Lecture 14: Unsupervised Learning

Cognitive Systems - Machine Learning

Part IV: Further Topics

k-means clustering, hierarchical clustering, competitive learning, SOMs

last change January 14, 2015

Outlook

- Unsupervised learning
- Similarity and distance measures
- k-means clustering
- Hierarchical clustering
- Self-organizing maps (SOMS)

Motivation

- Unsupervised learning: trying to find hidden structure in unlabeled data
  Knowledge Discovery
- Data can be categorized based on some similarity measure
- Typical domains of application
  - Data mining (categorize unlabelled data)
  - Data compression (represent data sets by their prototype)
  - Categorize data when labeling is costly, e.g. in speech recognition

Similarity Measures

- On metrical data (feature vectors): typically distance metrics
- On categorial features: contrast measures
- On structured data: extract features OR use structural similarity measures such as edit distance

Euklid:

$$d(\vec{x_i}, \vec{x_j}) = (\sum_{l=1}^{n}(x_{il} - x_{jl})^2)^{\frac{1}{2}}$$

City Block:

$$d(\vec{x_i}, \vec{x_j}) = \sum_{l=1}^{n}(x_{il} - x_{jl})$$

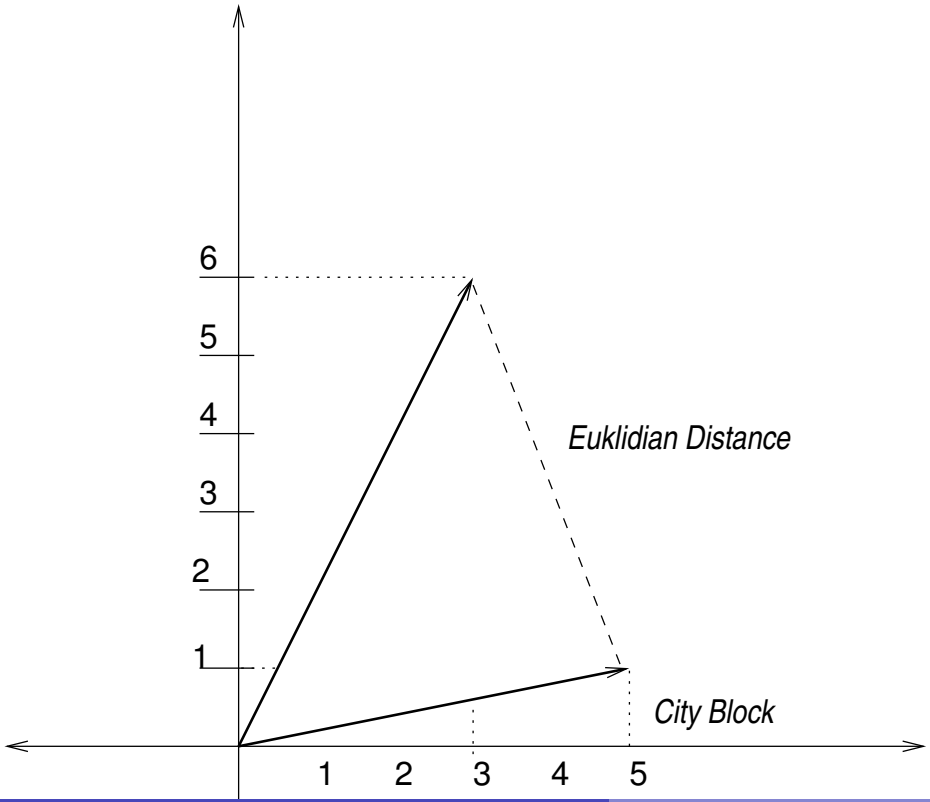
Distance Metric

- Distance Function/Measure
  - Non-negativity: $d(x, y) \geq 0$
  - Identity: $d(x, y) = 0$ iff $x = y$
  - Symmetry: $d(x, y) = d(y, x)$
- Metric
  - Triangle equality (subadditivity): $d(x, y) \leq d(x, z) + d(z, y)$
- Similarity Function/Measure
  - $s(x, y) \leq 1$
  - Identity: $s(x, y) = 1$ iff $x = y$
  - Symmetry: $s(x, y) = s(y, x)$
- Metric
  - $s(x, y) \geq s(x, z) \times s(z, y)$

Minkowski-Metric

$$d(\vec{x}_i, \vec{x}_j) = \left(\sum_{l=1}^{n} (x_{il} - x_{jl})^m\right)^{\frac{1}{m}}$$

- $m = 1$: City Block/Manhatten
- $m = 2$: Euklid
- $m = \infty$: Supremum



- City Block: $2 + 5$
- Supremum: 5

Mahalanobis Distance

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}$$

- S is the covarance matrix (symmetric, with variance in the diagonal)
- considers differences in scale (e.g. one feature has min=0 and max=1, another has min=0 and max=100)
- considers correlation of features
- If covariances are 1, Mahalanobis is equivalent to Euklidian distance

Measures for Categorial Variables

Jaccard similarity coefficient:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Jaccard distance:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Tanimoto similarity for binary variables:

$$f(A,B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

k-means Clustering

Iterative Algorithm:

1. Define the number of clusters k
2. Initialize clusters by
   - an arbitrary assignment of examples to clusters or
   - an arbitrary set of cluster centers (examples assigned to nearest centers)
3. Compute the sample mean of each cluster
4. Reassign each example to the cluster with the nearest mean
5. If the classification of all samples has not changed, stop, else go to step 3

Sample mean (S is the set of objects in one cluster)

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Distance between objects in same cluster should be minimal and distance between clusters maximal.

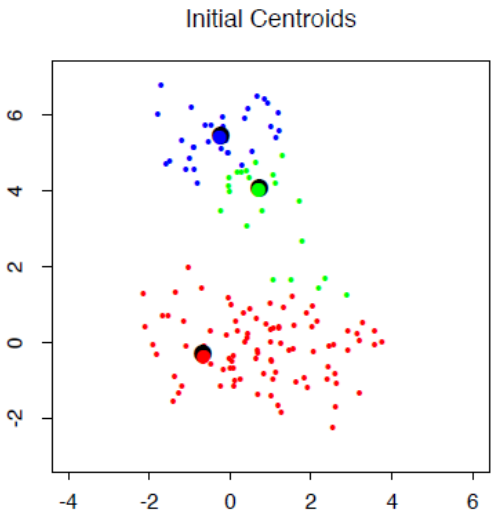
k-means Clustering

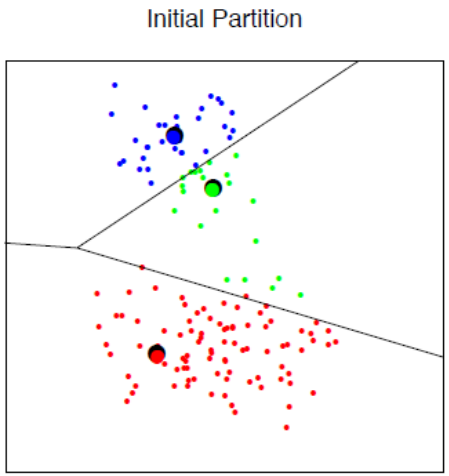
Criterion function (sum of squared errors):

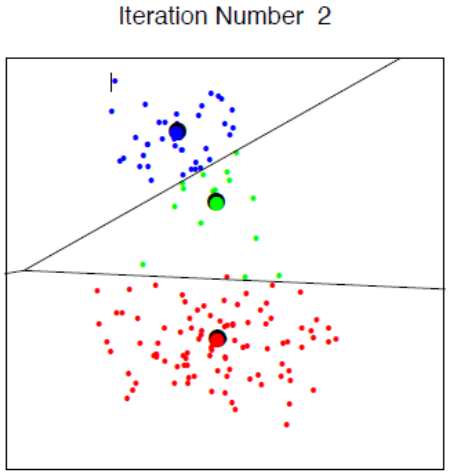
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

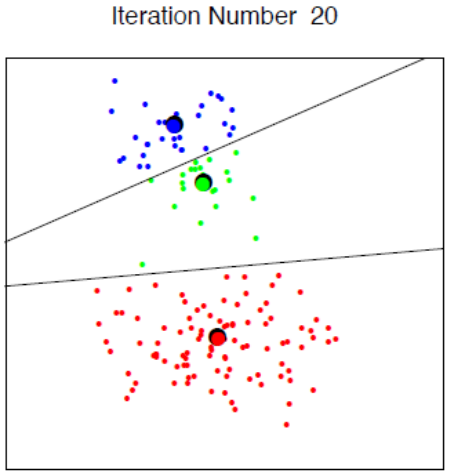
- Data points with largest distances greatest influence!
  k-means clustering is vulnerable to outliers!
- Effort: proportional to number of objects times number of clusters
- Danger of local optima: run several times with different starting points

Illustration









Determine Number of Clusters

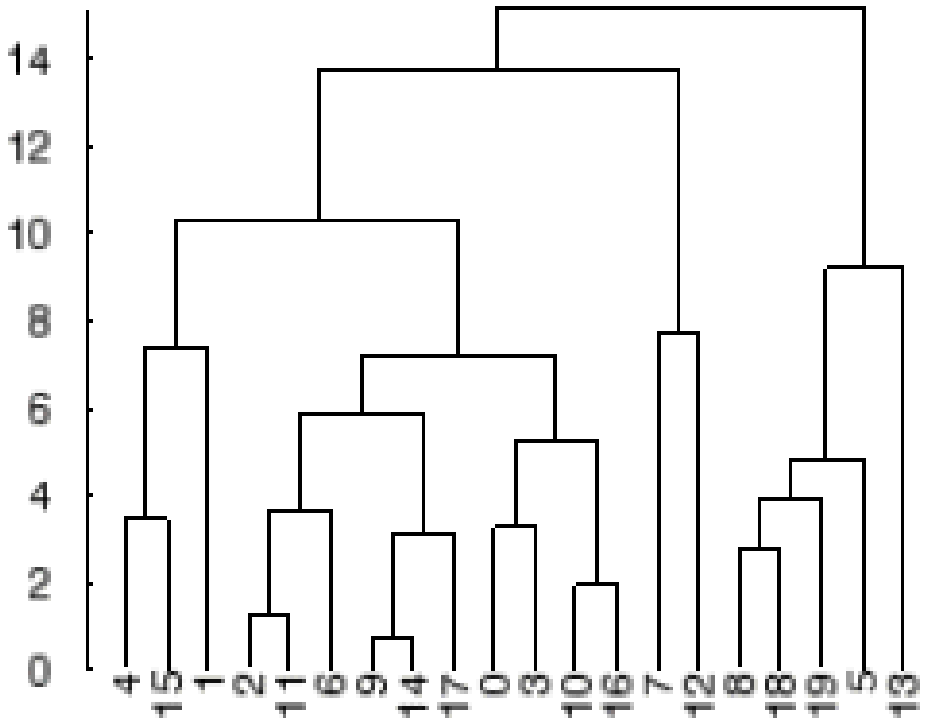
- Previous knowledge from the application domain
  e.g., classify plants in three groups reflecting their vitality
- Determine optimal $K*$ such that within-cluster distance is "optimal"

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(\vec{x_i}, \vec{x_{i'}})$$

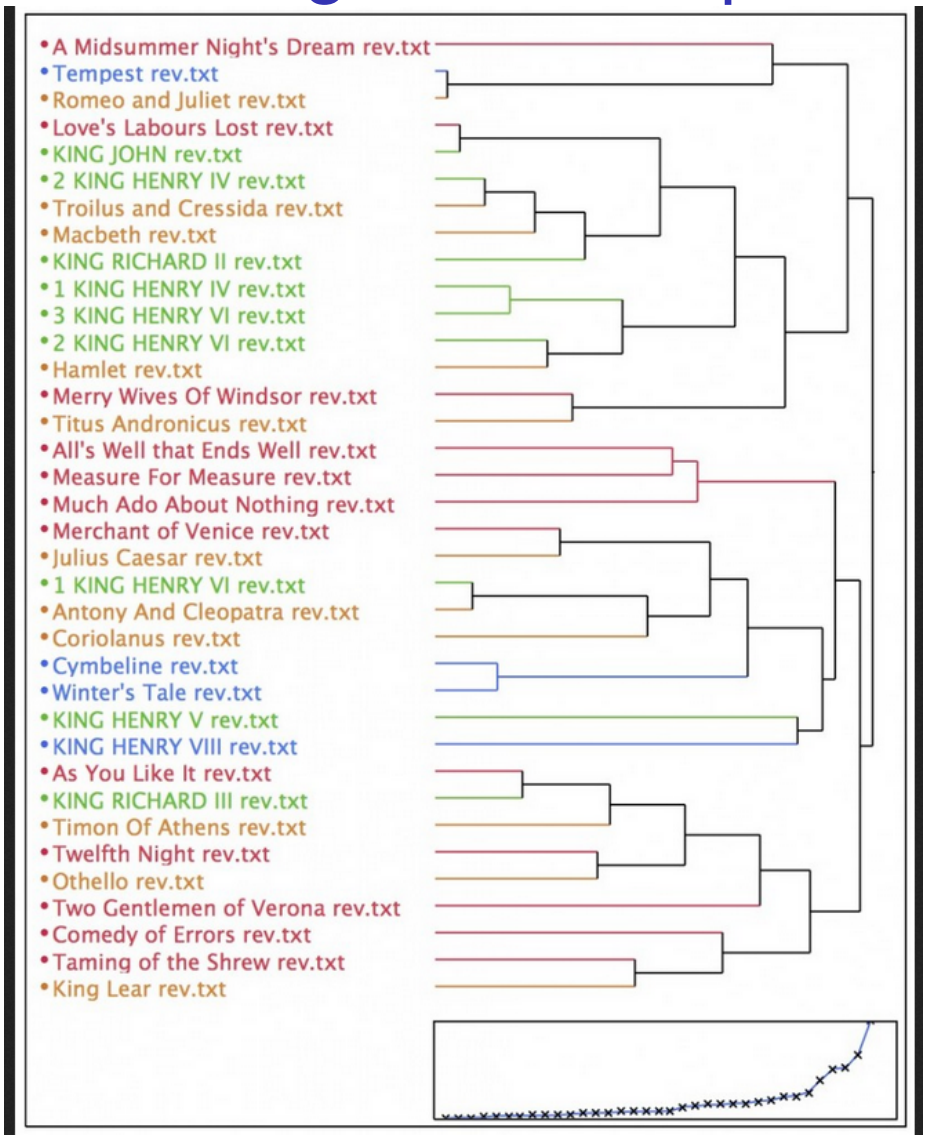
- Select K with minimal descend from $W_k - W_{k+1}$

Hierarchical Clustering

- Cluster are not on one level but constitute a taxonomy
- Lowest level: Objects, Top-Level: Single Cluster
- Each level contains cluster which subsume two clusters of the leve below
- There are top-down and bottom-up (agglomerative) approaches



Clustering of Shakespeare Plays



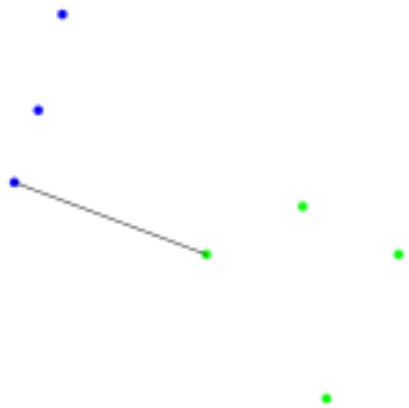
www.winedarksea.org, Michael Witmore, Published: November 29, 2009

Agglomerative Clustering

- User decides which level classifies data best
- Dis-similarity of clusters guides agglomeration: the two clusters with the least dis-similarity are joined
- Dis-similarity d between clusters G and H: distance between objects (e.g. feature vectors) in G and H
- Three important measures:
  - single linkage
  - complete linkage
  - average linkage

Single Linkage

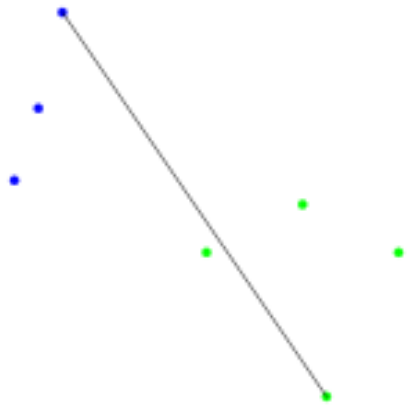
$$D_{SL}(A, B) := \min_{a \in A, b \in B} \{d(a, b)\}$$



- Dis-similarity determined by dis-similarity of nearest points
- Problem: Clusters with large dis-similarities, “chains”

Complete Linkage

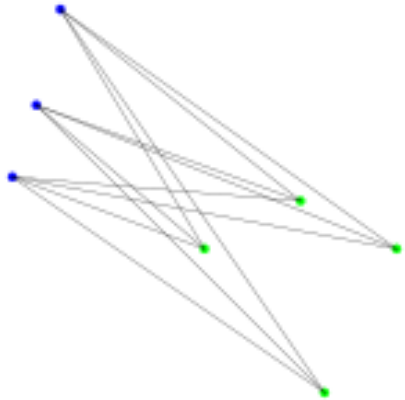
$$D_{CL}(A, B) := \max_{a \in A, b \in B} \{d(a, b)\}$$



- Dis-similarity determined by dis-similarity of farest points
- Compact clusters, but objects in a cluster might be more similar to objects in another cluster

Average Linkage

$$D_{AL}(A,B) := \frac{1}{|A||B|} \sum_{a \in A, b \in B} d(a,b)$$



- Good compromise: relatively compact clusters with relatively large "gaps" between them
- Problem: a strong monotonic transformation of the distance measure can significantly change the result

Self-organizing Maps (SOMs)

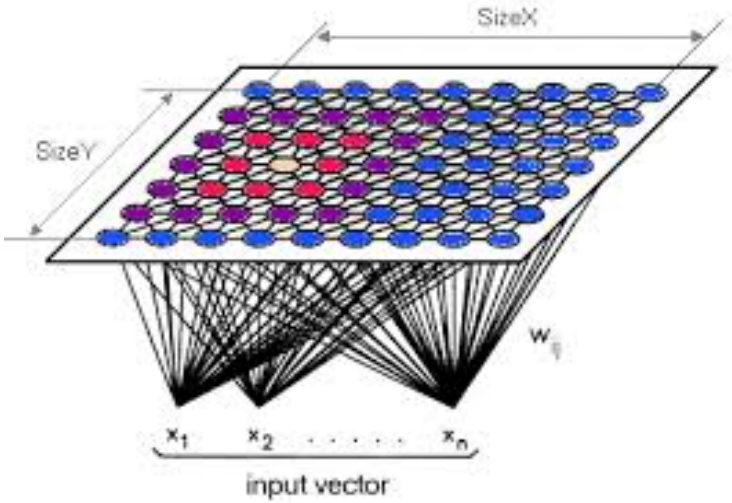
- Proposed by Kohonen, 1995
- An artificial neural net approach for unsupervised learning
- Can be applied for cluster analysis

Structure

- Input layer for n input values
- Completely connected to competitive layer
- In the competitive layer all neurons are inhibitorily connected

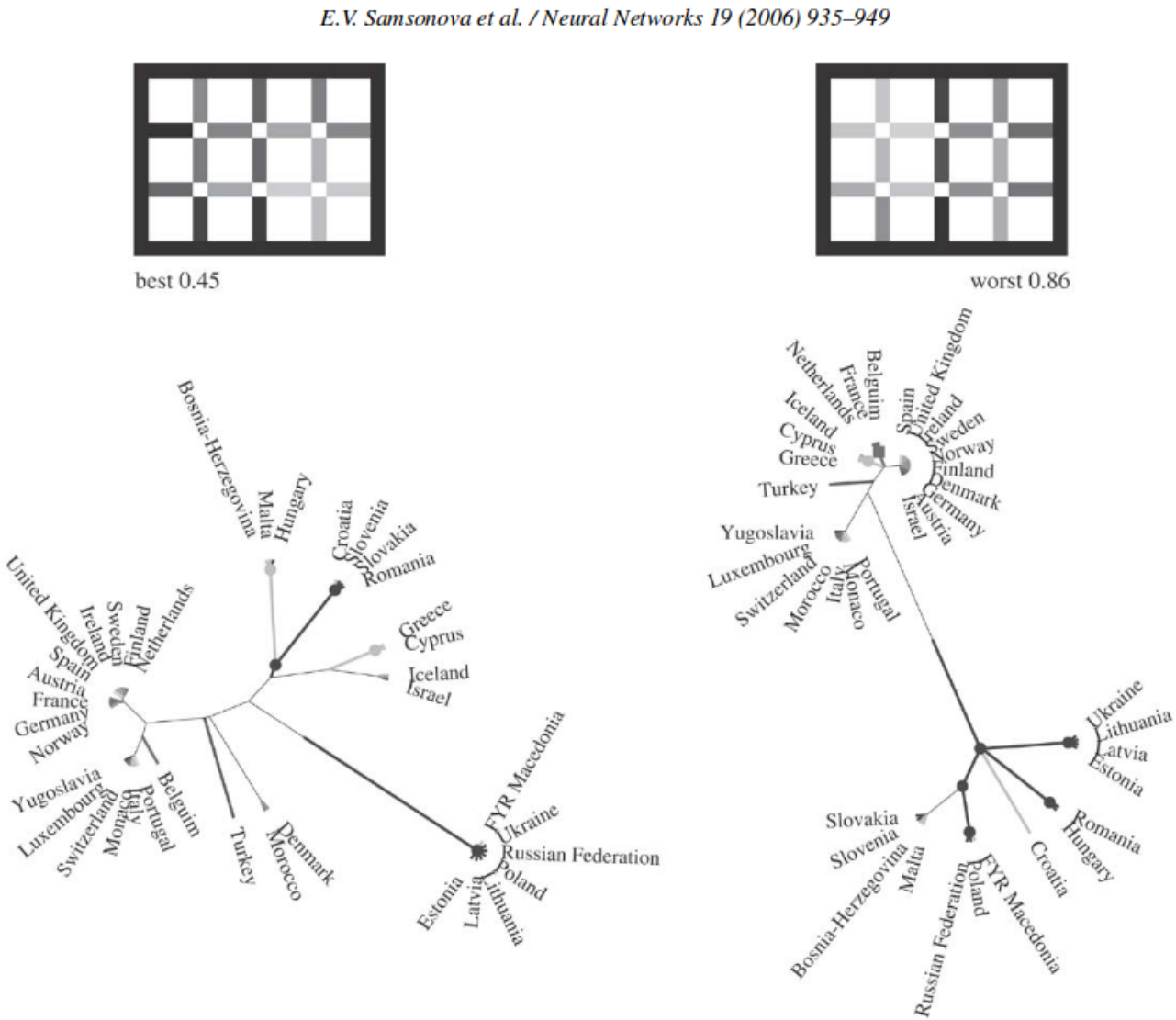
Learning

- Cause different parts of the network to respond similarly to certain input patterns
- Competitive learning
- Iterative optimization:
  - Best matching unit (BMU): neuron whose weight vector is most similar to the input
  - Weights of BMU and neighbours are adjusted towards input vector


SizeX
SizeY
w_{ij}
x_1 x_2 . . . . . x_n
input vector

Example: Voting Behavior

- A study of voting behavior of different countries during the yearly EuroVision Song Contests



Summary

- Unsupervised learning is discovery learning: given data are categorized into groups by similarity
- There are numerous similarity measures and distance measures for data sets with different characteristics
- k-means clustering is a standard approach to clustering by partitioning, it is related to k-nearest neighbor learning
- Hierarchical approaches allow to learn taxonomies
- A well-known neural network approach is SOMS

Learning Terminology

Clusteranalysis

Supervised Learning	**unsupervised learning**

Approaches:

Concept / Classification	Policy Learning
symbolic	**statistical / neuronal network**
inductive	analytical

Learning Strategy:	**learning from examples**
Data:	**categorial/metric features**
Target Values:	**concept/class**