Statistical Analysis of MicroArray Data

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

March 23, 2007

Research Groups

Machine Learning Group

- 7 researchers (1 prof, 6 PhD students), 4 graduate students).
- Research topics : Bioinformatics, Classification, Regression, Time series prediction, Sensor networks.
- Website : http://www.ulb.ac.be/di/mlg.
- Scientific collaborations in ULB : IRIDIA (Sciences Appliquées), Physiologie Molculaire de la Cellule (IBMM), Conformation des Macromolcules Biologiques et Bioinformatique (IBMM), CENOLI (Sciences), Microarray Unit (Hopital Jules Bordet), Service d'Anesthesie (Erasme).
- Scientific collaborations outside ULB : UCL Machine Learning Group (B), Politecnico di Milano (I), Universitá del Sannio (I), George Mason University (US).
- The MLG is part to the "Groupe de Contact FNRS" on Machine Learning and to CINBIOS (http://babylone.ulb.ac.be/Joomla/)

Research Groups

Functional Genomic Unit

- 7 researchers (1 prof, 3 postDocs, 6 PhD students), 3 technicians).
- Research topics : Genomic analyses and clinic studies.
- Website : http://www.bordet.be/en/services/medical/array/practical.htm.
- National scientific collaborations : ULB, Erasme, ULg, Gembloux
- International scientific collaborations : Genome Institute of Singapore, John Radcliffe Hospital, Karolinska Institute and Hospital, MD Anderson Cancer Center, Netherlands Cancer Institute, Swiss Institute for Experimental Cancer Research, NCI/NIH, Gustave-Roussy Institute, IDDI

Outline

1. Machine Learning
2. MicroArray Analysis Design
3. Unsupervised Learning
4. Supervised Learning
5. Graphs and Networks
6. Bioinformatics Softwares

Machine Learning

Definition

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience. [Mitchell, 1997]

Machine Learning

. . . and Applied Statistics

Reductionist attitude : *ML is a modern buzzword which equates to statistics plus marketing*

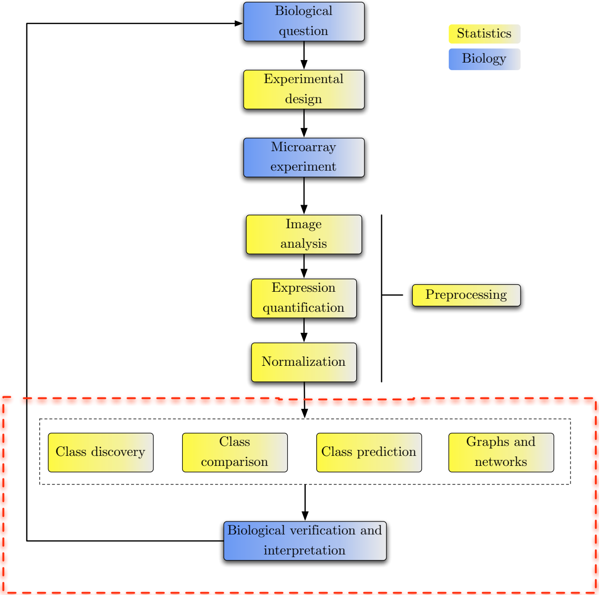
Positive attitude : ML paved the way to the treatment of real problems related to data analysis, sometimes overlooked by statisticians (nonlinearity, classification, pattern recognition, missing variables, adaptivity, optimization, massive datasets, data management, causality, representation of knowledge, parallelisation)

Interdisciplinary attitude : ML *should* have its roots on statistics and complements it by focusing on: algorithmic issues, computational efficiency, data engineering.

Notations

$\mathbf{x}$	random variable
x	realization of random variable $\mathbf{x}$
I	number of patients
J	number of genes
$X_i,\ i = 1, 2, \ldots, I$	gene expression profile of patient i
$x_{ij},\ j = 1, 2, \ldots, J$	expression of gene j of patient i
$D_{I \times J}$	dataset represented by a matrix of gene expressions with I lines (iid patients) and J columns (genes)

Microarray Analysis Design



Part I

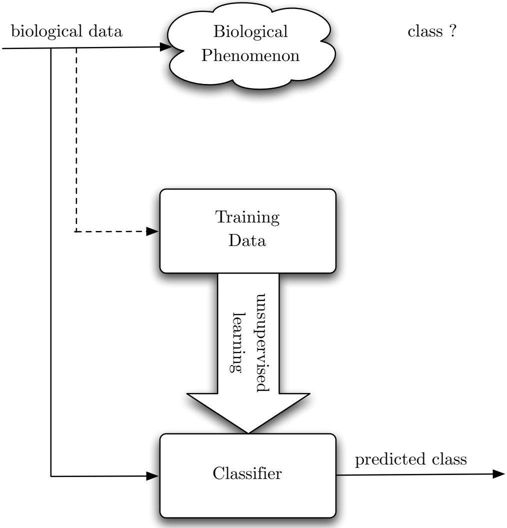
Unsupervised Learning

Unsupervised Learning

Outline

1. Introduction
2. Clustering
3. Distances
4. Clustering Methods
5. Multidimensional Scaling
6. Clustering Validity

Introduction



Class Discovery Example

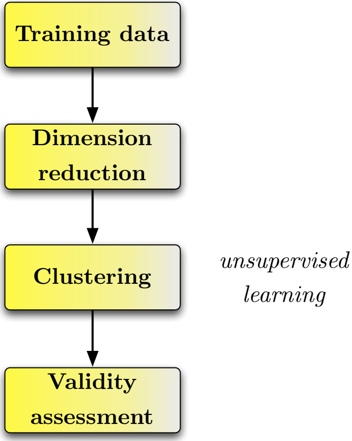
Breast Cancer Subtypes

- biological question : are there any natural gene patterns for different types of breast cancer ?

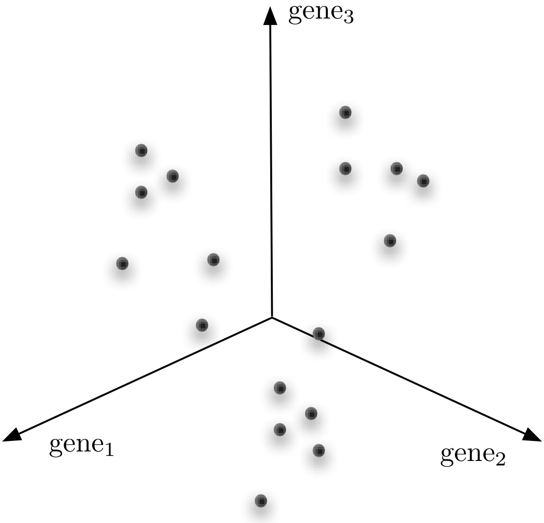
- input data : microarray data (let say 10.000 genes and 50 patients)
- output data : none

- result : a clustering of the 50 patients. Each group of patients exhibits distinct gene patterns

Design

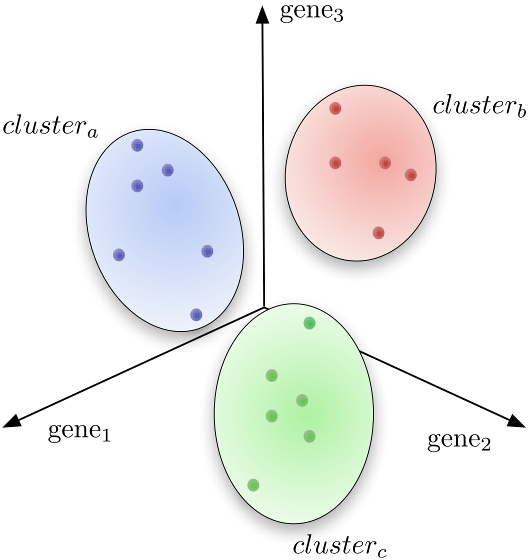

Training data
Dimension reduction
Clustering
Validity assessment
unsupervised learning

Clustering



Example of 17 samples drawn in a 3-diensional gene space

Clustering
cont'd



The samples can be divided into 3 clusters.

Clustering can be viewed as a mapping from the empirical ditribution of $X_1, X_2, \ldots, X_I$ to a sequence of k clusters.

Key Components of Clustering Algorithms

- Distance matrix : $J \times J$ symmetric matrix which quantifies the similarity of each pair of objects. An object could be a gene or a patient.
- Number of clusters : should be specified by the user or defined by an algorithm optimizing some criterion.
- Criterion : is often a continuous function of the cluster labels that measures how similar objects are within clusters ad how different objects are between clusters.
- Searching strategy : algorithm trying to find the clustering result that globally maximizes the criterion. Due to computational issues, heuristic search strategies, guaranteeing only the convergence to a local maximum, are often needed.

Classes of Clustering Algorithms

- Partitioning : partitioning methods map a collection of objects into $k \geq 2$
  - disjoint clusters maximizing a particular criterion. Example : k-means, partioning around medoids (PAM), . . .
  - overlapping clusters maximizing a particular criterion. Example : self-organizing map (SOM), c-means, model-based clustering, . . .
- Hierarchical : hierarchical methods involve constructing a tree of clusters in which the root is a single cluster containing all the objects and each leave contains only one object. Contruction of such a tree can be
  - divisive : build from the top down by recursively partitioning the objects. Example : diana
  - agglomerative : build from the bottom up by recursively combining the objects. Example : hclust and agnes

Distances

- Distances, metrics and similarities are related concepts.
- Definitions : a metric d needs to satisfy the following properties
    - non-negativity : $d(X_{i_1}, X_{i_2}) \geq 0$
    - symmetry : $d(X_{i_1}, X_{i_2}) = d(X_{i_2}, X_{i_1})$
    - identification mark : $d(X_{i_1}, X_{i_1}) = 0$
    - definiteness : $d(X_{i_1}, X_{i_2}) = 0 \text{ iff } X_{i_1} = X_{i_2}$
    - triangle inequality :$d(X_{i_1}, X_{i_2}) + d(X_{i_2}, X_{i_3}) \geq d(X_{i_1}, X_{i_3})$

Distances

cont'd

- Minkowski metric family $\left(\sum_{j=1}^{J} d(x_{i_1 j}, x_{i_2 j})^{\lambda}\right)^{\frac{1}{\lambda}}$:
    - Euclidian ($\lambda = 2$)

$$d_{euc}(X_{i_1}, X_{i_2}) = \sqrt{\sum_{j=1}^{J} (x_{i_1 j} - x_{i_2 j})^2}$$

    - Manhattan ($\lambda = 1$)

$$d_{man}(X_{i_1}, X_{i_2}) = \sum_{j=1}^{J} |x_{i_1 j} - x_{i_2 j}|$$

Distances

cont'd

- Correlation-based distance :
  - Pearson

$$d_{cor}(X_{i_1}, X_{i_2}) = 1 - \rho(x_{i_1 j}, x_{i_2 j})$$

  - Spearman

$$d_{spear}(X_{i_1}, X_{i_2}) = 1 - \frac{X_{i_1}' X_{i_2}}{\parallel X_{i_1} \parallel \parallel X_{i_2} \parallel}$$

  - Kendall's τ

$$d_{spear}(X_{i_1}, X_{i_2}) = 1 - \tau(X_{i_1}, X_{i_2})$$

- Distances depend on the scale of the data
- Standardization

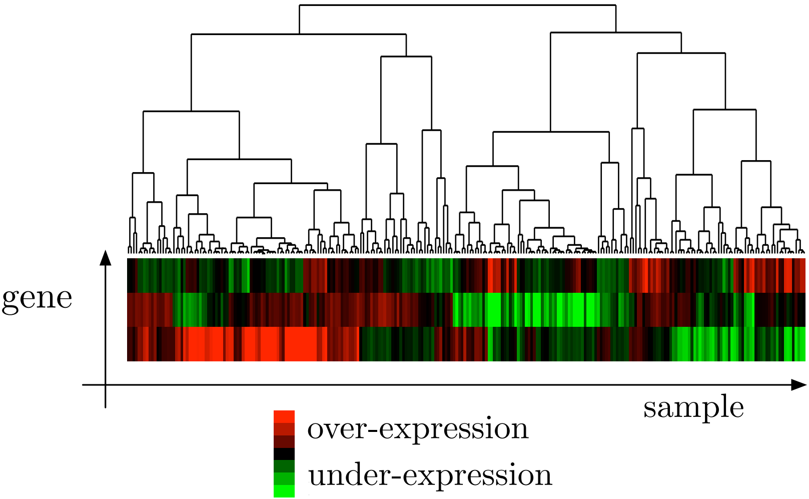
  $$\frac{x - center(x)}{scale(x)}$$

  may improve the comparison between objects but may also remove some interesting features in the data.
- We usually apply gene centering in one-channel microarray data, ie $x_{ij} = x_{ij} - center(x_{.j})$ with *center* being the median.

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**.
- Definition of a linkage method, ie the computation of a distance between a cluster (called A) and another object/cluster (called B)
  - complete : $\max\{d(x,y) : x \in A, y \in B\}$
  - single : $\min\{d(x,y) : x \in A, y \in B\}$
  - average : $\mathrm{mean}\{d(x,y) : x \in A, y \in B\}$
  - Ward : $\mathrm{var}\{A \cup B\} - \mathrm{var}\{A\}$

Hierarchical Clustering
Example



Dendrogram

- Apparent ease of interpretation but may be misleading
- Dendrogram corresponding to a given hierarchical clustering is not unique :
  - for each merge one needs to specify which subtree should go on the left and which on the right
  - so there are 2^{n-1} choices.
- Dendrogram imposes structure on the data, instead of revealing structure in these data.
  - such a representation will be valid only to the extent that the pairwise dissimilarities possess the hierarchical structure imposed by the clustering algorithm.
  - *Cophenetic* correlation coefficient tests if the hierarchical structure represents the pairwise distances.

Hierarchical Clustering

- Advantages :
  - no number of clusters to specify (full hierarchical binary tree)
  - deterministic
  - computationally efficient.
- Disadvantages :
  - impose a hierarchical structure for the clusters
  - dendrogram may be misleading
  - need to define a linkage method

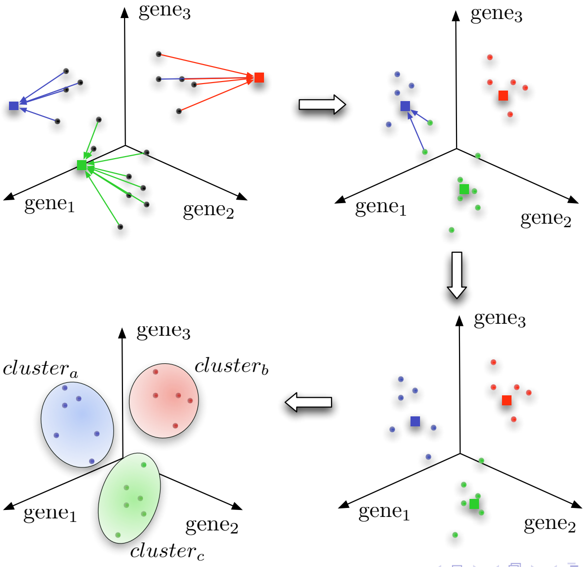
K-Means

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

- Advantage :
  - computationally efficient.
- Disadvantages :
  - need to specify the number of clusters
  - not deterministic

K-Means

Example : $k = 3$



K-Means Variants

- PAM : method introduced in [Kaufman and Rousseeuw, 1990].
  - it operates on the distance matrix only because of use of medoids instead of cluster centroids
  - it minimizes a sum of distances (instead of a sum of squared distances)
  - it uses the *silhouette* statistic to select the "good" number of clusters
- Fuzzy c-means : method introduced in [Dunn, 1973, Bezdek, 1981]
  - it allows objects to belong to 2 or more clusters
  - It minimizes the following objective funtion

$$\sum_{i=1}^{I} \sum_{k=1}^{K} u_{ik}^{m} \parallel X_i - C_k \parallel^2$$

where $1 \leq m < \infty$, u_{ik} is the degree of membership of X_i in the cluster k, C_k is the J-dimensional center of cluster k and $\parallel \cdots \parallel$ is any distance

Model-Based Clustering

- Approach consisting in using some models for clusters and attempting to optimize the fit between the data and the model
- Each cluster can be mathematically represented by a parametric distribution
- The entire dataset is therefore modeled by a mixture of these distributions
- Advantages
  - well-studied statistical inference techniques are available
  - flexibility in choosing the distribution
  - density estimation for each cluster
  - model for further classification
- Disadvantages
  - assumption about the distributions

- Example : mixture of Gaussians using an expectation-maxmimization (EM) algorithm to maximize the likelihood [Dempster et al., 1977]

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.

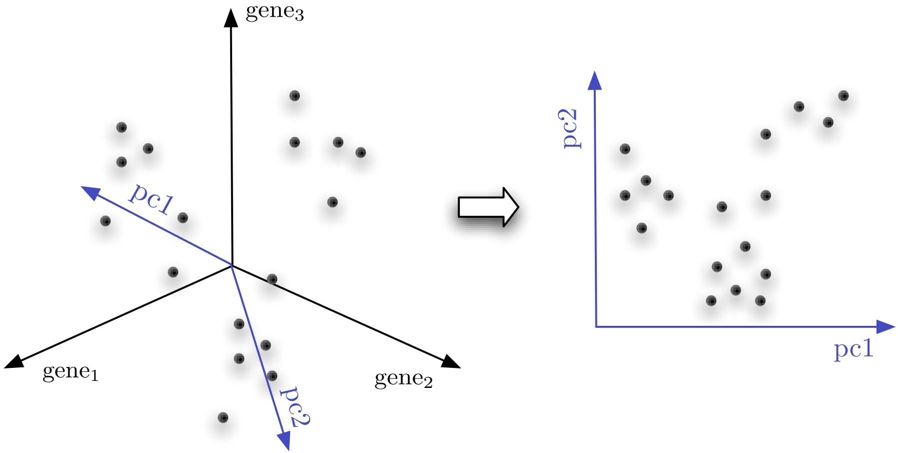
Multidimensional Scaling

- 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 (sometimes called *eigen-genes*)
  - 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 Validity

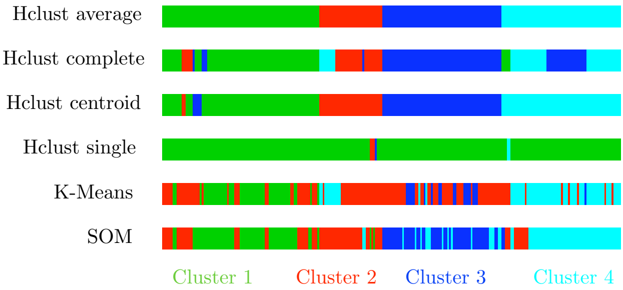
- Clustering algorithms always find structure in the data.
- Need of methods to assess the reliability of the discovered classes.

Procedure

1. Perturb the original dataset
   - by resampling the original dataset (eg in 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 validity 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

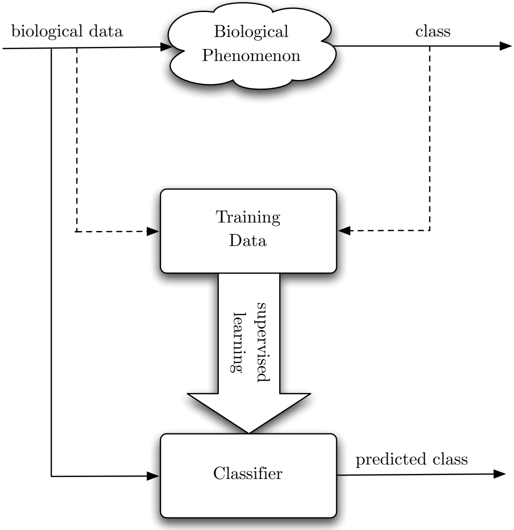
Supervised Learning

Supervised Learing

Outline

1. Introduction
2. Class Comparison
    1. (Non-)parametric Tests
    2. Multiple Testing Problem
3. Prediction
    1. Regression (Survival Analysis)
    2. Classification
        1. Classifiers
        2. Aggregation of Classifiers
    3. Feature Selection
    4. Accuracy Estimation

Introduction



Part III

Class Comparison

Class Comparison Example

Normal vs Tumor Tissue

- biological question : what are the genes differentially expressed between normal and tumor tissues ?

- input data : microarray data (let say 10.000 genes and 50 tissues)
- output data : label classes (25 normal tissues and 25 tumor tissues)

- result : a list of significantly differentially expressed genes

- Class comparison is often referred as "differential gene expression analysis"
- Classes are defined by phenotypes or experimental conditions and can have ≥ 2 levels
- Characteristics of microarray data influencing class comparison
  - gene expressions are noisy
  - number of genes much larger than number of patients
  - gene expressions are highly correlated, so they are not independent
- Analysis of joint distribution of genes is not feasible in practice

➡ We will focus on gene-by-gene analysis

Topics

Important topics in class comparison :

- parametric vs non-parametric methods : gene expression distributions ? power of the statistical test ?
- multiple testing correction : estimation of Type I error ?

Parametric vs Nonparametric Methods

- Nonparametric methods may be used when :
  - the sample size is small (no way to test the assumptions)
  - problems in measurement (nominal values, rankings, . . .).
- Parametric tests are more powerful when assumptions are not violated
- Hybrid: use a statistic from a parametric test (eg Student t) and estimate its null distribution by resampling

Nonparametric Methods

Overview

- Test differences between independent groups :
  - Wilcoxon Rank Sum/Mann-Whitney U (2 groups)
  - Kruskal-Wallis ($\geq$ 2 groups).
- Test differences between dependent groups :
  - Sign test
  - Wilcoxon's Matched Pairs test
  - McNemar's Chi-square (dichotomous variable).
- Relationship between variables, nonparametric equivalents to the standard correlation coefficient :
  - Spearman's ρ
  - Kendall's τ

AFFYMETRIX© Two-Chip Comparison

S-Score Algorithm

- Algorithm introduced in [Zhang et al., 2002]
- It allows for comparison of two AFFYMETRIX© chips
- How ?
  - classical tests work at probeset level, after summarization step (see preprocesing of AFFYMETRIX© data)
  - s-score algorithm works at the probe level
  - so the comparison 1 vs 1 is transformed in a paired comparison 20 vs 20 (1 probeset $\approx$ 20 probes)
  - be aware that the biological variation of a population is not taken into account (1 patient vs 1 patient)

Multiple Testing Problem

- The problem of multiple testing can be described as the potential increase in Type I error (α) that occurs when statistical tests are used repeatedly.

- If n independent comparisons are performed, the experiment-wide significance level $\alpha_{global} = 1 - (1 - \alpha)^n$.

Multiple Testing Problem

Example

- One might declare that a gene is differentially expressed if the statistical test leads to a p-value < 0.01.
- A multiple testing problem arises if one wanted to use this test (which is appropriate for testing the differential expression of one gene), to test the differential expression of many genes.
- Imagine if one wants to test 100 equally expressed genes by this method.
- Given that the probability to do a type I error is 0.01, see a differentially expressed gene would be a relatively likely event.
- The likelihood that all 100 equally expressed genes are identified as equally expressed by this test is $(1 - 0.01)^{100} = 0.366$.

➡ False positive rate must be controlled at the experiment-wide level.

Multiple Testing Problem

Bonferroni's Method

- n independent tests.
- Desired experiment-wide level of type I error is α.

➡ Test each hypothesis at level $\frac{\alpha}{n}$.

- Strong control of the family wise error rate (FWER), very stringent.

Multiple Testing Problem

Benjamini's and Hochberg's Method

- Method controlling the false discovery rate (FDR) of a set of predictions [Benjamini and Hochberg, 1995]
- FDR is the expected percent of false predictions in the set of predictions.
    - example : if the algorithm returns 100 differentially expressed genes with a false discovery rate of 0.3 then we should expect 70 of them to be correct.
- FDR is very different from a p-value, so a higher FDR an be tolerated
    - example : a set of 100 differentially expressed genes of which 70 are correct might be very useful, especially if there are thousands of genes on the array, most of which are equally expressed
    - in contrast p-value of 0.3 is generally unacceptable in any circumstance.

Benjamini's and Hochberg's Method

Formulation

- Let m be number of tested null hypotheses of which m_0 are true
- Let R denote the number of hypotheses rejected by a procedure
- Let V denote the number of null hypotheses erroneously rejected by a procedure
- Let $Q = \frac{V}{R}$ if $R > 0$, 0 otherwise
- then $FDR = E(Q)$

Benjamini's and Hochberg's Method

Estimation

- For each hypothesis H_j, a test statistic is computed with the corresponding p-value P_j
- Linear step-up procedure :
  - rank p-values $P_1 \leq P_2 \leq \cdots \leq P_m$
  - let q be the desired level of FDR
  - let $k = \max\{j : P_j \leq q \frac{j}{m}\}$
  - reject $H_1, H_2, \ldots, H_k$ if such a k exists

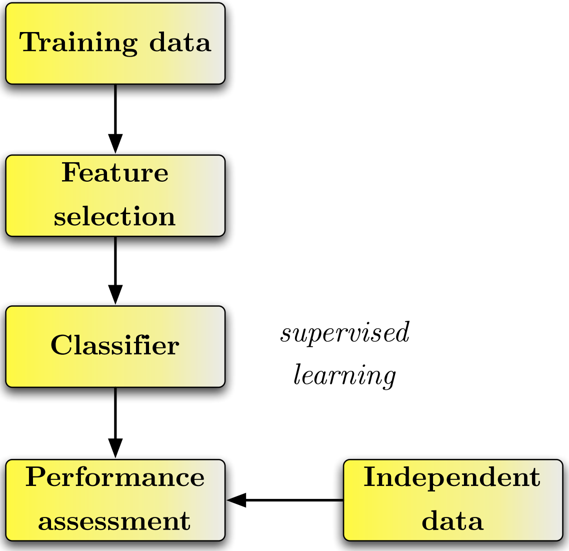
Conclusion

- Gene-by-gene class comparison is one of the most widely used analysis of microarray data
- However, due to intrinsic characteristics of microarray data, we must deal with :
  - the choice of an appropriate statistical test
  - the problem of multiple testing
- Because of low sample size, we are sometimes not able to reach significant results even in the presence of strong biological signal
- So we need to reduce the number of tested hypotheses by filtering the data or by focusing the analysis on a set of genes (eg selected by annotations)

Part IV

Prediction

Design



Training data

Feature selection

Classifier

supervised learning

Performance assessment

Independent data

Classification Example

Breast Cancer Prognostication

- biological question : can we build a gene classifier that discriminates patients at low and high-risk of relapse ?
- input data : microarray data (let say 10.000 genes and 100 patients)
- output data : survival data (25 events and 75 censored observations)

- result : a model using the gene expression of a set of genes (called gene signature) that is able to classify with high accuracy new patients.

Classifiers

- Classification
  - Linear : logistic regression, naive Bayes, linear disciminant analysis (LDA), . . .
  - Non-linear : k-nearest neighbors (KNN), Support Vector Machines (SVM), classification trees, . . .
- Regression
  - Linear : linear regression, Cox regression . . .
  - Non-linear : lazy learning, artificial neural networks (ANN or NNET), . . .

General Linear Models

- Family of regression models
- Outcome variable determines the choice of model

Outcome	Model
continuous	linear regression
counts	Poission regression
survival	Cox regression
binomial	logistic regression

- Applications :
  - estimate force of association between outcome and covariates
  - control of confounding
  - model building, risk prediction

Regression

Survival Analysis

- Survival analysis deals with (right-)censored data, ie time to event **t**
- Survivor function

$$S(t) = \Pr\{\mathbf{t} > t\}$$

  - $S(t)$ is the probability that an event time is greater than t

- Hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr\{t \leq \mathbf{t} < t + \Delta t \,|\, \mathbf{t} \geq t\}}{\Delta t}$$

  - $h(t)$ quantifies the instantaneous risk that an event will occur in the small interval between t and $t + \Delta t$

- Most famous model for is the semiparametric regression model proposed in [Cox, 1972] to estimate a hazard function given a dataset
  - no assumption about the probability distribution of survival times (proportional hazards model)
  - efficient estimation method (maximum partial likelihood).
- Cox model :

$$h_i(t) = \underbrace{\lambda_0(t)}_{\text{baseline hazard function}} \exp \underbrace{(\beta X_i)}_{\text{risk score}}$$

- $\exp(\beta_j)$ is the hazard ratio corresponding to feature x_j
- if x_j is binary, it is a summary of the difference between two survival curves, representing the reduction in the risk of event. For a continuous feature x_j, the same interpretation applies to a unit difference.

Classification

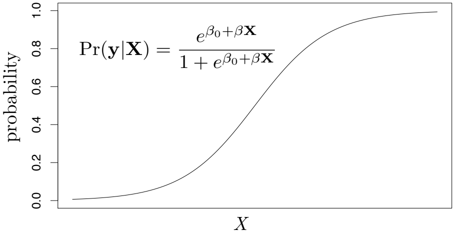
Logistic Regression

- Variation of ordinary regression introduced in [Agresti, 1990]
- Method uses when :
  - output is a dichotomous variable
  - input variables are continuous, categorical, or both

The form of the model is

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

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



- Simple probabilistic classifier with **y** a categorical variable having ≥ 2 levels
- Strong assumption of conditional independence of features
  - "naive" underlying probabilistic model
  - however, good accuracy in real-world situations
- Advantages :
  - simple model
  - inclusion of prior knowledge (class prior)
  - computationally efficient.
- Disadvantage :
  - over-simplified assumptions.

Classification

Naive Bayes : Problem Formulation

- Using Bayes' theorem

$$\Pr(\mathbf{y}|x_1, x_2, \ldots, x_J) = \frac{\Pr(\mathbf{y}) \Pr(x_1, x_2, \ldots, x_J|\mathbf{y})}{\Pr(x_1, x_2, \ldots, x_J)}$$

- The numerator is the joint probability

$$\begin{aligned} \Pr(\mathbf{y}, x_1, x_2, \ldots, x_J) &= \Pr(\mathbf{y}) \Pr(x_1, x_2, \ldots, x_J|\mathbf{y}) \\ &= \Pr(\mathbf{y}) \Pr(x_1|\mathbf{y}) \Pr(x_2, \ldots, x_J|\mathbf{y}, x_1) \\ & \ldots \end{aligned}$$

Naive Bayes : Problem Formulation (2)

- Under conditional independence assumption

$$\Pr(\mathbf{y}, x_1, x_2, \ldots, x_J) = \Pr(\mathbf{y}) \prod_{j=1}^{J} \Pr(x_j | \mathbf{y})$$

  where class prior $\Pr(\mathbf{y})$ and independent probability distributions $\Pr(x_j | \mathbf{y})$

- Finally, combine this model with a decision rule as *maximum a posteriori* (MAP)

$$\mathit{classify}(x_1, x_2, \ldots, x_J) = \mathrm{argmax}_j \Pr(\mathbf{y} = y) \prod_{j=1}^{n} \Pr(x_j | \mathbf{y} = y)$$

Classification

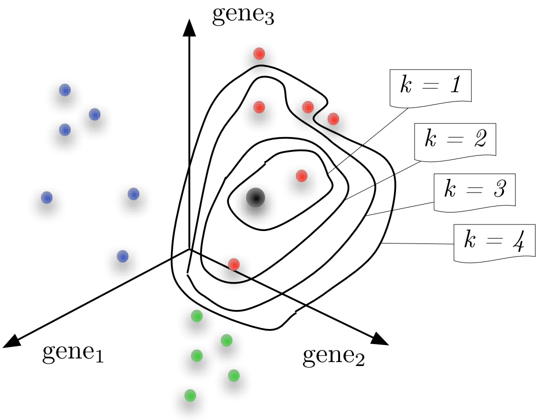
Naive Bayes : Classifier

- We have to estimate two terms
  - $Pr(\mathbf{y} = y)$: frequency of each class in training set
  - $\Pr(x_j|\mathbf{y} = y)$, $j = 1, 2, \ldots, J$:
    - ★ if x_i is discrete, just compute the observed frequency of x_i for a given class y
    - ★ else you can use non-parametric density estimation such as *kernel*

- Even if the assumption of independence is violates in microarray data, the naive Bayes classifier still exhibits very accuracy in real problems

Classification

K-Nearest Neighbors



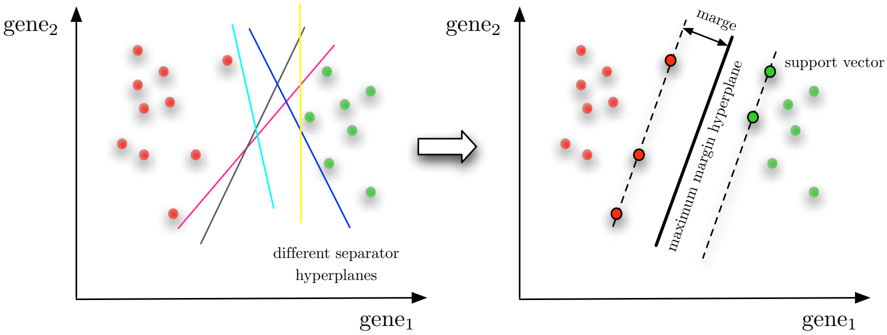
- *Lazy* model.
- Majority of voting over the k nearest neighbors.
- k is like a smoothing parameter.
- Distance to the query point can be used as weight in voting.

Classification

Support Vector Machines

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

1. linear classifier called the *maximum margin* hyperplane

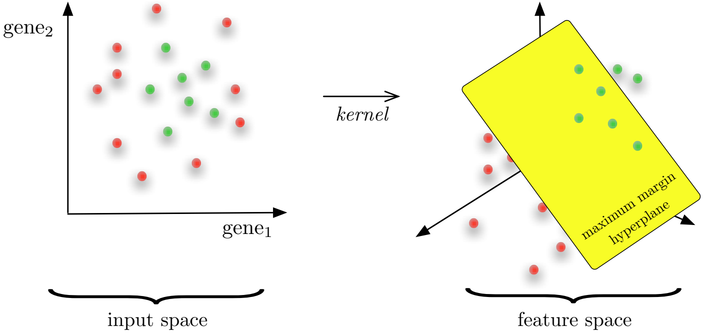


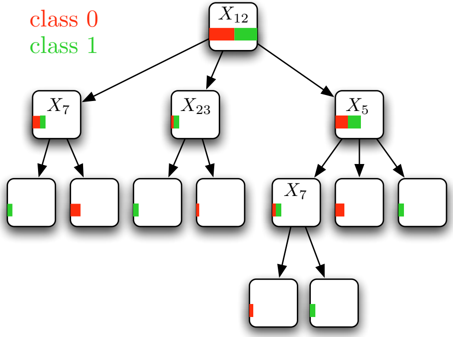
Classification

Support Vector Machines cont'd

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*





- Method introduced in [Breiman et al., 1984]
- Hierarchical tree-structured plan of a set of attributes to test in order to predict the output.
- Select recursively the attributes that maximize a criterion as
  - information gain
  - predictive accuracy

- Classification trees are flexible and easily interpretable
- Full classification trees are prone to overfitting and variability
  - may require to prune the tree
  - may require to aggregate several trees (as *boosting* or *bagging* see next slides about the aggregation of classifiers).

Aggregation of Classifiers

- Bagging (bootstrap aggregating) [Breiman, 1996] :
  - create k bootstrap samples $S_1, S_2, \ldots, S_k$
  - train distinct classifiers on each S_i
  - classify new instances by majority voting.
- Boosting [Freund and Schapire, 1996] :
  - generate a sequence of classifiers
  - each classifier put more weight on the instances of the training set that were not successfully classified previously
  - combine the different classifiers to derive one aggregated classifier
- Random forests for classification trees use the bagging for the samples and for the variables simultaneously.

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 accuracy.

Approaches to Feature Selection

- Filter methods : preprocessing methods attempting to assess the relevance of features from the data, ignoring the effects of the selected feature subset on the accuracy of the learning algorithm.
  - Example : ranking variables through variance or compression techniques like PCA
- Wrapper methods : assess subsets of variables according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function.
  - Example : forward, backward and stepwise feature selections
- Embedded methods : perform variable selection as part of the learning procedure and are usually specific to given learning machines.
  - Example : classification trees, regularization techniques like LASSO [Tibshirani, 1997]

Feature Selection Issues

Two main issues make the problem of feature selection a highly challenging task :

- Search in a high dimensional space : this is known to be a NP-hard problem.
- Estimation on the basis of a small set of samples : this is made difficult by the high ratio between the dimensionality of the problem and the number of measured samples.

Open issue : searching for the best subset in very large spaces is prone to overfitting, even if estimation relies on cross-validations.

- Important concern in microarray analysis is the small overlap observed between gene signatures dealing with the same biological question (eg breast cancer prognostication)
- Potential sources of variation :
  - sampling error (different patient cohorts)
  - different microarray technologies
  - different set of genes
  - different bioinformatics methods

  - or just noise discovery ? [Ioannidis, 2005]

Feature Selection in Microarray Analysis

Small Overlap of Gene Signatures cont'd

- Sampling error :
  - several recent studies investigate the variability of simple feature selection methods (filters)
    [Ein-Dor et al., 2005, Davis et al., 2006, Baker and Kramer, 2006]
  - such analyses attempt to estimate the variability due to sampling error but consider only one microarray technology
  - because of the high feature to sample ratio and the co-expression of many genes, it is expected to see different sets of genes providing similar accuracy
- Microarray technologies :
  - the microarray quality control (MAQC) project by the US Food and Drug Administration highlights a high concordance between technologies

Feature Selection in Microarray Analysis

Small Overlap of Gene Signatures cont'd

- A meta-analysis about the survival prediction of more than 2000 breast cancer patients will be published soon by Sotiriou *et al.* showing that :
  - Despite sampling error and microarray technologies, we can observe same relation with survival for some sets of genes
  - the proliferation set of genes seems to be the common denominator of many existing gene signatures
  - this set of genes recapitulating their predictive power [Sotiriou et al., 2006]

Accuracy Estimation

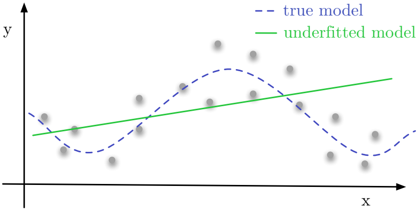
- Accuracy estimators :
  - specificity, sensitivity, PPV, NPV, (time-dependent) ROC curves, . . .
  - statistical test to compare the different groups of patients (eg hazard ratio in survival analysis)
- Classification accuracy can be estimated by resampling, eg bootstrap [Efron and Tibshirani, 1997] or cross-validation [Ambroise and McLachlan, 2002]
- Take into account feature selection and other training decisions in the accuracy estimation process (number of neighbors in KNN, kernel in SVMs, ...)
- Otherwise, accuracy estimates may be severely overly optimistic

➡ overfitting

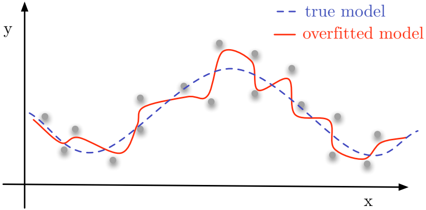
Accuracy Estimation

Overfitting

- Low model complexity (large bias)



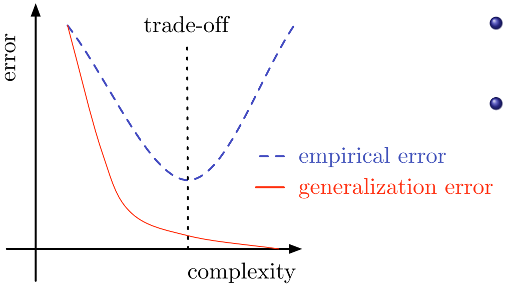
- High model complexity (large variance)



- A good model should achieve a trade-off between bias and variance to ensure low generalization error
- In classification, it is better to have small variance (even with large bias) [Friedman, 1996]

Accuracy Estimation

Overfitting cont'd



- Empirical error refers to estimated error on training set
- Generalization error refers to estimated error on test set

- As we can see, it is possible to have an empirical error ≈ 0 if the complexity of the model is high
- However, the generalization error increases when the model is too complex

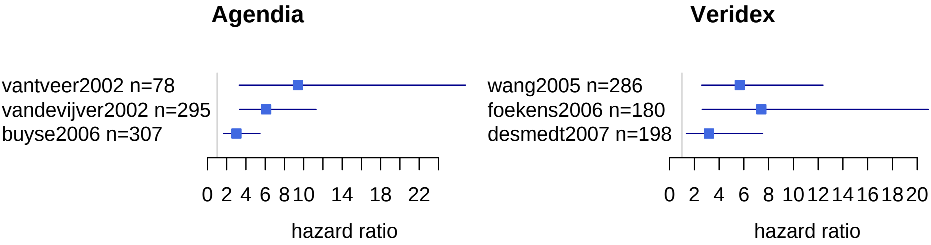
2 research groups have published key publications in the field of breast cancer prognostication :

- Agendia :
  - first article [van't Veer et al., 2002]
  - external validation [van de Vijver et al., 2002]
  - external and independent validation [Buyse et al., 2006]
- Veridex :
  - first article [Wang et al., 2005]
  - external validation [Foekens et al., 2006]
  - external and independent validation [Desmedt et al., 2007]

Accuracy Estimation in MicroArray Analysis

Breast Cancer Prognostication cont'd

Univariate hazard ratios for Agendia and Veridex related publications :



Accuracy Estimation in MicroArray Analysis

Breast Cancer Prognostication cont'd

Remarks about Agendia related publications :

- In [van't Veer et al., 2002], authors performed a gene ranking based on the correlation with a binary output (binarization of survival data) and applied a cross-validation afterwards
  - problem : all the patients are used for the feature selection ➟ biased accuracy estimates [Varma and Simon, 2006]
  - solution : include the feature selection in the global loop of cross-validation to have a better estimate of the generalization error
- In [van de Vijver et al., 2002], authors analyzed a new dataset containing a subset of patients from the previous publication
  - problem : the external validation series is not independent
  - solution : don't include patients used for model fitting, even if the aim is to have a more representative patient cohort
- In [Buyse et al., 2006] . . . no problem ☺

Accuracy Estimation in MicroArray Analysis

Breast Cancer Prognostication cont'd

Remarks about Veridex related publications :

- in [Wang et al., 2005], they split the data in training/test sets without specifying the exact composition
  - problem : impossible to reproduce the results
  - problem : is the test set a specially good case ?
  - solution : use thousands of random splits instead (in controlling the composition of both sets, see [Michiels et al., 2005])
  - solution : finally, report the distribution of the accuracy estimates (in the test sets)
- In [Foekens et al., 2006] . . . no problem
- In [Desmedt et al., 2007] . . . no problem ☺

Conclusion

- Pay attention to
  - use simple models in taking into account the model assumptions [Dudoit et al., 2002]
  - the impact of feature selection (maybe most important part of the analysis)
  - design the classifier to be able to validate its accuracy on different datasets as in [Loi et al., 2005] (article to be published)
  - be careful in doing the accuracy estimation
- Review and guidelines for microarray analysis [Simon et al., 2003, Dupuy and Simon, 2007]

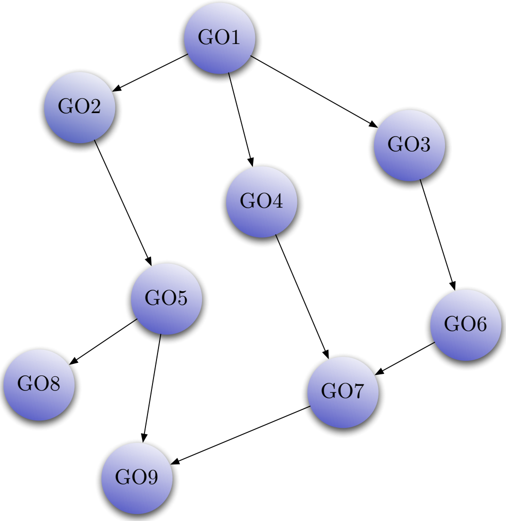
Part V

Graphs and Networks

Knowledge Representation

- Use of graphs to represent knowledge such as gene annotations
- Gene ontology (GO) [Ashburner et al., 2000] is frequently used in microarray analysis to interpret gene lists
  - it provides a structured vocabulary for the annotation of genes and proteins.
  - GO terms are structured in a hierarchy (directed acyclic graph), ranging from more general to more specific.
  - GO is structured in three ontologies, corresponding to biochemical function, cellular processes, and cellular components.
  - data can be annotated at varying levels, depending on the information available.
  - list of available GO tools : http://www.geneontology.org/GO.tools.shtml

Knowledge Representation
cont'd



Biological Networks Inference

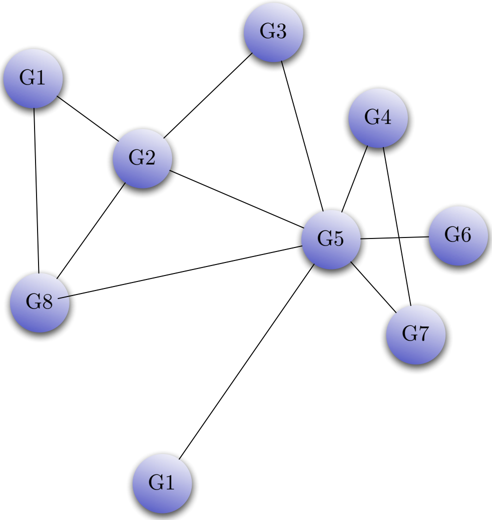
cont'd

- Use of graphs to represent interaction between genes
  - simplified biological dogma : gene ➟ RNA ➟ protein
  - a gene codes for a protein that blocks or activates another gene (interaction)

- A biological network is graph where
  - each node is a gene
  - each link represents an interaction between 2 genes

Biological Networks Inference

cont'd



Biological Networks Inference

cont'd

- We can infer such networks from microarray data in modeling the interactions using
  - information theory [Butte et al., 2000]
  - correlation [de la fuente et al., 2004]
  - ...
- MLG applied successfully such methods on yeast genome [Kontos and Bontempi, 2006, Meyer et al., 2007]

Part VI

Bioinformatics Software

Bioinformatics Softwares

- **R** is a widely used open source language and environment for statistical computing and graphics
  - Software and documentation are available from http://www.r-project.org

- **Bioconductor** is an open source and open development software project for the analysis and comprehension of genomic data
  - Software and documentation are available from http://www.bioconductor.org

Bioinformatics Softwares

cont'd

- **Java Treeview** is an open source software for clustering visualization
  - Software and documentation are available from http://jtreeview.sourceforge.net
- **Cluster3** is a open source clustering software with GUI
  - Software and documentation are available from `http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv`

Bioinformatics Softwares

cont'd

- **BRB Array Tools** is a software suite for microarray analysis working as an Excel macro
  - Software and documentation are available from http://linus.nci.nih.gov/BRB-ArrayTools.html
- **TIGR** is a java-based software suite for microarray analysis
  - Software and documentation are available from http://www.tm4.org

Links

- Master in Bioinformatics at ULB and other belgian universities : http://www.bioinfomaster.ulb.ac.be/
- Personal homepage : http://www.ulb.ac.be/di/map/bhaibeka/
- This presentation : http://www.ulb.ac.be/di/map/bhaibeka/papers/haibekains2007iddistats.pdf

Thank you for your attention.

For Further Reading I

Agresti, A. (1990).
Categorical data analysis.
Wiley.

Akaike, H. (1974).
A new look at the statistical model identification.
IEEE-AC, 19(6):716–723.

Allen, D. M. (1974).
The relationship between variable and data augmentation and a method of prediction.
Technometrics, 16:125–127.

Ambroise, C. and McLachlan, G. (2002).
Selection bias in gene extraction on the basis of microarray gene-expression data.
Proc. Natl. Acad. Sci. USA, 99(6562-6566).

For Further Reading II

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwoght, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000).
  Gene ontology: tool for the unfication of biology. the gene ontology consortium.
  Nat Genet, 25:25–29.

- Baker, S. G. and Kramer, B. S. (2006).
  Identifying genes that contributes most to good classification in microarrays.
  BMC Bioinformatics, 7(407).

For Further Reading III

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.

Benjamini, Y. and Hochberg, Y. (1995).
Controlling the false discovery rate: a practical and powerful approach to multiple testing.
Journal of the Royal Statistical Society Series B, 57:289–300.

Bezdek, J. C. (1981).
Pattern Recognition with Fuzzy Objective Function Algoritms.
Plenum.

Bontempi, G. (1999).
Local Learning Techniques for Modeling, Prediction and Control.
PhD thesis, IRIDIA- Université Libre de Bruxelles.

For Further Reading IV

Breiman, L. (1996).
Bagging predictors.
Machine Learning, 24(123-140).

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984).
Classification and Regression Trees.
Chapman and Hall, New-York.

Butte, A. J., Tamayo, P., slonim, D., Golubb, T. R., and Koshane, I. S. (2000).
Discovering functional relationships between rna expression and chemotherapeutic susceptibilty using relevance networks.
PNAS, 97(22):12182–12186.

For Further Reading V

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., and Piccart, M. (2006).
Validation and clinical utility of a 70-gene prognostic signature for patients with node-negative breast cancer.
Journal of National Cancer Institute, 98:1183–1192.

Cox, D. R. (1972).
Regression models and life tables.
Journal of the Royal Statistical Society Series B, 34:187–220.

For Further Reading VI

- Davis, C. A., Gerick, F., Hintermair, V., Friedel, C. C., Fundel, K., Kuffner, R., and Zimmer, R. (2006).
  Reliable gene signatures for microarray classification: assessment of stability and performance.
  Bioinformatics, 22(19):2356–2363.

- de la fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004).
  Discovery of meaningful associations in genomic data using partial correlation coefficients.
  Bioinformatics, 20(18):3565–3574.

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).
  Maximum likelihood from incomplete data via the EM algorithm.
  Journal of the Royal Statistical Society, B, 39(1):1–38.

For Further Reading VII

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M., Buyse, M., and Sotiriou, C. (2007).
Strong time-dependency of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multi-centre independent validation series.
Clinical Cancer Research.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002).
Comparison of discrimination methods for the classification of tumors using gene expression data.
Journal of the American Statistical Association, 97(457):77–87.

For Further Reading VIII

Dunn, J. C. (1973).
A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.
Journal of Cybernetics, 3:32–57.

Dupuy, A. and Simon, R. (2007).
Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting.
Journal of National Cancer Institute, 99:147–157.

Efron, B. and Tibshirani, R. (1997).
Improvements on cross-validation: the .632+ bootstrap method.
Journal of American Statistical Asscoiation, 92(348-360).

Ein-Dor, L., Kela, I., , Getz, G., and Domany, E. (2005).
Outcome signature genes in breast cancer: Is there a unique set?
Bioinformatics, 21:171–178.

For Further Reading IX

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

Foekens, J. A., Atkins, D., Zhang, Y., Sweep, F. C., Harbeck, N., Paradiso, A., Cufer, T., Sieuwerts, A. M., Talantov, D., Span, P. N., Tjan-Heijnen, V. C., Zito, A. F., Specht, K., Hioefler, H., Golouh, R., Schittulli, F., Schmitt, M., Beex, L. V., Klijn, J. G., and Wang, Y. (2006).
Multicenter validation of a gene expression–based prognostic signature in lymph node–negative primary breast cancer.
Journal of Clinical Oncology, 24(11).

For Further Reading X

Freund, Y. and Schapire, R. E. (1996).
Experiments with a new boosting algorithm.
In Saitta, L., editor, *Machine Learning: Proceedings of the Thirteen International Conference*, pages 148–156, San Francisco. Morgan Kaufmann.

Friedman, J. H. (1996).
On bias, variance, 0/1 loss and the curse of dimensionality.
Data Mining and knowledge Discovery, pages 564–569.

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.

For Further Reading XI

Ioannidis, J. P. (2005).
Microarrays and molecular research: noise discovery?
Lancet, 365:454–455.

Kaufman, L. and Rousseeuw, P. J. (1990).
Finding groups in Data.
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.

For Further Reading XII

- Kontos, K. and Bontempi, G. (2006).
  Scale-free paradigm in yeast genetic regulatory network inferred from microarray data.
  In *Proceedings of AISB'06: Adaptation in Artificial and Biological Systems*, volume 3, pages 133–144.

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

- Loi, S., Piccart, M., Haibe-Kains, B., Desmedt, C., Harris, A., Bergh, J., Ellis, P., Miller, L., Liu, E., and Sotiriou, C. (2005).
  Prediction of early distant relapses on tamoxifen in early-stage breast cancer (BC): a potential tool for adjuvant aromatase inhibitor (AI) tailoring.

For Further Reading XIII

In *Proceedings of the American Society of Clinical Oncology Meeting, Orlando 2005, abstract 509*.

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.

Meyer, P. E., Kontos, K., and Bontempi, G. (2007).
Biological network inference using redundancy analysis.
In *1st International Conference on Bioinformatics Research and Development*.

For Further Reading XIV

Michiels, S., Koscielny, S., and Hill, C. (2005).
Prediction of cancer outcome with microarrays: a multiple radom validation strategy.
Lancet, 365:488–492.

Mitchell, T. (1997).
Machine Learning.
McGraw.

Schwarz, G. (1978).
Estimating teh dimension of a model.
Annals of Statistics, 6:461–464.

For Further Reading XV

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003).
Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.
J. Natl Cancer Inst, 95(1):14–18.

Sotiriou, C., Haibe-Kains, B., Desmedt, C., Wirapati, P., Durbecq, V., Harris, A., Larsimont, D., Bontempi, G., Buyse, M., Delorenzi, M., and Piccart, M. (2006).
Comprehensive molecular analysis of several prognostic signatures using molecular indices related to hallmarks of breast cancer: proliferation index appears to be the most significant component of all signatures.
In Springer, editor, *Breast Cancer Research and Treatment*, volume 100, page S86.

For Further Reading XVI

- Tibshirani, R. (1997).
  The lasso method for variable selection in the cox model.
  Statistics in Medecine, 16:385–395.

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

- van de Vijver, M. J., He, Y. D., van't Veer, L., Dai, H., Hart, A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002).
  A gene expression signature as a predictor of survival in breast cancer.
  The new England Journal of Medecine, 347(25):1999–2009.

For Further Reading XVII

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhiven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002).
Gene expression profiling predicts clinical outcome of breast cancer.
Nature, 415:530–536.

Vapnik, V. N. (1998).
Statistical Learning Theory.
Springer.

Varma, S. and Simon, R. (2006).
Bias in error estimation when using cross-validation for model selection.
BMC Bioinformatics, 7(91):1471–2105.

For Further Reading XVIII

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., van Gelder, M. E. M., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D., and Forekens, J. A. (2005).
Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.
Lancet, 365(671-679).

Zhang, L., Wang, L., Ravindranathan, A., and Miles, M. F. (2002).
A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions.
Journal of Molecular Biology, 317(2):225–235.

Part VII

Appendix

Self-Organizing Maps

- Method introduced in [Kohonen, 1997].
- Use of self-organizing neural networks (map) to reduce dimension.
- Most SOMs uses a map of 2 dimensions but user can increase the dimensions (in losing nice visualization)

- 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
  - computationally intensive.

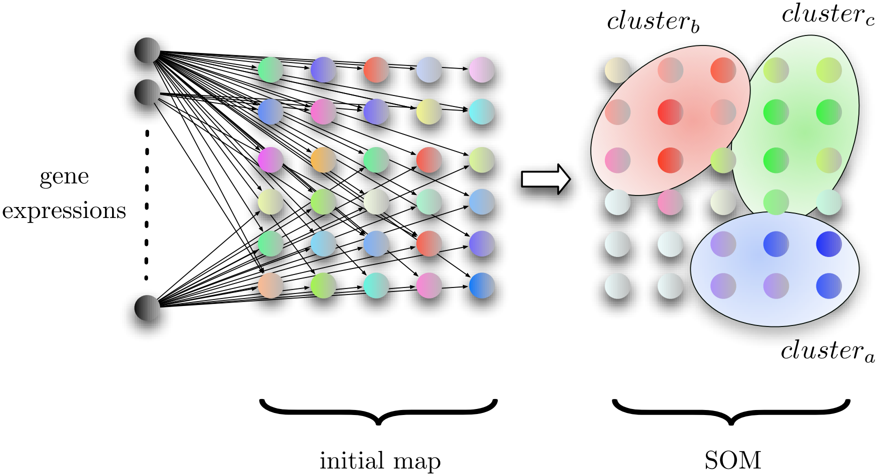
Self-Organizing Maps

Procedure

- Initialize a map, ie a matrix of nodes.
- Set the update parameter t to 1 and decrease it to 0 by small amounts
- Randomly select a sample
- Get the closest feature in the map
  - if there are several equally distant features, pick one randomly
- Update the feature and its neighborhood
  - Methods to define the neighborhood and update features can be tuned by the user
- Continue till t equals 0 (no more update)

Self-Organizing Map

Procedure cont'd



Criterion for Number of Clusters

Cophenetic Correlation Coefficient

- Cophenetic correlation coefficient can be used to measure how well the hierarchical structure from the dendrogram represents the actual distances.
- This measure is defined as the correlation between the $\frac{n(n-1)}{2}$ pairwise dissimilarities between observations and their cophenetic dissimilarities from the dendrogram, i.e., the between cluster dissimilarities at which two observations are first joined together in the same cluster.

Criterion for Number of Clusters

PAM : Silhouette

- Each cluster is represented by one *silhouette*
- Consider any object X_i of the dataset and let A be the cluster to which it is assigned

$$m_A(X_i) = \mathrm{mean}\{d(X_i, X_A)\}$$

  where X_A are all objects assigned to A
- Consider any cluster C different from A

$$m_C(X_i) = \mathrm{mean}\{d(X_i, X_C)\}$$

  where X_C are all objects assigned to C
- let B be the cluster such that

$$m_B(X_i) = \min\{m_C(X_i)\}$$

- *silhouette* statistic is

$$s(X_i) = \frac{m_B(X_i) - m_A(X_i)}{\max\{m_B(X_i), m_A(X_i)\}}$$

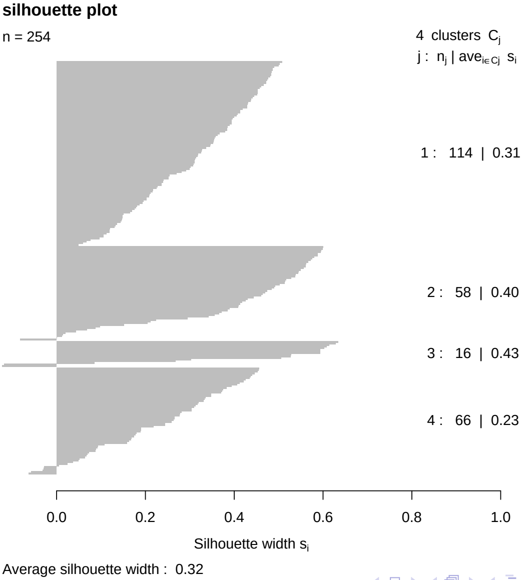
Criterion for Number of Clusters

PAM : Silhouette cont'd

- $s(X_i)$ lies betwen -1 and $+1$
  - $s(X_i) = 1$: within distance $m_A(X_i)$ is much smaller than the smallest between distance. In other words, object X_i has been assigned to an appropriate cluster. The second best cluster B is not nearly as close as the actual cluster A.
  - $s(X_i) = 0$: $m_A(X_i)$ and $m_B(X_i)$ are approximately equal. Hence, it is not clear whether X_i should be assigned to A or B. It can be considered as an intermediate case.
  - $s(X_i) = -1$: object X_i is badly classified. When s is close to -1, the object is poorly classified. Its distance with other objects in its cluster is much greater than its distance with objects in the nearest cluster.
- A *silhouette* of a cluster is a barplot of ranked $s(X_i)$
- The mean of all *silhouette* statistics may be used to select the "good" number of clusters

Criterion for Number of Clusters

PAM : Silhouette Plot



Criterion for Number of Clusters

Bayesian information criterion

- Bayesian information criterion (BIC) [Schwarz, 1978]

- Akaike's information criterion (AIC) [Akaike, 1974]

- The hypothesis must be stated in mathematical/statistical terms that make it possible to calculate the probability of possible samples assuming the hypothesis is correct.
  - example : *the mean response to treatment being tested is equal to the mean response to the placebo in the control group*.
- A test statistic must be chosen that will summarize the information in the sample that is relevant to the hypothesis (also called sufficient statistic).
  - example : the numerical difference between the two sample means, $mean_1 - mean_2$

Introduction

Statistical Hypothesis Testing cont'd

- The distribution of the test statistic is used to calculate the probability sets of possible values.
  - assumptions about the distribution of the test statistic : **parametric statistic**
  - no assumptions about the distribution of the test statistic : **nonparametric statistic**
- Among all the sets of possible values, we must choose one that we think represents the most extreme evidence against the hypothesis (critical region of the test statistic).

		Truth	
		H_0	H_1
Decision	H_0	correct acceptance	type II error (β)
	H_1	type I error (α)	correct rejection

Statistical Hypothesis Testing cont'd

- If the test statistic is inside the critical region, then our conclusion is one of the following :
  - the hypothesis is incorrect, therefore reject the null hypothesis
  - an event of probability less than or equal to α has occurred.
- If the test statistic is outside the critical region, the only conclusion is that there is not enough evidence to reject the hypothesis.
- This is not the same as evidence in favor of the hypothesis.
  - lack of evidence against a hypothesis is not evidence for it.
- On this basis, statistical research progresses by eliminating error, not by finding the truth.

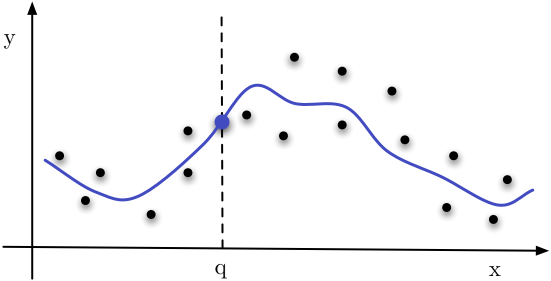
Linear Regression

- We assume that the relation between the outcome variable **y** and the features **X** is linear
- The form of the model is

$$\mathbf{y} = \beta_0 + \beta X + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$
- To fit the unknown parameters of the model, ie β_0, β and σ, we can use either the least squares or the maximum likelihood methods

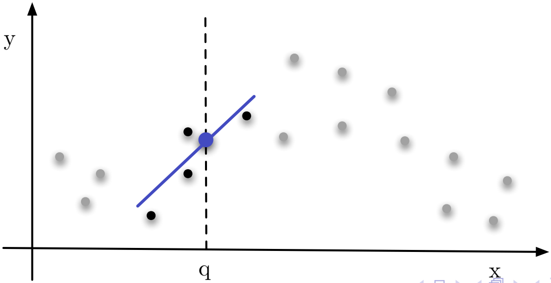
- Traditional approach to supervised learning is global modeling which describes the relationship between the input and the output with an analytical function over the whole input domain



Regression

Lazy Learning cont'd

- Lazy learning [Bontempi, 1999] uses a local modeling approach where a complex problem is divided into simpler problems
- Solutions of these simpler problems can be combined to yield a solution to the original problem
- The algorithm is lazy because all computations are made at query only
- It fits local linear models using cross-validation, via the PRESS statistic [Allen, 1974], to select the best model



Information Gain

Suppose $\mathbf{x}$ can have one of m values $v_1, v_2, \ldots, v_m$ such that

$$\Pr(\mathbf{x} = v_1) = p_1, \Pr(\mathbf{x} = v_2) = p_2, \ldots, \Pr(\mathbf{x} = v_m) = p_m$$

- Entropy : $H(x) = -\sum_{j=1}^{m} p_j \log p_j$
  - "High Entropy" means x is from a uniform (flat) distribution
  - "Low Entropy" means x is from varied (peaks and valleys) distribution
- Conditional entropy : $H(\mathbf{y}|x) = \sum_{j=1}^{m} \Pr(\mathbf{x} = v_j) H(\mathbf{y}|\mathbf{x} = v_j)$
- Information gain : $IG(\mathbf{y}|x) = H(\mathbf{y}) - H(\mathbf{y}|x)$