

## Web Mining — Lecture 20090402

© 2007-2009 Mauro Brunato and Elisa Cilia

Facoltà di Scienze  
Università di Trento

Academic Year 2008-2009, second semester  
Last revision: April 6, 2009

### Top-down Clustering

Iteratively refine the assignment of documents to a preset number of clusters.

### k-Means with “hard” assignment

- ▶ Initial configuration: arbitrary (or chosen by a heuristic) grouping of documents into  $k$  groups
- ▶ Computation of the  $k$  corresponding centroids

#### k-Means Algorithm

initialize centroids to arbitrary vectors

**while** further improvement is possible

**for** each document  $d$

    find cluster  $c$  whose centroid is **most similar** to  $d$   
    assign  $d$  to the cluster  $c$

**for** each  $c$

    recompute the centroid of  $c$

### k-Means with “soft” assignment

- ▶ no explicit assignment of documents to clusters
- ▶ each cluster is represented by a vector  $\mu_c$  (not necessarily the centroid)

#### Goal

Find  $\mu_c$  for each  $c$  that minimizes

$$\sum_d \min_c |d - \mu_c|^2 (\text{quantization error})$$

## Iterative Algorithm

Basic idea:

bring the mean vectors closer to docs that they are closest to.

### Repeat

for each document  $d$

compute  $\Delta\mu_c = \sum_d \begin{cases} \eta(d - \mu_c), & \text{if } \mu_c \text{ is closest to } d \\ 0 & \text{otherwise} \end{cases}$

$\mu_c \leftarrow \mu_c + \Delta\mu_c$

### Alternative update rules

$$\Delta\mu_c = \eta \frac{1/|d - \mu_c|^2}{\sum_\gamma 1/|d - \mu_\gamma|^2} (d - \mu_c)$$
$$\Delta\mu_c = \eta \frac{\exp(-|d - \mu_c|^2)}{\sum_\gamma \exp(-|d - \mu_\gamma|^2)} (d - \mu_c)$$

## Geometric Embedding Approaches

Self Organizing Maps (SOMs)

Multidimensional scaling

- ▶ FastMap

Latent Semantic Indexing (LSI)

## Running Time

- ▶ bottom-up approaches can be used for the  $k$ -means algorithm initialization:
  - ▶ randomly select  $O(kn)$  documents and subject them to bottom-up clustering  $\rightarrow O(kn \log n)$
- ▶  $n$  document compared against  $k$  centroids at each round  $\rightarrow O(kn)$
- ▶ the number of rounds may be consider fixed (it is not strongly dependent on  $n$  or  $k$ )

### Overall time complexity

$O(kn \log n)$

## Multidimensional scaling

- ▶  $k$ -means and Kohonen maps require document placement in (vector) space, mutual distances are not enough.
- ▶ What if only distance (or similarity) is available?
- ▶ Useful for incorporating user feedback (“ $i$  is quite similar to  $j$  but not to  $k$ ”).

### Given data

A matrix of mutual distances  $d_{ij}$  between documents.

### Goal

Embed documents in a low-dimensional space (just like Kohonen maps) so that mutual distances  $\hat{d}_{ij}$  differ as little as possible from those specified in the original matrix.