CSCE 633: Machine Learning

Lecture 35: Unsupervised Learning: K-means and Hierarchical Clustering

Texas A&M University
Bobak Mortazavi

Outline

- (Re-)Introducing unsupervised learning
- K-Means clustering
- Gaussian Mixture Models
- Training with Expectation Maximization

- Some slides adapted from Introduction to Statistical Learning, 2nd edition, James, Witten, Hastie, Tibshirani

Unsupervised learning

- Find patterns/structure/sub-populations in data “knowledge discovery”
- Training data does not contain outputs
- Less well-defined problem with no obvious error metrics
- Examples: topic modeling, market segmentation, handwritten digits, news stories, etc.

K-Means Clustering

- Input D = $\{x_i\}_{i=1}^{N}, where\ x_i \in \mathbb{R}^p$
- Output: Clusters $\mu_1, \dots, \mu_K$
- Decision: Define cluster membership, provide a cluster id assigned to each sample x

$$A(x_i) \in \{1, \dots, K\}$$

- Evaluation Metric: Distortion Measure

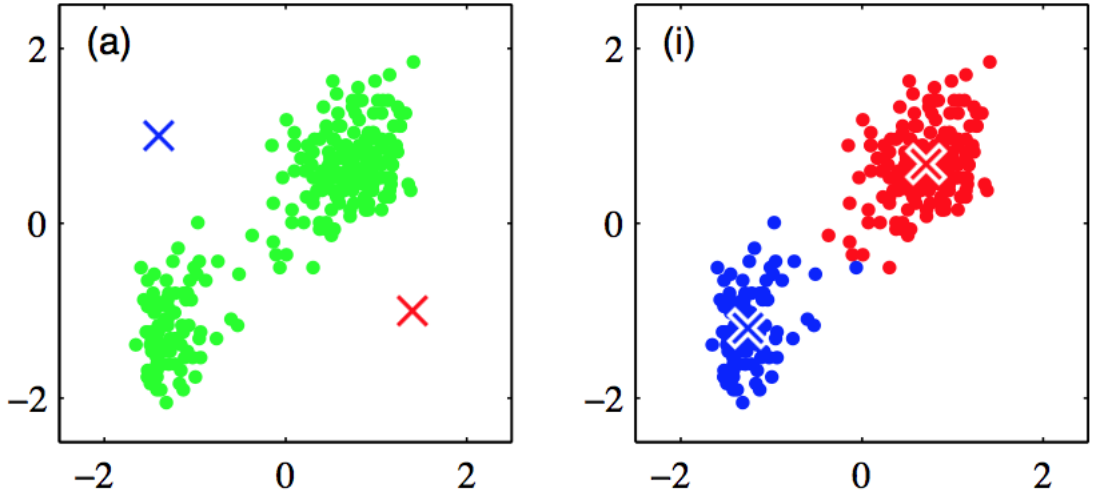
$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_i - \mu_k\|_2^2$$

Where $r_{nk} = 1$ if $A(x_i) = k$

- Intuition: Data points get assigned to cluster k if they are close to centroid μ_k

K-Means Clustering

- Input D = $\{x_i\}_{i=1}^{N}, where\ x_i \in \mathbb{R}^p$
- Output: Clusters $\mu_1, \ldots, \mu_K$
- Decision: Define cluster membership, provide a cluster id assigned to each sample x
- Evaluation Metric: Dis



Where $r_{nk} = 1$ if $A(x_i)$

- Intuition: Data points get assigned to cluster k if they are close to centroid μ_k

K – means Algorithm

- Optimization Function

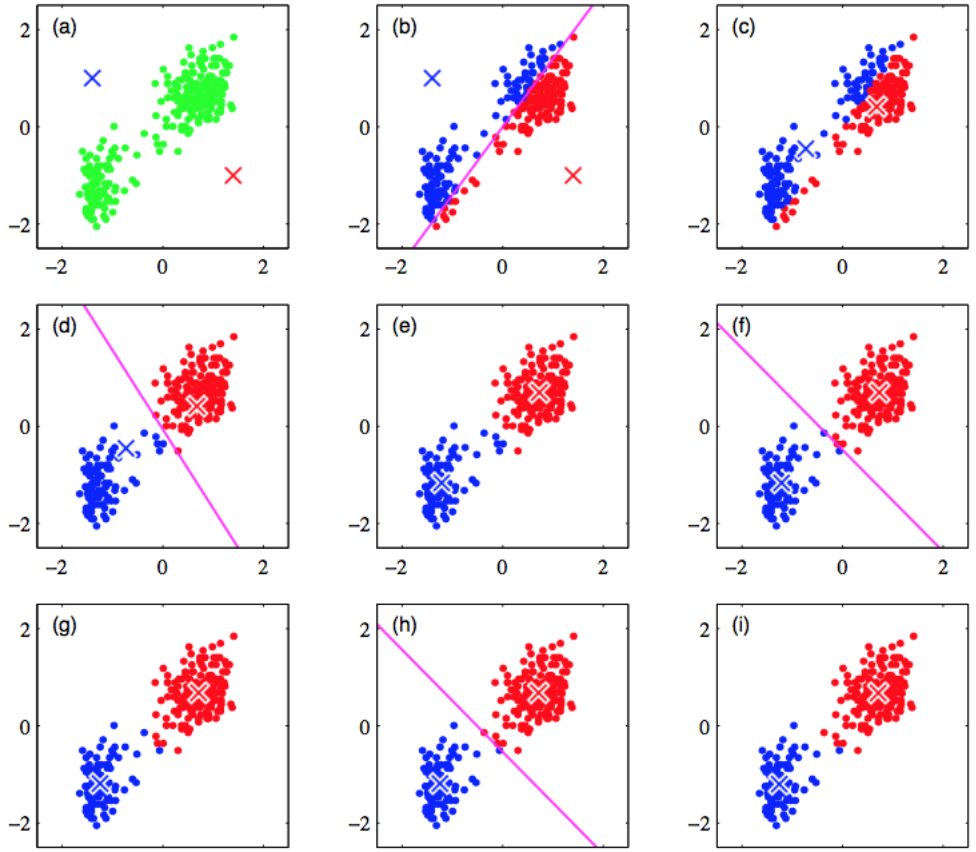
$$\min_{r_{nk}} J = \min_{r_{nk}} \sum_{i=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_i - \mu_k\|_2^2$$

- Step 0: Initialize μ_k to some random values
- Step 1: Assume the current μ_k is fixed, minimize J over r_{nk} which leads to cluster assignment
- Step 2: Assume the fixed cluster assignment, update the cluster centroids as in

$$\mu_k = \frac{\sum r_{nk}\, x_n}{\sum r_{nk}}$$

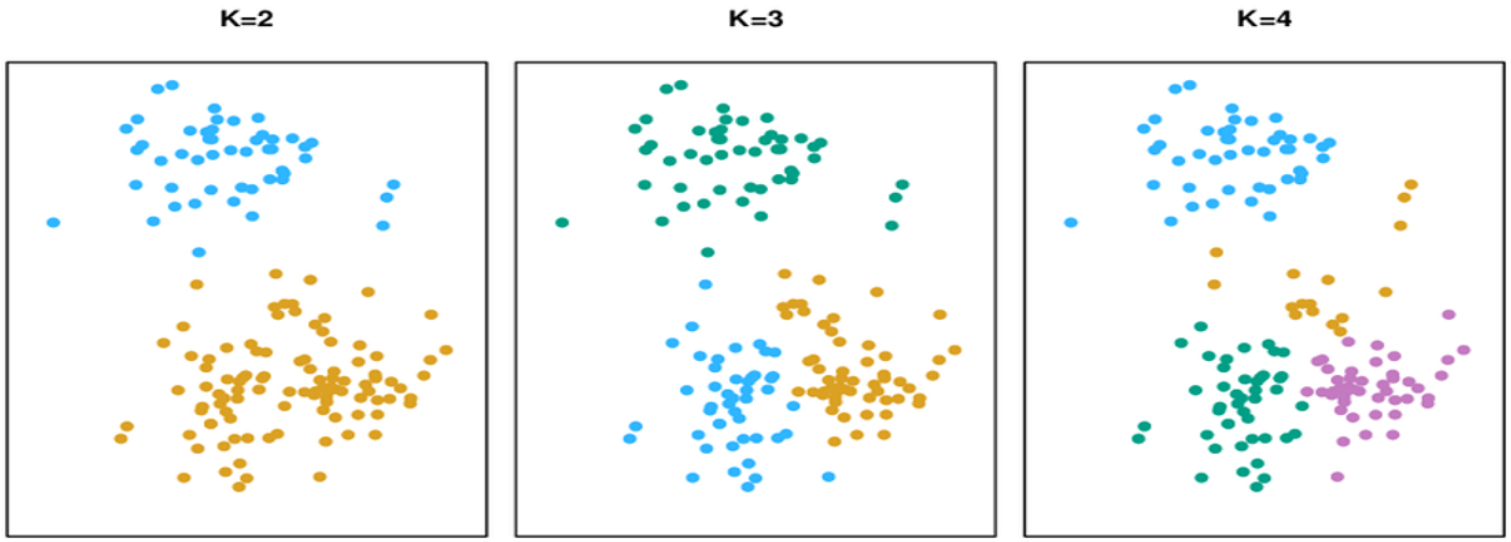
- Step 3: decide to stop or return to step 1

Visualization of K-Means

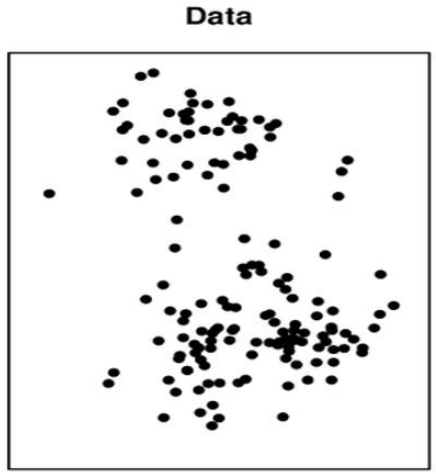


Ways to Measure K-Means Fit

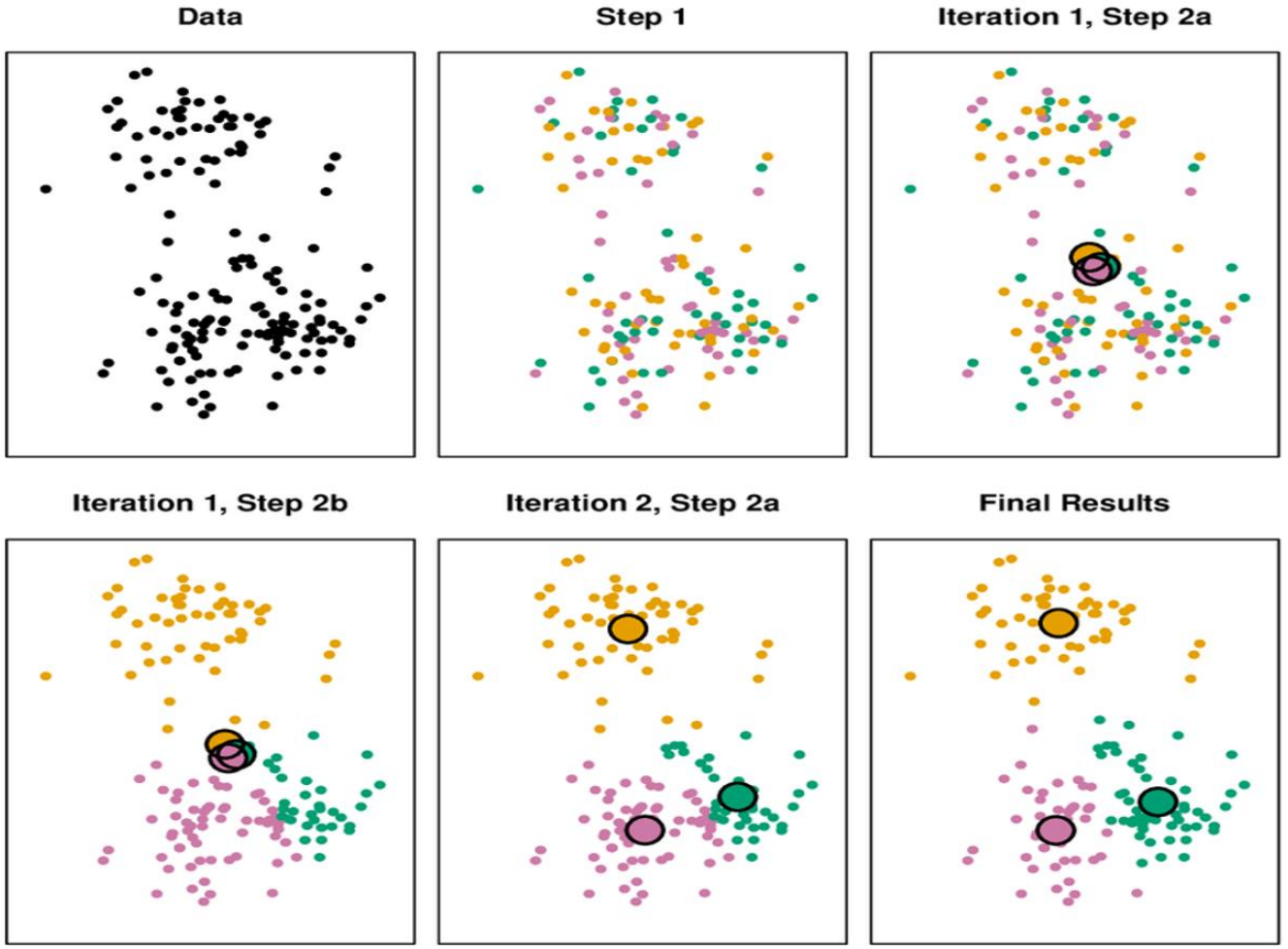
- Ideally, a good k-means clustering has small within-cluster variation.
- In other words, all elements within the cluster should be very similar.
- No guarantee a finished clustering is the optimal -> random restarts of the algorithm help identify.
- How do you know how many components, k, to use?



Another example



Another example



Another example: different starts

Discussion

- What are common examples of unsupervised learning you have interacted with?
- What are limitations of k-means clustering?
- Let’s brainstorm ways around these limitations!

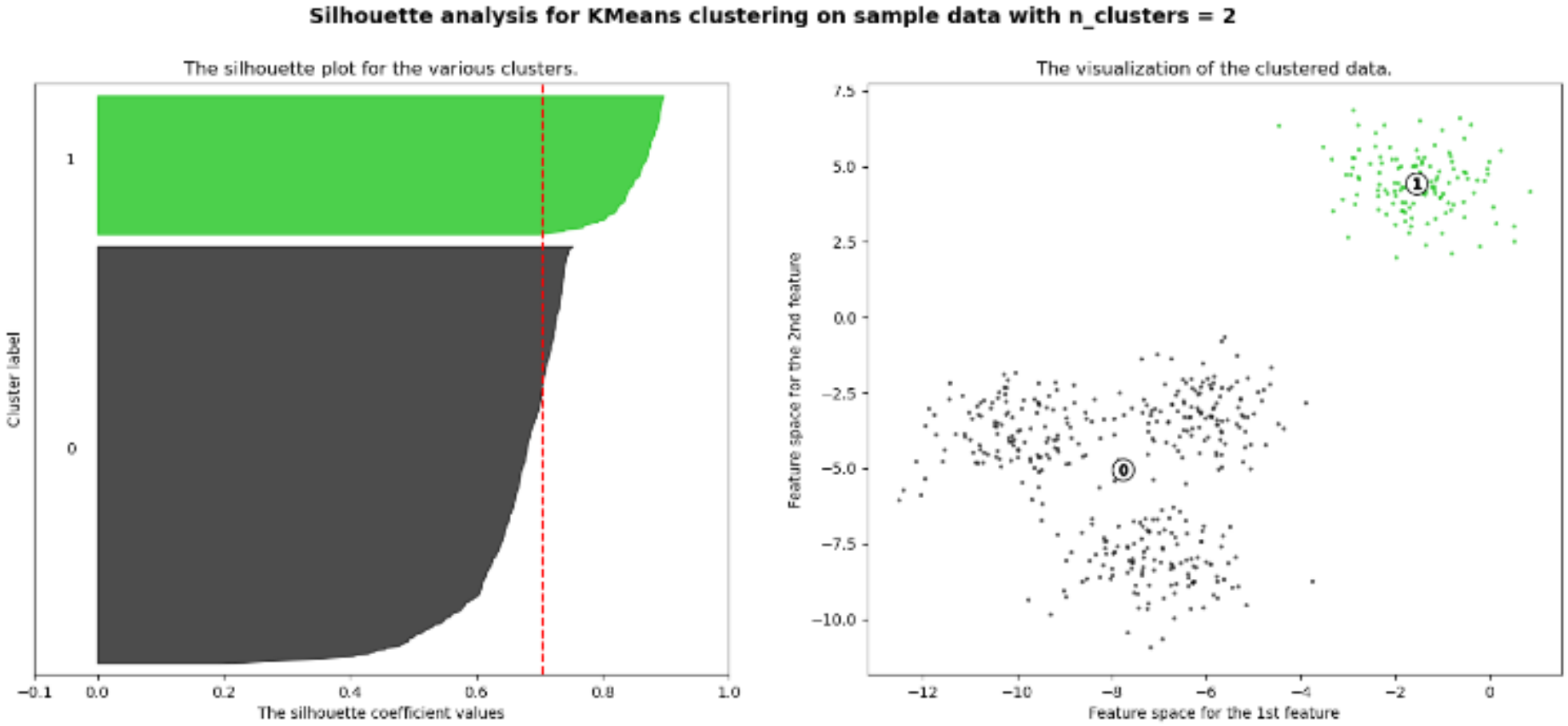
K-means limitations

- K-means requires us to pre-specify the number of clusters k.
- This can be a disadvantage if we choose the wrong k
- What if there were an alternative approach that does not pre-specify this?
- What would that look like? Would it be top-down or bottom up?

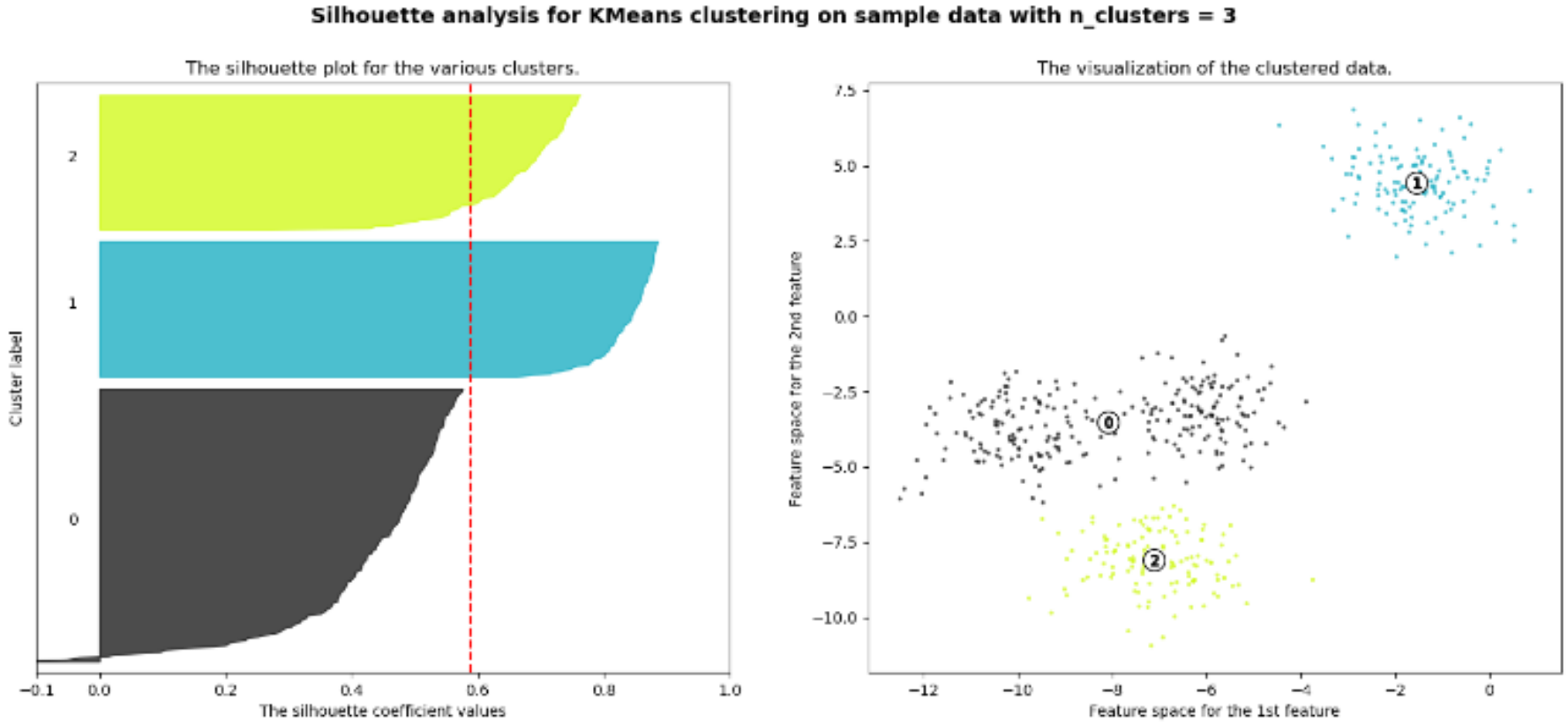
Number of Clusters

- Silhouette Score
  - For data point i in cluster C_I
    - $a(i) = \frac{1}{|C_I|-1}\sum_{j \in C_I, i \neq j} d(i,j)$
    - $\mathrm{b}(i) = \min_{J \neq I} \frac{1}{|C_J|}\sum_{j \in C_J} d(i,j)$
    - $s(i) = \frac{\mathrm{b(i)-a(i)}}{\max(a(i),b(i))}$

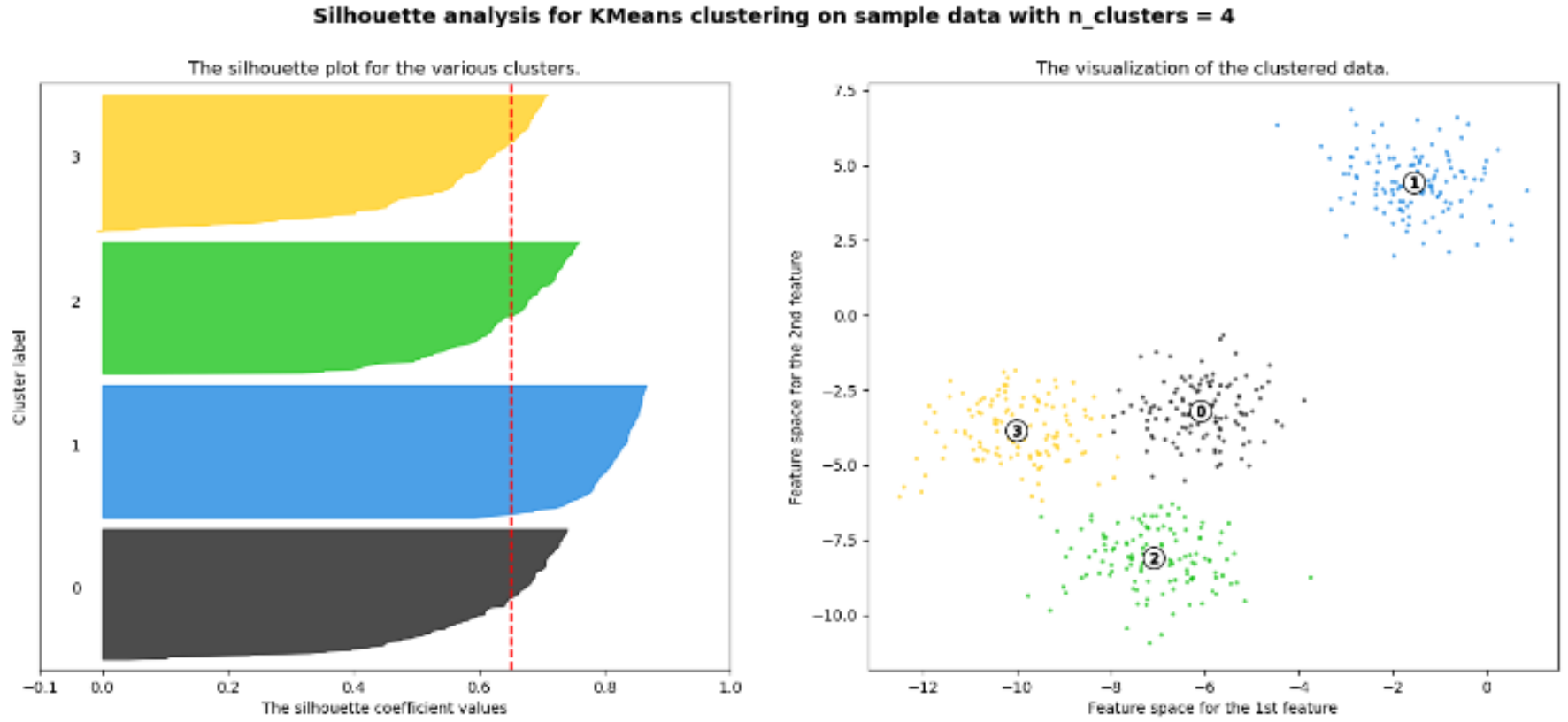
Number of Clusters



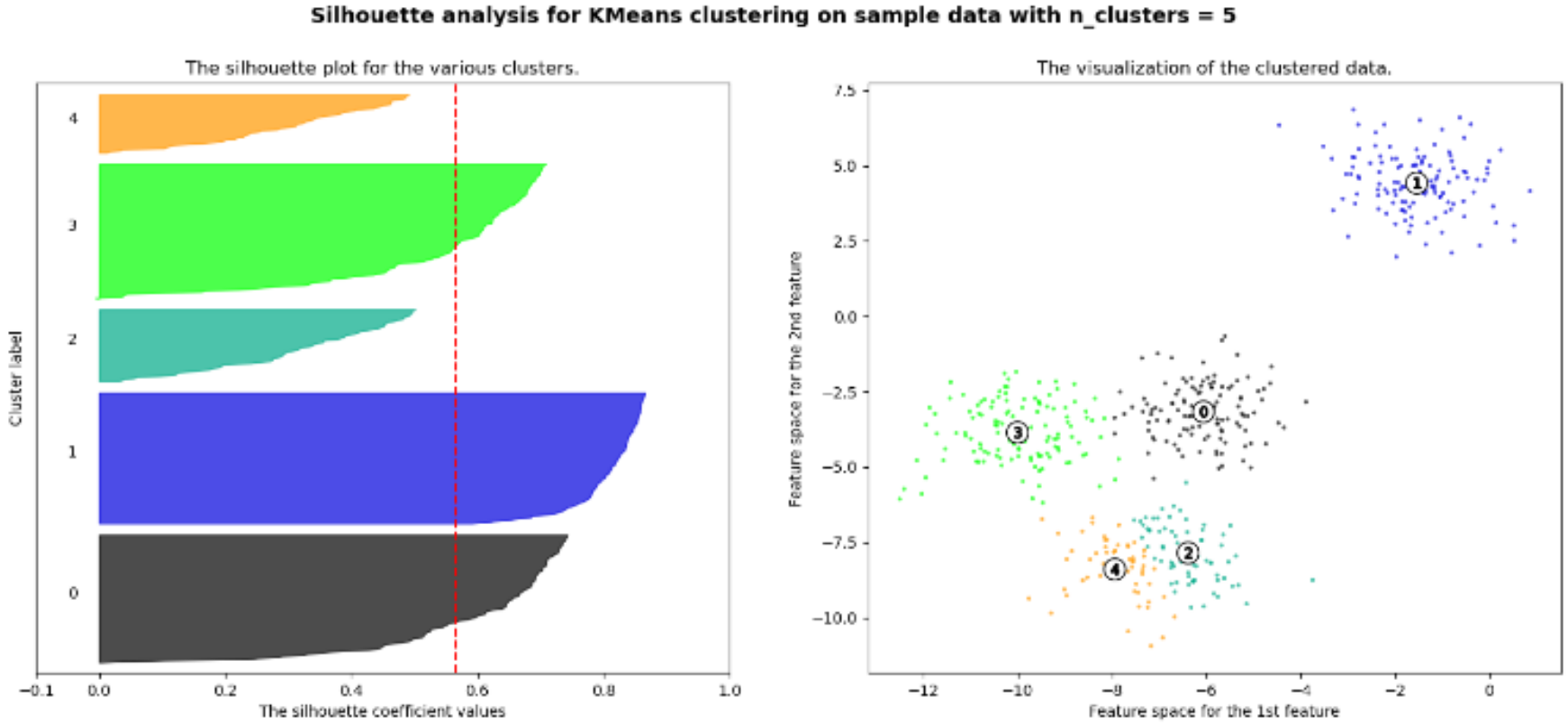
Number of Clusters



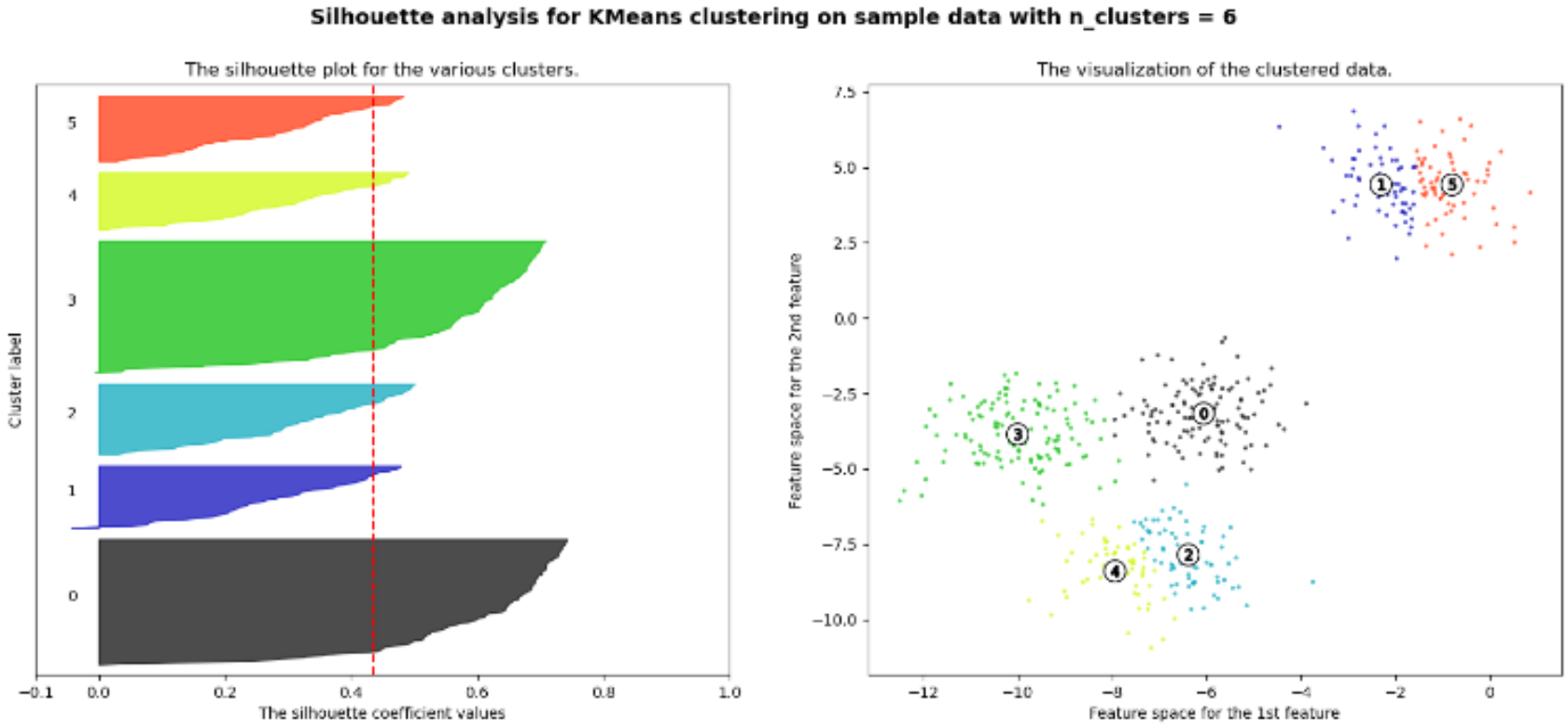
Number of Clusters



Number of Clusters



Number of Clusters



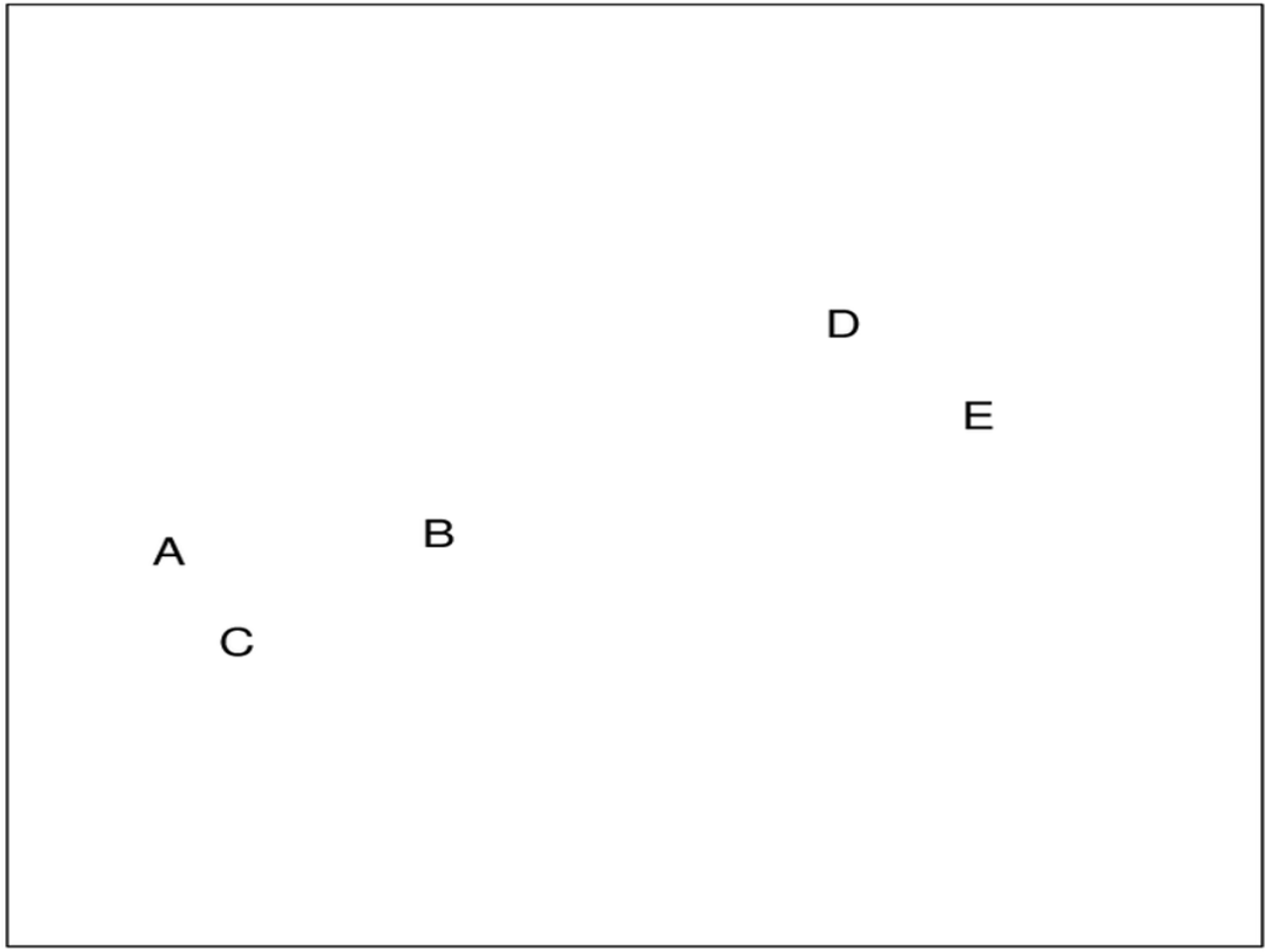
Number of Clusters

- Silhouette score
  - N Clusters = 2: 0.705
  - N Clusters = 2: 0.588
  - N Clusters = 2: 0.651
  - N Clusters = 2: 0.566
  - N Clusters = 2: 0.436

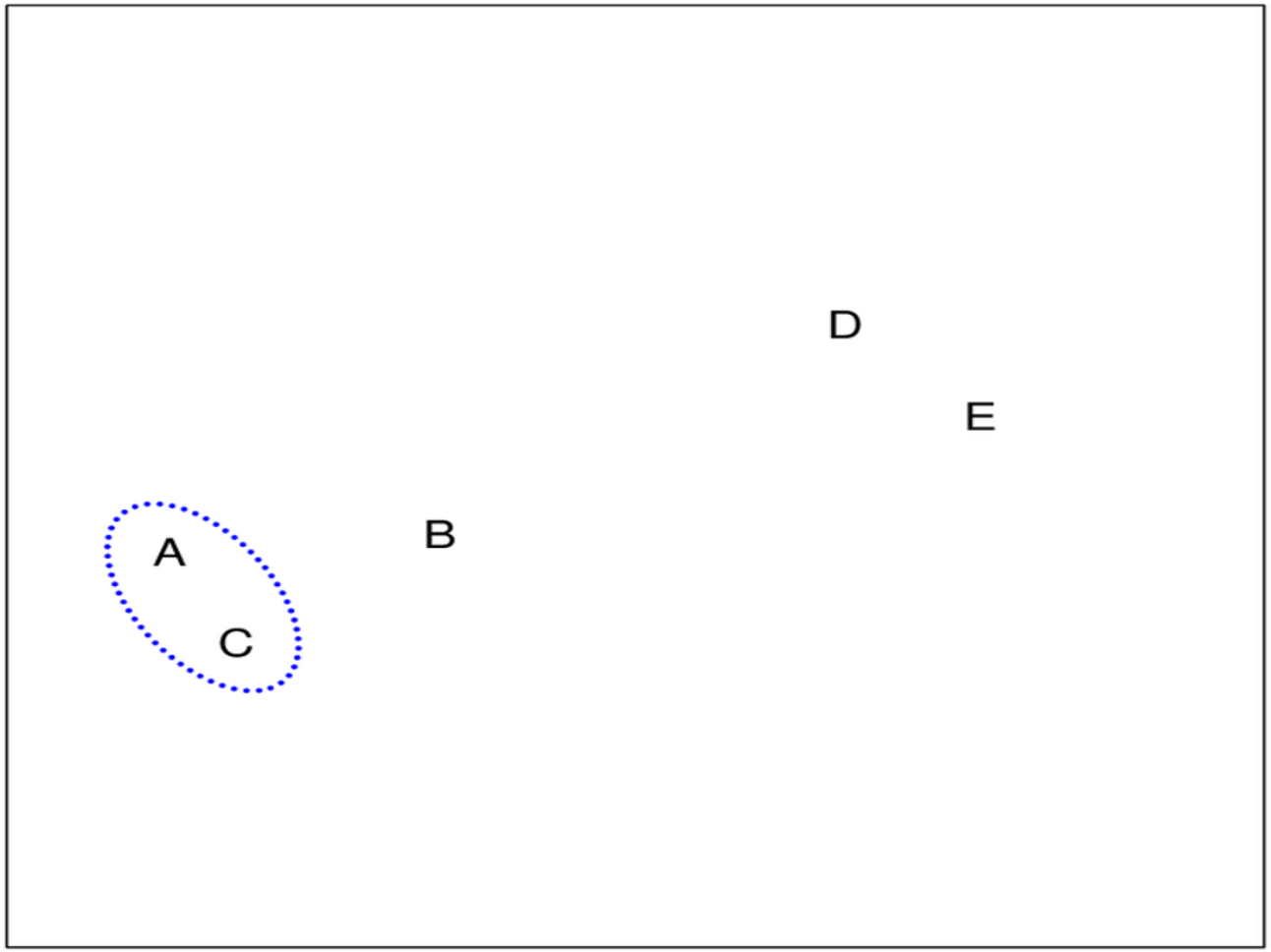
Hierarchical Clustering

- Bottom up or agglomerative clustering
- Hierarchical clustering is the most common type of this form of clustering
- Every element begins in its own cluster, and we take steps to group similar items together
- The algorithm ends when everything is one giant cluster

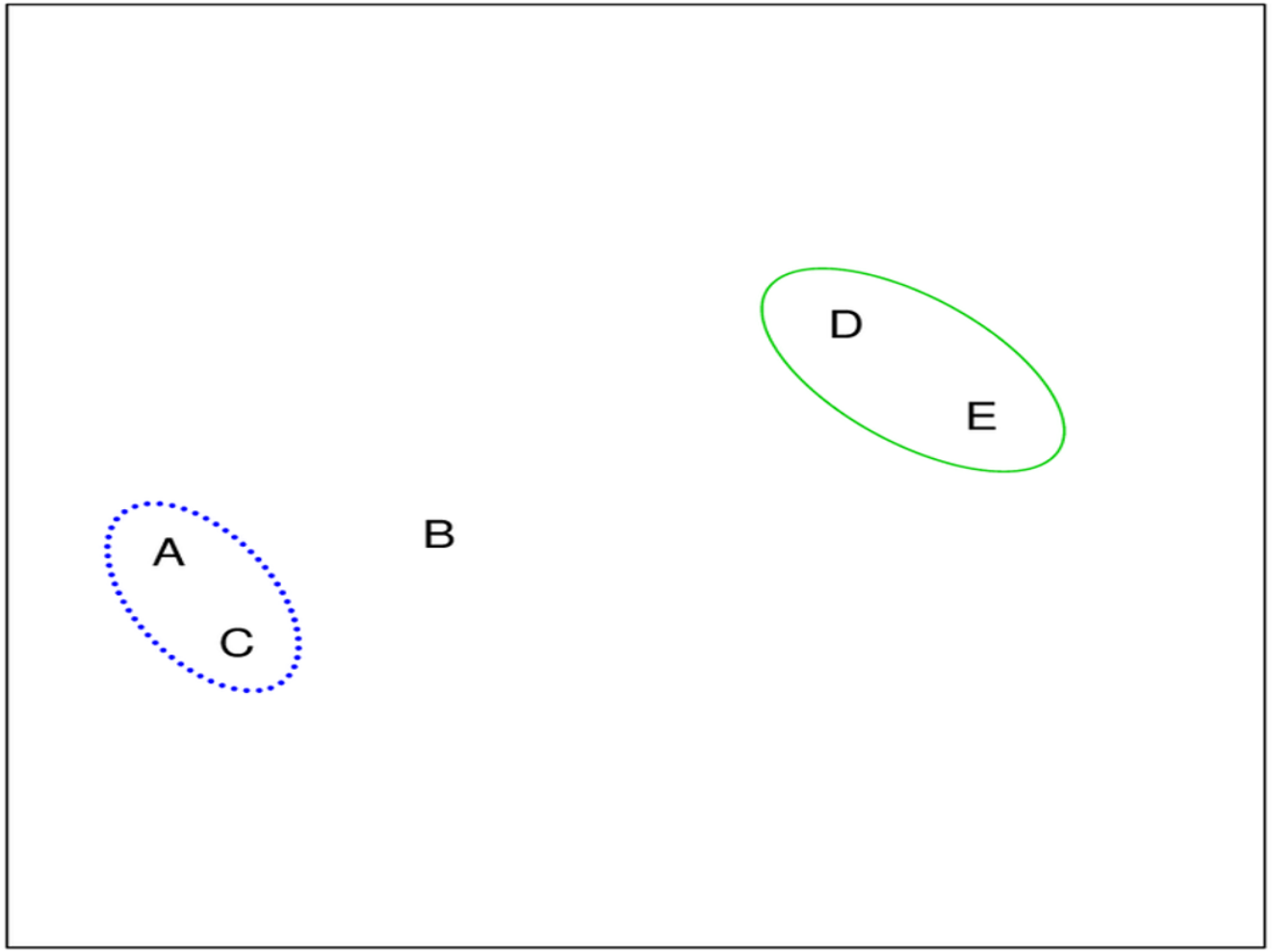
Hierarchical Clustering: Visualization



Group first two closest elements

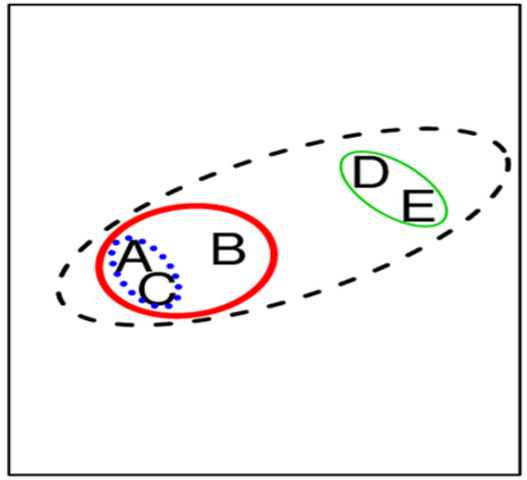


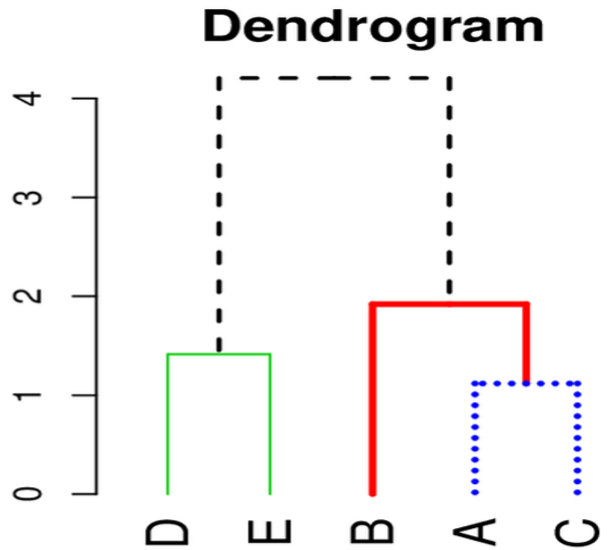
Then the next two



Hierarchical Clustering

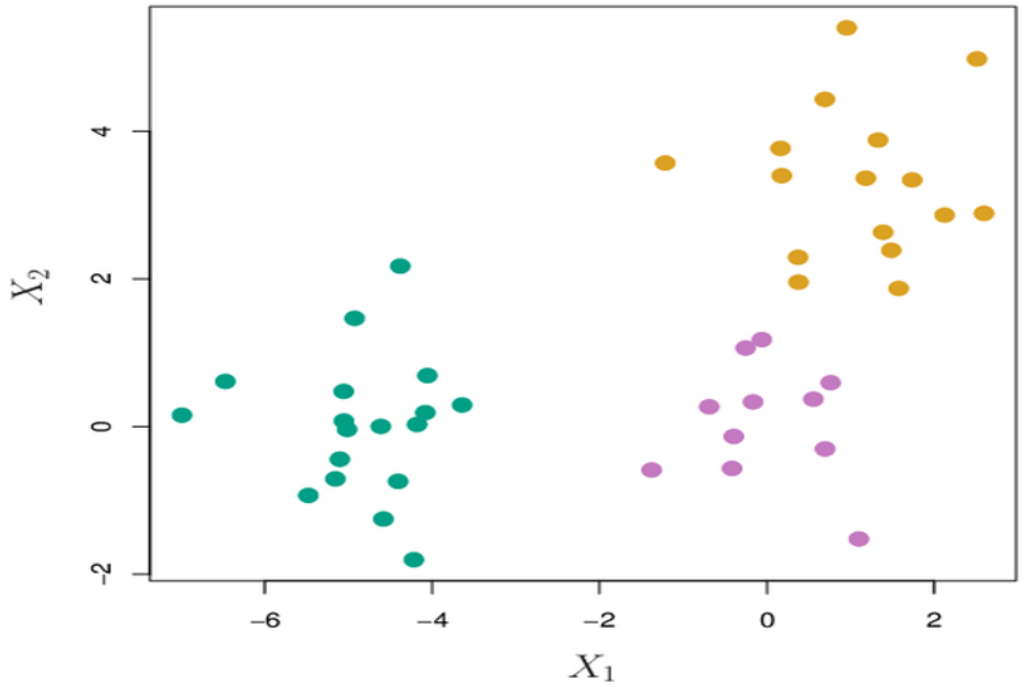
- Start with each point in its own cluster
- Identify the closest two clusters and merge them
- Repeat this process
- End when all points are in a single cluster
- Then define from Dendogram what ideal number of clusters you wish to keep





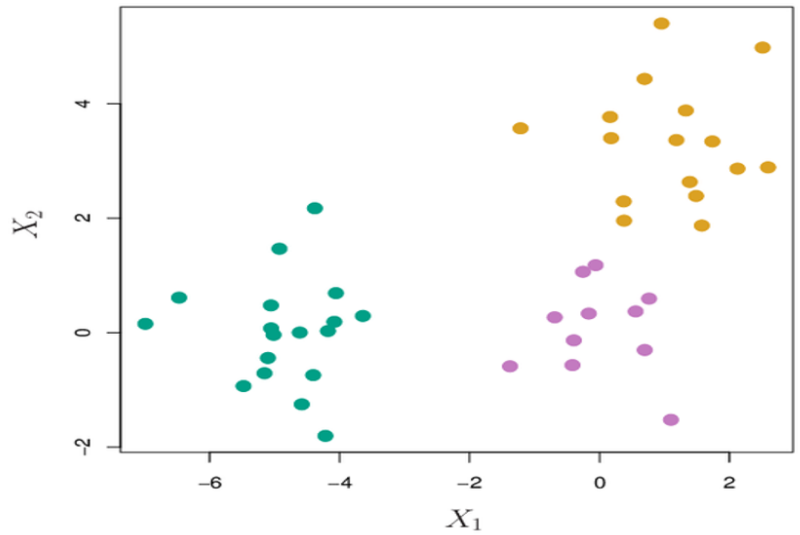
Another example

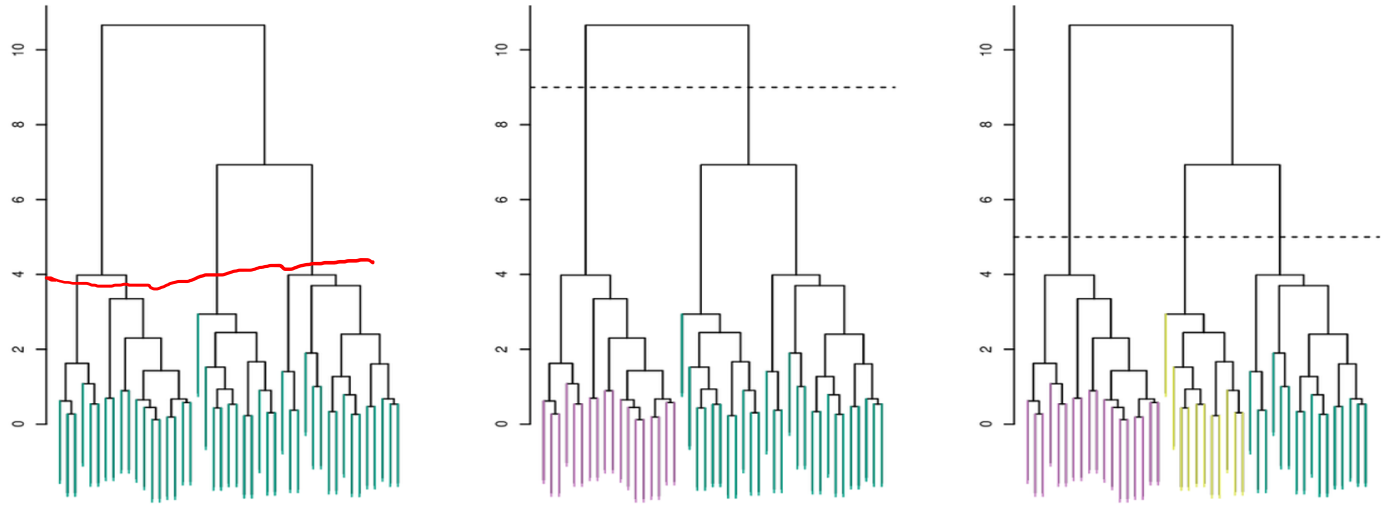
- Take the following example
- In reality it was created with three distributions
- However, you can see how a clustering algorithm will vary when you hide this information



Hierarchical Clustering Rather than K-Means

- This application still naturally illustrates 3 clusters
- However, you can also decide to view it as 2 clusters and as more clusters
- What would the within-cluster variance look if we used k-means?





Hierarchical Clustering: Methods of defining “closest”

- Methods by which we group things are called linkage
- Complete:
  - Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, record the largest of these dissimilarities

Hierarchical Clustering: Methods of defining “closest”

- Methods by which we group things are called linkage
- Complete:
  - Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, record the largest of these dissimilarities
- Single:
  - Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, record the smallest of these dissimilarities

Hierarchical Clustering: Methods of defining “closest”

- Methods by which we group things are called linkage
- Complete:
  - Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, record the largest of these dissimilarities
- Single:
  - Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, record the smallest of these dissimilarities
- Average:
  - Mean inter-cluster dissimilarity
- Centroid:
  - Dissimilarity between the centroid for cluster A (a mean vector), and the centroid for cluster B.

Practical Issues

- We have talked about these as Euclidean distances, what if we use other distance metrics?
- Scaling matters, do we need to standardize features before we cluster?
- How many clusters do we chose?
- What features should we select specifically to drive clusters?

Takeaways

- Goals:
  - Understanding why we use unsupervised learning
  - Understanding the key differences between k-Means, hierarchical clustering, and GMM
  - Next Time: Understanding the Expectation Maximization approach to training