Machine Learning for Data Science (CS4786)
Lecture 12

Clustering + Linkage Clustering

Clustering

- Grouping sets of data points s.t.
  - points in same group are similar
  - points in different groups are dissimilar
- A form of unsupervised classification where there are no predefined labels

Some Notations

- Kary clustering is a partition of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ into K groups
- For now assume the magical K is given to use
- Clustering given by $C_1, \ldots, C_K$, the partition of data points.
- Given a clustering, we shall use $c(\mathbf{x}_t)$ to denote the cluster identity of point $\mathbf{x}_t$ according to the clustering.
- Let n_j denote $|C_j|$, clearly $\sum_{j=1}^{K} n_j = n$.

How do we formalize?

Say dissimilarity$(\mathbf{x}_t, \mathbf{x}_s)$ measures dissimilarity between $\mathbf{x}_t$ & $\mathbf{x}_s$

Given two clustering $\{C_1, \ldots, C_K\}$ (or c) and $\{C'_1, \ldots, C'_K\}$ (or c')

How do we decide which is better?

- points in same cluster are not dissimilar
- points in different clusters are dissimilar

Clustering Criterion

- Minimize total within-cluster dissimilarity

$$M_1 = \sum_{j=1}^{K} \sum_{s,t \in C_j} \text{dissimilarity}(x_t, x_s)$$

- Maximize between-cluster dissimilarity

$$M_2 = \sum_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

- Maximize smallest between-cluster dissimilarity

$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

- Minimize largest within-cluster dissimilarity

$$M_4 = \max_{j \in [K]} \max_{s,t \in C_j} \text{dissimilarity}(x_t, x_s)$$

Clustering Criterion

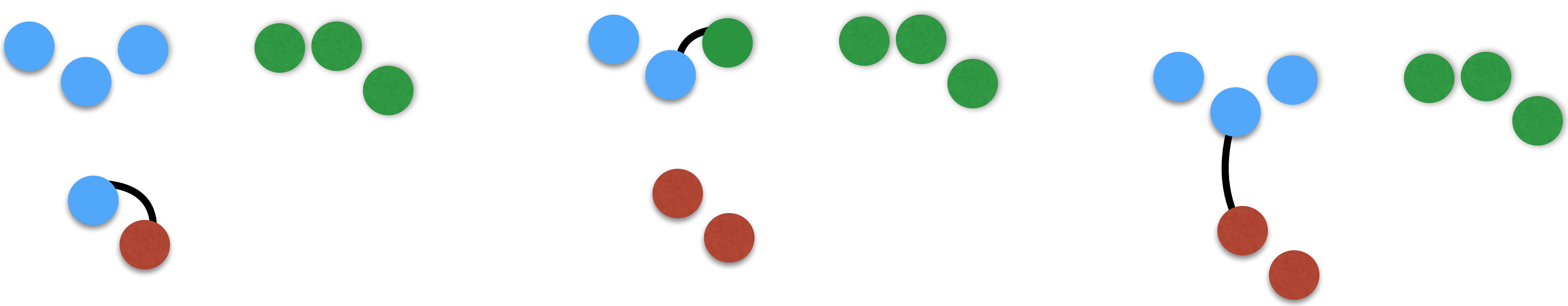
- minimizing $M_1 \equiv$ maximizing M_2

Clustering

- Multiple clustering criteria all equally valid
- Different criteria lead to different algorithms/solutions
- Which notion of distances or costs we use matter

Lets Build an Algorithm

$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

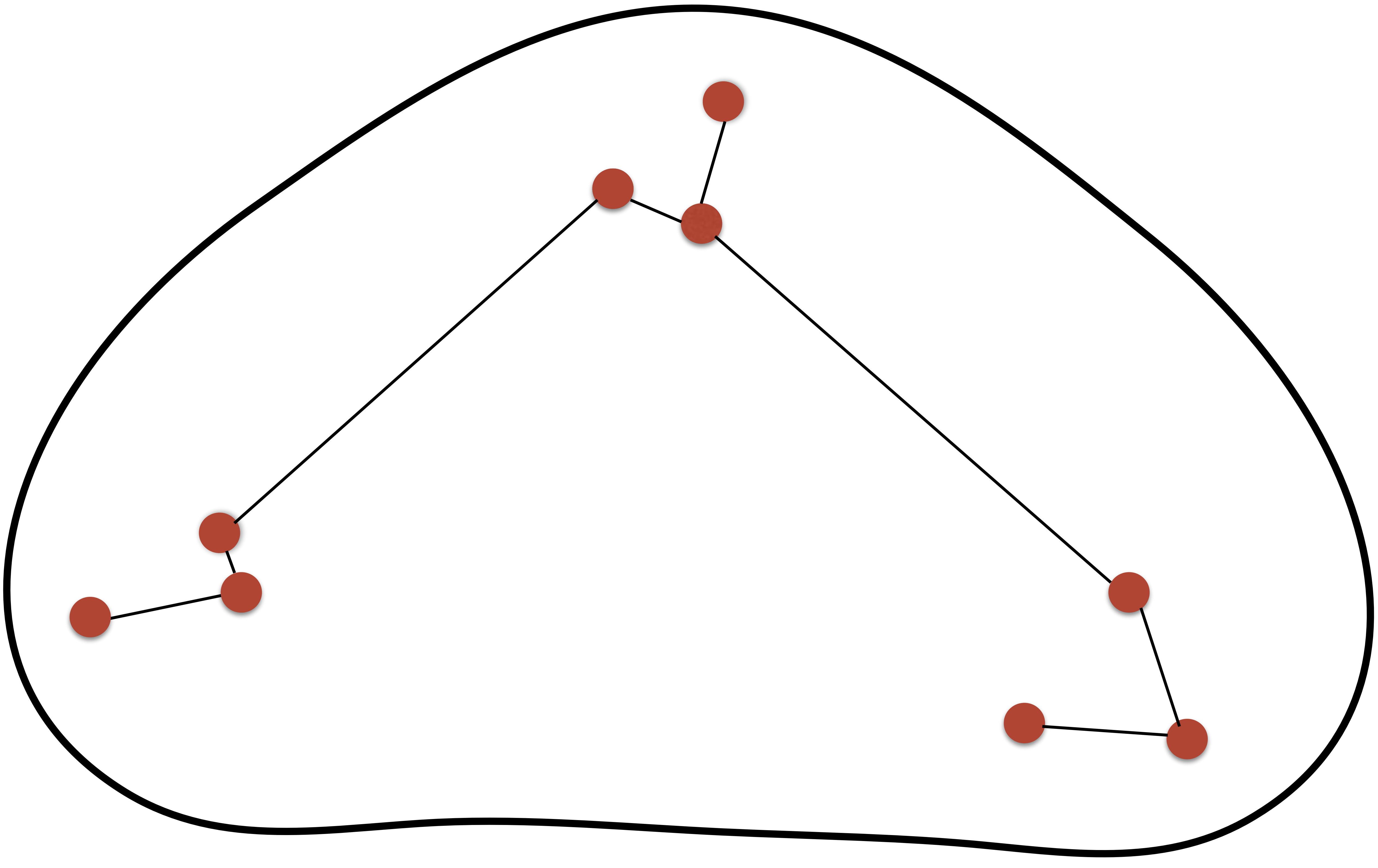


Single Link Clustering

- Initialize n clusters with each point $\mathbf{x}_t$ to its own cluster
- Until there are only K clusters, do
  1. Find closest two clusters and merge them into one cluster

$$\text{dissimilarity}(C_i, C_j) = \min_{t \in C_i, s \in C_j} \text{dissimilarity}(\mathrm{x}_t, \mathrm{x}_s)$$

Demo



Single Link Objective

Objective for single-link:

$$M_3 = \min_{\mathbf{x}_s, \mathbf{x}_t : c(\mathbf{x}_s) \neq c(\mathbf{x}_t)} \text{dissimilarity}(x_t, x_s)$$

Single link clustering is optimal for above objective!

Single Link Objective

Proof:

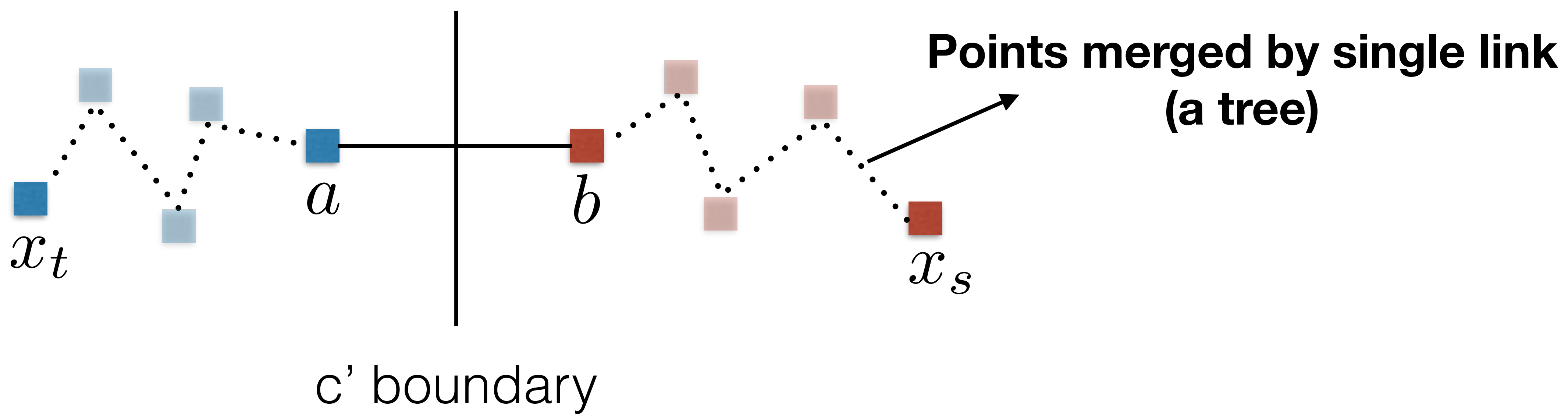
Say c is solution produced by single-link clustering

Key observation:

$$\min_{t,s:c(x_i)\neq c(x_j)} \text{dissimilarity}(x_i, x_j) >$$ Distance of points merged (on the tree)

Say $c' \neq c$ then,

$$\exists\ t, s \text{ s.t. } c'(x_t) \neq c'(x_s) \text{ but } c(x_t) = c(x_s)$$



Linkage Clustering

- Start with each point being its own cluster
- Merge the closest two clusters
  - Changing the meaning of what makes two cluster closest yield different linkage algorithms
- Single link is the only one provable optimal
- Linking based on average distance works best in practice

Clustering Criterion

- Minimize average dissimilarity within cluster

$$M_6 = \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{s \in C_j} \text{dissimilarity}\left(\mathbf{x}_s, C_j\right)$$

$$= \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{s \in C_j} \left(\sum_{t \in C_j, t \neq s} \text{dissimilarity}\left(\mathbf{x}_s, \mathbf{x}_t\right) \right)$$

$$= \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{s \in C_j} \left(\sum_{t \in C_j, t \neq s} \left\| \mathbf{x}_s - \mathbf{x}_t \right\|_2^2 \right)$$

- Minimize within-cluster variance: $\mathbf{r}_j = \frac{1}{n_j} \sum_{\mathbf{x} \in C_j} \mathbf{x}$

$$M_5 = \sum_{j=1}^{K} \sum_{t \in C_j} \left\| \mathbf{x}_t - \mathbf{r}_j \right\|_2^2$$

Clustering Criterion

- minimizing $M_5 \equiv$ minimizing M_6