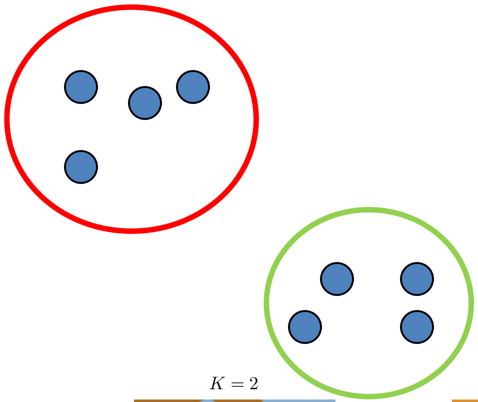
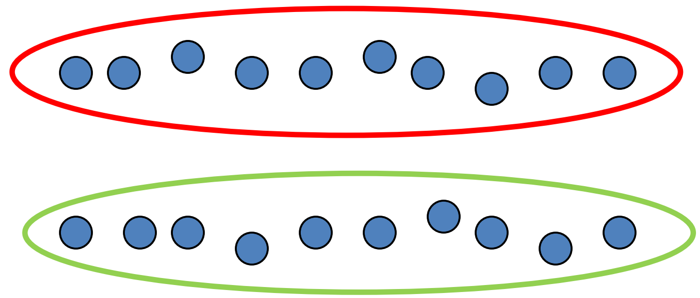
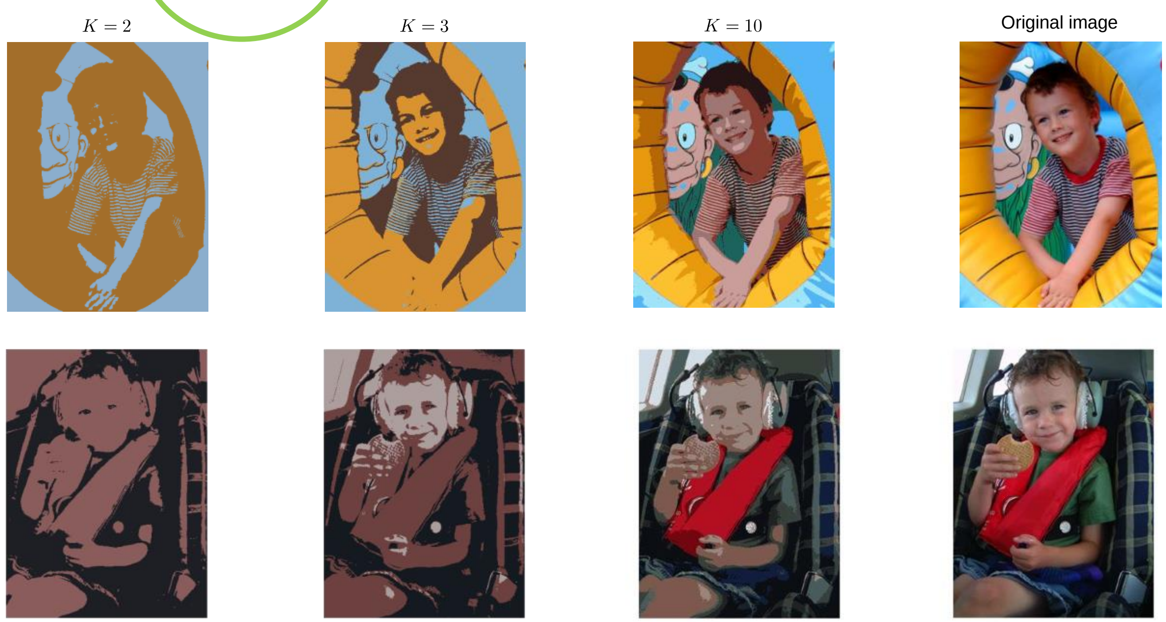
CSE 446
Clustering

Clustering

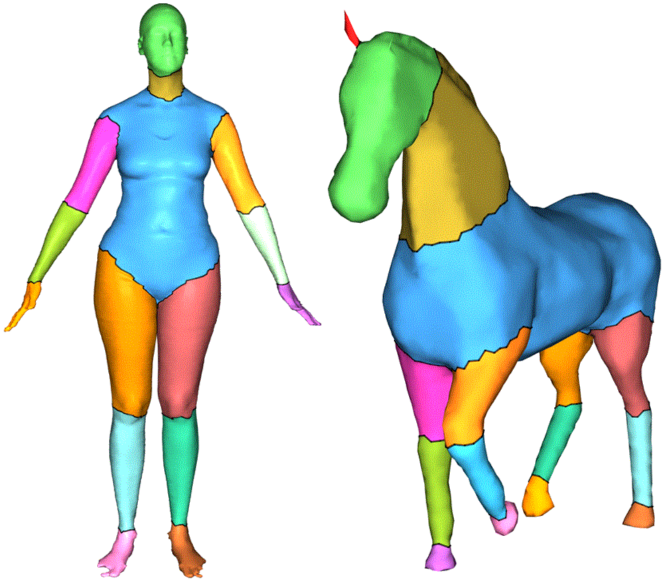
- Basic idea: group together similar instances
- Example: 2D point patterns

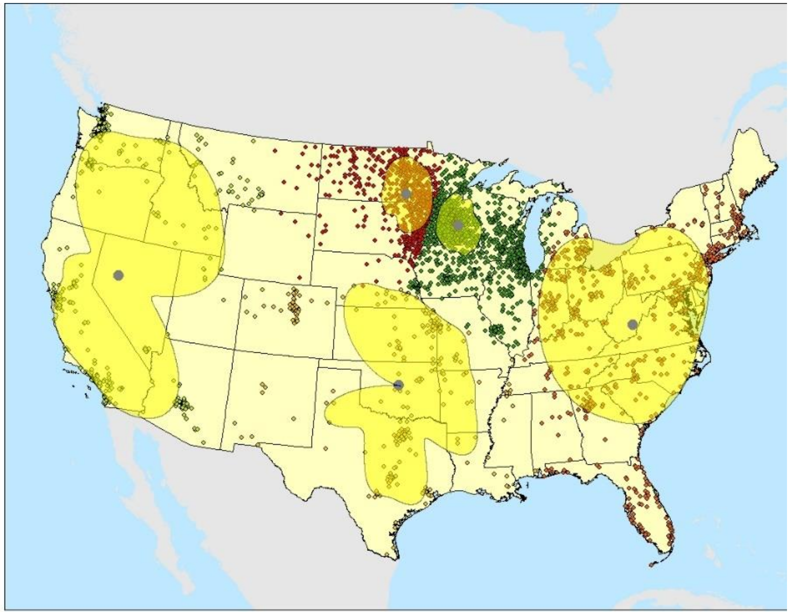


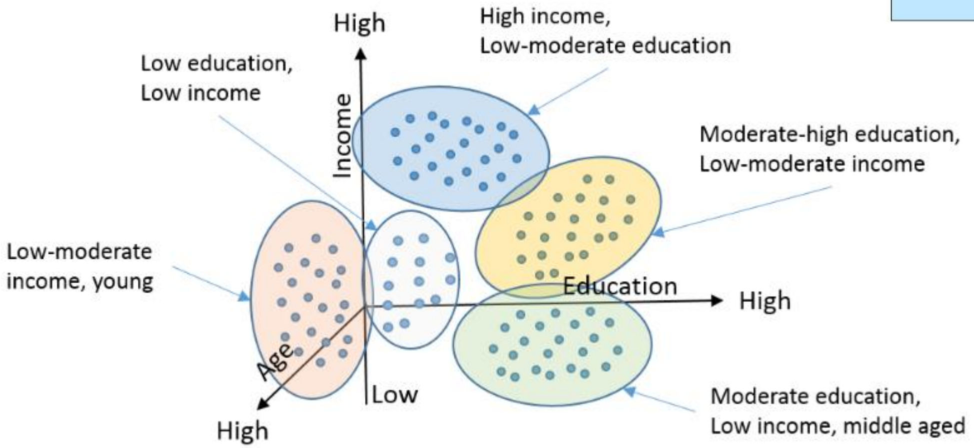




More Clustering Examples

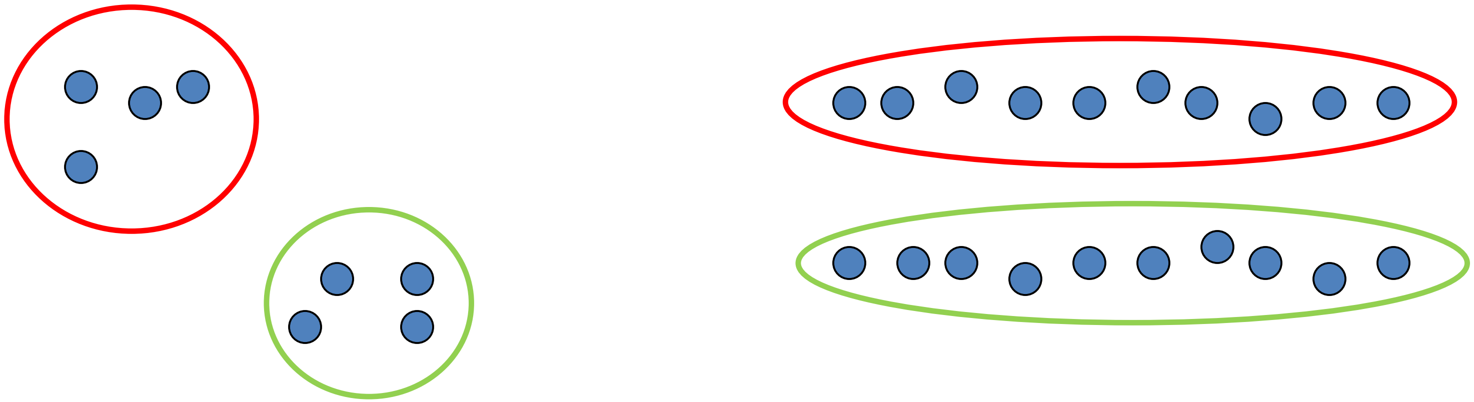






Clustering

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



- What could "similar" mean?
  - One option: small (squared) Euclidean distance

$$\mathcal{L}(C_1,\ldots,C_K)=\sum_{k=1}^{K}\frac{\sum_{\mathbf{x}_i\in C_k,\mathbf{x}_j\in C_k}\frac{1}{2}\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\|\{i:\mathbf{x}_i\in C_k\}\|}$$

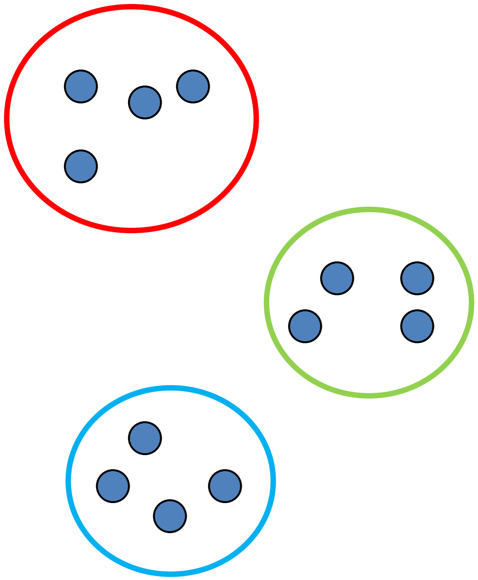
Hypothesis Space?

clusters: $C_1, \ldots, C_K$

points: $\mathbf{x}_1, \ldots, \mathbf{x}_N$

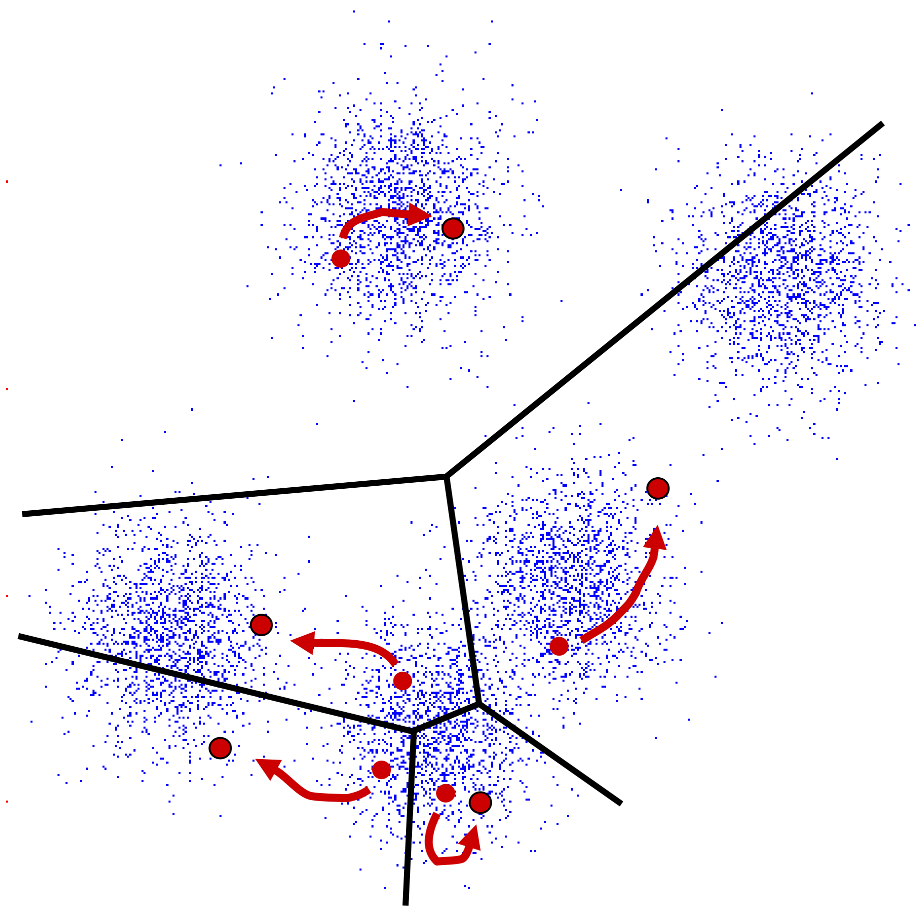
cluster labels: $y_1, \ldots, y_N$

$$y_i = j \Leftrightarrow \mathbf{x}_i \in C_j$$

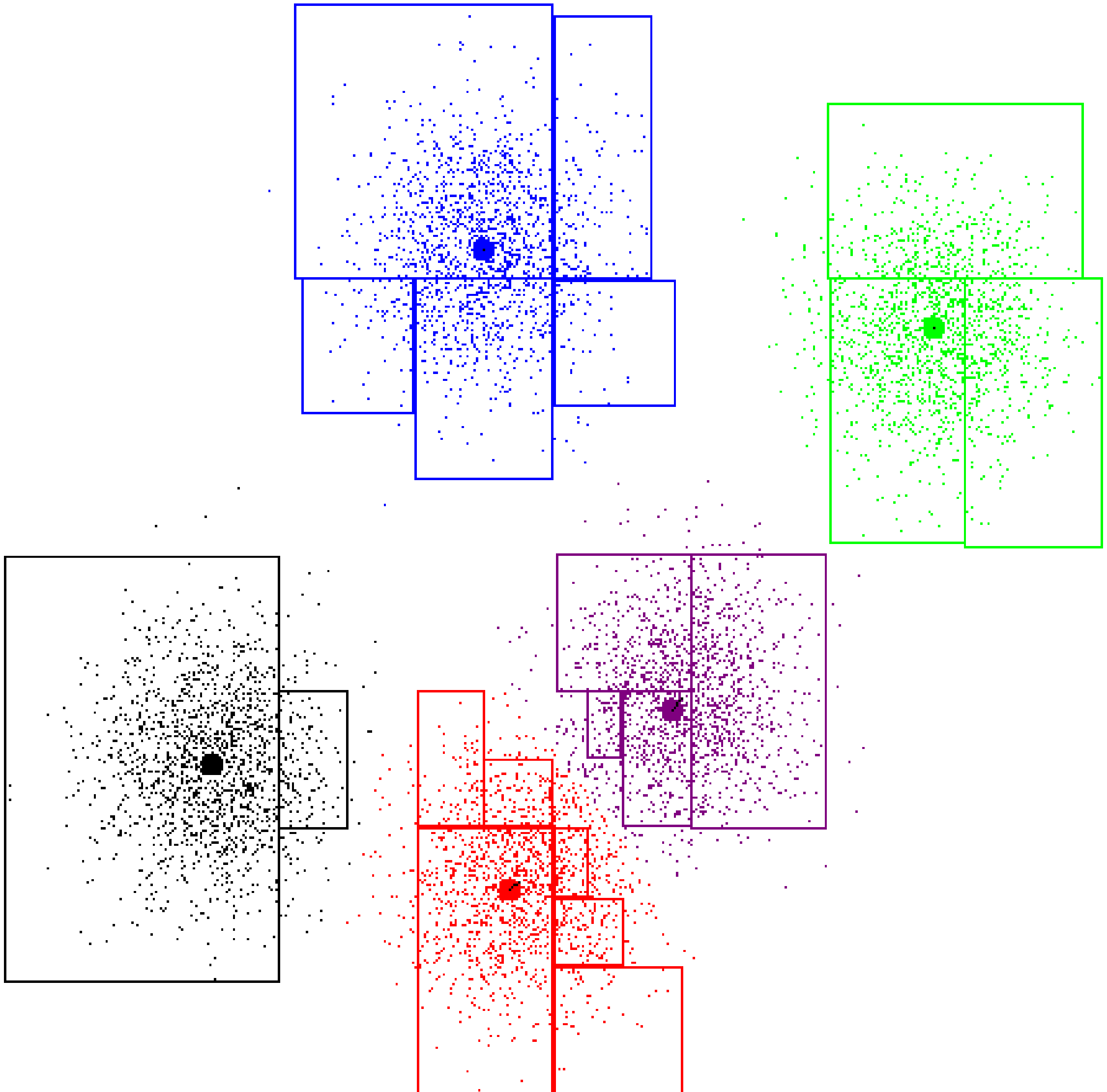


K-Means

- An iterative clustering algorithm
  - Assign points to clusters randomly
  - Alternate:
    - Set each mean c^j to the average of its assigned points
    - Assign each example x^i to the mean c^j that is closest to it
  - Stop when no points' assignments change



K-Means Example



K-Means

Algorithm 1 K-means clustering

Initialize cluster assignments y_i with random integers in $\{1, \ldots, K\}$
while not converged **do**
  $\mathbf{c}_k \leftarrow \frac{1}{\|i : y_i = k\|} \sum_{i:y_i=k} \mathbf{x}_i$ (average all points with $y_i = k$)
  $y_i \leftarrow \arg\min_k \|\mathbf{x}_i - \mathbf{c}_k\|^2$ (assign each point to nearest cluster)
end while

K-Means

Algorithm 1 K-means clustering

Initialize cluster assignments y_i with random integers in $\{1, \ldots, K\}$
while not converged **do**
    $\mathbf{c}_k \leftarrow \frac{1}{\|i:y_i=k\|} \sum_{i:y_i=k} \mathbf{x}_i$ (average all points with $y_i = k$)
    $y_i \leftarrow \arg\min_k \|\mathbf{x}_i - \mathbf{c}_k\|^2$ (assign each point to nearest cluster)
end while

original objective: $$\mathcal{L}(C_1, \ldots, C_K) = \sum_{k=1}^{K} \frac{\sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_k} \frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\|\{i : \mathbf{x}_i \in C_k\}\|}$$

K-means objective: $$\hat{\mathcal{L}}(y_1, \ldots, y_N, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i:y_i=k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

K-Means

Algorithm 1 K-means clustering

1: Initialize cluster assignments y_i with random integers in $\{1, \ldots, K\}$
2: **while** not converged **do**
3: $\quad \mathbf{c}_k \leftarrow \frac{1}{\|i:y_i=k\|} \sum_{i:y_i=k} \mathbf{x}_i$ (average all points with $y_i = k$)
4: $\quad y_i \leftarrow \arg\min_k \|\mathbf{x}_i - \mathbf{c}_k\|^2$ (assign each point to nearest cluster)
5: **end while**

$$\hat{\mathcal{L}}(y_1, \ldots, y_N, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i:y_i=k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

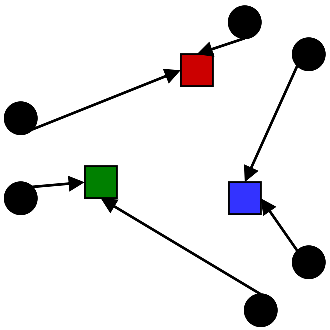
$$\frac{d\hat{\mathcal{L}}}{d\mathbf{c}_k} =$$

K-Means

Algorithm 1 K-means clustering

Initialize cluster assignments y_i with random integers in $\{1, \ldots, K\}$
while not converged **do**
    $\mathbf{c}_k \leftarrow \frac{1}{\|i:y_i=k\|} \sum_{i:y_i=k} \mathbf{x}_i$ (average all points with $y_i = k$)
    $y_i \leftarrow \arg\min_k \|\mathbf{x}_i - \mathbf{c}_k\|^2$ (assign each point to nearest cluster)
end while

$$\hat{\mathcal{L}}(y_1, \ldots, y_N, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i:y_i=k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$



What about original objective?

original objective: $$\mathcal{L}(C_1, \ldots, C_K) = \sum_{k=1}^{K} \frac{\sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_k} \frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\|\{i : \mathbf{x}_i \in C_k\}\|}$$

K-means objective: $$\hat{\mathcal{L}}(y_1, \ldots, y_N, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i : y_i = k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

$$\sum_{k=1}^{K} \frac{\sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_k} \frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\|\{i : y_i = k\}\|} =$$

What about original objective?

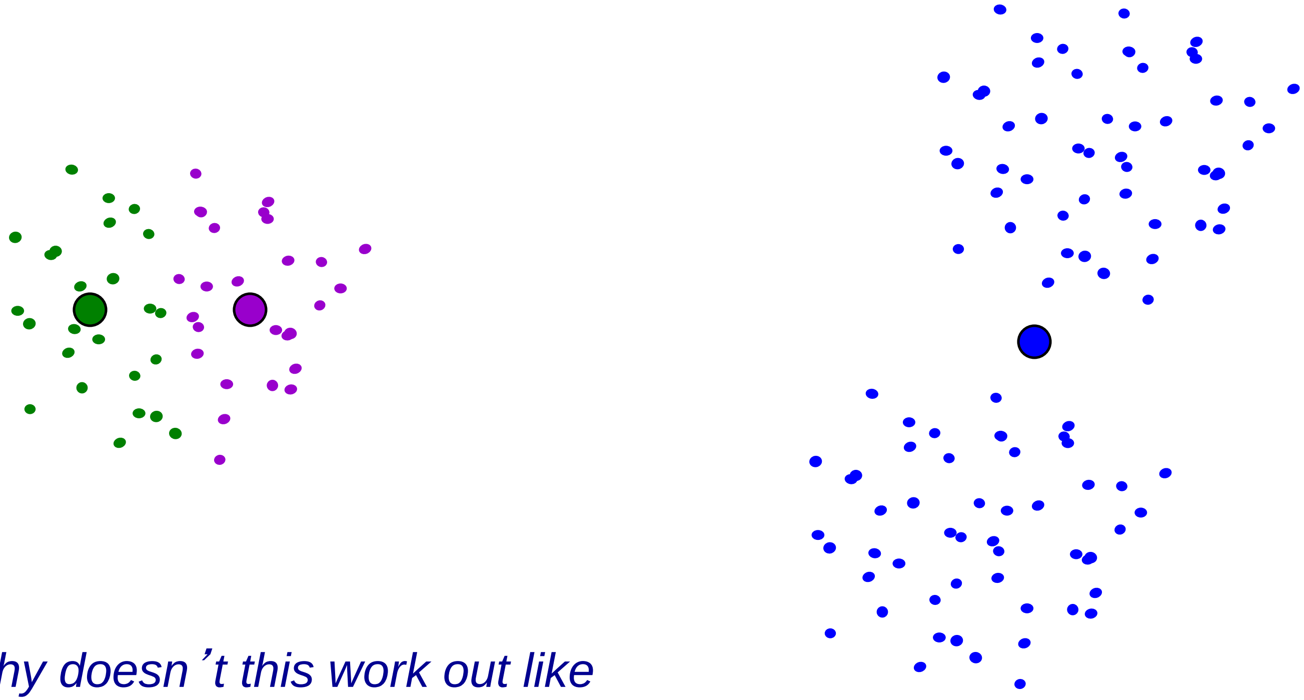
original objective: $$\mathcal{L}(C_1, \ldots, C_K) = \sum_{k=1}^{K} \frac{\sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_k} \frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\|\{i : \mathbf{x}_i \in C_k\}\|}$$

K-means objective: $$\hat{\mathcal{L}}(y_1, \ldots, y_N, \mathbf{c}_1, \ldots, \mathbf{c}_K) = \sum_{k=1}^{K} \sum_{i:y_i=k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

$$\sum_{k=1}^{K} \frac{\sum_{\mathbf{x}_i \in C_k, \mathbf{x}_j \in C_k} \frac{1}{2} \|\mathbf{x}_i - \mathbf{x}_j\|^2}{\|\{i : y_i = k\}\|} = \sum_{k=1}^{K} \sum_{i:y_i=k} \|\mathbf{x}_i - \mathbf{c}_k\|^2$$

K-Means Getting Stuck

- A local optimum:



Why doesn’t this work out like the earlier example, with the purple taking over half the blue?

K-Means Questions

- Will K-means converge?
  - To a global optimum?
- Will it always find the true patterns in the data?
  - If the patterns are very very clear?
- Will it find something interesting?

K-Means in Practice

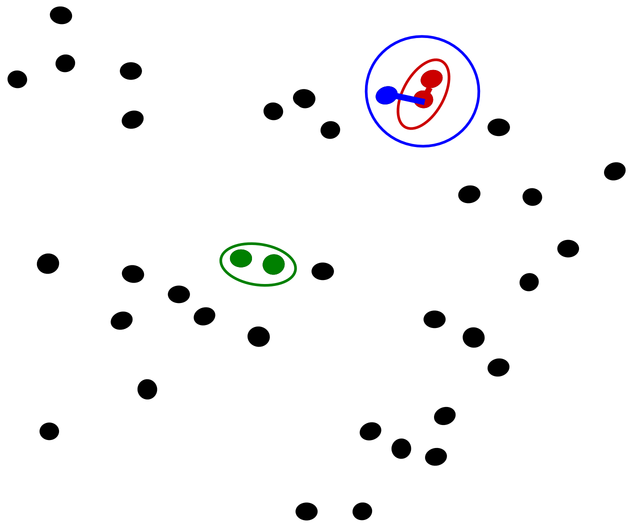
- Avoiding getting stuck
  - Random restarts
  - Take restart with best objective value
- Better initialization
  - Kmeans++
    - 1 Choose first centroid to be a random datapoint x
    - 2 For each datapoint, compute distance to nearest centroid so far
    - 3 Choose next center randomly among the datapoints, but weight the choice by the squared distance to the nearest centeroid
    - 4 Repeat steps 2 and 3 until all centeroids are chosen

K-Means Questions

- Will K-means converge?
  - To a global optimum?
- Will it always find the true patterns in the data?
  - If the patterns are very very clear?
- Will it find something interesting?
- How to choose number of clusters?

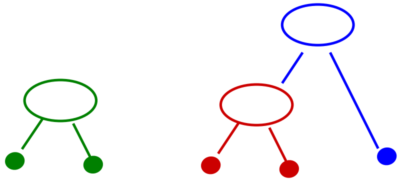
Agglomerative Clustering

- 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

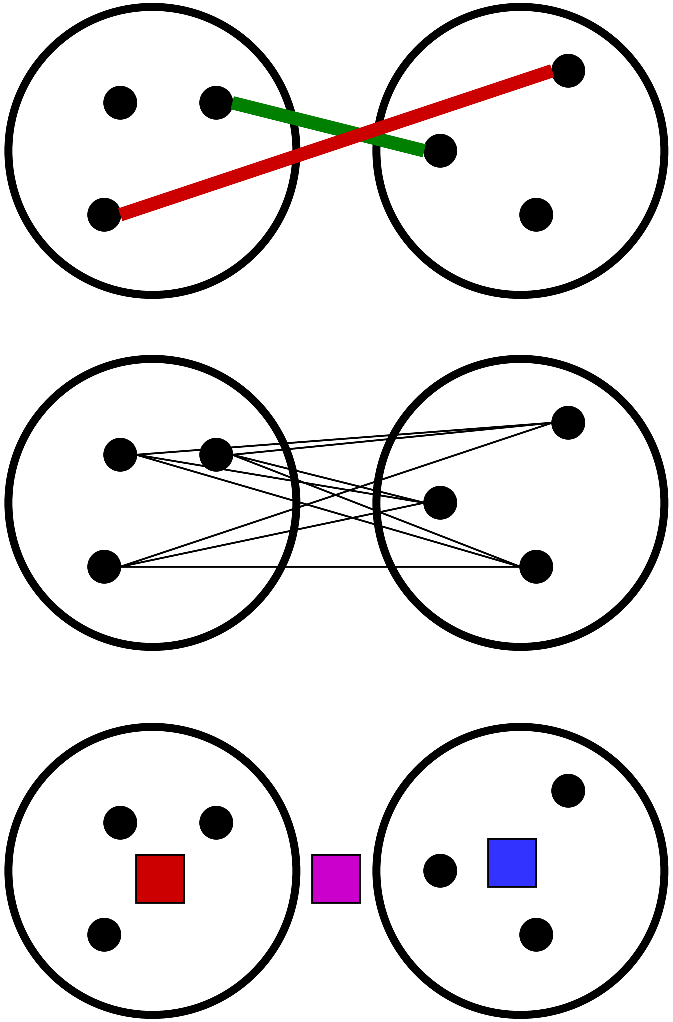


Agglomerative Clustering

- 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 Questions

- Will agglomerative clustering converge?
  - To a global optimum?
- Will it always find the true patterns in the data?
  - If the patterns are very very clear?
- Will it find something interesting?