

# Introduction to Symbolic and Statistical NLP in Scheme

Damir Ćavar  
dcavar@unizd.hr

ESLLI 2006, Malaga

July/August 2006

© 2006 by Damir Ćavar

# Vector Space Modeling

- Contextual vectors (distributional model):

$$\mathcal{X} = \begin{bmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,d} \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,d} \\ \vdots & & & \\ \mathbf{x}_{k,1} & \mathbf{x}_{k,2} & \cdots & \mathbf{x}_{k,d} \end{bmatrix}$$

# Optimization Clustering

- Given a clustering criterion
  - How to find a partition into  $n$  groups that optimizes the criterion?
- Find all possible partitions and calculate their value of the given criterion.
- Choose the partition with the optimal value.

# Optimization Clustering

- K-means generates
  - $k$  number of disjoint clusters (non-hierarchical)
  - globular clusters (spherical, elliptical, convex)
- properties:
  - numerical
  - unsupervised
  - iterative

# Optimization Clustering

- K-means
  - $k$  clusters
  - At least one element per cluster
  - No overlapping clusters
  - Non-hierarchical

# Optimization Clustering

- K-means
  - Every member of a cluster is closer to its cluster than to any other cluster
  - Procedure

# Optimization Clustering

- K-means
  - Initial partitioning of data set into  $k$  clusters
  - For each data point: calculate distance to each cluster
  - If one data point is closer to another cluster, relocate it
  - Repeat until no further relocations possible

# Optimization Clustering

- K-means advantages
  - For large number of variables it is faster than hierarchical algorithms (for small  $k$ 's)
  - Tighter clusters than hierarchical clustering, if cluster are globular



# Optimization Clustering

- K-means disadvantages
  - Initial set of  $k$  clusters can affect the result
  - Does not work well with non-globular clusters

# Optimization Clustering

- K-means example

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

# Optimization Clustering

- Initial 2 clusters on the basis of the most distant individuals:

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

# Optimization Clustering

- Initial clustering of all remaining individuals:
  - For every other individual:
    - \* Calculate Euclidean distance to the centroid of every cluster
    - \* Assign individual to cluster
    - \* Recalculate centroid for every cluster

# Optimization Clustering

- Mean vector or centroid (with coordinates  $x_1$  to  $x_n$ ) with equal weight coordinates:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

# Optimization Clustering

- Mean vector or centroid example for  $x = \{(3, 5), (7, 9)\}$ , i. e.  $n = |x| = 2$ :

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^2 x_i}{2} = \frac{(3, 5) + (7, 9)}{2} = \\ &= \frac{(3 + 7, 5 + 9)}{2} = \left(\frac{10}{2}, \frac{14}{2}\right) = (5, 7)\end{aligned}$$

# Optimization Clustering

- Initial clustering of all remaining individuals:

	Group 1		Group 2	
	Individual	Mean Vector	Individual	Mean Vector
Step 1	1	(1.0, 1.0)	4	(5.0, 7.0)
Step 2	1, 2	(1.3, 1.5)	4	(5.0, 7.0)
Step 3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
Step 4	1, 2, 3	(1.8, 2.3)	4, 5	(4.3, 6.0)
Step 5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
Step 6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

# Optimization Clustering

- Initial partitions and clustering criterion:

	Individual	Mean Vector	Sum of SQR error
Group 1	1, 2, 3	(1.8, 2.3)	6.84
Group 2	4, 5, 6, 7	(4.1, 5.4)	5.38
total			12.22



# Optimization Clustering

- Error = for every point distance to centroid
  - Criterion: the smaller the sum of square errors, the better the cluster
- Two dimensional Euclidean distance:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

# Optimization Clustering

- Error = for every point distance to centroid
- $N$ -dimensional Euclidean distance, with  $p_i$  and  $q_i$  the coordinates for  $p$  and  $q$  in dimension  $i$ :

$$\sqrt{\sum_{i=1}^N (p_i - q_i)^2} \quad (3)$$

# Optimization Clustering

- Optimization Iteration:
  - Compare each individual's distance to its own mean with distance to the opposite group mean.
  - If distance to the mean in opposite group is smaller, relocate the individual.
  - Calculate the sum of square errors, if smaller than before, this is an improvement.

# Optimization Clustering

- Distance to means:

Individual	distance to mean 1	distance to mean 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.8
7	2.8	1.1

# Optimization Clustering

- Subsequent partitions and new clustering criterion:

	Individual	Mean Vector	Sum of SQR error
Group 1	1, 2	(1.3, 1.5)	0.63
Group 2	3, 4, 5, 6, 7	(3.9, 5.1)	7.9
total			<b>8.53</b>

- Decrease of clustering criterion (from 12.22 to 8.53).

# Optimization Clustering

- Remember:
  - *k-means* or *k-nearest neighbors* is a fast and efficient algorithm.
  - You have to know how many clusters you are looking for.
  - Specific cluster shapes will not be discovered.