

The Curse of dimensionality

- There are two problems that come from the dimensionality of a dataset
 - The computational cost of processing the data (scalability of the algorithms)
 - The quality of the data (more probability of bad data)
- There are two elements that define the dimensionality of a dataset
 - The number of examples
 - The number of attributes
- Usually the problem of having too many examples can be solved using sampling.
- Attribute reduction has different solutions

Reducing attributes

- Usually the number of attributes of the dataset has an impact on the performance of the algorithms:
 - Because their poor scalability (cost is a function of the number of attributes)
 - Because the inability to cope with irrelevant/noisy/redundant attributes
- There are two main methodologies to reduce the number of attributes of a dataset
 - Transforming the data to a space of less dimensions preserving somewhat the original data (*dimensionality reduction*)
 - Eliminating the attributes that are not relevant for the goal task (*feature subset selection*)

Dimensionality reduction

- We are looking for a new dataset that preserves the information of the original dataset but has less attributes
- Many techniques have been developed for this purpose
 - Projection to a space that preserve the statistical model of the data (PCA, ICA)
 - Projection to a space that preserves distances among the data (Singular Values Decomposition, Multidimensional Scaling, Random Projection, Nonlinear Scaling)

Projections

We transform the dataset to another feature space

- **Principal Component Analysis:**

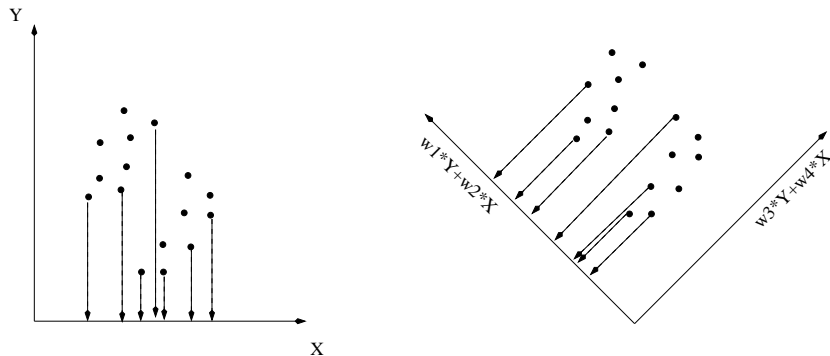
- We assume that attributes follow a gaussian distribution
- Data is projected to a set of orthogonal dimensions (components) that are linear combination of the original attributes. Global variance is preserved.
- The new dimensions are uncorrelated and can be ordered by the original information they preserve.
- We can keep the subset that preserves the most information

- **Independent Component Analysis:**

- We assume non gaussian data.
- Transforms the dataset projecting the data to a set of variables statistically independent (all statistical momentums are independent).

Principal Component Analysis

We look for a projection of the original space to a space with orthogonal dimensions (linearly independent)

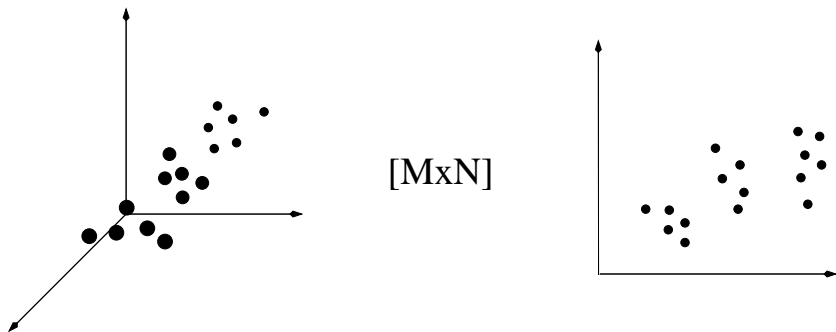


Multidimensional Scaling

- **Multidimensional Scaling:** Projects a dataset to a space with less dimensions preserving the distances among the data
- A projection matrix is obtained by optimizing a function of the pairwise distances (stress function)
- There are different methods for linear (least squares, Sammong mapping, classical scaling, ...) and non linear transformation (ISOMAP, LLE, ...)

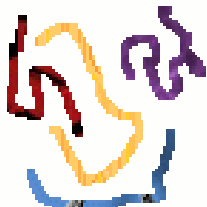
Projections

- **Multidimensional scaling:** Project the data set directly to a space of lower dimensionality preserving relative distances among the data (many techniques for linear and non-linear transformations)

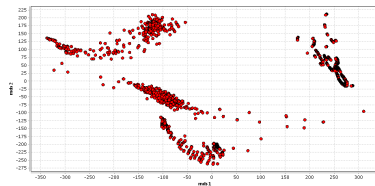


Projections - Example

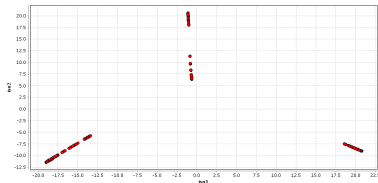
Data



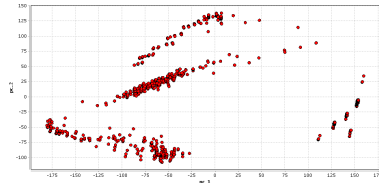
Classical MDS



ISOMAP



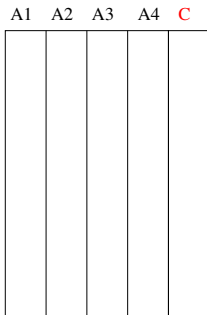
PCA



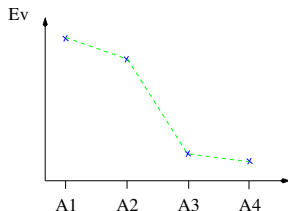
Attribute selection - Filters

Filters:

- We assume that we have an evaluation measure that allows for each attribute to assess its relevance respect with the target concept
- A ranking of the attributes is computed from the individual relevance to the target (computationally cheap)
- From the ranking a cutting point is decided and the firsts in the ranking are selected
- Examples: Entropy (ID3), χ^2 test, Relief



$$Ev(A1,C) > Ev(A2,C) > Ev(A3,C) > Ev(A4,C)$$



Attribute selection - Filter - Relief-F

Procedure: Relief-F

Input: W vector of feature weights initialized to 0

$X \leftarrow$ random sample of the dataset

foreach $x \in X$ **do**

$NH \leftarrow k$ nearest neighbors of x of different class (near hit)

$NM \leftarrow k$ nearest neighbors of x of the same class (near miss)

foreach $n \in NH$ and all features i **do**

if $n_i \neq x_i$ **then**

 decrease w_i value

foreach $n \in NM$ and all features i **do**

if $n_i \neq x_i$ **then**

 increase w_i value

Attribute selection - Wrappers

Wrappers

- Subsets of the set of features are evaluated until the more adequate is found (computationally expensive)
- A learning algorithm is used to evaluate the quality of each
- Exhaustive search is unfeasible (Hill-climbing, Simulated Annealing, Best First, Beam Search, Genetic algorithms, ...)
- Two strategies for greedy algorithms: Forward Selection, Backward Elimination

