

- Given a training set of n unlabeled samples - x(i)
- What can we learn from the samples?
  - Estimate the overall distribution of the data without knowing their label. — EM algorithm
  - Figure out the groupings of the samples (if any). — Clustering (K-means)
  - Identify some features that may be more important than others. — Feature selection

Finding clusters

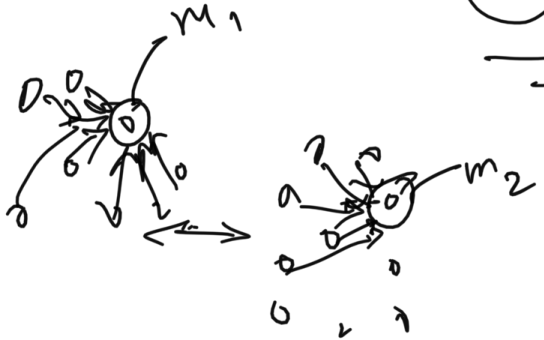
- How to represent the clusters?
  - Centroid (mean of all samples)
- Which cluster a sample should be assigned to (e.g., membership)?
  - Some similarity measure like Euclidean dist.
- What similarity measure to use?
  - Different measures – Euclidean dist.
    – Cosine similarity.

Clustering Objective function

- The sum-of-squared-error criterion/cost-

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

(centroid of i^{th} cluster.)



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

Reduce the intra-cluster distance
Increase the inter-cluster distance.

K-Means Clustering

Given: n samples, a number k. → No. of clusters.

Begin

initialize $\mu_1, \mu_2, \ldots, \mu_k$ (randomly selected)

do classify n samples according to nearest μ_i ⟹ Membership Assignment

recompute μ_i

until no change in μ_i

return $\mu_1, \mu_2, \ldots, \mu_k$

End

K-Means Clustering - Some scenarios

- What happens if the distance of a point is same from more than one centroid?

  Randomly assigned one of the clusters.

  - Hard assignment

- What happens in case of outliers?

  Sensitive to outliers.

K-Means Clustering - Some scenarios

- Different initializations?

  - Might lead to sub-optimal results.

- Different number of clusters than k?

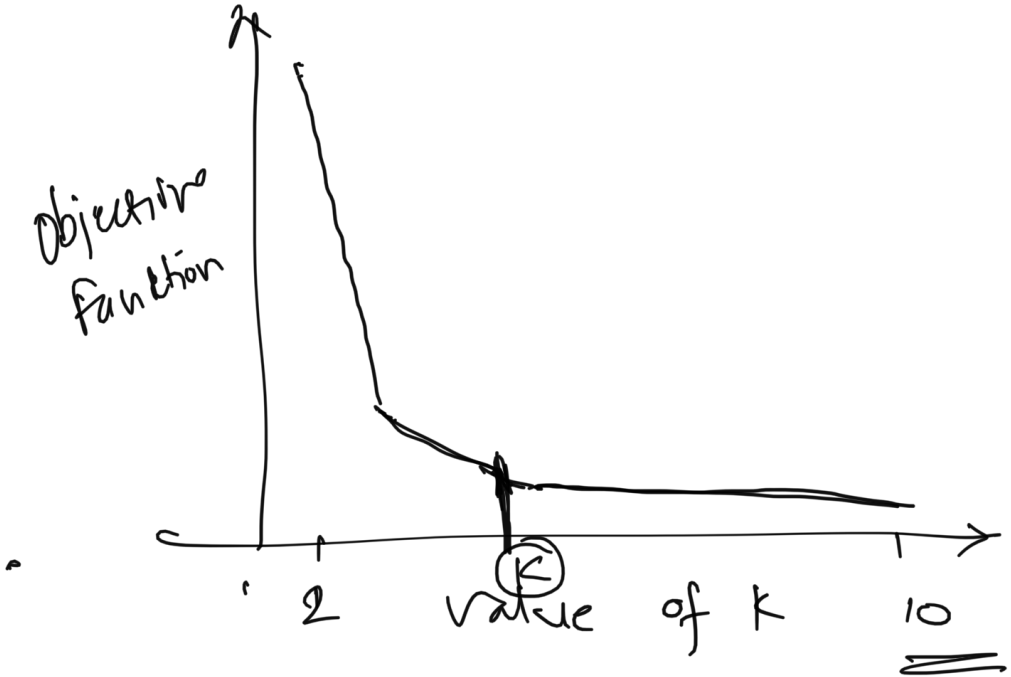
How to choose the k?

- Elbow method.

10,000 $k=1$

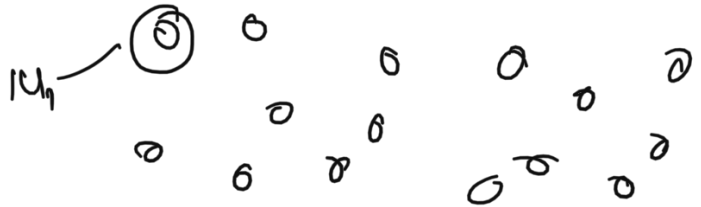
1-9999 $k=n$

$1 < k < n$



How to deal with sensitivity to initialization problem?

1. Run the algo for few epochs & take the average result
2. first centroid randomly. – Sensitive to outliers.
3. Kmeans++ ⇒ Smart initialization



For detailed explanation of how K-Means++ works, refer https://www.youtube.com/watch?v=HatwtJSsj5Q

K-Means Clustering - Pros and Cons

Pros

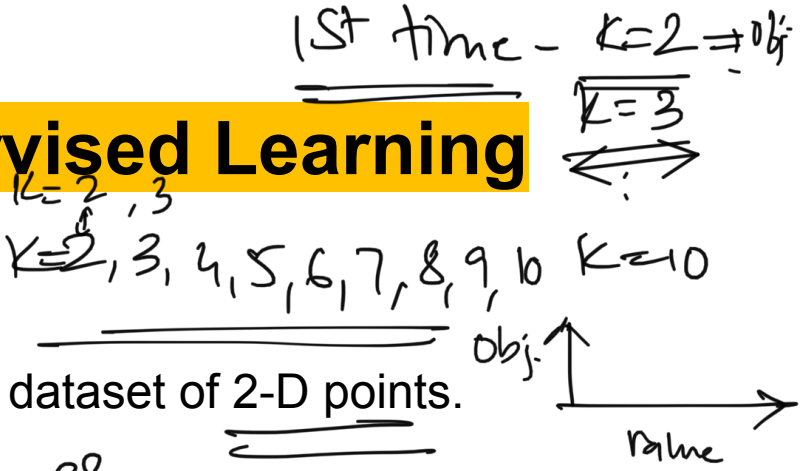
1. Easy to implement
2. fast.

Cons

1. Sensitive to initialization
2. " " outliers.
3. Hard Assignment of data points.
4. Choose no. of clusters manually.

Project Part 2 - Unsupervised Learning (K-means)

- Apply k-means algorithm the given dataset of 2-D points.
- Initialization strategies:
  - Strategy 1: randomly pick the initial centers from the given samples.
  - Strategy 2: pick the first center randomly; for the i-th center (i>1), choose a sample (among all possible samples) such that the average distance of this chosen one to all previous (i-1) centers is maximal.



Project Part 2 - Unsupervised Learning (K-means)

- Objective function-

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

- Number of clusters (k) - from 2-10
- Plot the objective function value vs. the number of clusters k.

Project Part 2 - Deliverables

— Due on 26 March (Arizona time) 11:59 PM

- Well-commented code — python / MATLAB / Jupyter notebook (.ipynb)
- A report that summarizes the results and includes all the plots. — .pdf format (typed, not handwritten)
- Please submit the code and the report as separate files on Canvas. **Do not zip them.**

.py
.py
.pdf

Questions?