Microarray Data Generation and Analysis : Class Discovery and Class Prediction from Human Cancer Microarray Datasets

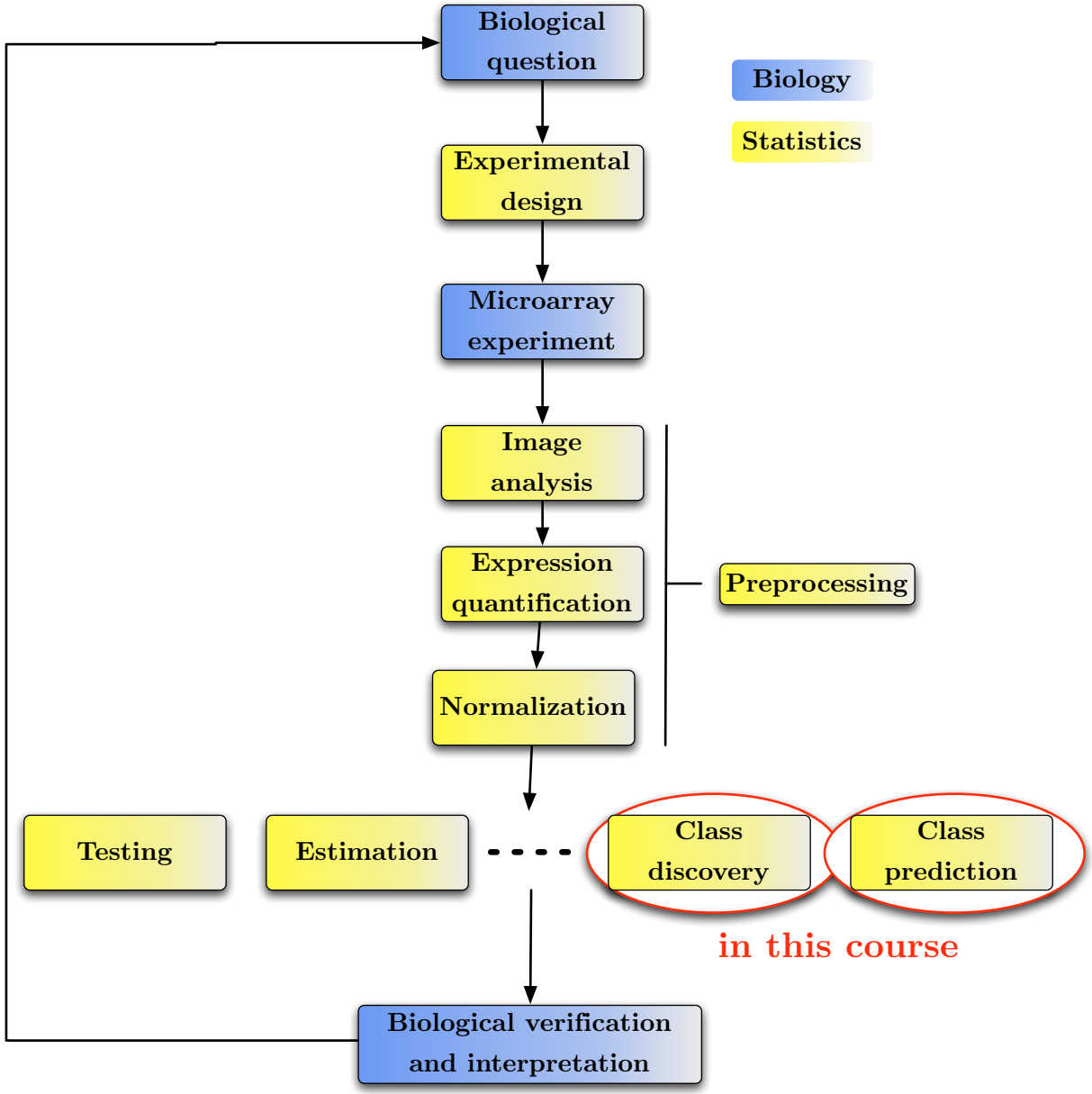
Haibe-Kains B[1,2] Sotiriou C[1] Bontempi G[2]

[1]Unité Microarray, Institut Jules Bordet

[2]Machine Learning Group, Université Libre de Bruxelles

February 28, 2006

Microarray Analysis Design



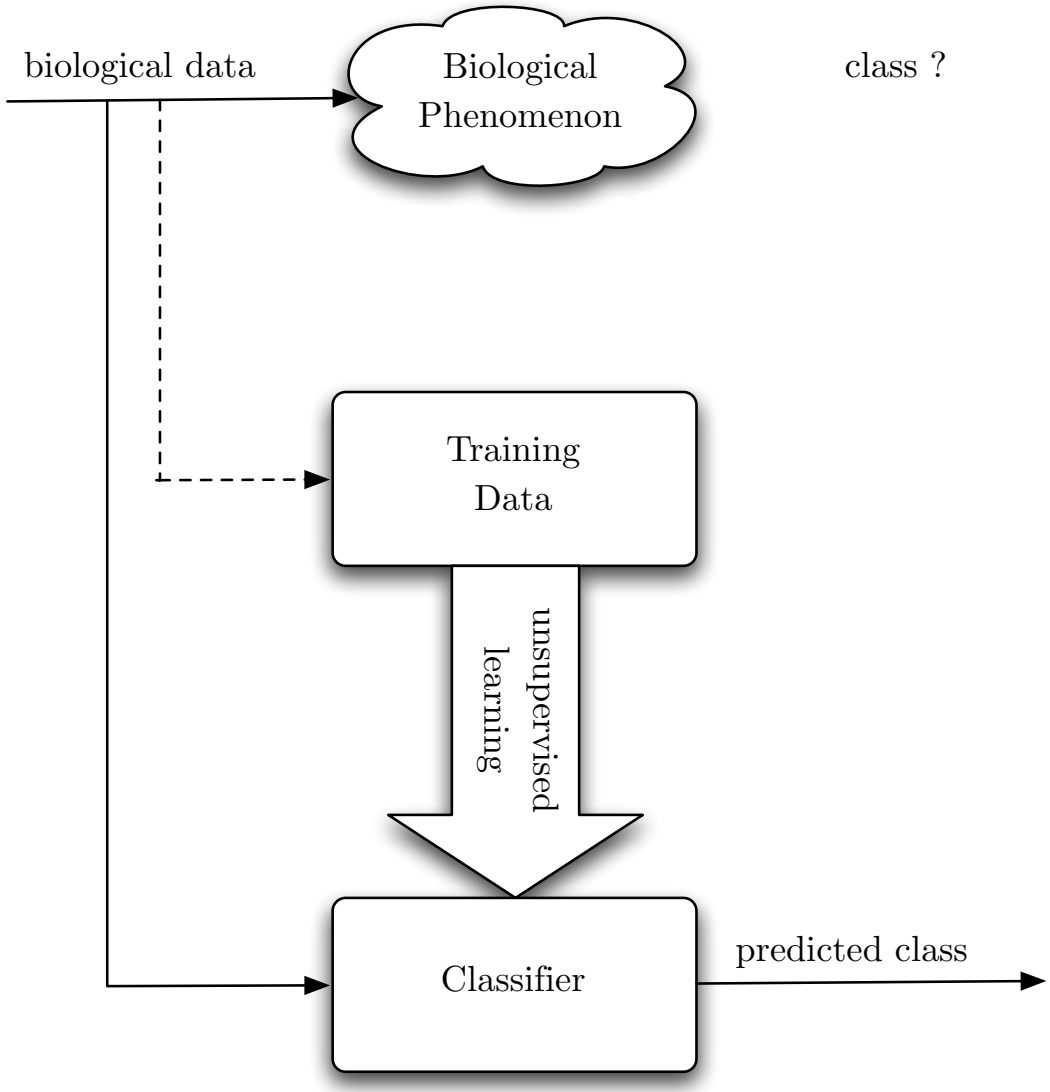
Part I

Class Discovery

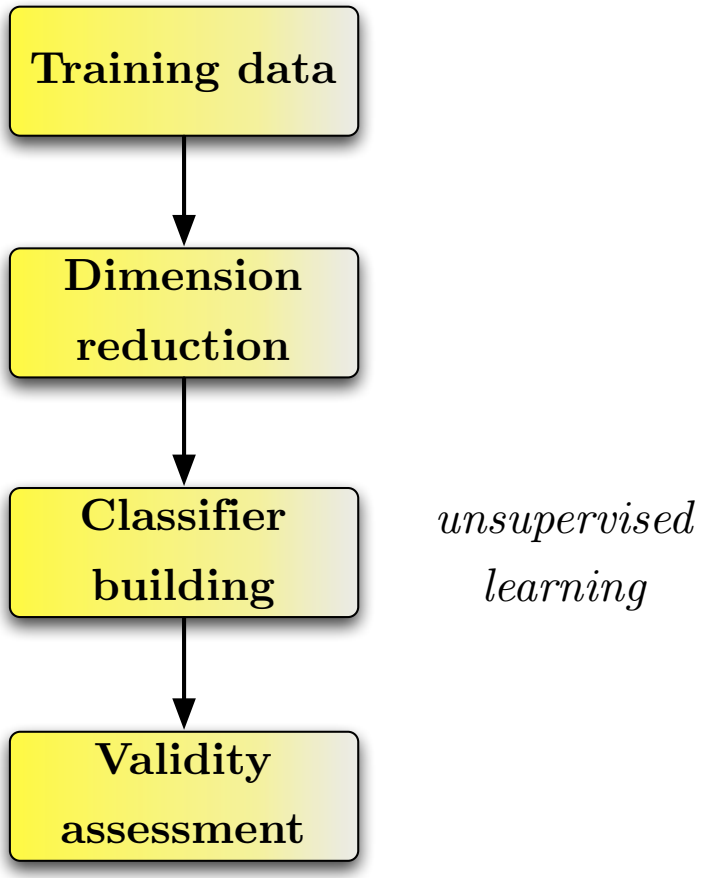
Class Discovery

Table of Contents

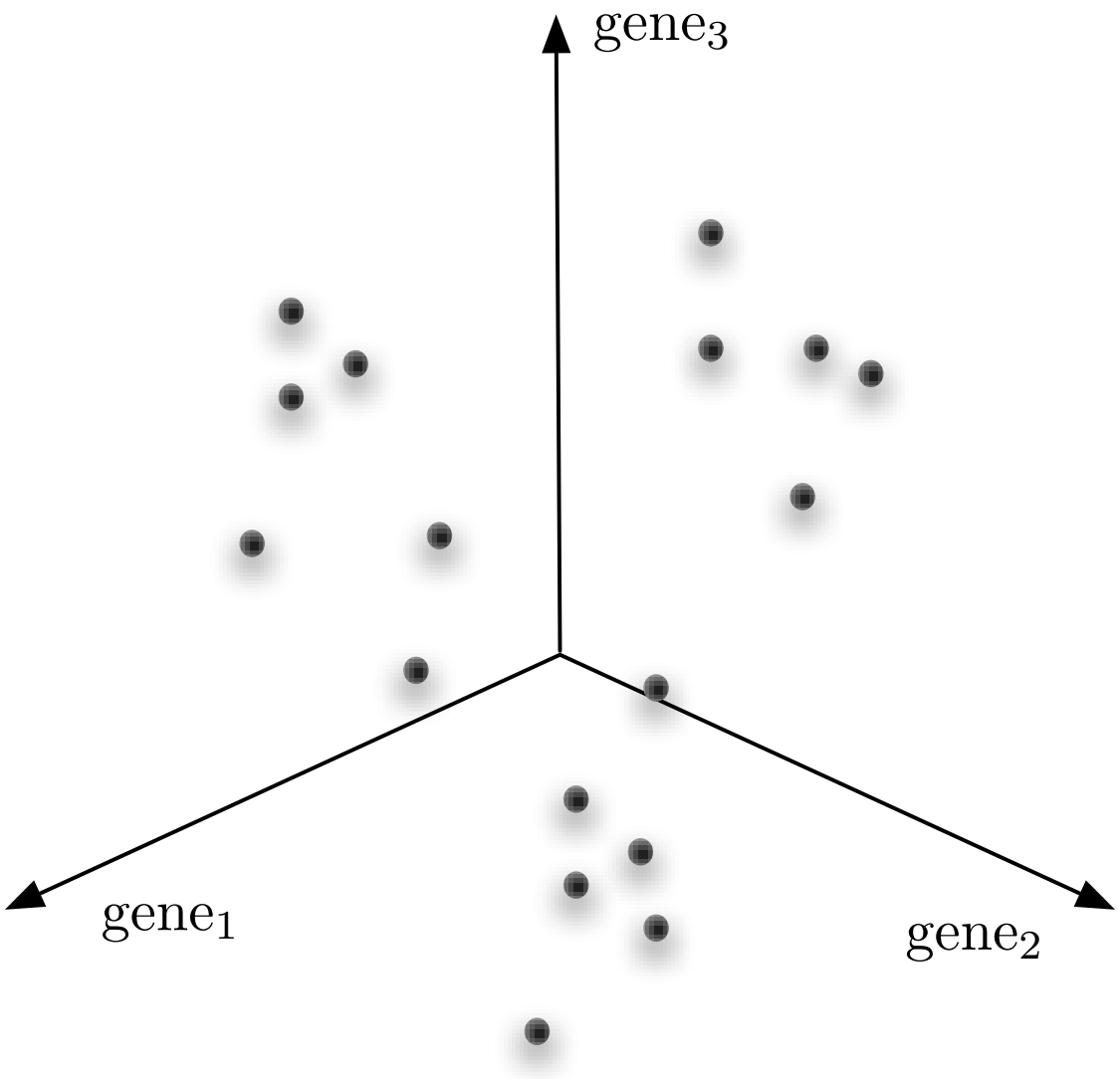
Introduction



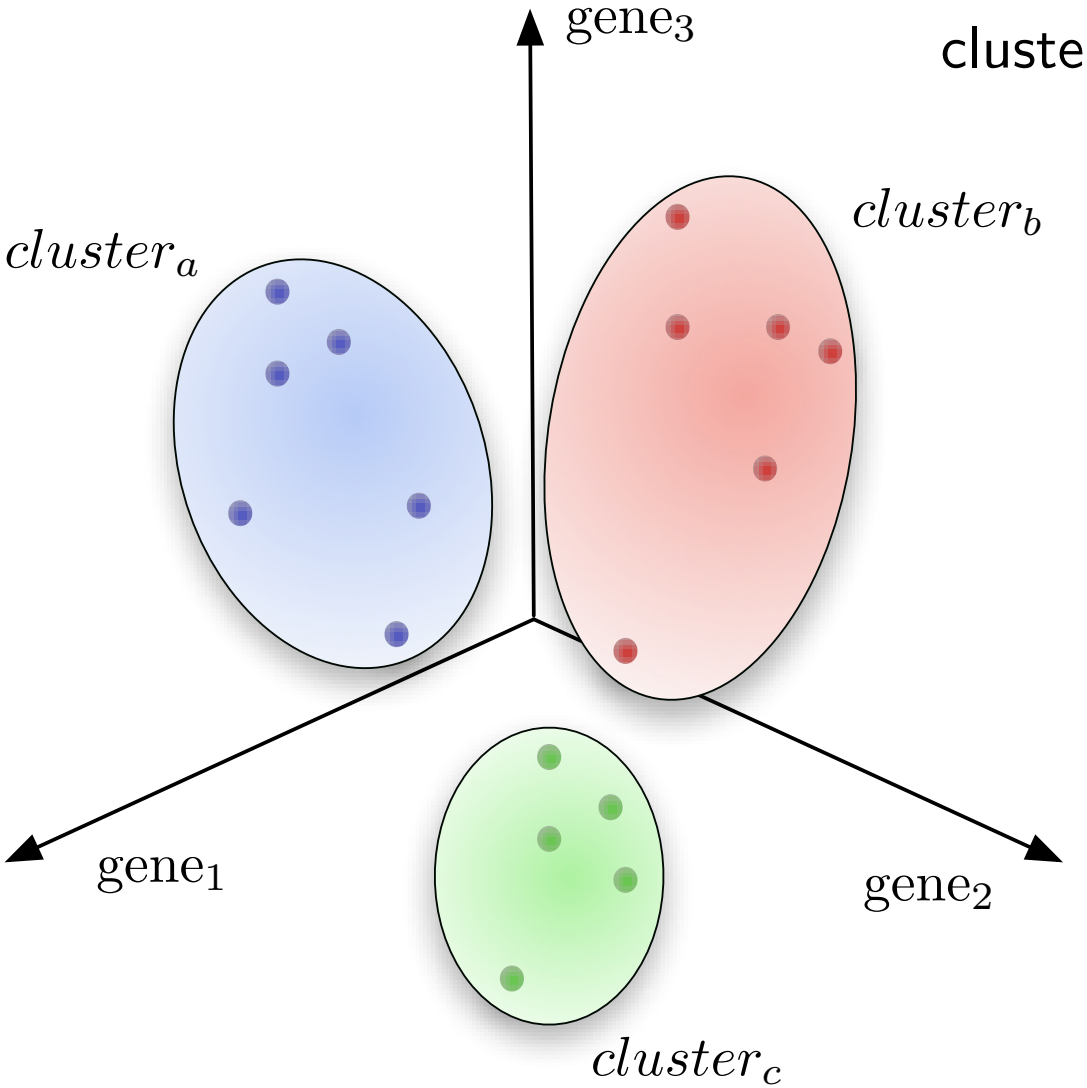
Class Discovery Design



Clustering



Example of samples drawn in the gene space



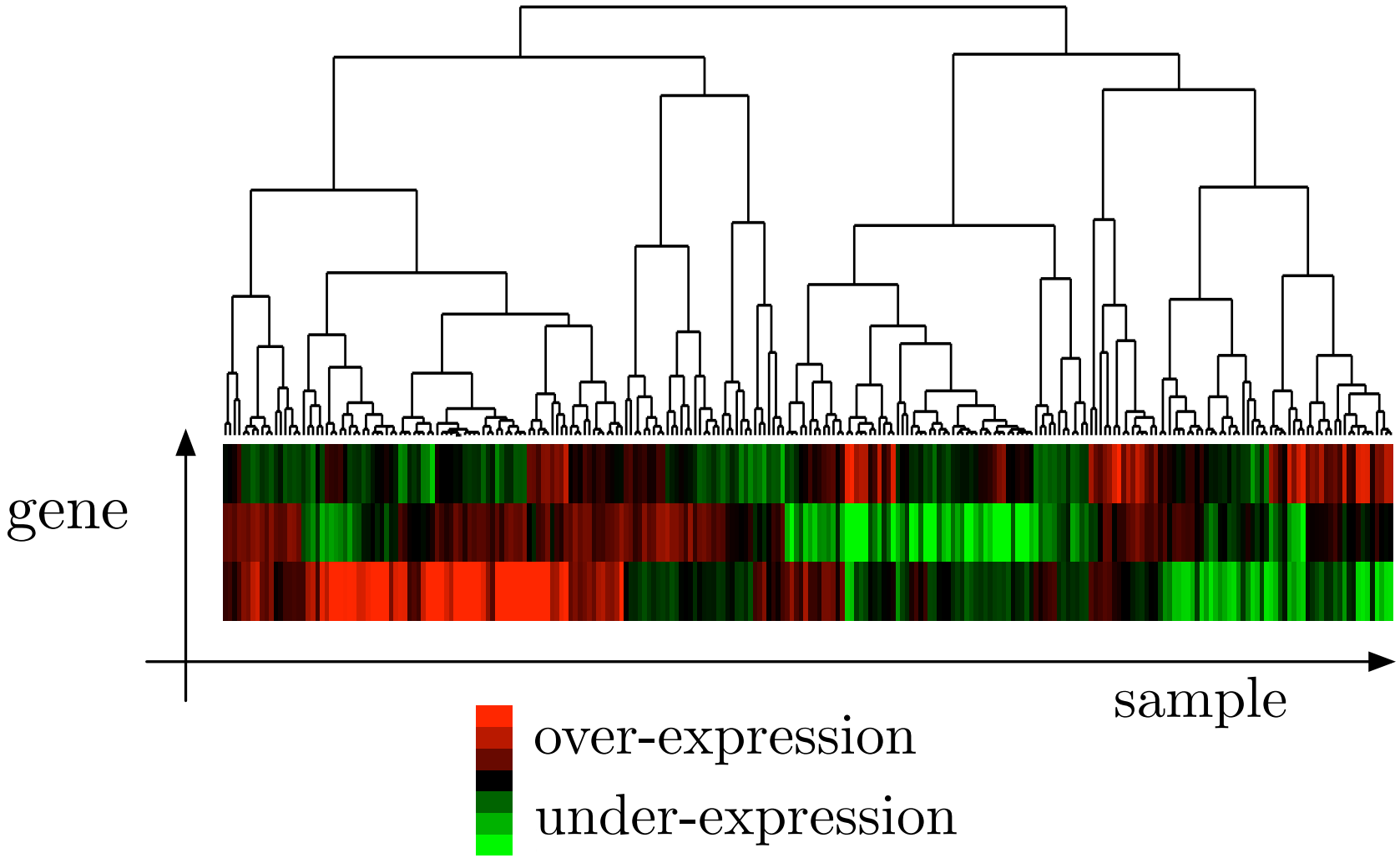
The samples can be divided into 3 clusters.

Hierarchical Clustering

- Widely used clustering method [Hartigan, 1975, Eisen et al., 1998].
- Organizing objects in a hierarchical binary tree (**dendrogram**) based on their **degree of similarity**.
  - distance : 1 - *uncentered Pearson correlation*, *Kendall's tau*, *Euclidean*
  - linkage : *complete*, *single*, *average*, *centroid*
- Advantages :
  - no number of clusters to specify (full hierarchical binary tree)
  - deterministic
  - computationally efficient.
- Disadvantages :
  - dendrogram may be misleading
  - need to define a metric of similarity and a linkage
  - need complete data.

Hierarchical Clustering

Example



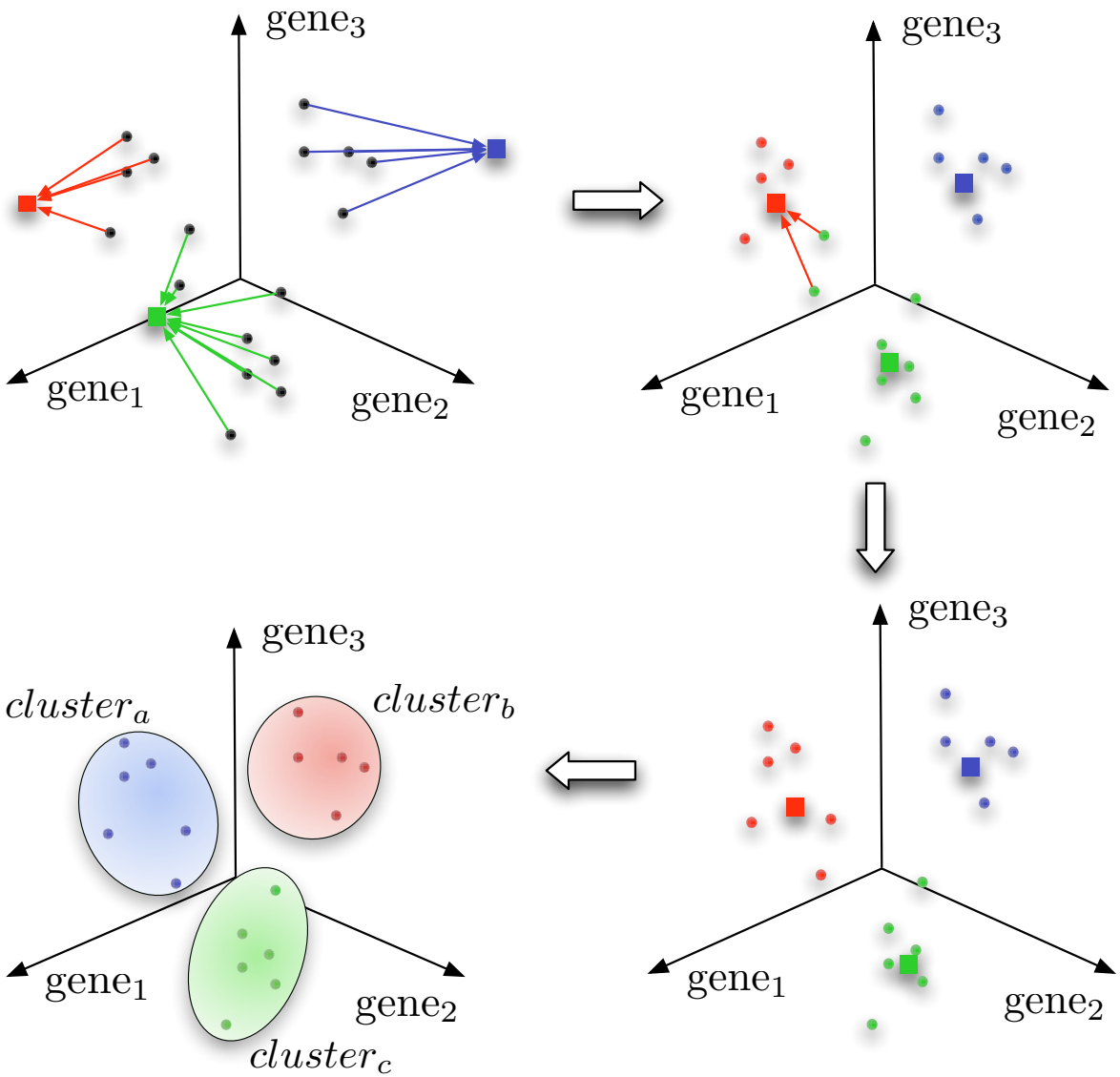
K-Means

- Method introduced in [MacQueen, 1967].
- Partitioning objects in k disjoint subsets.
- Minimization of the distance between the samples and the cluster centroids.

- Advantage :
  - computationally efficient.
- Disadvantages :
  - need to specify the number of clusters
  - need to define a distance
  - not deterministic
  - need complete data.

K-Means

Example : $k = 3$



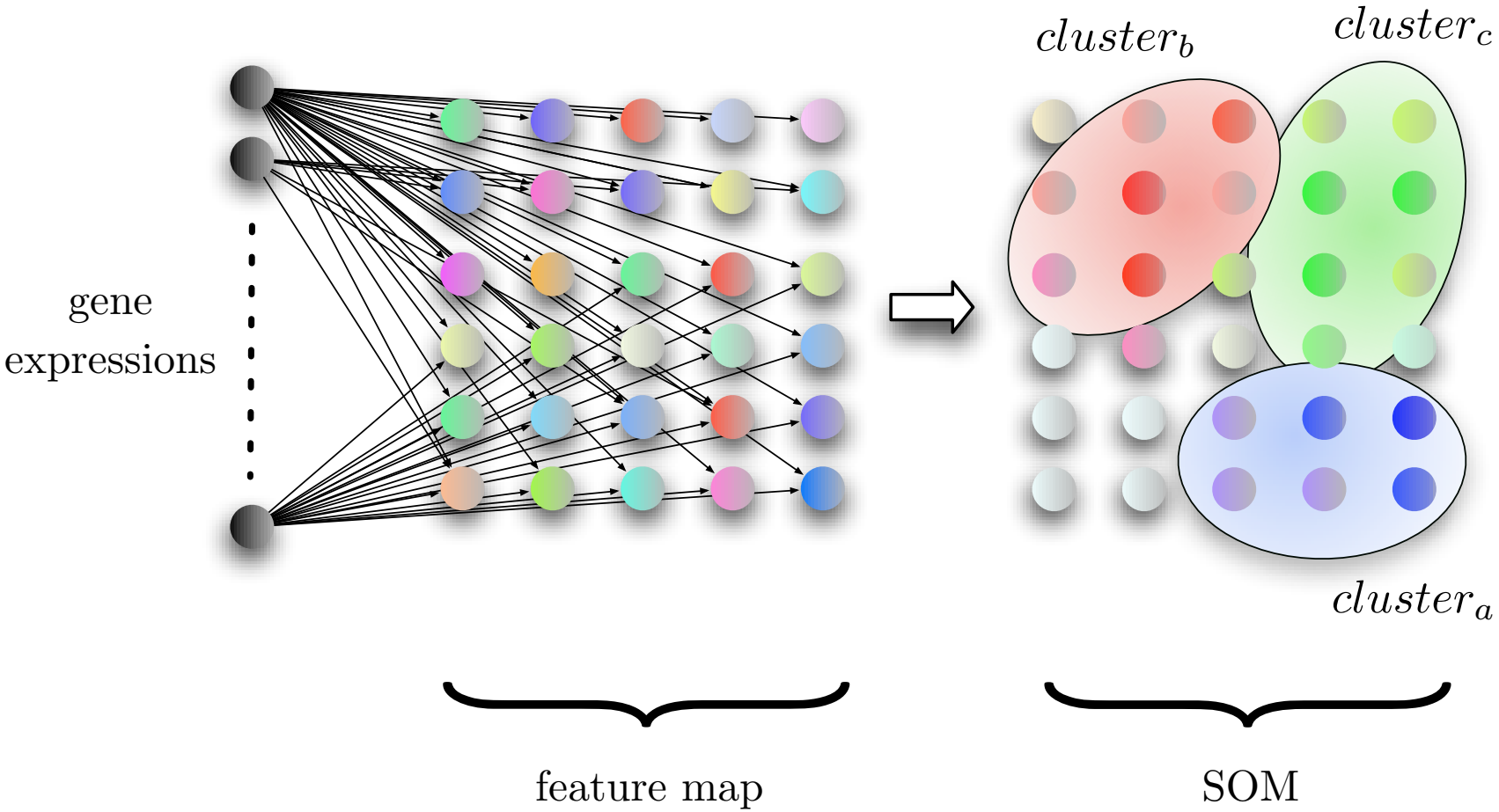
Self-Organizing Maps

- Method introduced in [Kohonen, 1997].
- Use of self-organizing neural networks to reduce dimension.

- Advantages :
  - no number of clusters to specify
  - display similarities.
- Disadvantages :
  - need to define the size of the feature map
  - need to define the neighborhood and update functions
  - not deterministic
  - need complete data
  - computationally intensive.

Self-Organizing Map

Suite



Dimension Reduction

- Microarray experiments generate a huge amount of data (thousands of probes).
- Microarray data are noisy.
- Common practice is to reduce dimension of the data because
  - most of the probes are non-informative
  - in removing these probes, we remove noise.
- Widely used methods :
  - filtering based on variance
  - multidimensional scaling.

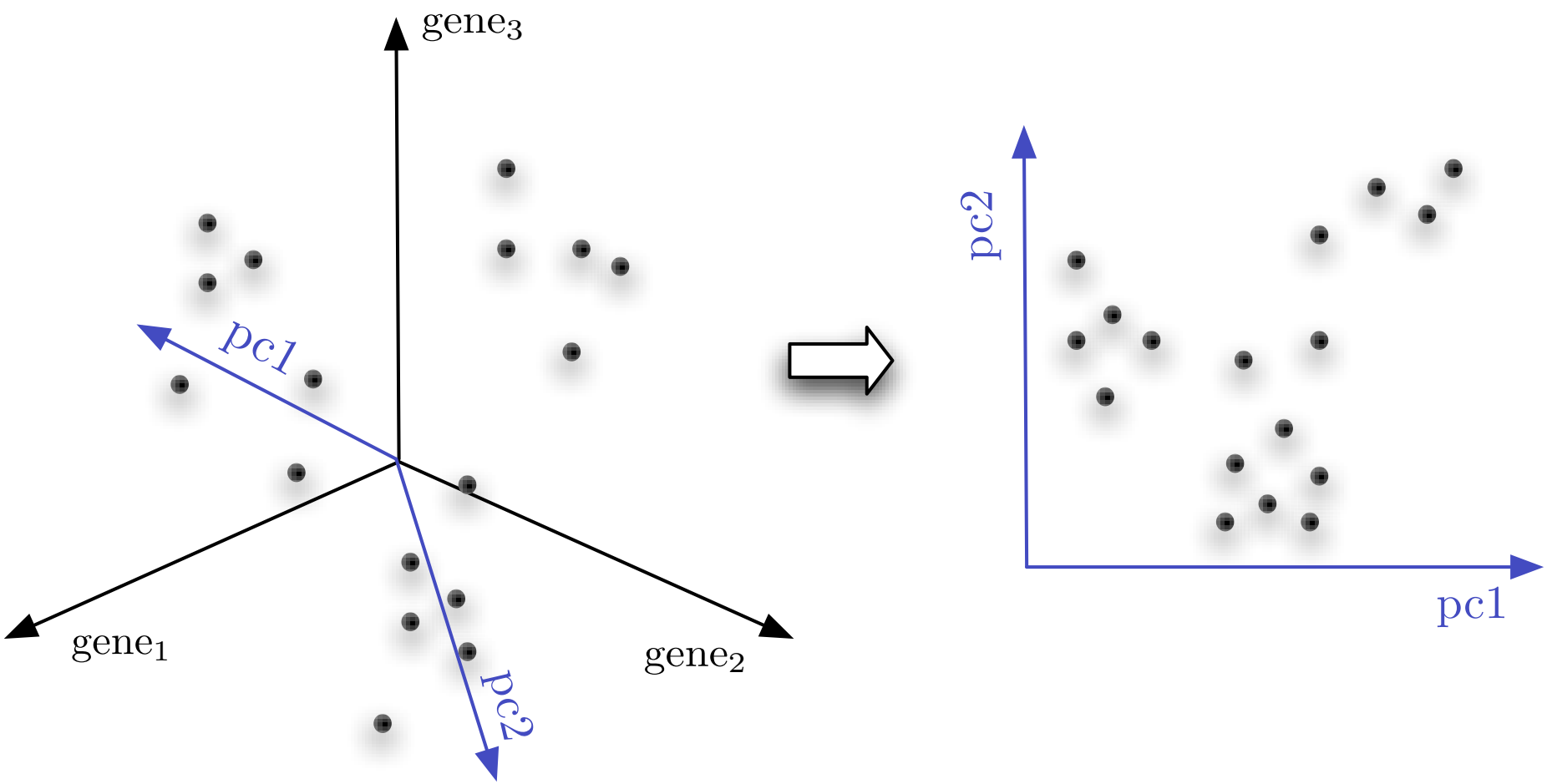
Multidimensional Scaling

- Provide a low (e.g. 2 or 3) dimensional representation of the distances which conveys information on the relationships between the objects [Kruskal and Wish, 1978].
- MDS with Euclidean distance = principal component analysis (PCA)
  - rotation of the original variable maximizing the variance
  - new axes = principal components
  - principal components are orthogonal.

- Advantages :
  - deterministic
  - computationally efficient.
- Disadvantages :
  - need to select the number of principal components
  - need complete data
  - new dimensions are complex to interpret.

Multidimensional Scaling

Example : Reduction from 3 to 2 Dimensions Using PCA



Clustering Stability

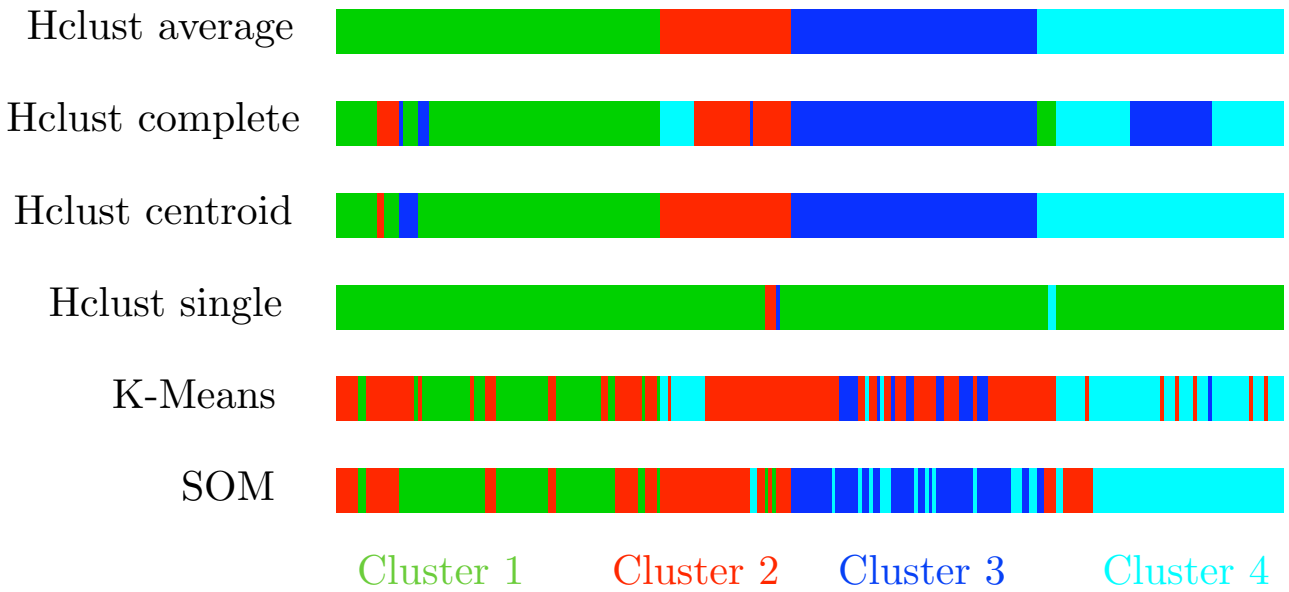
- Clustering algorithms always find structure in the data.
- Need of methods to assess the reliability of the discovered classes.

Procedure

1. Perturb the original dataset
   - by resampling the original dataset using jackknife [Ben-Hur et al., 2002]
   - by randomized projections in lower dimensional subspaces preserving approximately the distances between samples [Valentini, 2006].
2. Generate several clusterings.
3. Compute statistics assessing the reliability of clusters
   - single individual clusters inside a clustering
   - overall clustering (estimate of the "optimal" number of clusters)
   - confidence by which object may be assigned to each cluster.

Conclusion

- Pay attention to
  - select a meaningful distance depending of the problem under study
  - the impact of feature selection before clustering (also called *semi-supervised clustering*)
  - look at the stability of the clustering w.r.t. the dimension reduction and the dataset.
- Keep in mind that all these methods may give different results :



Example with only 3 probes (KI67, ESR1 and PGR) and 254 Tamoxifen treated patients.

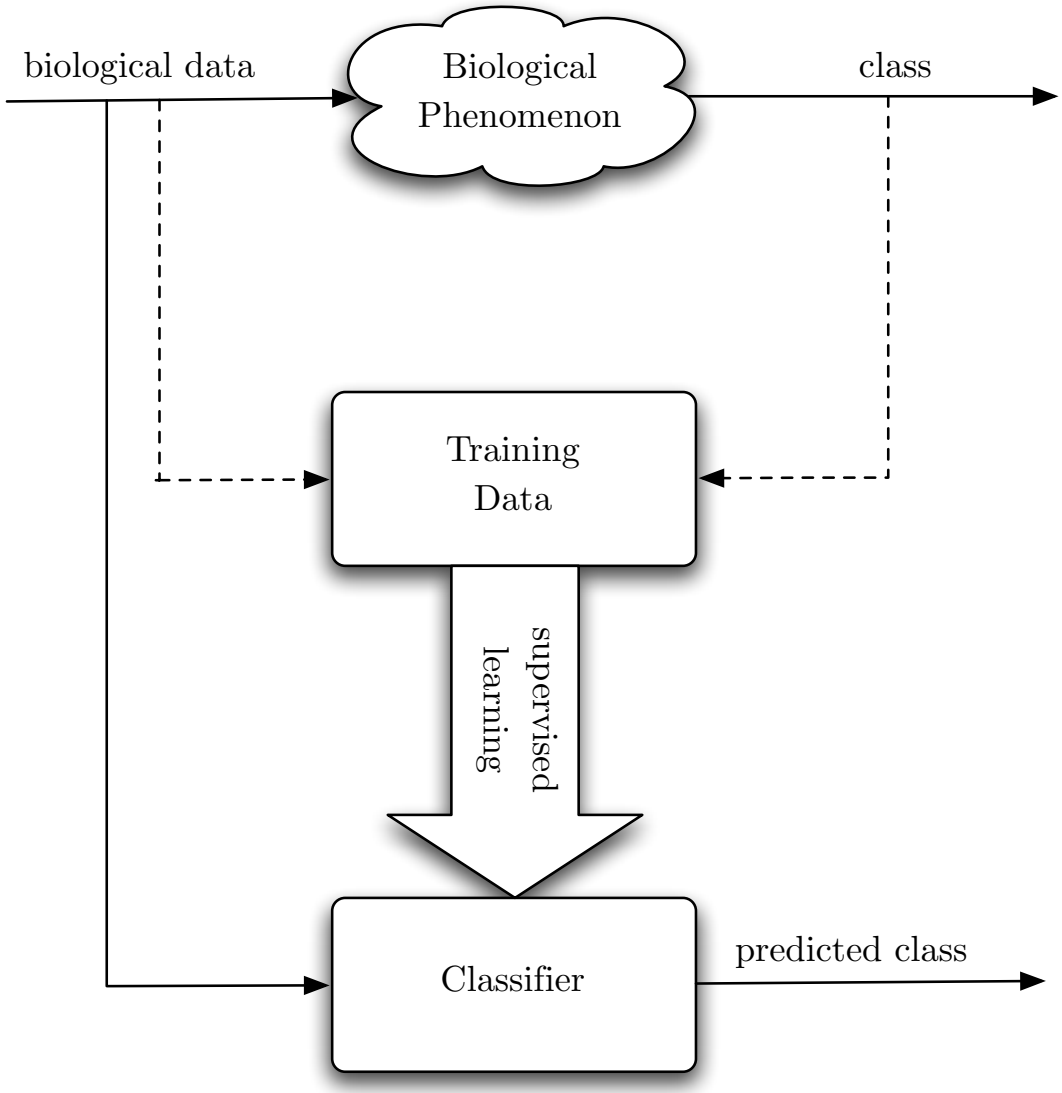
Part II

Class Prediction

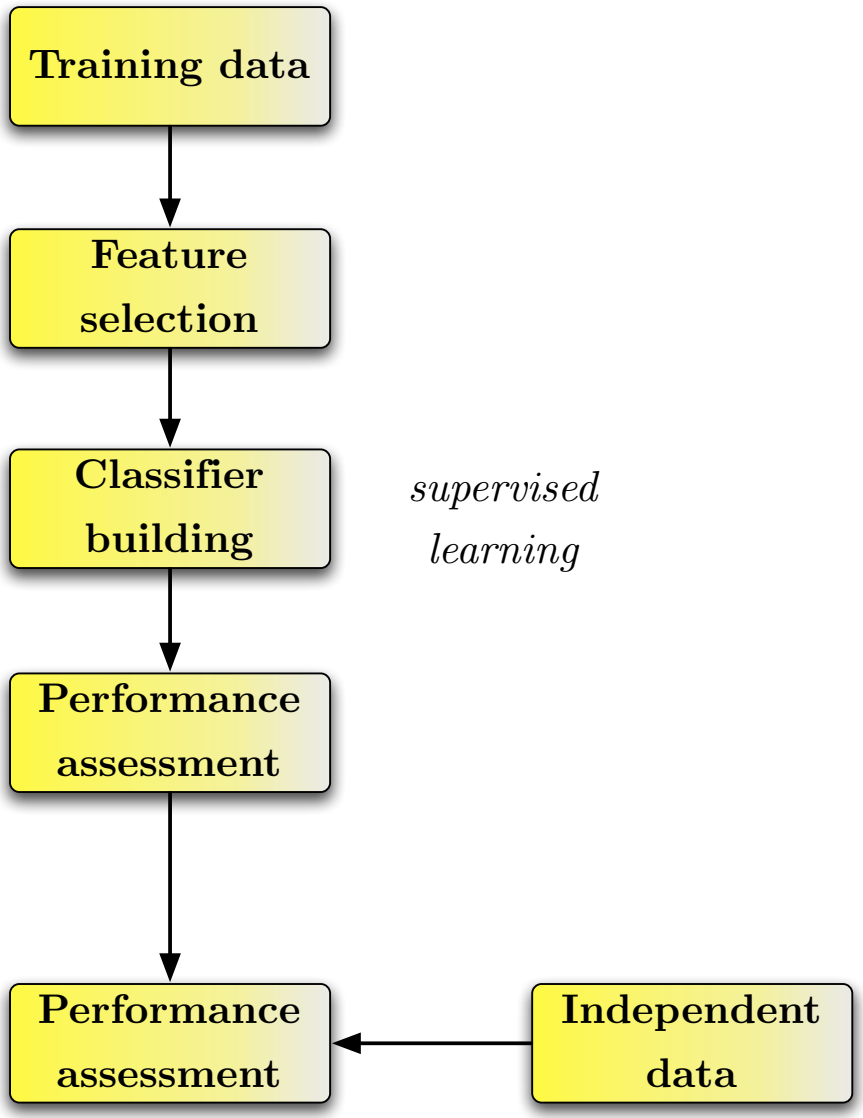
Class Prediction

Table of Contents

Introduction



Class Prediction Design



Classifiers

- Example of Classifiers
  - Linear : **Logistic Regression**, Naive Bayes, Linear and Quadratic Disciminant Analysis, ...
  - Non-linear : **K-Nearest Neighbors**, **Support Vector Machines**, Classification Trees, Artificial Neural Networks, ...
- Some classifiers can not deal output multiple classes directly
  - all pairwise classifications with a voting scheme
  - one against all classifications

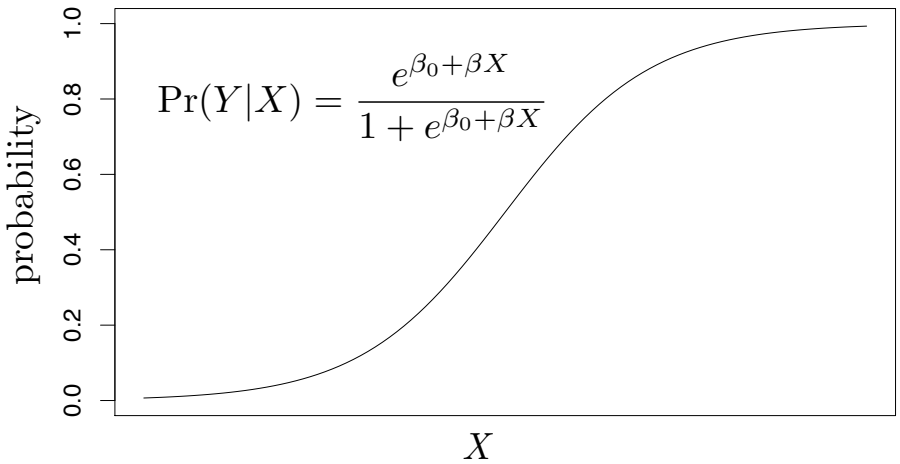
Logistic Regression

- Variation of ordinary regression used when :
  - output is a dichotomous variable
  - input variables are continuous, categorical, or both

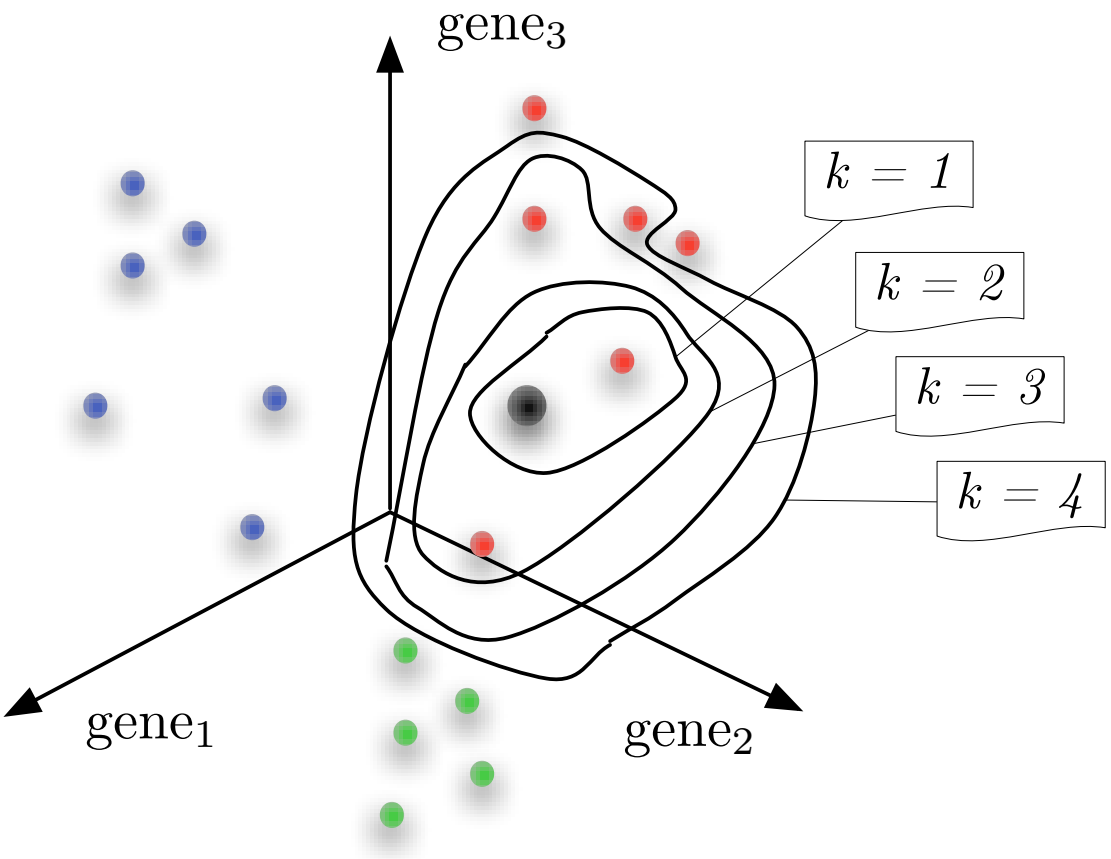
The form of the model is

$$\Pr(Y|X) = \frac{e^{\beta_0+\beta X}}{1+e^{\beta_0+\beta X}}$$

$$\underbrace{\log\left(\frac{p}{1-p}\right)}_{logit} = \beta_0+\beta_1 X_1+\cdots+\beta_k X_k$$



K-Nearest Neighbors



- *Lazy* model.
- Majority of voting over the k nearest neighbors.
- k is like a smoothing parameter.
- Distance to the query point can be used as weight in voting.

K-Nearest Neighbors

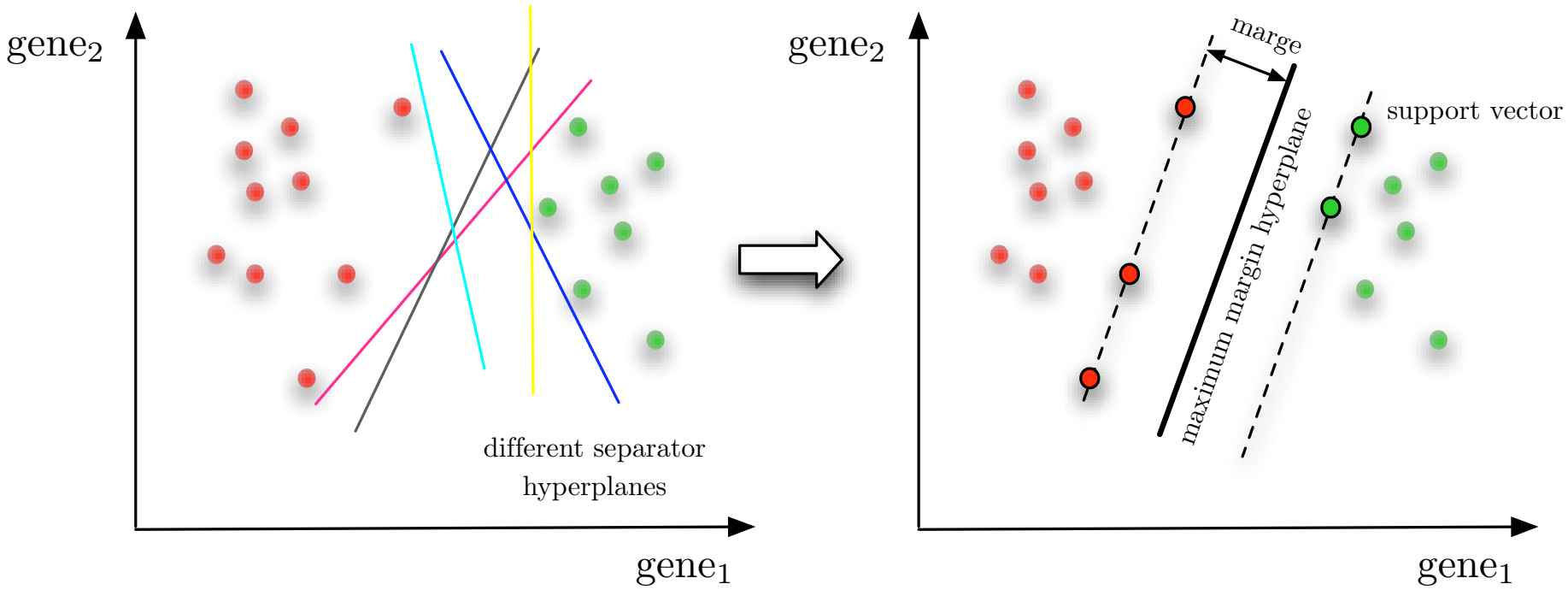
Suite

- Advantages :
  - can also be used for regression
  - computationally efficient.
- Disadvantages :
  - need to store all data points
  - need to specify the number of neighbors
  - need complete data.

Support Vector Machines

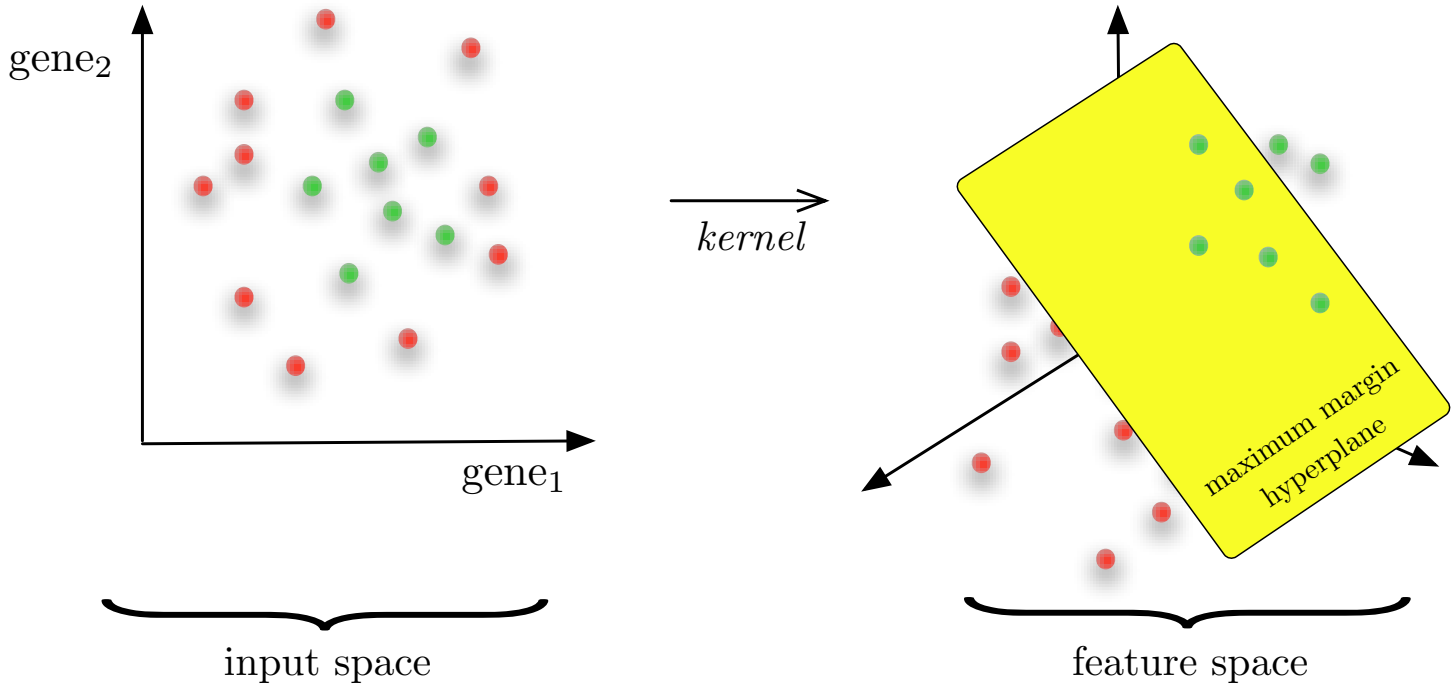
SVMs [Vapnik, 1998] are composed of 2 parts:

1. linear classifier called the *maximum margin* hyperplane



SVMs [Vapnik, 1998] are composed of 2 parts:

1. linear classifier called the *maximum margin* hyperplane
2. (non)linear transformation of the input space called the *kernel function*



Support Vector Machines

Suite (2)

- Advantages :
  - can perform linear and non-linear classification depending of the kernel function
  - computationally efficient.
- Disadvantages :
  - need to select the kernel function
  - need complete data.

Feature Selection

- Microarray data deal with a very large number n of variables (thousands of probes) and comparably few samples (dozens or hundreds of patients).
- Microarray data deal with highly correlated variables.
- In these cases, it is common practice to adopt feature selection algorithms to improve the generalization accuracy [Guyon and Elisseeff, 2003, Kohavi and John, 1997].
- There are many potential benefits of feature selection :
  - facilitating data visualization and data understanding
  - reducing the measurement and storage requirements
  - reducing training and utilization times
  - defying the curse of dimensionality to improve prediction performance.

Performance Assessment

- Performance estimators :
  - specificity, sensitivity, PPV, NPV, ...
  - statistical test to compare the different groups of patients (e.g. hazard ratio in survival analysis)
- Classification performance can be estimated by resampling, e.g. bootstrap or cross-validation.
- Take into account feature selection and other training decisions in the performance estimation process (number of neighbors in KNN, kernel in SVMs, ...)
- Otherwise, performance estimates may be severely biased upward, i.e. overly optimistic.

Conclusion

- Pay attention to
  - use simple models in taking into account the model assumptions
  - the impact of feature selection (maybe most important part of the analysis)
  - be careful in doing the performance assessment
  - think about the classifier validation on different datasets as in [Ioi, 2005].

Part III

Human Cancer Microarray Datasets

Differences between Datasets

- Different types of microarray technology
  - cDNA (dual-channel)
  - oligonucleotide
    - ★ short oligos (e.g. AFFYMETRIX, single-channel)
    - ★ long oligos (e.g. AGILENT, dual-channel, CODELINK, single channel)

- Be careful when comparing different datasets
  - mapping of the probes through annotations (e.g. gene ids, unigene cluster)
  - do meta-analysis to consider several datasets in one study as in [Shen et al., 2004, Rhodes et al., 2004, Sotiriou et al., 2006].

Databases

- Databases of microarray datasets
  - gene expression omnibus (GEO) from NCBI : http://www.ncbi.nlm.nih.gov/geo/
  - array express (AE) from EBI : http://www.ebi.ac.uk/arrayexpress/
  - oncomine : http://www.oncomine.org
- Databases for mapping
  - Cleanex from SIB : http://www.cleanex.isb-sib.ch
  - Adapt from Paterson Institute for Cancer Research : http://bioinformatics.picr.man.ac.uk/adapt

Thank you for your attention.

Part IV

Bibliography

(2005).
Prediction of early distant relapses on tamoxifen in early-stage breast cancer (BC): a potential tool for adjuvant aromatase inhibitor (AI) tailoring.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002).
A stability based method for discovering structure in clustered data a stability based method for discovering structure in clustered data.
Proc. Symp. Biocomput., 7:6–17.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998).
Cluster analysis and display of genome-wide expression patterns.
PNAS, 95:14863–14868.

Guyon, I. and Elisseeff, A. (2003).
An introduction to variable and feature selection.
Journal of Machine Learning Research, 3:1157–1182.

Hartigan, J. A. (1975).
Clustering Algorithms.
Wiley.

Kohavi, R. and John, G. (1997).
Wrappers for feature subset selection.
AIJ, 97(1-2):273–324.

Kohonen, T. (1997).
Self-Organizating Maps.
Springer-Verlag.

Kruskal, J. B. and Wish, M. (1978).
Multidimensional Scaling.
Beverly Hills, California.

MacQueen, J. B. (1967).
Some methods for classification and analysis of multivariate observations.
In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Rhodes, D. R., Yu, J., Shanker, K., Desphande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., and Chinnaiyan, A. M. (2004).

Latge-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *PNAS*, 101(25):9309–9314.


Shen, R., Ghosh, D., and Chinnaiyan, A. M. (2004).
Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data.
BMC Genomics, 5(94).


Sotiriou, C., Wirapati, P., Loi, S. M., Desmedt, C., Durbecq, V., Harris, A., Bergh, J., Smeds, J., Haibe-Kains, B., Larsimont, D., Cardoso, F., Buyse, M., Delorenzi, M., and Piccart, M. (2006).
Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis.
JNCI, 98:262–272.


Valentini, G. (2006).
Clusterv: a tool for assessing the reliability of clusters.
Bioinformatics Applications Note, 22(3):369–370.


Vapnik, V. N. (1998).

Statistical Learning Theory.
Springer.