Clustering
and
Subspace Clustering

Outline

- **Clustering** and the **Curse of Dimensionality:**
  - Discovering local structures: **Subspace Clustering**
- **Clustering** and its **ill-posed nature:**
  - **Clustering Ensembles**
  - **Semi-supervised clustering**

Clustering

- Goal: Grouping a collection of objects (data points) into subsets or "clusters", such that those within each cluster are more closely related to one other than objects assigned to different clusters.

- Fundamental to all clustering techniques is the choice of ***distance or dissimilarity measure*** between two objects.

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary

Similarity is hard to define, but…
"We know it when we see it"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Slide by E. Keogh

Dissimilarities based on Features

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iq})^T \in \Re^q,\ \ i = 1, \cdots, N$$

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{q} d_k(x_{ik}, x_{jk})$$

$$d_k(x_{ik}, x_{jk}) = (x_{ik} - x_{jk})^2$$

$$\Rightarrow D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{q} (x_{ik} - x_{jk})^2$$ Squared Euclidean distance

Clustering

- Fundamental to all clustering techniques is the choice of distance measure between data points;

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{q} (x_{ik} - x_{jk})^2$$ Squared Euclidean distance

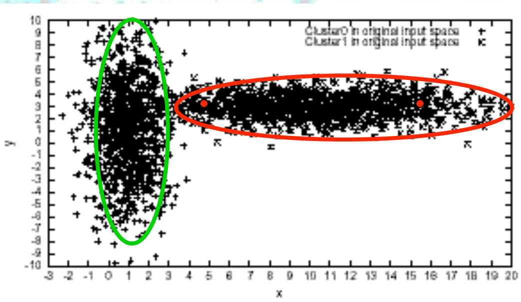
- Assumption: All features are **equally important**;

- Such approaches fail in high dimensional spaces

Clustering: The Curse of Dimensionality

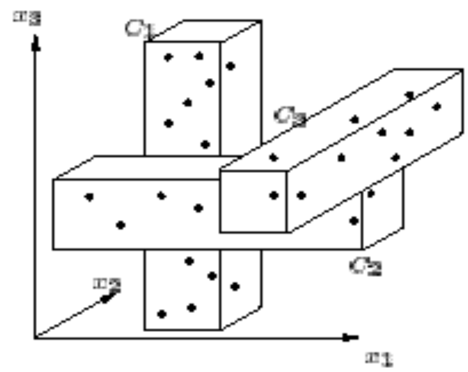
- A full-dimensional distance is often irrelevant, as the farthest point is expected to be almost as close as the nearest point;
- In high dimensional spaces, it is likely that, for any given pair of points within the same cluster, there exist at least a few dimensions on which the points are far apart from each other.

Example



Clustering

- Clusters may exist in different subspaces, comprised of different combinations of features:



Each dimension is relevant to at least one cluster

Global Dimensionality Reduction

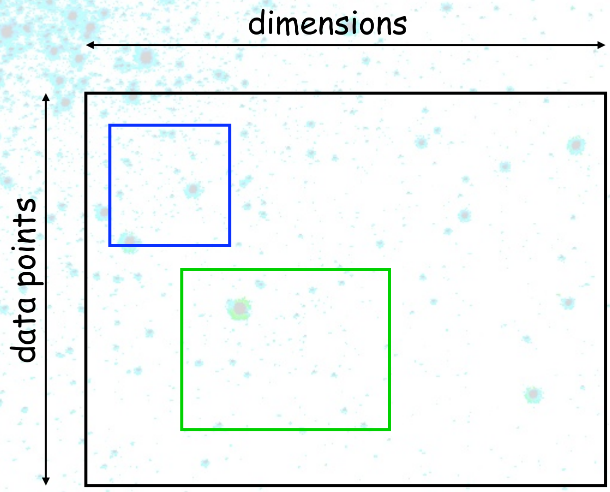
- We cannot prune off dimensions without incurring a loss of crucial information;
- Global dimensionality reduction techniques, e.g. PCA, do not handle well situations where different clusters are dense in different subspaces;
- The data presents **local structure**

Local Dimensionality Reduction

- To capture the local correlations of data, a proper feature selection procedure should operate locally;
- A local operation would allow to embed different distance measures in different regions;

Subspace clustering

Simultaneous clustering of both row and column sets in a data matrix



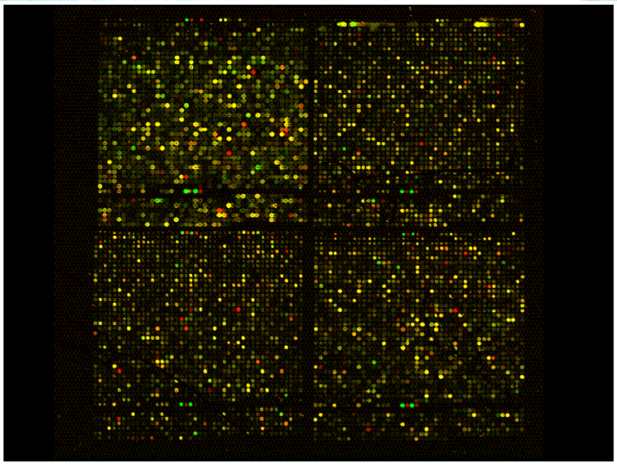
Subspace clustering

Other terms used:

1. Biclustering
2. Coclustering
3. Box clustering
4. Projective clustering
5. …

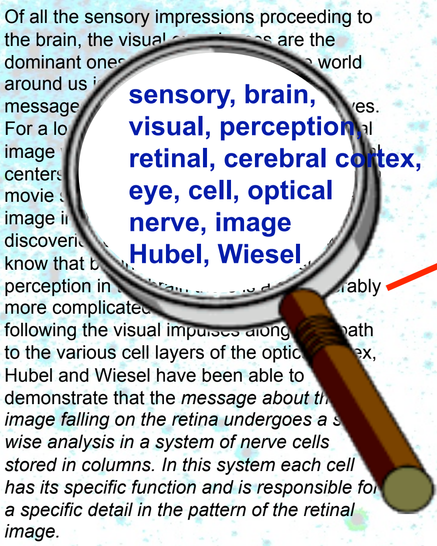
Subspace clustering

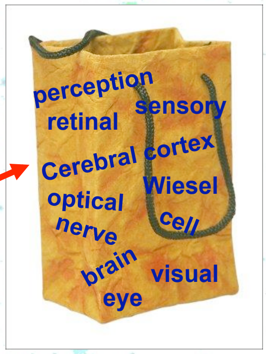
- Important problem in practice
- Real life problems:
  - Are high dimensional
  - Present local structure



Clustering of Microarray data:

- Different conditions may have different importance for a given set of genes;
- The relevance of one condition may vary from gene to gene





Bag-of-words representation of a document

Text classification: Different words may have different degrees of relevance for a given category of documents;
A single word may have a different importance across different categories.

Example: clustering query results

- Query: "Bush"
- Returns documents on the president of the United States as well as information on landscaping.
- Clustering using BOW representation
  - Documents on the president and documents on landscaping are related to different sets of features (i.e., terms)

Approaches to Subspace Clustering

- **Bottom-up** Finds dense regions in low dimensional spaces, and combines them to form clusters.

- **Top-down** Finds an initial clustering in the full space, and evaluates the subspaces of each cluster, iteratively improving the results.

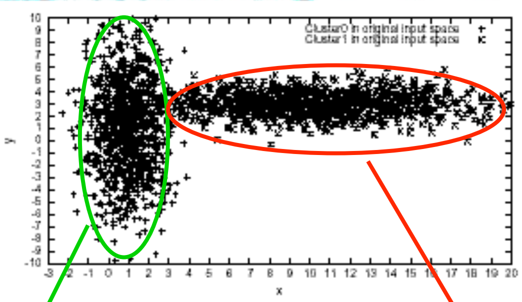
Approaches to Subspace Clustering

- Most methods provide "hard" clustering solutions at data level.
- In each subspace typically features are equally weighted.
- More recently: "soft subspace clustering" and weighted subspace clustering approaches.

Locally Adaptive Clustering (LAC)

- We wish to ***learn*** from the data the relevant features for each cluster.
- Idea: Develop a ***soft*** feature selection procedure
  - Assign (local) weights to features according to the strength with which the feature participates to the cluster.

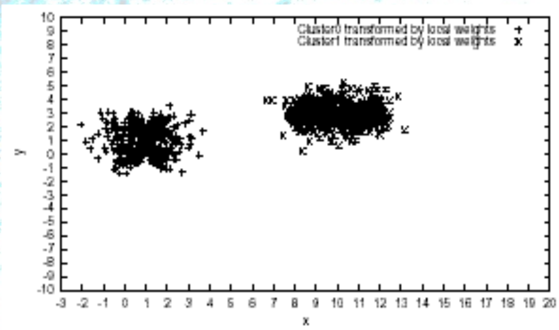
Locally Adaptive Clustering: Example



$(w_{1x}, w_{1y}),\ w_{1x} > w_{1y}$

$(w_{2x}, w_{2y}),\ w_{2y} > w_{2x}$

Locally Adaptive Clustering: Example



Within-cluster distances between points are computed using the respective local weights

Categorization and Keyword Identification of Unlabeled Documents

The Overall Idea

- The result of LAC is twofold:
  - It achieves a ***clustering*** of the documents;
  - It achieves the identification of ***cluster-dependent keywords*** via a continuous term-weighting mechanism.

Data set: 20 Newsgroups

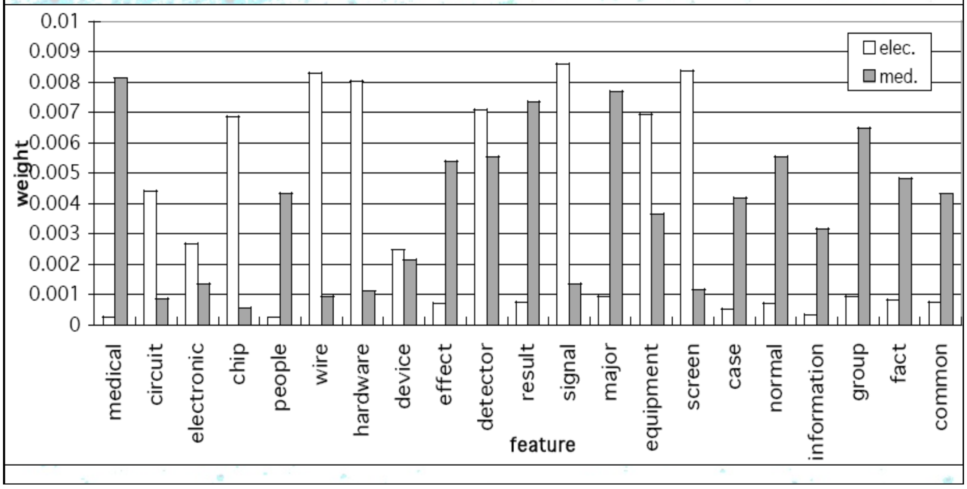
- 20 Newsgroups: messages collected from 20 different netnews newsgroups;
- Two class classification problem: electronics (981) and medical (990) classes;
- The original size of the dictionary is 24546.

NewsGroups [Electronic:981 – Medical:990]

S	Q	q	n	Ave Err	Min Err	K-means
1%	6217	1359	1971	11.5 (2.4)	9.5	49.6
2%	6217	583	1971	18.1 (11.8)	13.5	49.7
3%	6217	321	1971	21.0 (9.5)	16.8	49.6
4%	6217	201	1971	21.8 (0.4)	20.8	49.7
5%	6217	134	1971	29.1 (7.5)	23.3	49.6

Newsgroups (electronics-medical)

Words receive largest weights **within** the representative class



Results

- Selected keywords are representative of the underlying categories;
- The subspace clustering technique is capable of sifting the most relevant words, while discarding the spurious ones;
- Relevant keywords, combined with the associated weight values can be used to provide short summaries for clusters and to automatically annotate documents (e.g. for indexing purposes).

Clustering: An ill-posed Problem

- Document clustering: Based on content? Based on style? Based on authorship?
- Given a data set, different clustering algorithms are likely to produce different results.
- Given a data set, the same algorithm with different parameter settings is likely to produce different results. E.g.: k-means with different random initialization.
- What do we do?

Clustering: An ill-posed Problem

- Solutions:
  - CLUSTERING ENSEMBLES
  - SEMI-SUPERVISED CLUSTERING