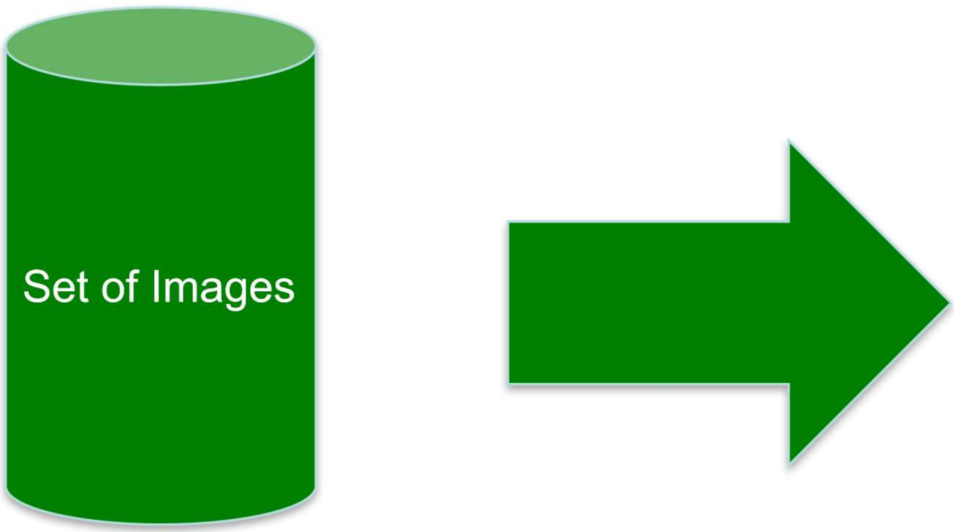
Unsupervised Learning

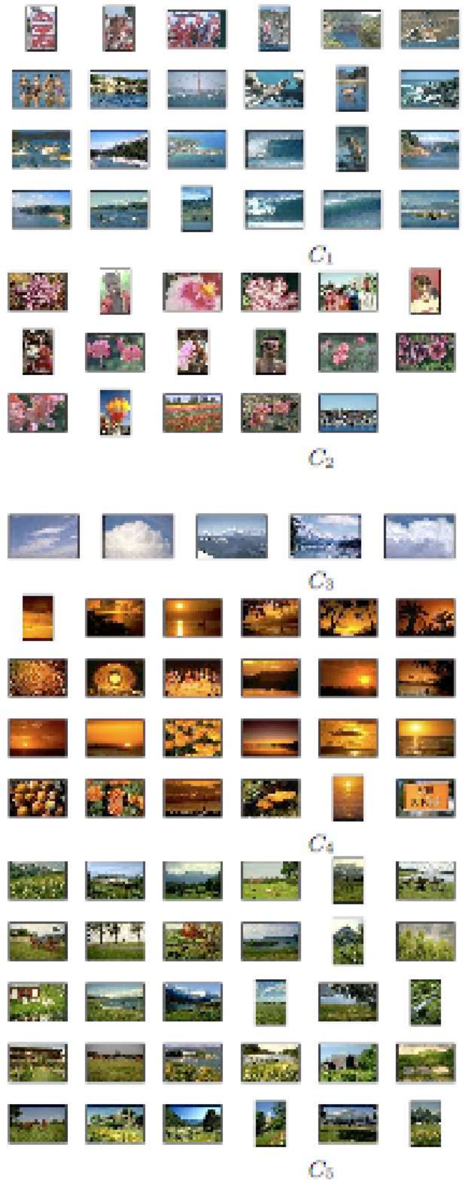
ECE 449

Unsupervised Learning

- Clustering: group similar things
  - Finding structure in data
  - Applications: group search results or customers, find anomalies
- Feature projections: dimensionality reduction
  - Feature reduction that preserves structure in data
  - Applications: improved learning, visualizing data
- General theme: no labels

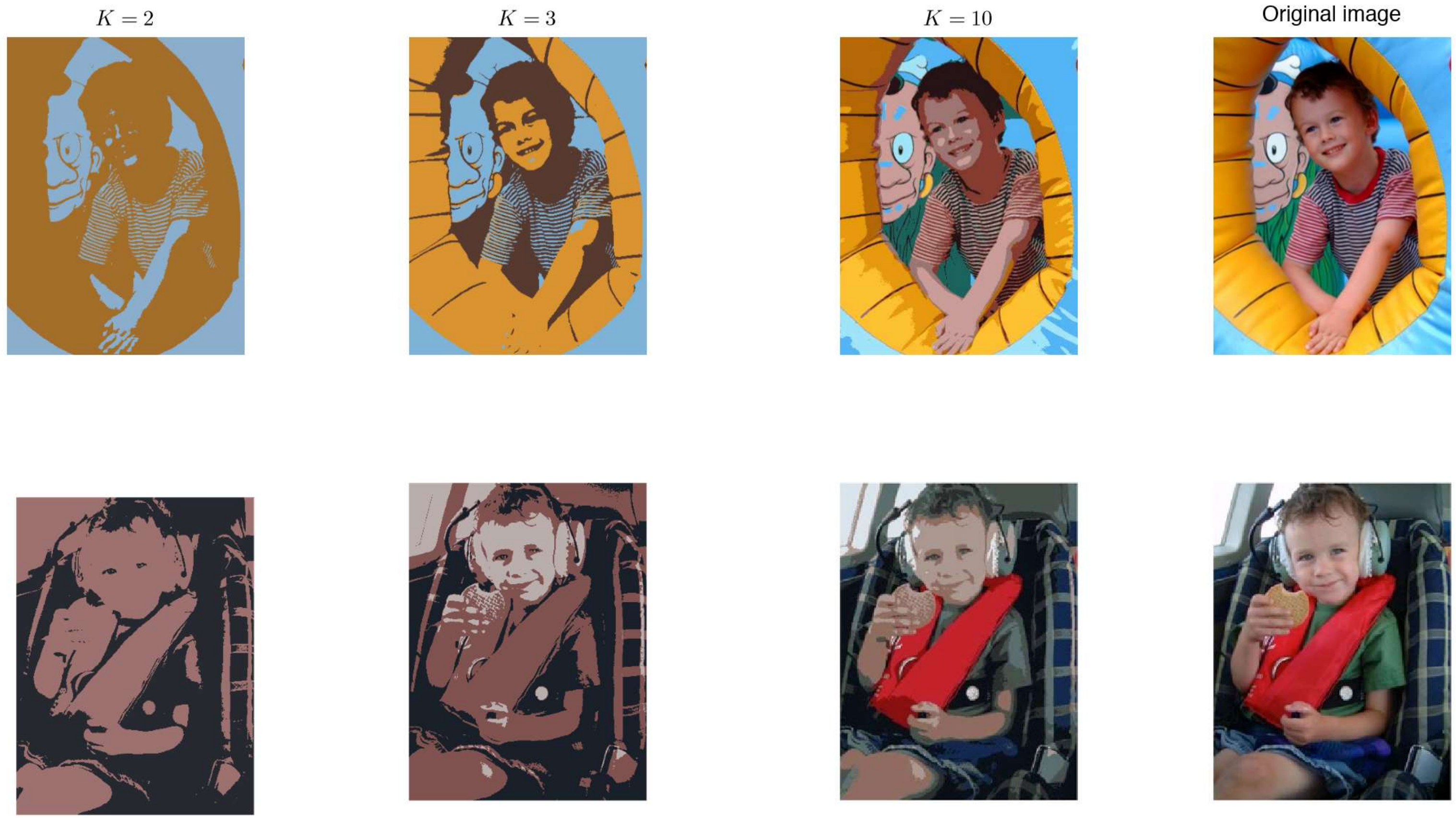
Clustering Images





[Goldberger et al.]

Clustering for Segmentation

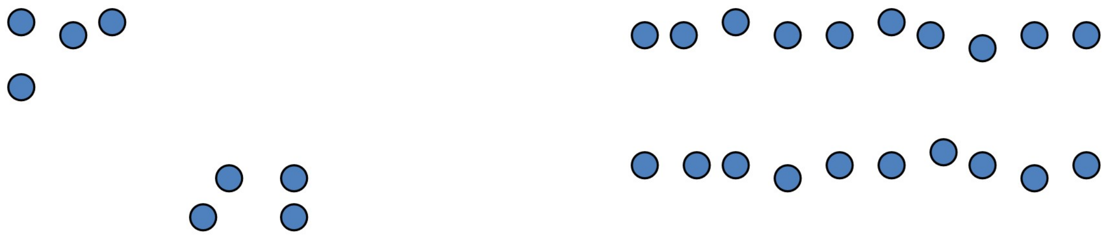


Unsupervised Learning: Clustering

- (Dis)similarity
- Hierarchical clustering
- K-means (partitioning) and its variants
- Soft clustering (EM for GMM)
- Density based clustering
- Evaluating clusters
- Principle components analysis (PCA)

Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



- What could “similar” mean?
  - For two different samples
  - Between a sample and a cluster

What could Similar Mean

- With respect to another sample
  - Small distance
    - Euclidean distance (L2), city block (L1)
  - High match
    - Correlation, cosine distance (equivalent to L2)
    - Feature overlap
- With respect to the cluster
  - Close to (all, some, avg) members of the group

Other Clustering Questions

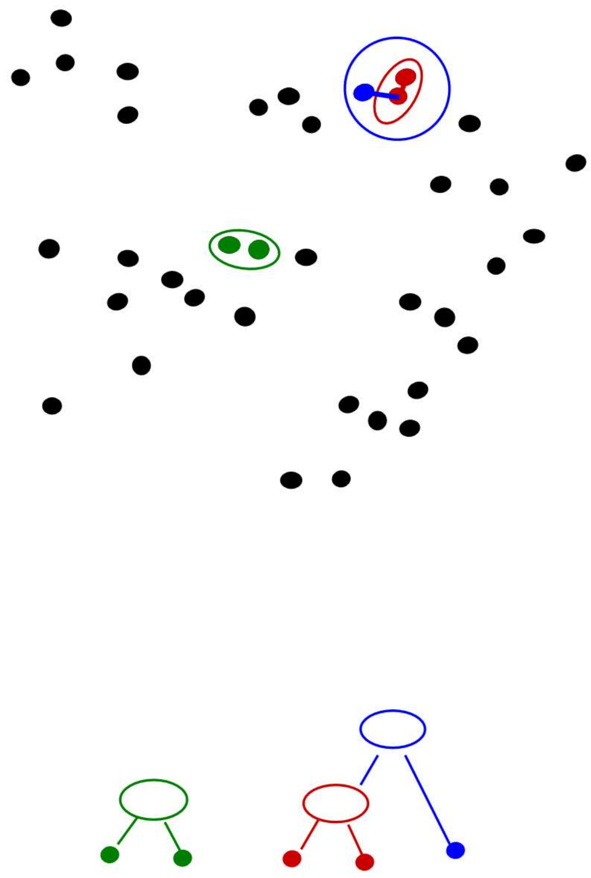
- Will the algorithm converge?
- Will it find the true patterns in the data?
- How many clusters to pick?
- How good are the clusters?

Hierarchical Clustering

- Agglomerative
  - Iteratively group samples
  - Good for different linkages, but expensive
- Divisive clustering
  - Iteratively divide samples (iterative K-means)
  - Better decisions for a small number of clusters
- Both produce a hierarchy –good for when you don’t know the # of clusters

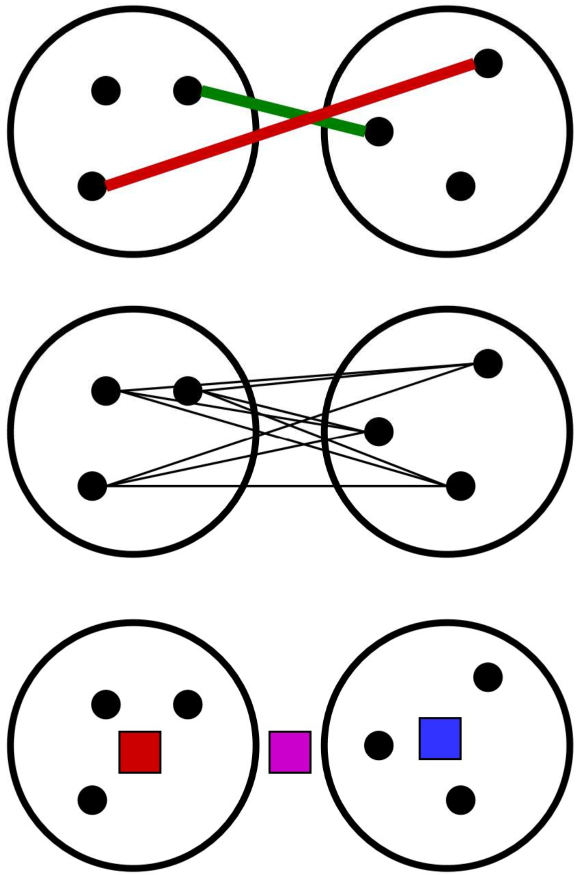
Agglomerative Clustering

- Agglomerative clustering
  - First merge very similar instances
  - Incrementally build larger clusters out of smaller clusters
- Algorithm
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
- Stop when there's only one cluster left
- Produces not one clustering, but a family of clusterings represented by a dendrogram

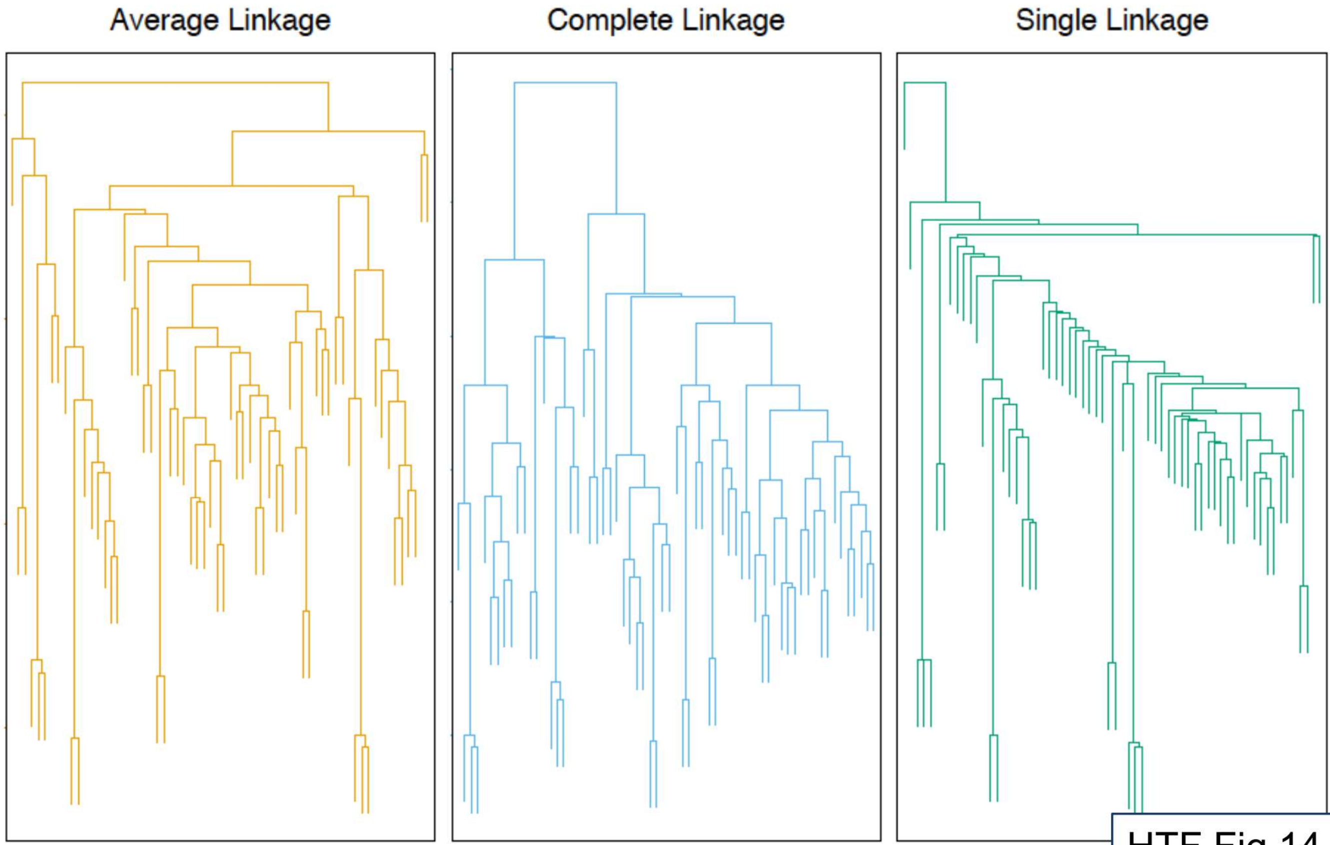


Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?
- Many options:
  - Closest pair (single-link clustering)
  - Farthest pair (complete-link clustering)
  - Average of all pairs
  - Ward’s method (min variance, like k-means)
- Different choices create different clustering behaviors



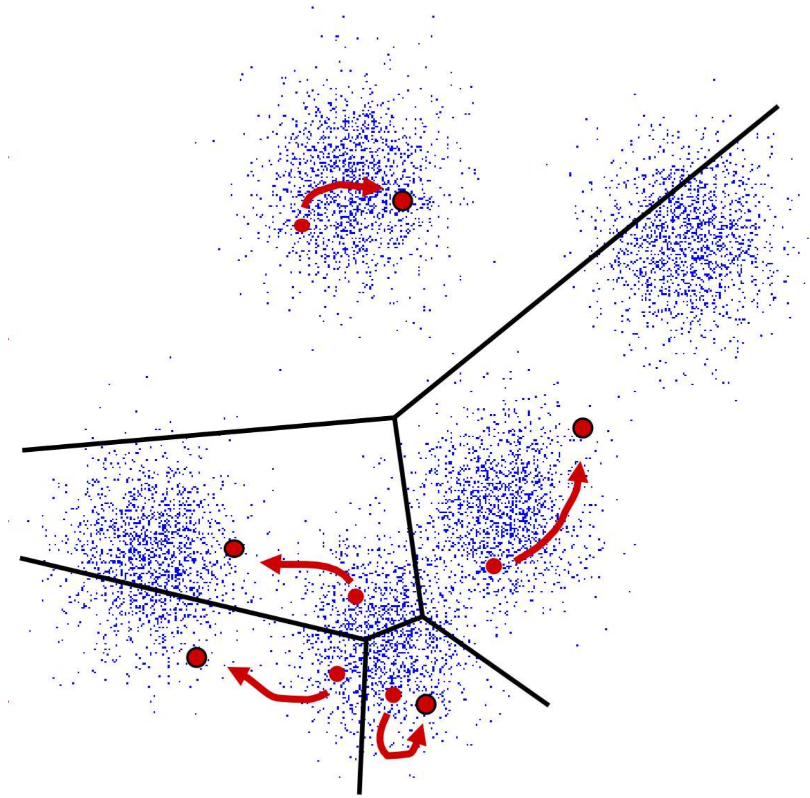
Agglomerative Clustering: Diff Objectives



HTF Fig 14.13

K-Means

- An iterative clustering algorithm
  - Pick K random points as cluster centers (means), $c^1 \ldots c^k$
  - Alternate:
    - Assign each example x^i to the mean c^j that is closest to it
    - Set each mean c^j to the average of its assigned points
  - Stop when no points' assignments change



K-Means

- Data: $\{x^j \mid j=1..n\}$
- An iterative clustering algorithm
  - Pick K random cluster centers, $c^1 \ldots c^k$
  - For t=1..T: [or, stop if assignments don't change]
    - for j = 1.. n: [recompute cluster assignments]

$$a^j = \arg\min_i \operatorname{dist}(x^j, c^i)$$

    - for j= 1…k: [recompute cluster centers]

$$c^j = \frac{1}{|\{i|a^i = j\}|} \sum_{\{i|a^i=j\}} x^i$$

K-Means as Optimization

- Consider the total distance to the means

$$\underbrace{L\big(\{x^i\}}_{\text{points}},\underbrace{\{a^i\}}_{\text{assignments}},\underbrace{\{c^k\}}_{\text{means}}\big) = \sum_i \text{dist}(x^i, c^{a^i})$$

- Two stages each iteration:
    - Update assignments: fix means c, change assignments a
    - Update means: fix assignments a, change means c

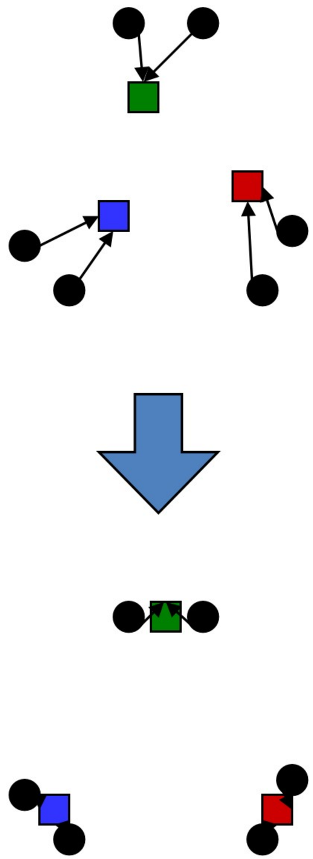
Phase I: Update Assignments

- For each point, re-assign to closest mean

$$a^i = \arg\min_j \operatorname{dist}\left(x^i, c^j\right)$$

- Can only decrease total distance L

$$L\left(\{x^i\}, \{a^i\}, \{c^k\}\right) = \sum_i \operatorname{dist}(x^i, c^{a^i})$$



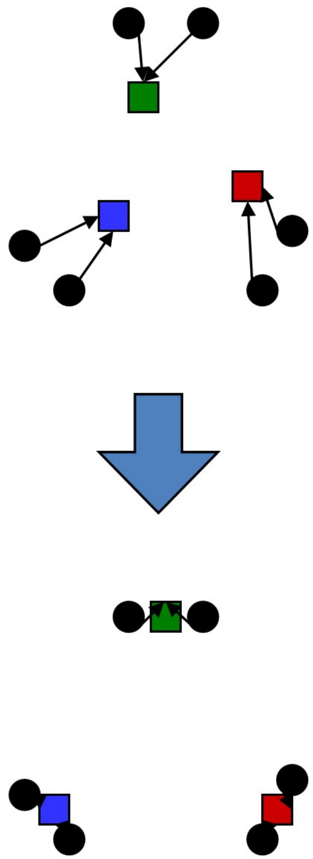
Phase II: Update Means

- Move each mean to the average of its assigned points

$$c^j = \frac{1}{|\{i|a^i = j\}|} \sum_{\{i|a^i=j\}} x^i$$

- Also can only decrease total distance

- Fun fact: the point y with minimum squared Euclidean distance to a set of points {x} is their mean

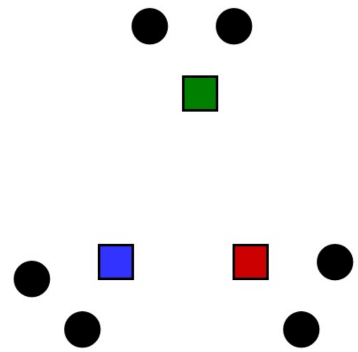


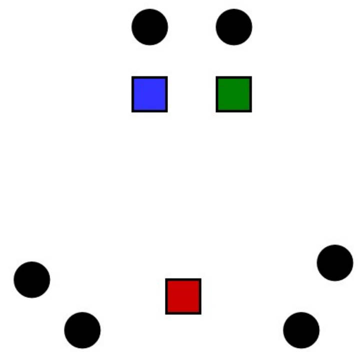
Questions

- Optimal solution?

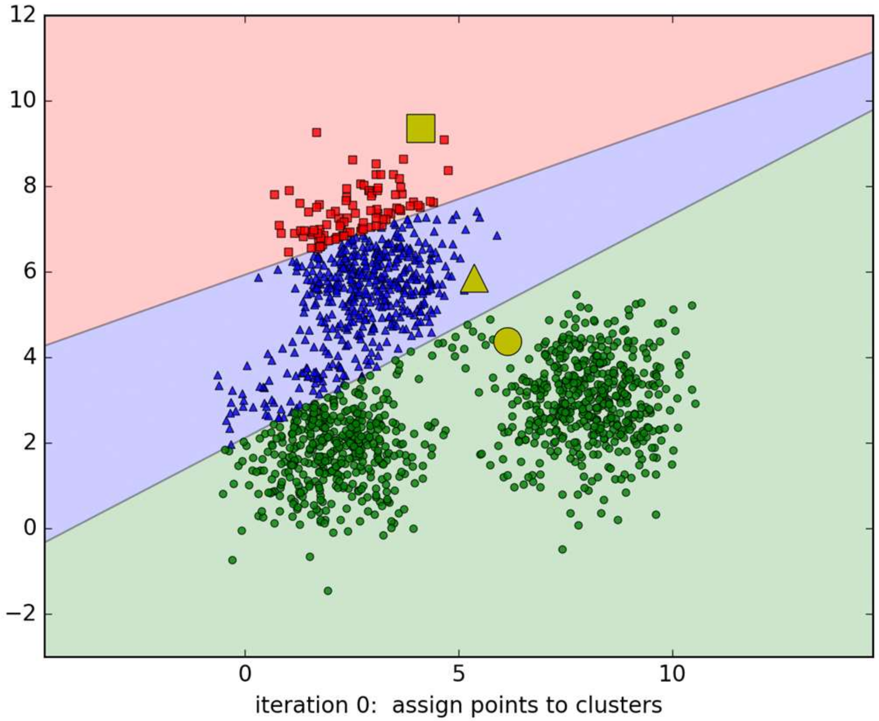
Initialization

- K-means is non-deterministic
  - Requires initial means
  - It does matter what you pick!
- Various schemes for preventing this kind of thing:
  - Multiple random starts
  - Divisive clustering

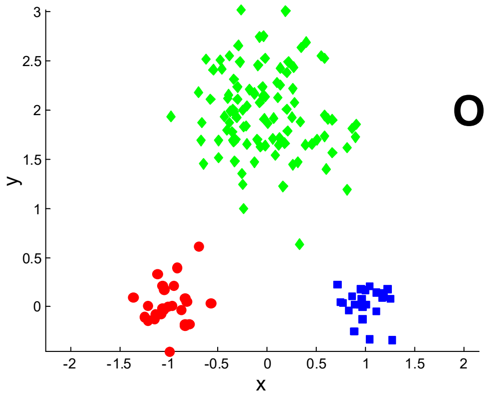




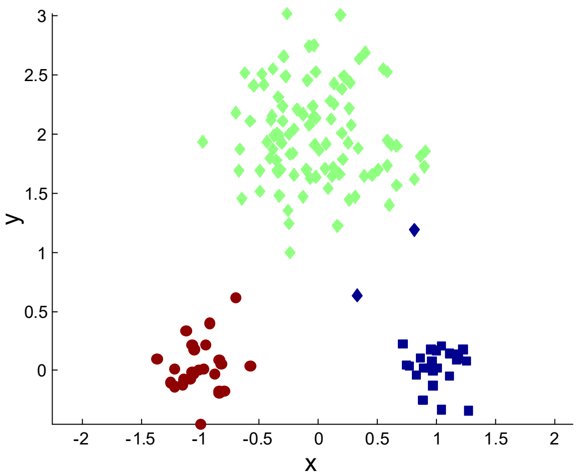
Examples



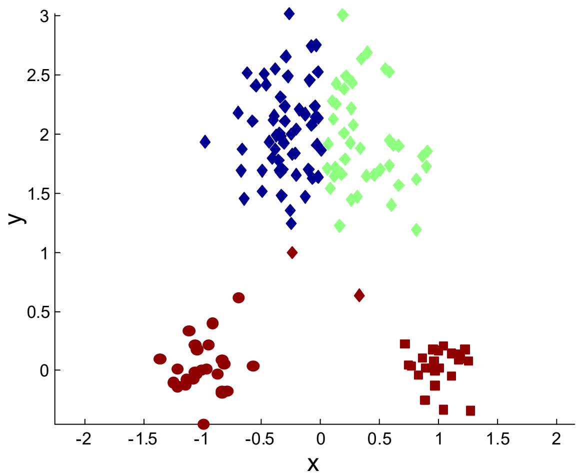
Importance of Choosing Initial Centroids



Original Points

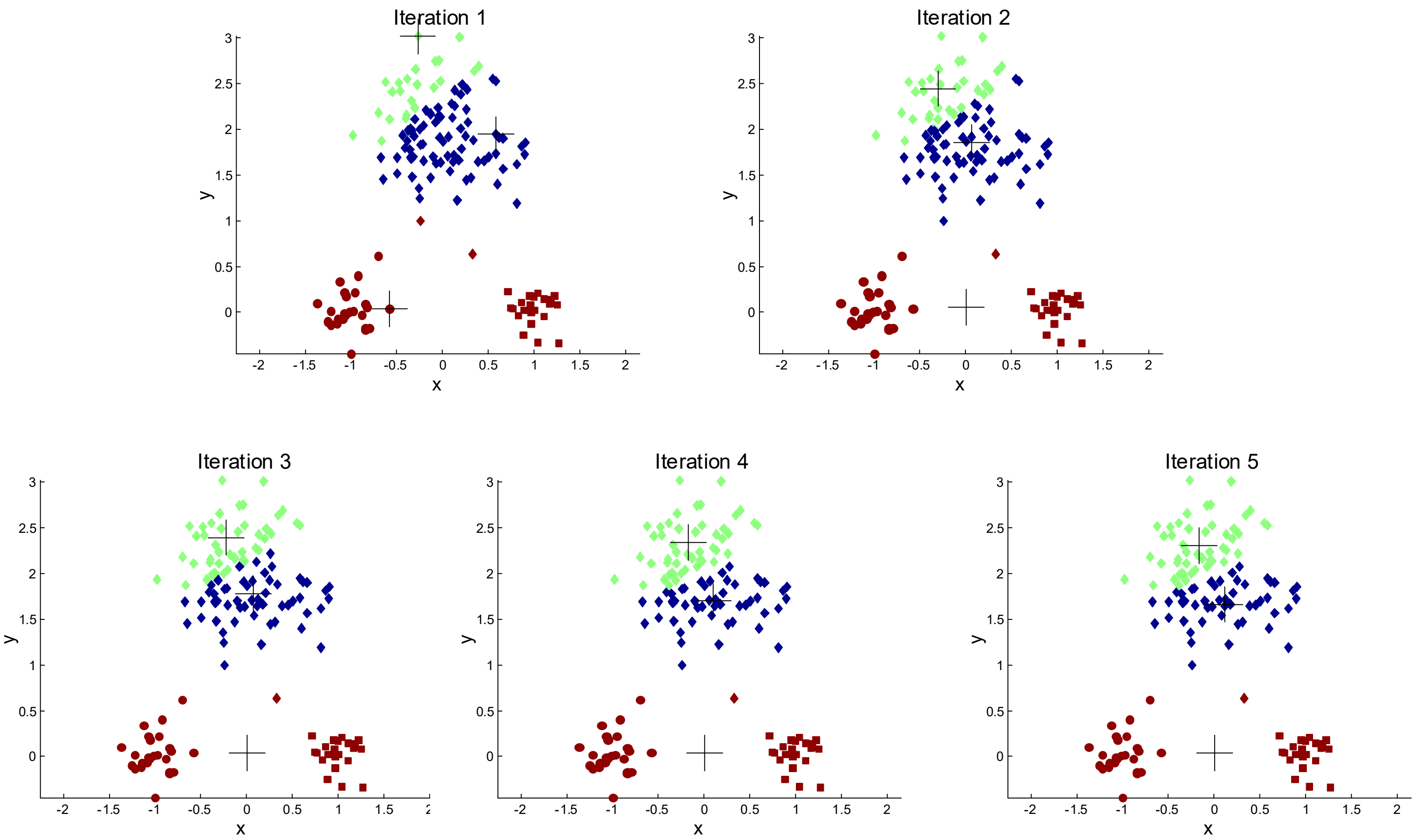


Optimal Clustering



Sub-optimal Clustering

Importance of Choosing Initial Centroids



Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Use some strategy to select the k initial centroids and then select among these initial centroids
  - Select most widely separated
    - K-means++ is a robust way of doing this selection
  - Use hierarchical clustering to determine initial centroids

K-means++

- This approach can be slower than random initialization, but very consistently produces better results in terms of SSE
- To select a set of initial centroids, C, perform the following

Select an initial point at random to be the first centroid

For k – 1 steps

For each of the N points, x_i, $1 \leq i \leq N$, find the minimum squared distance to the currently selected centroids, $C_1, \ldots, C_j,$ $1 \leq j < k$, i.e., $\min_j d^2(C_j, x_i)$

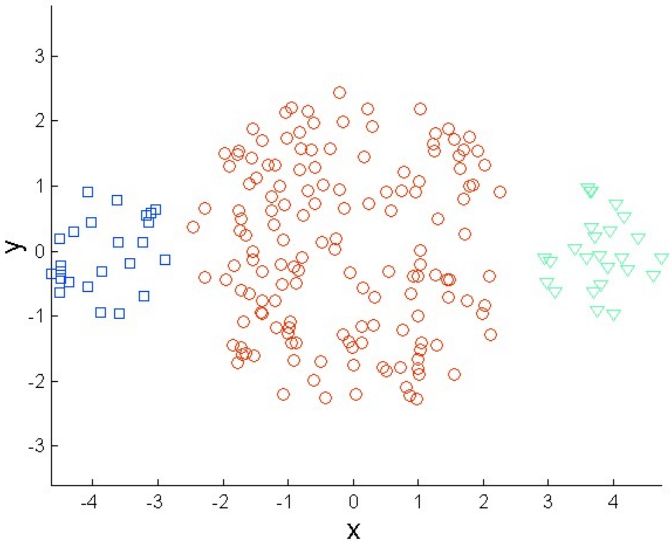
Randomly select a new centroid by choosing a point with probability proportional to $\frac{\min_j d^2(Cj, xi)}{\sum_i \min_j d^2(Cj, xi)}$

End For

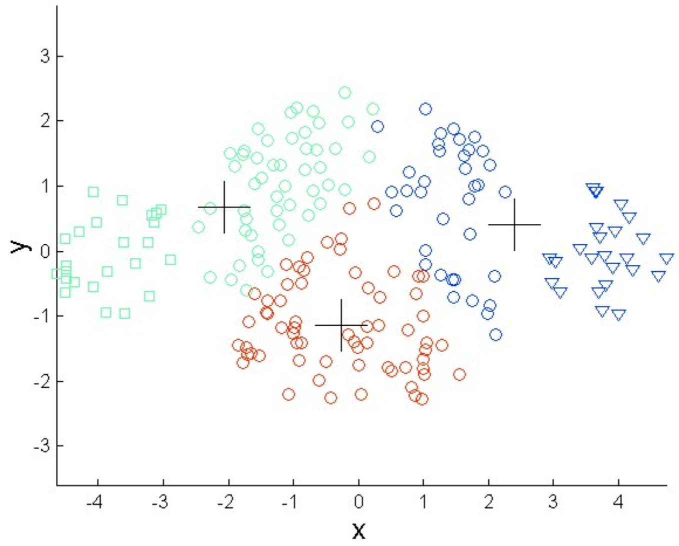
Limitations

- K-means has problems when clusters are of differing
    - Sizes
    - Densities
    - Non-globular shapes

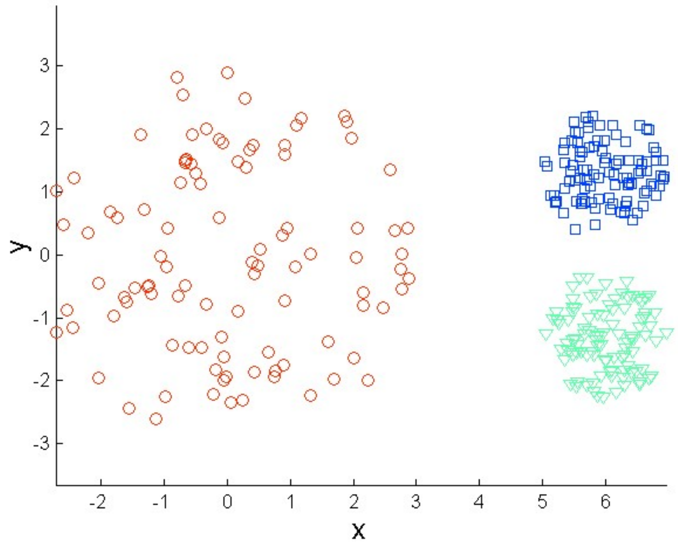
- K-means has problems when the data contains outliers.



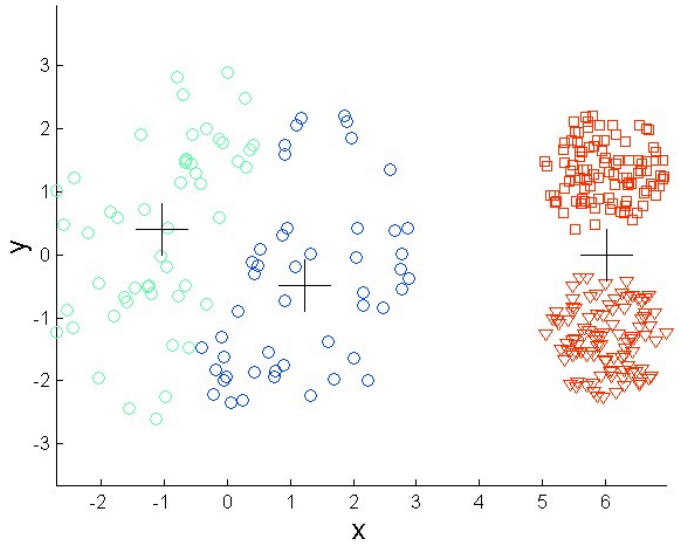
Original Points



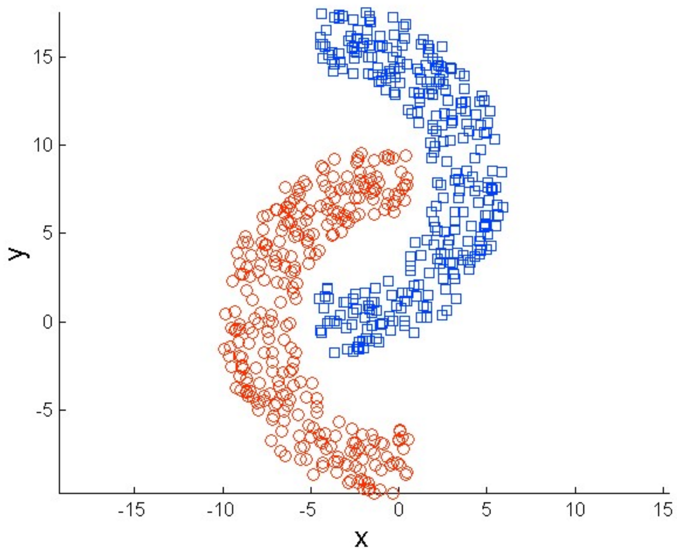
K-means (3 Clusters)



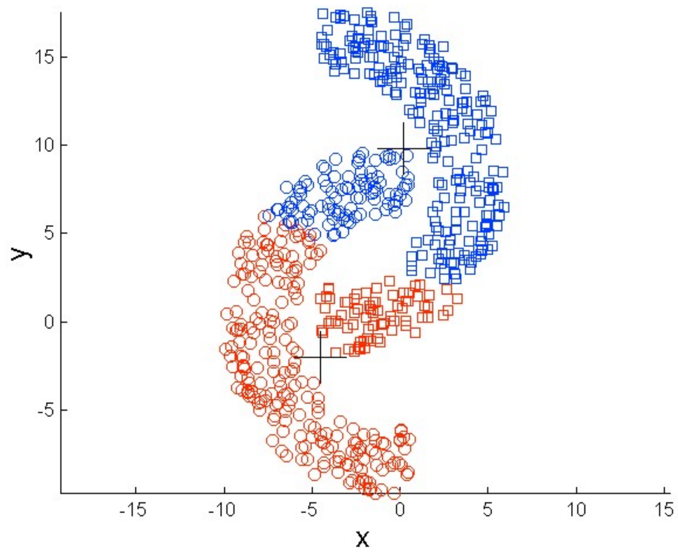
Original Points



K-means (3 Clusters)



Original Points



K-means (2 Clusters)