

4.0 - Data Mining

Sébastien Lemieux

Elitra Canada Ltd.

lemieuxs@iro.umontreal.ca

slemieux@elitra.com



Overview

- What is it!
- Dimensionality
- Normalization
- Dimensionality reduction
- Clustering
 - Metrics;
 - K-mean;
 - Hierarchical clustering.
- Deriving significance.

What is data mining?

- Discover hidden facts in databases. Uses a combination of machine learning, statistical analysis, modeling techniques and database technology.
- Establishes relationships and identify patterns.
- “A technique geared for the user who typically does not know exactly what he’s searching for.”

Normalization

- Why normalize?
- Most used approach for spotted array data: loess.
 - Local regression is performed on a “sliding window” over the data (next slide).
- The Durbin-Rocke normalization:
 - A variance-stabilizing transformation for gene-expression microarray data (2004) *Bioinformatics*, **18**:S105-S110

$$g(y) = \ln \left(y - \alpha + \sqrt{(y - \alpha)^2 + c} \right)$$

Loess normalization

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Dimensionality

- Data mining algorithms work on a set of labeled or unlabeled data points.
 - Labeled: supervised learning (neural network, HMM).
 - Unlabeled: unsupervised learning (k-mean, SOM).
- Each data point is a vector in a continuous or discrete space.
 - Dimensionality of a dataset is the length of those vectors.
- The algorithm tries to identify some parameters to model the distribution of data points.
 - The dimensionality of the parameter's space is necessarily proportional to the dimensionality of the dataset.
 - Unicity of the ideal parameters require that a number of independent data points equal or greater than the number of parameters.

Dimensionality (cont.)

- The choice of how to convert your data to vectors will have great impact on your chance of success:
 - Too many dimensions: will require too many parameters and thus more data points that you may have. The algorithm may *over-train*.
 - Too few dimensions: you may lose some information that can be essential to identify a feature (pattern, class, etc.).
- The case of microarray data, data mining on:
 - samples,
 - genes,
 - significantly over/under expressed genes...
- Curse of dimensionality... (Bellman, 1961)

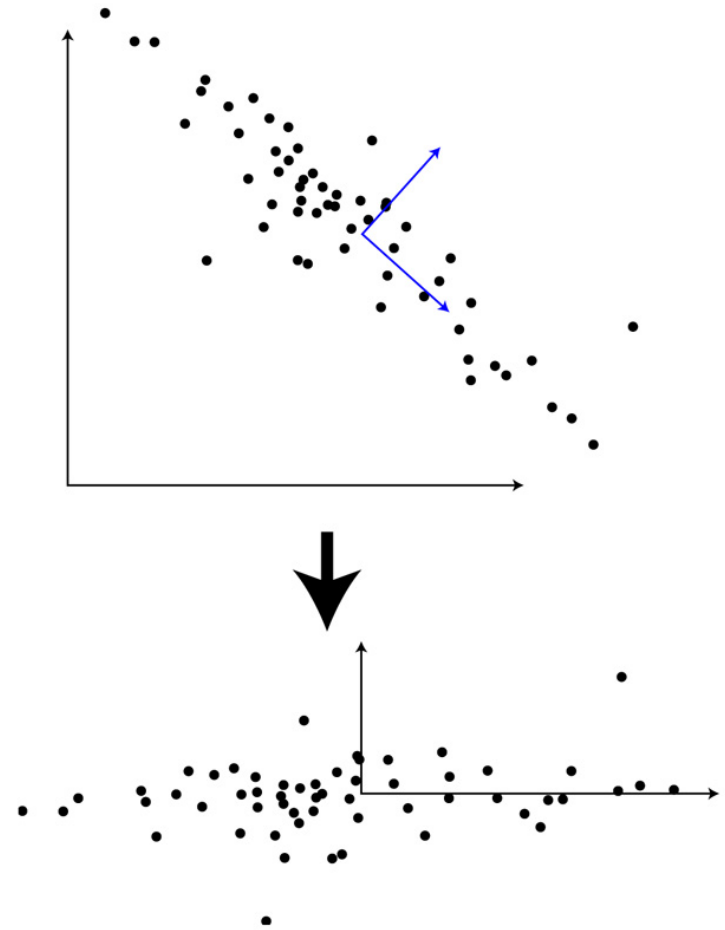
Dimensionality reduction

- Different methods:
 - Keeping just the relevant dimensions: guesswork.
 - Keeping just the relevant genes!
 - Principal component analysis (PCA):
 - Rotation of the dataset that maximizes the dispersion across the first axes.
 - Singular value decomposition (SVD):
 - Press *et al.*, p. 59.
 - Neural networks: map input vector onto itself, hidden units correspond to the reduced vector.

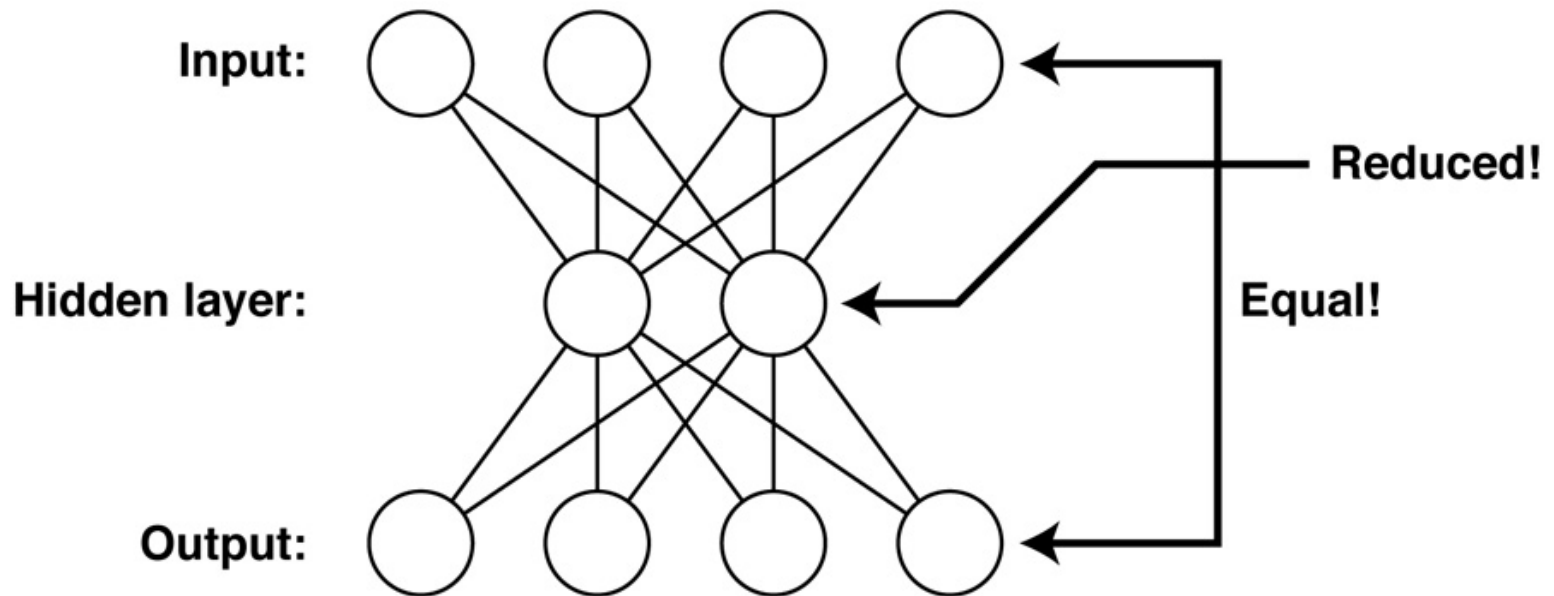
Principal Component Analysis

- Subtract the mean (centers the data).
- Compute the covariance matrix, Σ .
- Compute the eigenvectors of Σ , sort them according to their eigenvalues and keep the M first vectors.
- Project the data points on those vectors.
- Also called the Karhunen-Loeve transformation.

$$\Sigma = \sum_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$



Neural networks for dimensionality reduction



- Neural Networks for Pattern Recognition (Bishop, 1995), p. 314.

Clustering

- Unsupervised learning algorithm:
 - No information is assumed to be known for the different samples.
- Identifies classes within the data provided:
 - Genes that seem to react the same way in different conditions;
 - Different conditions that seem to induce the same response.
- Applications:
 - Parameters of the discovered classes can be used to classify new data points.
 - Parameters by themselves may be viewed as the measurement of a data feature.
 - Exploratory step.

Metrics

- A metric quantifies the differences between to elements of a data set.
 - Similarity metrics;
 - Dissimilarity or distance metrics.

- Example of metrics:

- Manhattan:
$$D(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n |a_i - b_i|$$

- Euclidean:
$$D(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

K-mean

- Input: a set of vector $\{ x_i \}$, a metric D , and the number of classes to identify, n .
- 1. Each vector receives a random label in $[1, n]$.
- 2. Compute the average of each class: μ_j
- 3. For each vector x_i update its label to the closest class center according to $D(x_i, \mu_j)$.
- 4. If at least one vector changed its label, repeat 2-4.
- Output: each vector is labeled and each class is defined by its center.

K-mean *(cont.)*

- *Should* not be sensitive to random initialization.
 - The algorithm can be repeated with different initialization to provide a distribution of solution. (beware of “cluster swapping”!)
- Perform a linear segmentation of the data space.
 - Similar to a Voronoï diagram:
 - Clusters should be linearly separable.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

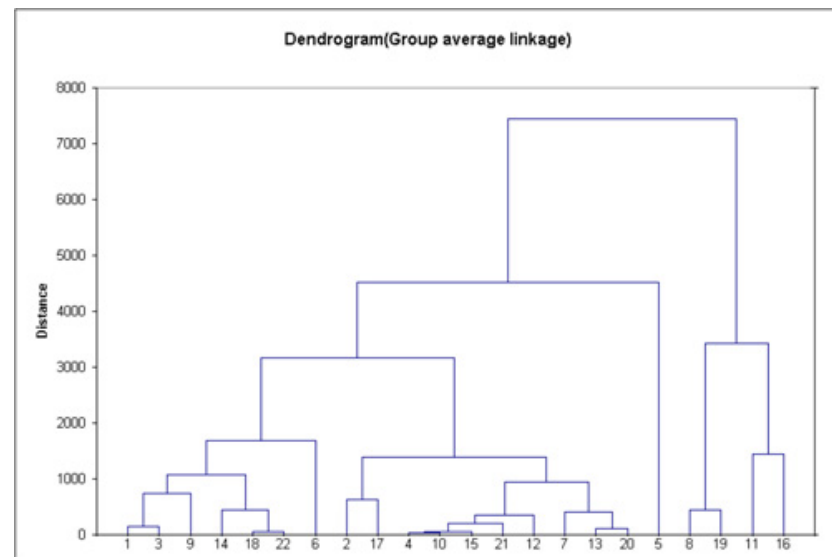
Hierarchical clustering

- Input: a set of vector $\{ x_i \}$, a metric D , and a group metric F .
 1. Create a set of groups $G = \{ g_i \}$ such that each $g_i = (x_i)$.
 2. Remove the two closest groups g_i and g_j from G and insert $(D(g_i, g_j), g_i, g_j)$.
 3. If $|G| > 1$, repeat 2-3.
- Output: a tree structure.

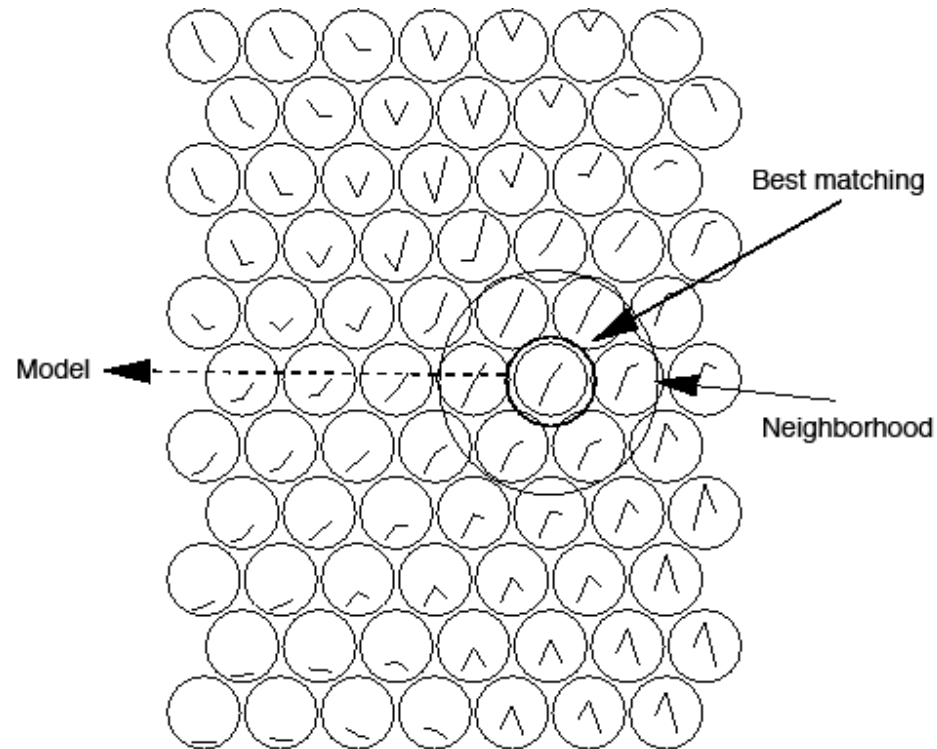
$F(a, b)$ typically returns the $\max(D(x_i, x_j))$ where x_i in a and x_j in b .

Hierarchical clustering (cont.)

- The resulting may not be unique.
- The formation of group is very sensitive to noise in the data.
- The separation of two data points in two cluster does not mean they are necessarily distant!
- There is a “hidden” metric used to compute the distance between two groups:
 - minimum;
 - maximum;
 - average.



Self organizing maps (SOM)



Timo Honkela (Description of Kohonen's Self-Organizing Map)

Self organizing maps (SOM) *(cont.)*

- Input: a set of vector $\{ x_i \}$, and an array of random model vectors $m_i(0)$.
- 1. Randomly initialize each model vector.
- 2. For each x_i :
 - a) Identify the model, $m_i(t)$, that is closest according to the metric.
 - b) $m_w(t + 1) \leftarrow m_i(t) + \alpha(t)[x_i - m_w(t)]$, for each $i \in N_c(t)$
 - c) Compute: $E = \sum_{x_i} \|x_i - m_c(t)\|$, where $m_c(t)$ is the best-matching model for x_i .
- 3. Repeat 2. until convergence of E .
- T. Kohonen, *Self-Organizing Maps*, Springer, 1995.

Deriving significance

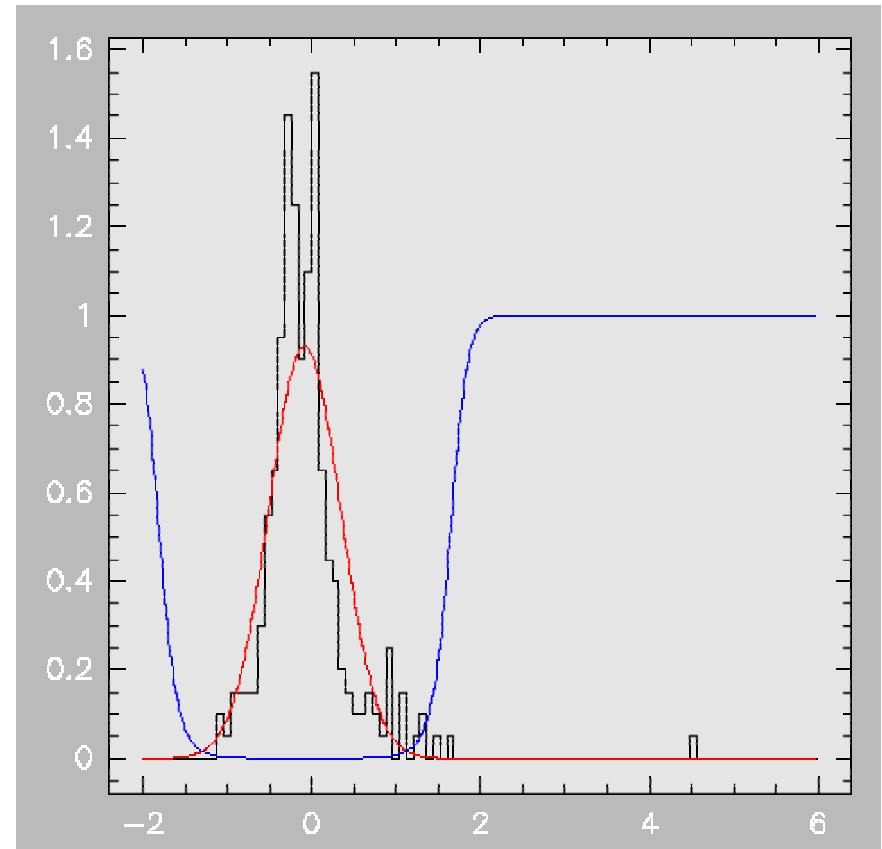
- Traditionally, a 2-fold difference in spot intensities between two treatments was assumed to be “significant”.
- What is significant?
 - Did not happen at random;
 - Is not an artifact.
- All methods involve modeling a random process and using this model to quantify the likelihood that an observed value was generated by that random process.
 - Corresponds to the classic hypothesis test in statistics.

Deriving significance *(cont.)*

- Modeling the random process is often done by assuming normality:
 - Microarray intensities are not normally distributed.
 - Ratios are never normally distributed.
 - log-ratios are close enough to a normal distribution.
- Outliers are not produced by the random process and they will affect the parameterization of the selected distribution.
 - They should be removed or accounted for in some way.
 - Hard cutoffs may work, but they introduce biases...
- Information about this random process requires a significant amount of data!
 - Can be done on a gene or sample basis.

Deriving significance *(cont.)*

- An example:
 - using the log-ratio for one specific gene across 800 experiments.
 - EM algorithm (a variant of the k-mean) is used to smoothly remove the outliers.
 - The resulting normal distribution is an approximation of the expected variation on that gene.



Other approaches...

- EM algorithm:
 - Used to approximate an unknown distribution;
 - Can be used to return a probabilistic classifier.
- Neural networks
- HMM:
 - Could be used for time-series applications.
- Support Vector Machines:
 - Powerful supervised learning technique;
 - Known to work well in high-dimension space;
 - Theory is “scary”! And thus is often used as a black box.
- Statistics!
 - Avoid a dogmatic approach...

References

- C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
- R. Schalkoff, *Pattern recognition: statistical, structural and neural approaches*, John Wiley & Sons, 1992.
- N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, 2000.
- W. H. Press et al., *Numerical Recipes in C - The Art of Scientific Computing*, Cambridge University Press, 1992.