Advanced Topics: An Overview

Topics

- Clustering
- Subspace clustering
- Ensembles of classifiers and clusterings
- Semi-supervised clustering
- Learning Metrics

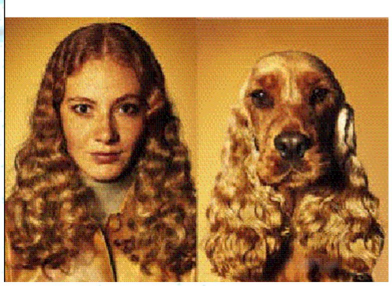
Clustering

- **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

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

$$D(x_i, x_j) = \sum_{k=1}^{q} d_k(x_{ik}, x_{jk})$$

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

$$\Rightarrow D(x_i, x_j) = \sum_{k=1}^{q} (x_{ik} - x_{jk})^2$$

Squared Euclidean distance

Combinatorial Algorithms

- These algorithms work directly on the observed data, without regard to a probability model describing the data.
- Commonly used in data mining, since often no prior knowledge about the process that generated the data is available.

Combinatorial Algorithms

$\boldsymbol{x}_i \in \Re^q,\ i = 1,\cdots,N$

Prespecified number of clusters $K,\ k \in \{1,\cdots,K\}$

Each data point $\boldsymbol{x}_i$ is assigned to one, and only one cluster

Goal : Find a partition of the data into K clusters that achieves a required objective, defined in terms of a dissimilarity function $D(\boldsymbol{x}_i, \boldsymbol{x}_k)$

Usually, the assignment of data to clusters is done so as to **minimize** a "loss" function that measures the degree to which the clustering goal is **not** met

Combinatorial Algorithms

Since the goal is to assign close points to the same cluster, a natural loss function would be :

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i \in C_k}\sum_{j \in C_k} D(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Within cluster scatter

Then, clustering becomes straightforward *in principle* : Minimize W over all possible assignments of the N data points to K clusters

Combinatorial Algorithms

Unfortunately, such optimization by complete enumeration is feasible only for very small data sets.

The number of distinct partitions is:

$$S(N,K) = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^N$$

For example:

$S(10,4) = 34{,}105$ $\quad$ $S(19,4) \approx 10^{10}$

We need to limit the search space, and find in general a good suboptimal solution

Combinatorial Algorithms

- **Initialization**: a partition is specified.
- **Iterative step**: the cluster assignments are changed in such a way that the value of the loss function is improved from its previous value.
- **Stop criterion**: when no improvement can be reached, the algorithm terminates.

Iterative greedy descent.

Convergence is guaranteed, but to local optima.

K-means

- One of the most popular iterative descent clustering methods.
- Features: quantitative type.
- Dissimilarity measure: Euclidean distance.

K-means

The "***within cluster point scatter***" becomes:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j\in C_k}\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2$$

$W(C)$ can be rewritten as:

$$W(C) = \sum_{k=1}^{K}|C_k|\sum_{i\in C_k}\left\|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k\right\|^2$$

(obtained by rewriting $(\boldsymbol{x}_i - \boldsymbol{x}_j) = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_k) - (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_k)$)

where

$\overline{\boldsymbol{x}}_k = \frac{1}{|C_k|}\sum_{i\in C_k}\boldsymbol{x}_i$ is the mean vector of cluster C_k

$|C_k|$ is the number of points in cluster C_k

K-means

The objective is:

$$\min_{C} \sum_{k=1}^{K} |C_k| \sum_{i \in C_k} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k\|^2$$

We can solve this problem by noticing:

for any set of data S

$$\bar{\boldsymbol{x}}_S = \arg\min_{\boldsymbol{m}} \sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{m}\|^2$$

(this is obtained by setting $\frac{\partial \sum_{i \in S} \|\boldsymbol{x}_i - \boldsymbol{m}\|^2}{\partial \boldsymbol{m}} = 0$)

So we can solve the enlarged optimization problem:

$$\min_{C, \boldsymbol{m}_k} \sum_{k=1}^{K} |C_k| \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2$$

K-means: The Algorithm

1. Given a cluster assignment C, the total within cluster scatter

$\sum_{k=1}^{K} |C_k| \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2$ is minimized with respect to the $\{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_K\}$

giving the means of the currently assigned clusters;

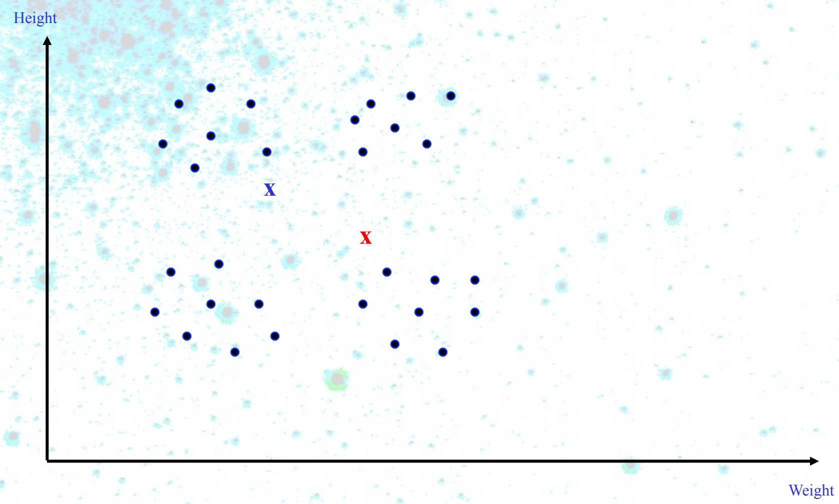
2. Given a current set of means $\{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_K\}$,

$\sum_{k=1}^{K} |C_k| \sum_{i \in C_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|^2$ is minimized with respect to C

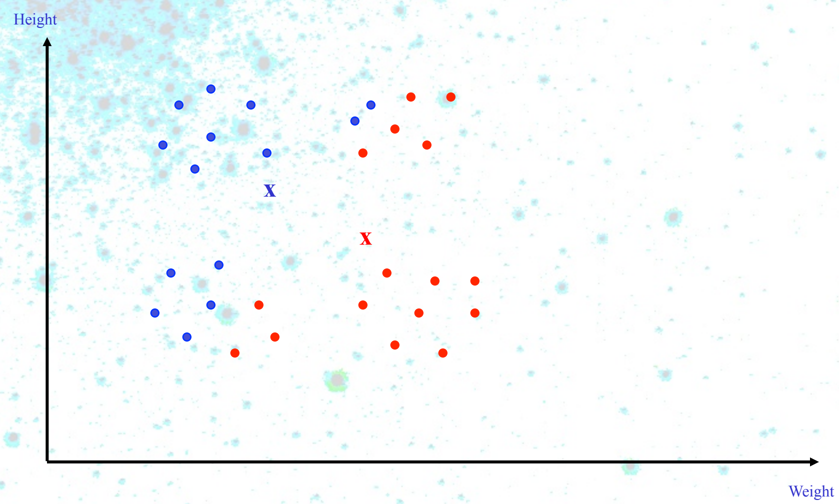
by assigning each point to the closest current cluster mean;

3. Steps 1 and 2 are iterated until the assignments do not change.

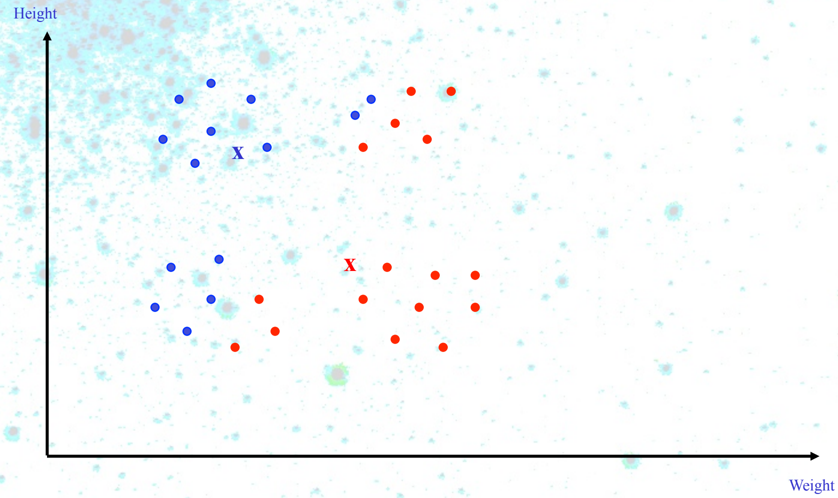
K-means Example (K=2)
Initialize Means



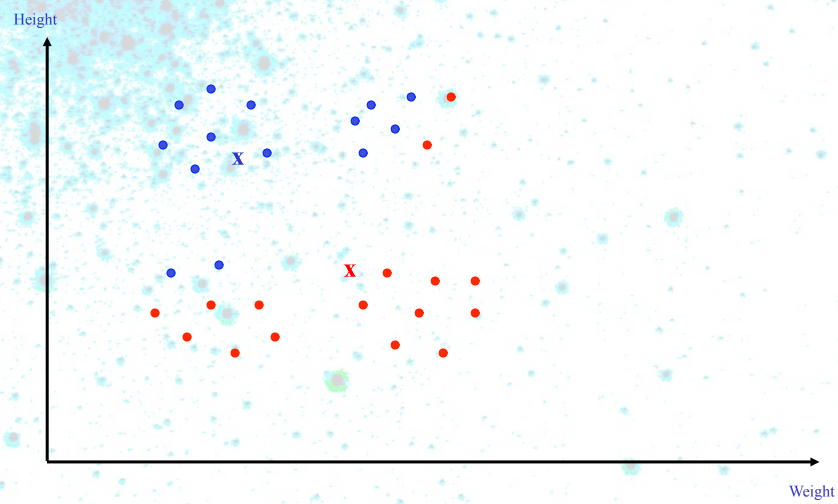
K-means Example
Assign Points to Clusters



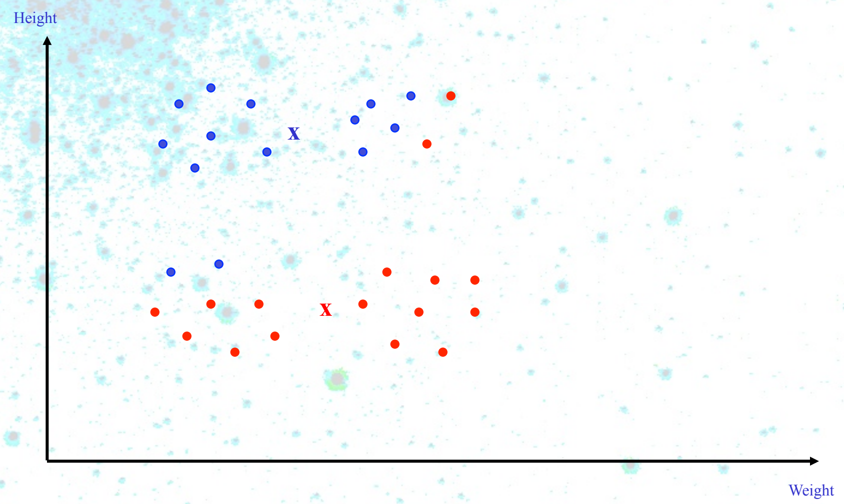
K-means Example
Re-estimate Means



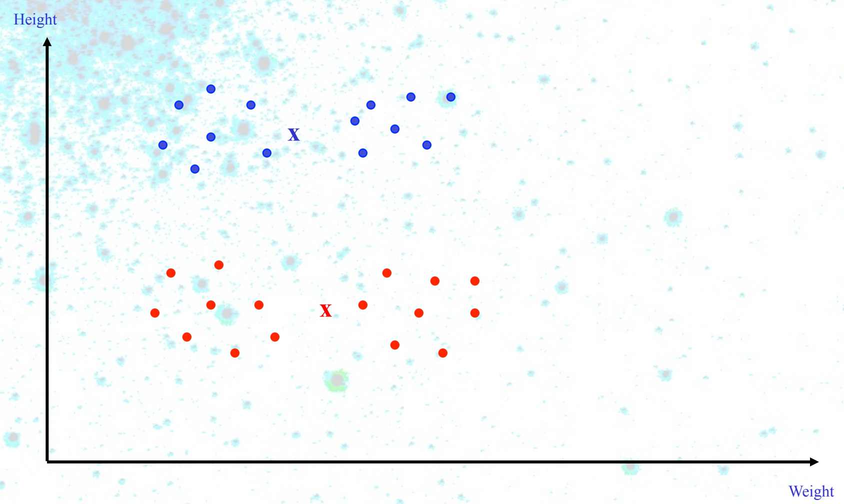
K-means Example
Re-assign Points to Clusters



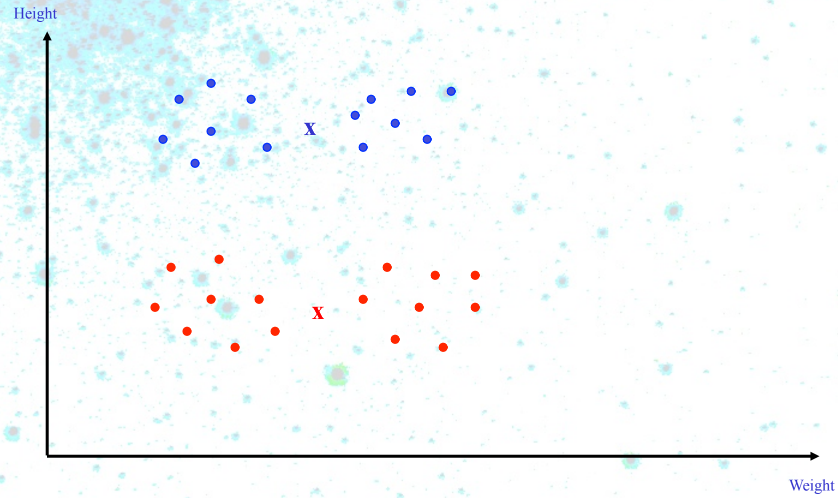
K-means Example
Re-estimate Means



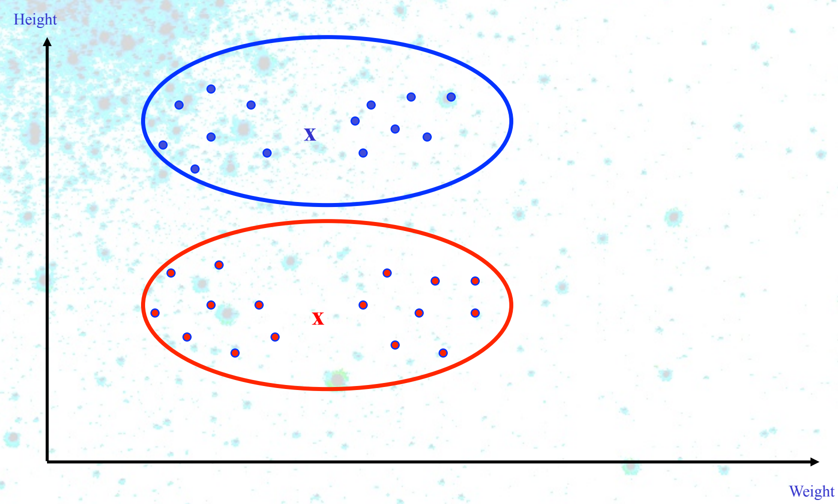
K-means Example
Re-assign Points to Clusters



K-means Example
Re-estimate Means and Converge



K-means Example
Convergence



K-means: Properties and Limitations

- The algorithm converges to a local minimum
- The solution depends on the initial partition
- One should start the algorithm with many different random choices for the initial means, and choose the solution having smallest value of the objective function

K-means: Properties and Limitations

- The algorithm is sensitive to outliers
- A variation of K-means improves upon robustness (K-medoids):
  - Centers for each cluster are restricted to be one of the points assigned to the cluster;
  - The center (*medoid*) is set to be the point that minimizes the total distance to other points in the cluster;
  - K-medoids is more computationally intensive than K-means.

K-means: Properties and Limitations

- The algorithm requires the number of clusters K;
- Often K is unknown, and must be estimated from the data:

We can test $K \in \{1,2,\cdots,K_{\max}\}$

Compute $\{W_1,W_2,\cdots,W_{\max}\}$

In general: $W_1 > W_2 > \cdots > W_{\max}$

K^* = actual number of clusters in the data,

when $K < K^*$, we can expect $W_K >> W_{K+1}$

when $K > K^*$, further splits provide smaller decrease of W

Set $\hat{K}^*$ by identifying an "elbow shape" in the plot of W_k

Clustering

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

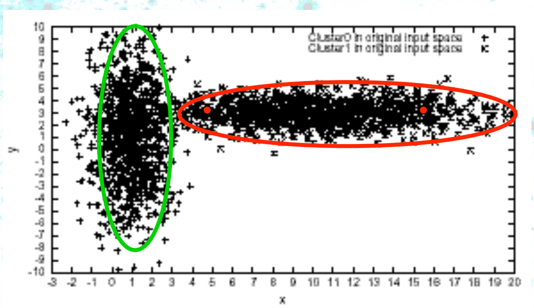
$$D(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sum_{k=1}^{q}\left(x_{ik} - x_{jk}\right)^2$$ Squared Euclidean distance

- Assumption: All features are **equally important**;
- Such approaches fail in high dimensional spaces

Clustering: The Curse of Dimensionality

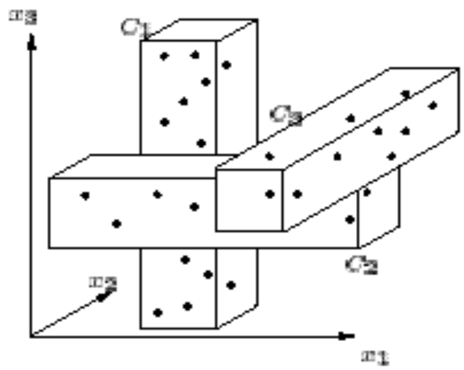
- A full-dimensional distance is often irrelevant, as the farthest point is expected to be almost as close as the nearest point;
- In high dimensional spaces, it is likely that, for any given pair of points within the same cluster, there exist at least a few dimensions on which the points are far apart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

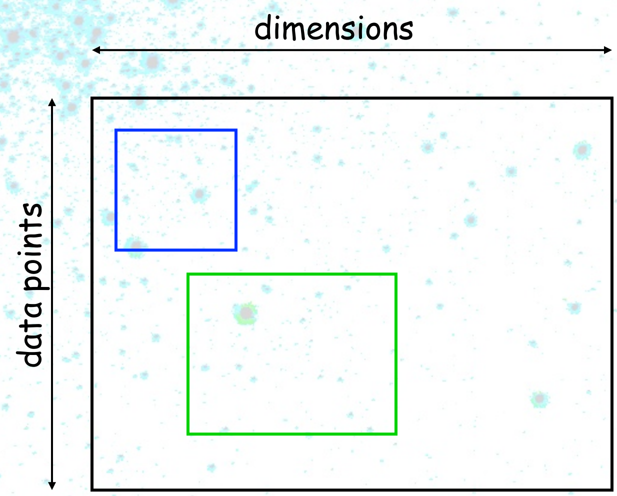
- We cannot prune off dimensions without incurring a loss of crucial information;
- Global dimensionality reduction techniques, e.g. PCA, do not handle well situations where different clusters are dense in different subspaces;
- The data presents **local structure**

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



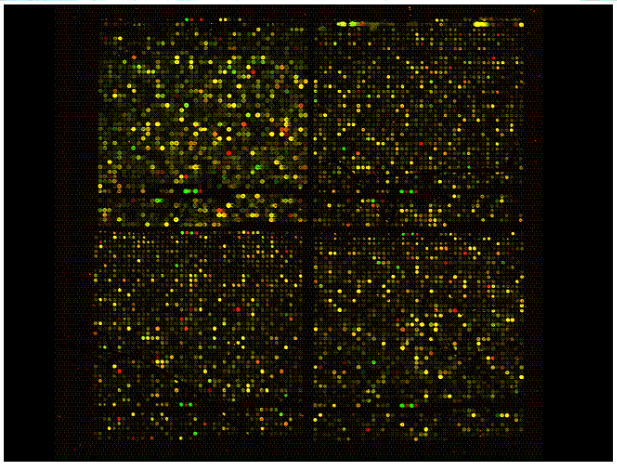
Subspace clustering

Other terms used:

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

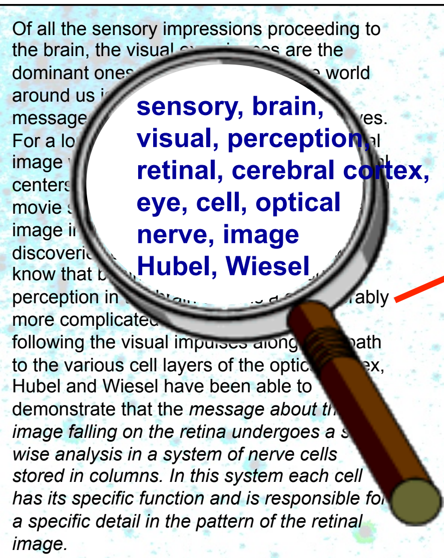
Subspace clustering

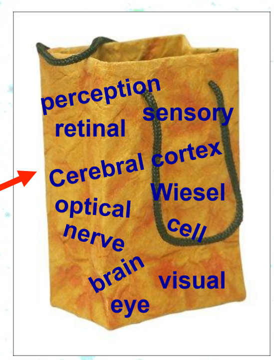
- Important problem in practice
- Real life problems:
  - Are high dimensional
  - Present local structure



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.

Approaches to Subspace Clustering

- Most methods provide "hard" clustering solutions at data level.
- In each subspace typically features are equally weighted.

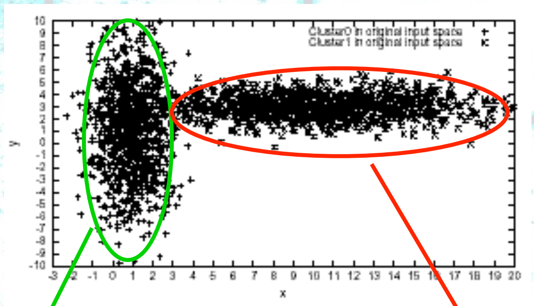
- More recently: "soft subspace clustering" and weighted subspace clustering approaches.

Locally Adaptive Clustering (LAC)

- Task: ***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.

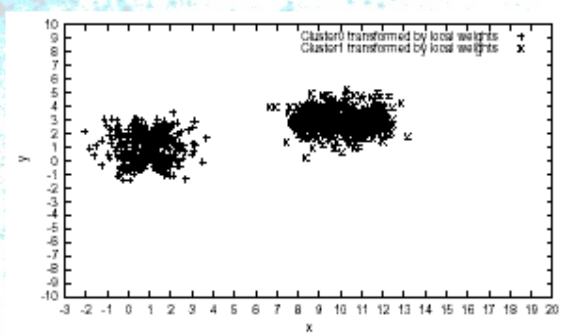
Locally Adaptive Clustering: Example



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

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

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

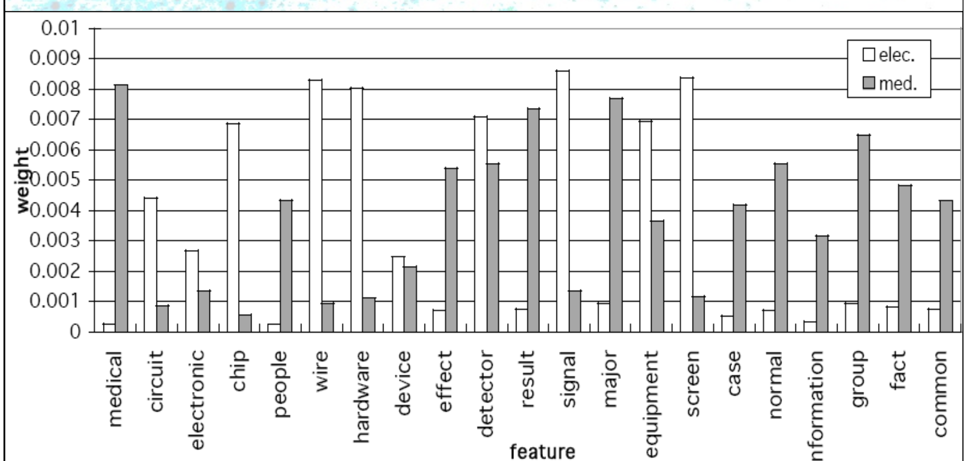
- The result of LAC is twofold:
  - It achieves a ***clustering*** of the documents;
  - It achieves the identification of ***cluster-dependent keywords*** via a continuous term-weighting mechanism.

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 (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

Ensembles of Classifiers and Clusterings

- How to construct effective ensembles
- Bagging and Boosting
- Analysis in term of bias and variance
- Tradeoff between diversity and accuracy
- Subspace clustering ensembles
- …

Semi-supervised learning

Two fundamental approaches:

- Learning distance functions
- Modify objective function to enforce constraints

Learning Metrics

- Supervised vs. unsupervised methods
- Local vs. global methods