TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

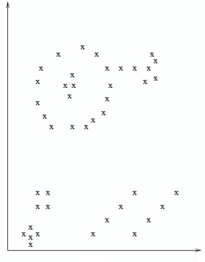
9.3 Hierarchical Clustering

Machine Learning 1: Foundations

Marius Kloft *(TUK)*

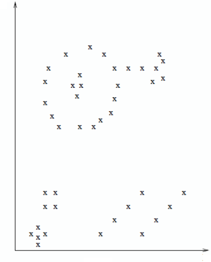
Disadvantage

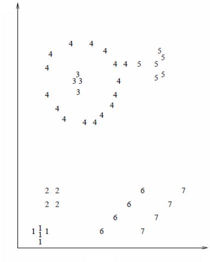
Big problem in k-means: do not know the number of clusters in advance



Disadvantage

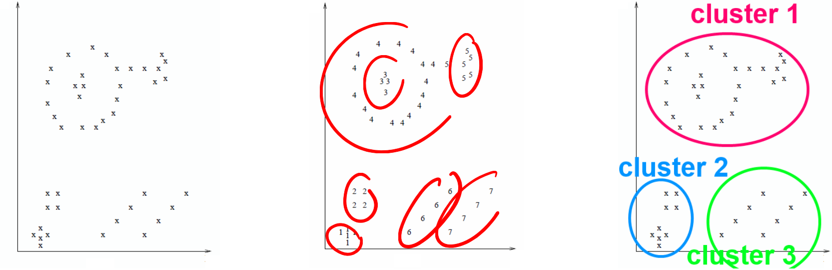
Big problem in *k*-means: do not know the number of clusters in advance





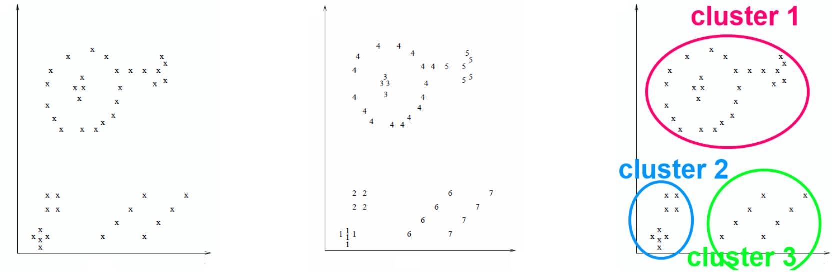
Disadvantage

Big problem in *k*-means: do not know the number of clusters in advance



Disadvantage

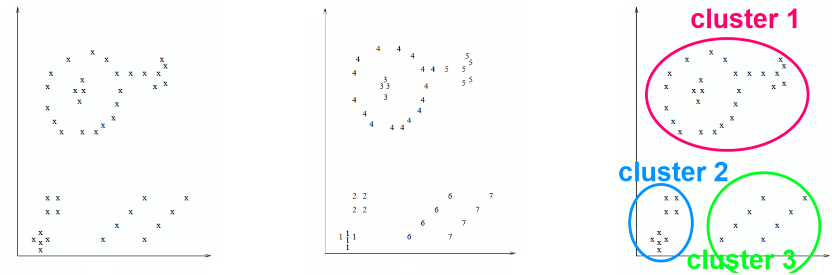
Big problem in k-means: do not know the number of clusters in advance



7 clusters vs. 3 clusters

Disadvantage

Big problem in *k*-means: do not know the number of clusters in advance



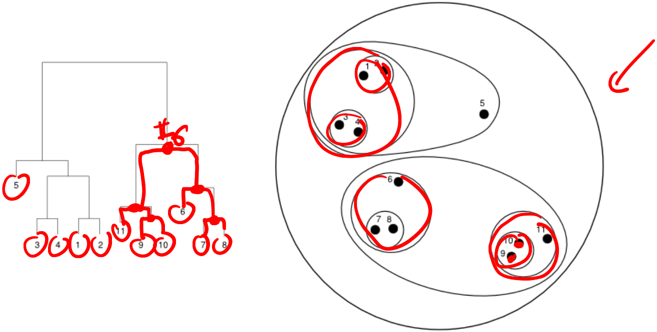
7 clusters vs. 3 clusters

Quiz: what to do?

Hierarchical Clustering

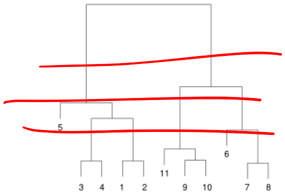
Hierarchical Clustering

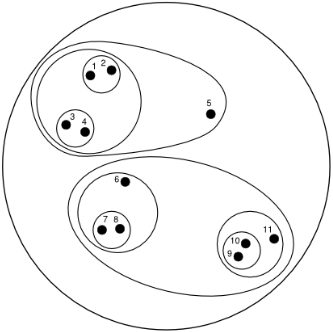
- Generates a tree (“hierarchy”) of clusters



Hierarchical Clustering

- Generates a tree ("hierarchy") of clusters





  - do not need to specify number of clusters

Hierarchical Clustering

1: **function** HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)

Hierarchical Clustering

function HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
    assign each input to a cluster

Hierarchical Clustering

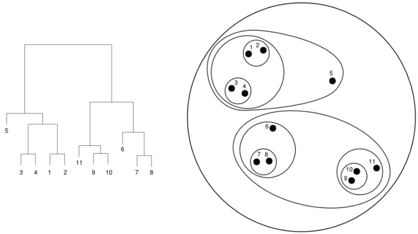
```
function HIERARCHICALCLUSTERING(inputs x_1, ..., x_n ∈ ℝ^d)
    assign each input to a cluster
    repeat
        link the two clusters with minimal distance
```

Hierarchical Clustering

function HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
    assign each input to a cluster
    repeat
        link the two clusters with minimal distance
    until finished

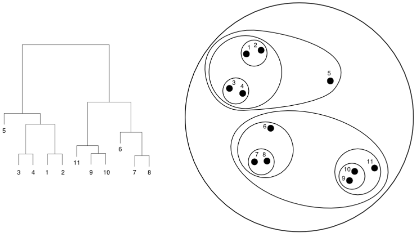
Hierarchical Clustering

function HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
    assign each input to a cluster
    repeat
        link the two clusters with minimal distance
    until finished
    return tree of cluster linkages
end function



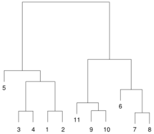
Hierarchical Clustering

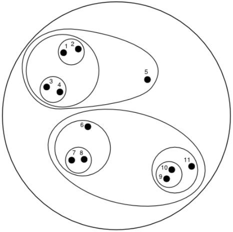
function HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
    assign each input to a cluster
    repeat
        link the two clusters with minimal distance
    until only a single root cluster left
    return tree of cluster linkages
end function



Hierarchical Clustering

1: **function** HIERARCHICALCLUSTERING(inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$)
2: assign each input to a cluster
3: **repeat**
4: link the two clusters with minimal distance
5: **until** only a single root cluster left
6: **return** tree of cluster linkages
7: **end function**





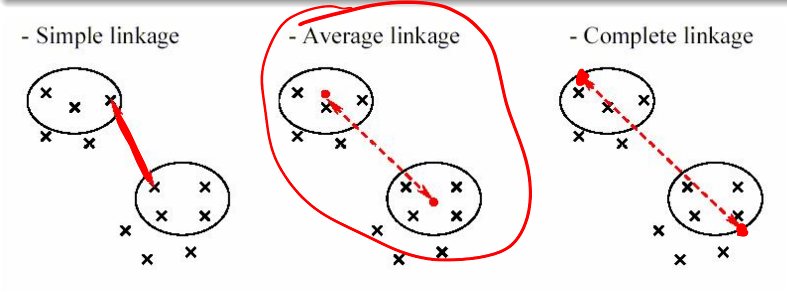
How to measure distance between two clusters?

Hierarchical Clustering

Quiz: How to measure distance $d(i, j)$ between two clusters i and j?

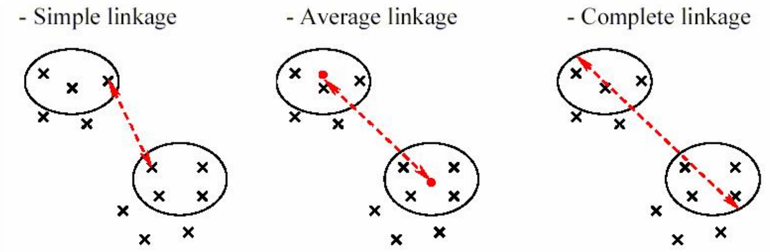
Hierarchical Clustering

Quiz: How to measure distance $d(i,j)$ between two clusters i and j?



Hierarchical Clustering

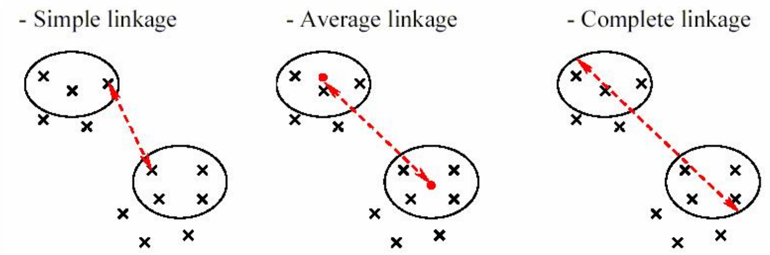
Quiz: How to measure distance $d(i, j)$ between two clusters i and j?



- Let $S_j \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of inputs contained in the jth cluster

Hierarchical Clustering

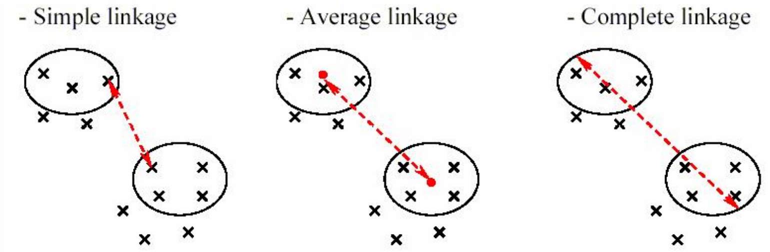
Quiz: How to measure distance $d(i, j)$ between two clusters i and j?



- Let $S_j \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of inputs contained in the jth cluster
- **Simple linkage**: $d(i, j) := \min_{\mathbf{x} \in S_i, \tilde{\mathbf{x}} \in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$

Hierarchical Clustering

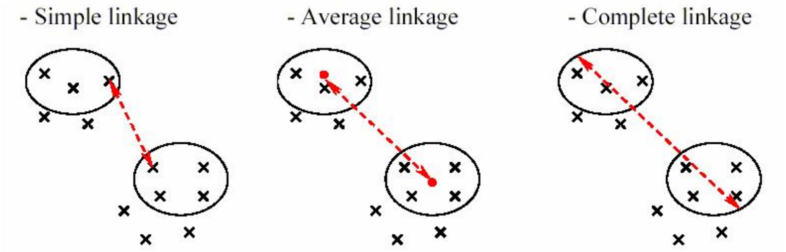
Quiz: How to measure distance $d(i,j)$ between two clusters i and j?



- Let $S_j \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of inputs contained in the jth cluster
- **Simple linkage**: $d(i,j) := \min_{\mathbf{x} \in S_i, \tilde{\mathbf{x}} \in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$
- **Average linkage**: $d(i,j) := \text{mean}_{\mathbf{x} \in S_i, \tilde{\mathbf{x}} \in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$

Hierarchical Clustering

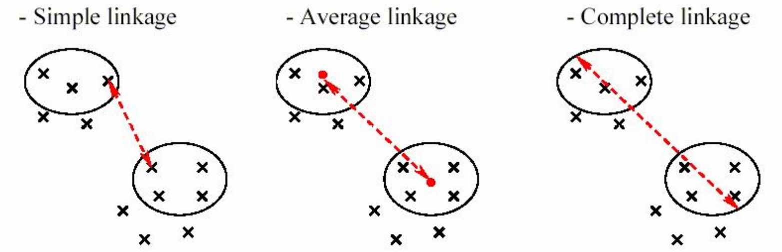
Quiz: How to measure distance $d(i,j)$ between two clusters i and j?



- Let $S_j \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of inputs contained in the jth cluster
- **Simple linkage**: $d(i,j) := \min_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$
- **Average linkage**: $d(i,j) := \text{mean}_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$
- **Complete linkage**: $d(i,j) := \max_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|$

Hierarchical Clustering

Quiz: How to measure distance $d(i,j)$ between two clusters i and j?

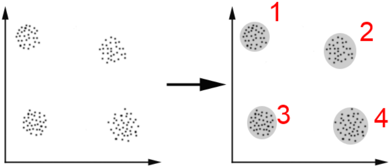


- Let $S_j \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the set of inputs contained in the jth cluster
- **Simple linkage**: $d(i,j) := \min_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2$
- **Average linkage**: $d(i,j) := \text{mean}_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2$
- **Complete linkage**: $d(i,j) := \max_{\mathbf{x}\in S_i, \tilde{\mathbf{x}}\in S_j} \|\mathbf{x} - \tilde{\mathbf{x}}\|^2$
- Note: again, all this can be kernelized...

Conclusion

Clustering:

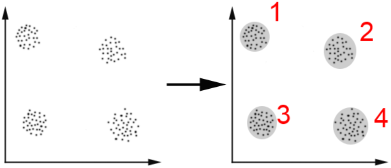
- Organizing data into groups



Conclusion

Clustering:

- Organizing data into groups



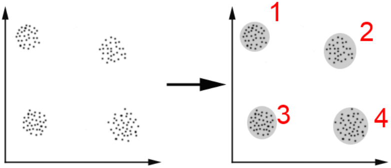
k-means:

- Alternatingly, assign inputs to closest cluster center and re-compute centers

Conclusion

Clustering:

- Organizing data into groups



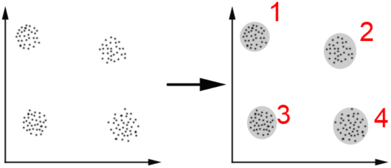
k-means:

- Alternatingly, assign inputs to closest cluster center and re-compute centers
- Can be kernelized
- Can be deepified using transfer learning

Conclusion

Clustering:

- Organizing data into groups



k-means:

- Alternatingly, assign inputs to closest cluster center and re-compute centers
- Can be kernelized
- Can be deepified using transfer learning

Hierarchical clustering:

- Consider clusters at various scales
- Helpful when the number of clusters is unknown

Refs I