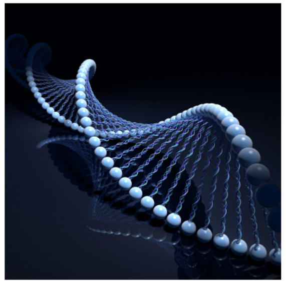
LGBIO2010: Large scale gene expression analysis

Pierre Dupont



UCL – ICTEAM

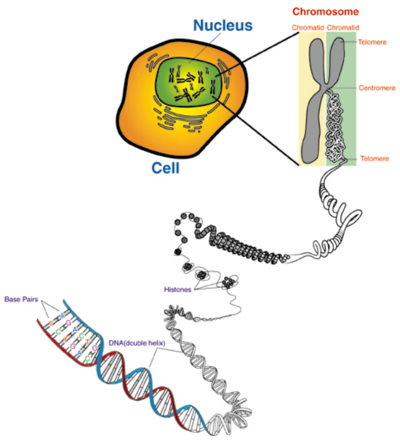
Outline

1. Introduction
2. Preprocessing
3. Unsupervised gene selection
4. Supervised gene selection
   - Filters
     - Non-specific filtering
     - Fold changes
     - t-Test
     - Mutual information
     - Multivariate filters
   - Wrappers
   - Embedded Methods
     - Classification

Outline

1. Introduction
2. Preprocessing
3. Unsupervised gene selection
4. Supervised gene selection

Gene expression



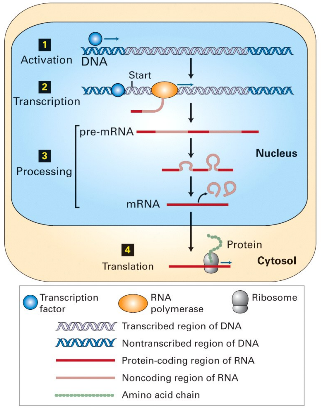
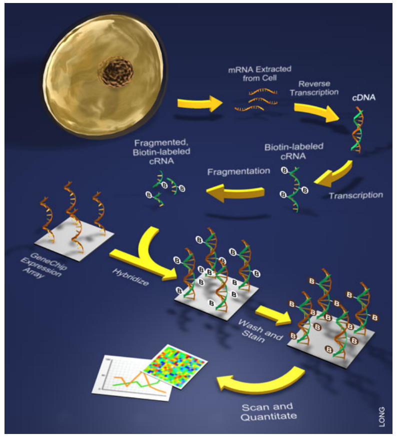


Illustration from Molecular Cell Biology, 5e (© WHFreeman 2004).

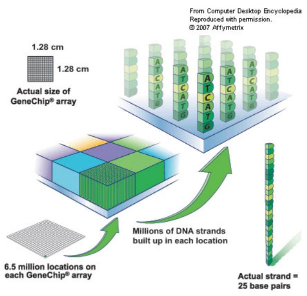
DNA Microarrays

DNA Microarrays measure the level of expression of all genes in a single experiment

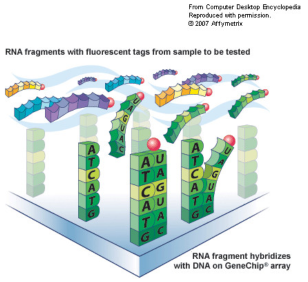
1. Data measurements
2. Preprocessing and sample normalization
3. Gene selection and sample classification
4. Diagnosis, prognosis or prediction of the response to a treatment



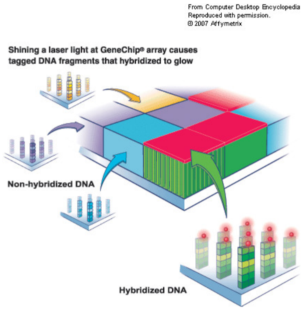
Affymetrix® technology


1.28 cm
1.28 cm
Actual size of GeneChip® array
6.5 million locations on each GeneChip® array
Millions of DNA strands built up in each location
Actual strand = 25 base pairs

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2007 Affymetrix


RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip® array

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2007 Affymetrix


Shining a laser light at GeneChip® array causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2007 Affymetrix

Alternative measurement technologies

- Other companies sell DNA chips (Agilent®, . . .)
- Multiplex qPCR (Applied Biosystems®, . . .)
  - larger dynamic range than microarrays
  - limited to ≈ 100 genes
- RNAseq (Illumina®, Ion Torrent®, . . .)
  - fastly evolving
  - scaling effects influence per sample cost

Example: diagnosis

Biomarkers for an early diagnosis of rheumatoid infections

Prediction problem: multi-class feature selection

- Rheumatoid arthritis
- Lupus
- Psoriatic rheumatism
- Microcristalline arthritis
- Inflammatory osteoarthritis

RHEUMAGENE research project with Prof. Lauwerys (UCL/IREC/RUMA)

Example: prognosis

Biomarkers to predict the risk of allergies of newborns

- More than 30% of children are allergic in industrial countries
- Predicting who is more likely to become allergic is a path to prevention and possible treatment

CRISTALL research project with Profs. Sokal and Smets (UCL/IREC/PEDI)

Example: response to treatment prediction

Gene profiling for cancer treatment

Objective

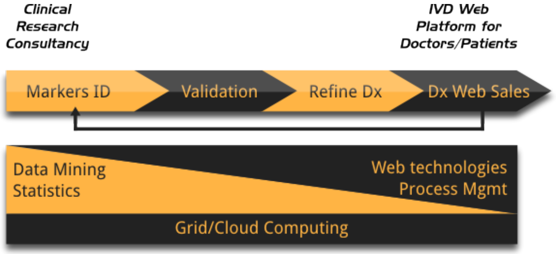
Identify biomarkers for predicting patient response to MAGE-A3 immuno-therapy against melanoma before treatment

In collaboration with GSK Biologicals - WO/2010/029174 (patent).

DNAlytics

Data Mining & IT Solutions for Personalized Medicine

A UCL spin-off



www.dnalytics.com

Supervised selection

	gene 1	gene 2	...	gene p	response
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$	y_1
...	...	...	...	...	...
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	y_n
test sample	x_1	x_2	...	x_p	?

- The number p of input dimensions (probes, probesets or genes) may be very large ($10^4 \dots 10^6$)
- The number n of samples is typically much smaller ($\approx 50 \dots 100$)
- Each sample is characterized by a vector $\boldsymbol{x}$ of p measurements
- Each training sample has a known response: class label y ($y \in \{-1, 1\}$ or $y \in \mathbb{N}$) or $y \in \mathbb{R}$

Gene selection

Find a small subset of genes, (a.k.a features, attributes or input variables), to predict the response or class y of new samples

Gene selection

Objectives

- Insight into the data and the predictive model
- Link between data analysis and medical expert
- Biological validation on a few genes rather than thousand ones
- Reduction of the financial cost of a diagnosis/prognosis kit (technological constraints)

Difficulties

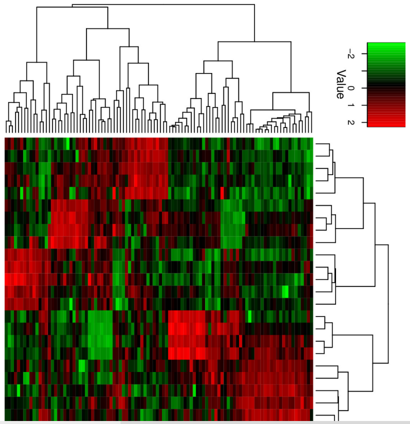
- Measurements are noisy
- Gene expression varies due to many factors (gender, cell type, growth of the organism, chemical environment of the cell, . . .) often not related to the response to be predicted
- Financial cost: $500 \ldots 1,000$ €/experiment
- Small n (e.g. 50), large p (e.g. $50,000$) problems

Unsupervised selection

	gene 1	gene 2	...	gene p	cluster
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$	?
...	...	...	...	...	...
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	?
cluster	?	?	...	?	

Objective

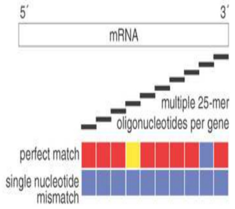
Find clusters of genes and/or samples that share a similar profile: up or down regulated genes across the same samples

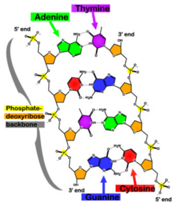


Outline

1. Introduction
2. Preprocessing
3. Unsupervised gene selection
4. Supervised gene selection

Summarization


5′
3′
mRNA
multiple 25-mer oligonucleotides per gene
perfect match
single nucleotide mismatch



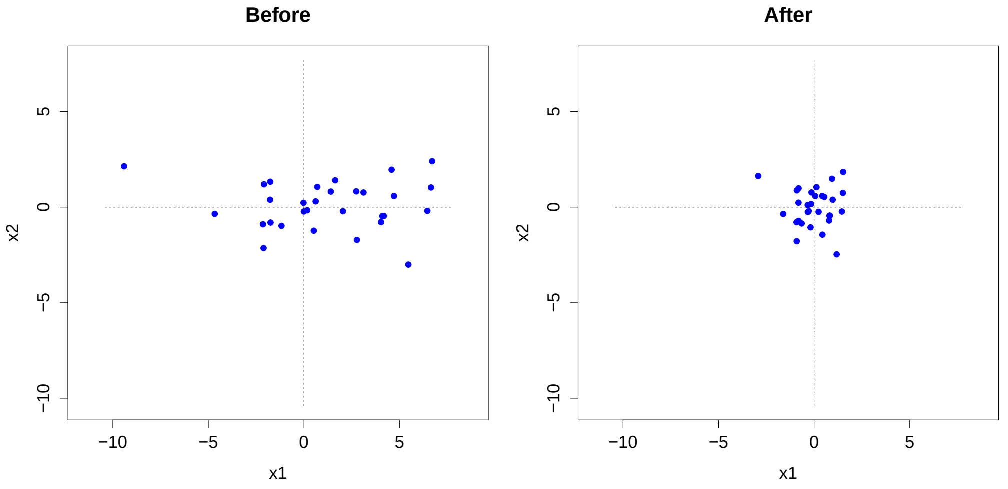
- Define a single probeset expression level from the various probe intensities
- Popular techniques: MAS 5.0, RMA, **GC-RMA**
  1. background adjustment: optical noise correction, probe affinity adjustment (influenced by the GC content), RMA ignores the MM probes
  2. sample normalization: quantiles should be stable across samples, after conversion to log intensities for (GC-)RMA
  3. summarization: median polish

Feature normalization

	gene 1	gene 2	...	gene p
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$
...	...	...	...	...
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$

- Make sure that each gene (probeset) has roughly the same expression range across all samples
- Z-score normalization
  Replace $x_{i,j}$ by $\frac{x_{i,j}-\mu_j}{s_j}$ with μ_j the mean level of expression of probeset j over the training samples and s_j its standard deviation

Feature normalization example



Distance between expression values

	gene 1	gene 2	...	gene p
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$
...	...	...	...	...
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$

Euclidean distance

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| = \sqrt{\sum_{i=1}^{n}(x_{i,1} - x_{i,2})^2}$$

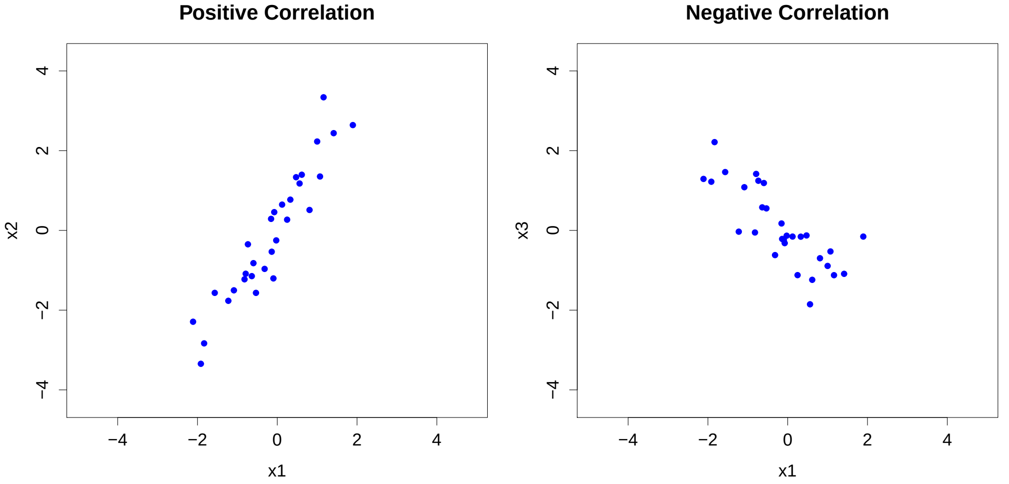
Correlation based distance

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1 - \frac{1}{2}(1 + \mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2))$$

Correlation between expression values

Are both genes over/under expressed on the same samples?

Is one gene over-expressed when the other is under-expressed?



Pearson correlation

For two random vectors (*e.g.* gene expression values) $\boldsymbol{x}_1, \boldsymbol{x}_2$ measured over n samples

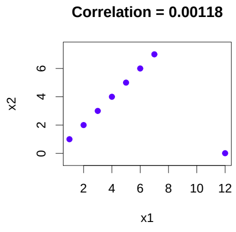
$$\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i,2} - \bar{x}_2)^2}}$$

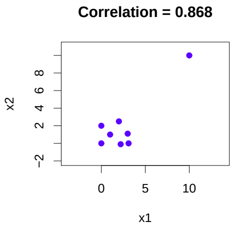
- $\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \pm 1$ if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are perfectly linearly correlated
- $\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$ if they are not linearly correlated
- whenever $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and normalized to zero mean and unit variance:

$$\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{n} x_{i,1} x_{i,2} = \boldsymbol{x}_1^{\top} \boldsymbol{x}_2$$

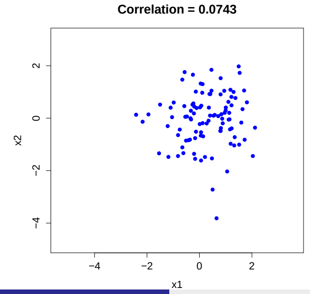
Pitfalls with correlation measures

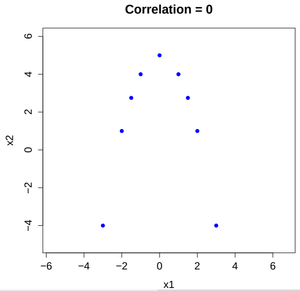
- Correlation is very sensitive to outliers





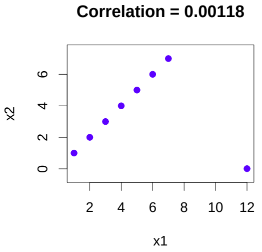
- Correlation measures linear dependence

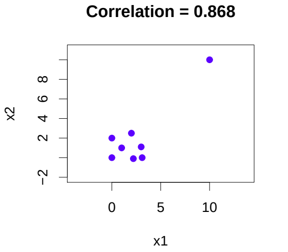


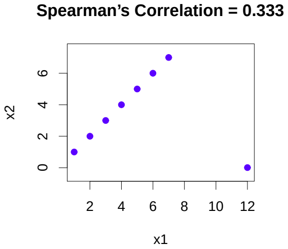


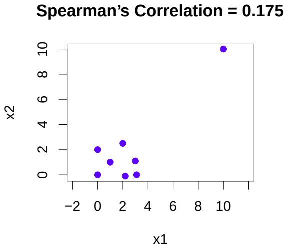
Spearman's rank correlation: less sensitive to outliers

1. Replace feature value by feature value rank across observations
2. Compute Pearson correlation between rank vectors

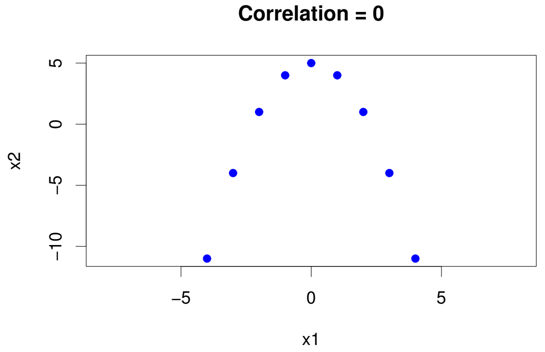








Uncorrelated features are not necessarily independent



- $\mathrm{corr}(\boldsymbol{x}_1, \boldsymbol{x}_2) = 0$ (both Pearson and Spearman correlations)
- $P(\boldsymbol{x}_2 | \boldsymbol{x}_1) \neq P(\boldsymbol{x}_2)$

Outline

1 Introduction

2 Preprocessing

3 Unsupervised gene selection

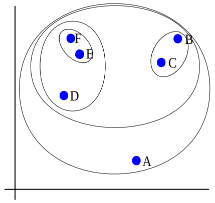
4 Supervised gene selection

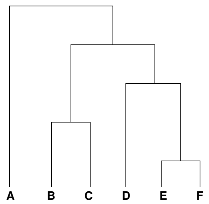
Agglomerative Hierarchical clustering

Each observation represents

either a sample across genes

or a gene across samples





```
Algorithm AGGLOMERATIVEHIERARCHICALCLUSTERING
Input: $D$ a set of observations $\vec{x}_1, \ldots, \vec{x}_m$ ; $d(\vec{x}, \vec{y})$ a distance measure between
         observations
Output: A tree $T$ of subsets of $D$
// Initialize a set $\mathcal{D}$ of clusters $D_1, \ldots, D_m$
$\mathcal{D} \leftarrow \{\{\vec{x}_1\}, \ldots, \{\vec{x}_m\}\}$        // Initial clusters are tree leaves
$T \leftarrow$ a partial tree whose leaves are the $\vec{x}_i$'s
while $|\mathcal{D}| > 1$ do
    Choose pair of clusters $(D_i, D_j)$ in $\mathcal{D}$ such that $Distance(D_i, D_j, d)$ is minimal
    Define a new cluster $D_k = D_i \cup D_j$
    $\mathcal{D} \leftarrow \mathcal{D} \cup D_k - \{D_i, D_j\}$
    Add $D_k$ as parent node of $D_i$ and $D_j$ in the tree $T$
return $T$
```

Distance measure between clusters

Single-link or nearest neighbor rule

$$Distance(D_i, D_j, d) = \min_{\vec{x} \in D_i, \vec{y} \in D_j} d(\vec{x}, \vec{y})$$

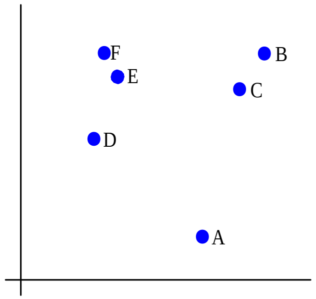
Complete-link or farthest neighbor rule

$$Distance(D_i, D_j, d) = \max_{\vec{x} \in D_i, \vec{y} \in D_j} d(\vec{x}, \vec{y})$$

Average-link rule

$$Distance(D_i, D_j, d) = \frac{1}{|D_i|.|D_j|} \sum_{\vec{x} \in D_i, \vec{y} \in D_j} d(\vec{x}, \vec{y})$$

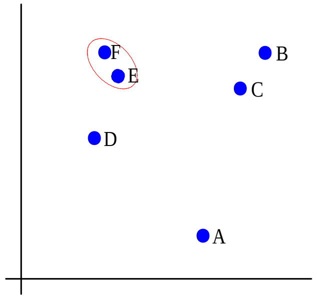
Hierarchical clustering example

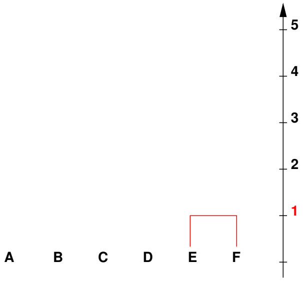


A B C D E F

	A	B	C	D	E	F
A						
B	8					
C	6	2				
D	5	8	6			
E	7	6	4	3		
F	8	6	5	4	1	

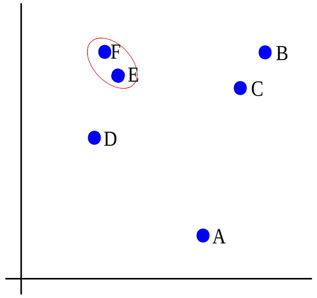
Hierarchical clustering example

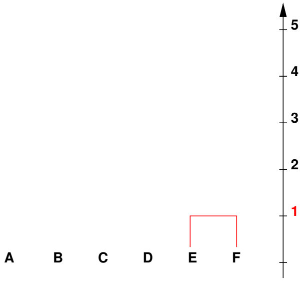




	A	B	C	D	E	F
A						
B	8					
C	6	2				
D	5	8	6			
E	7	6	4	3		
F	8	6	5	4	1	

Hierarchical clustering example

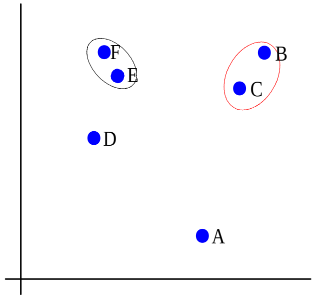


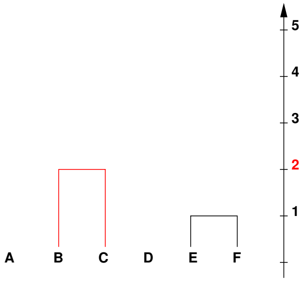


Single-link rule

	A	B	C	D	EF	
A						
B	8					
C	6	2				
D	5	8	6			
E F	7	6	4	3		
	8	6	5	4	1	

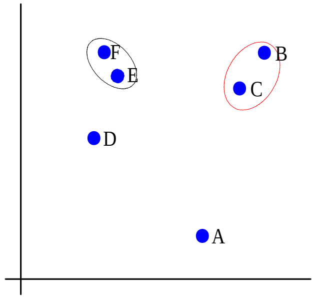
Hierarchical clustering example

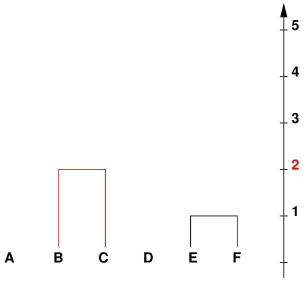




	A	B	C	D	EF
A					
B	8				
C	6	2			
D	5	8	6		
EF	7	6	4	3	

Hierarchical clustering example

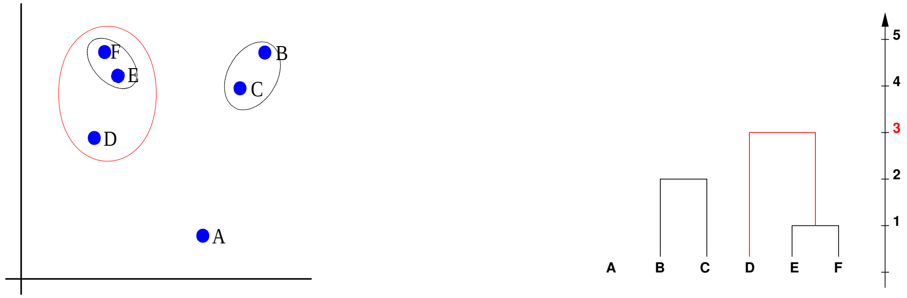




Single-link rule

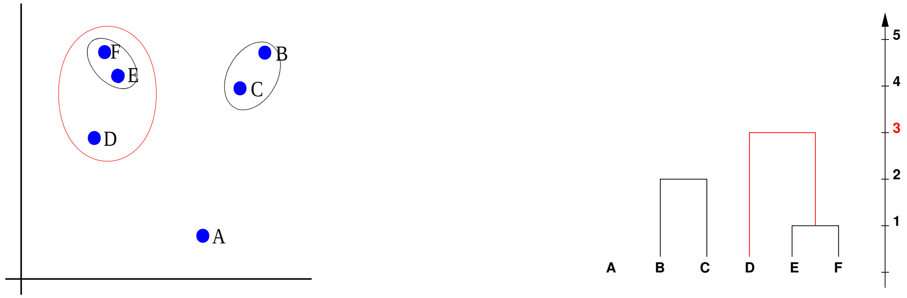
	A	BC		D	EF
A					
BC	8				
	6	2			
D	5	8	6		
EF	7	6	4	3	

Hierarchical clustering example



	A	BC	D	EF
A				
BC	6			
D	5	6		
EF	7	4	3	

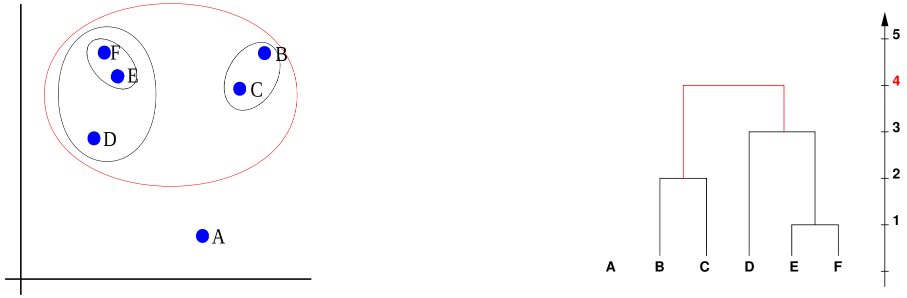
Hierarchical clustering example



Single-link rule

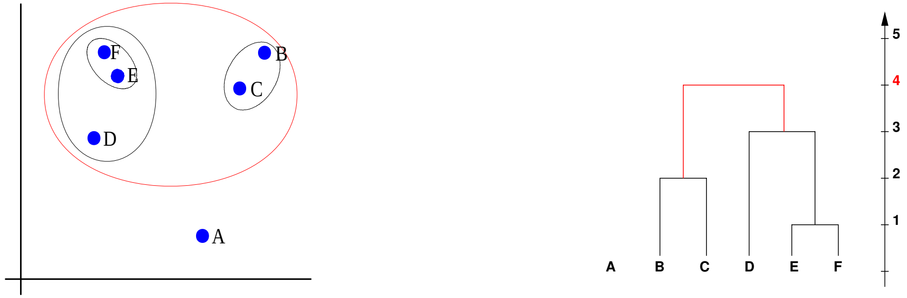
	A	BC	DEF	
A				
BC	6			
DEF	5	6		
	7	4	3	

Hierarchical clustering example



	A	BC	DEF
A			
BC	6		
DEF	5	4	

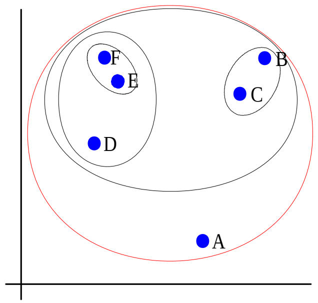
Hierarchical clustering example

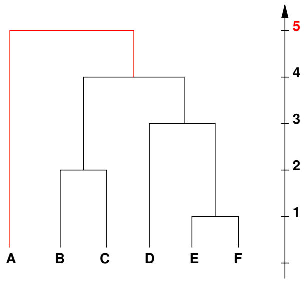


Single-link rule

	A	BCDEF	
A			
BC DEF	6		
	5	4	

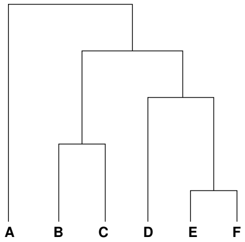
Hierarchical clustering example





	A	BCDEF
A		
BC DEF	5	

A note on phylogeny



This hierarchical clustering algorithm, with the average-link rule, is known as the UPGMA algorithm used in phylogeny

- The observations are (fragments) of sequences representative of some species, called taxa
- The pairwise distance measure is based on alignment scores, generally corrected according to an evolutionary model (*e.g.* Kimura)
- The final tree is interpreted as a phylogenetic tree and the branch length as representative of time

Outline

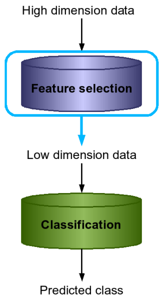
1. Introduction
2. Preprocessing
3. Unsupervised gene selection
4. Supervised gene selection
   - Filters
   - Wrappers
   - Embedded Methods

Supervised gene selection

	gene 1	**gene 2**	**...**	**gene p**	**class**
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$	+
sample 2	...	...	...	...	+
...	...	...	...	...	...
sample n-1	...	...	...	...	-
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	-
test sample	x_1	x_2	...	x_p	?

- we discuss binary classification first:
  e.g. responders (+ or class 1) vs non-responders (- or class 2)
- samples can be indexed by their class label
  - means and variances can be computed on samples of a given class
- find a subset of most discriminating genes for the prediction of the class of any new sample

Feature selection: filters



High dimension data

Feature selection

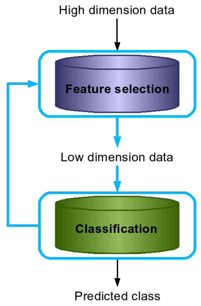
Low dimension data

Classification

Predicted class

- Use only the training data + class labels during the feature selection step
- Standard techniques: fold changes, t-Test, mutual information, . . .
- Train a single classifier taking the selected features as inputs
- The simplest and less computing intensive approach

Feature selection: wrappers



High dimension data

Feature selection

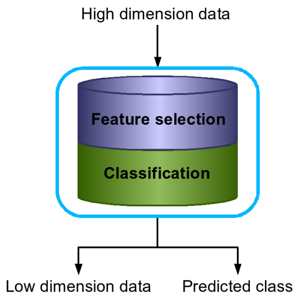
Low dimension data

Classification

Predicted class

- Train a classifier on several subsets of all possible features
  - Exhaustive evaluation of all possible subsets is unfeasible Note: there are $\mathcal{O}(2^p)$ subsets with $p \geq 10,000$
  - Typical solutions: use feature ranking or forward/backward selection
- Select the feature set that optimizes the performance of the trained classifier

Feature selection: embedded approaches



High dimension data

Feature selection

Classification

Low dimension data

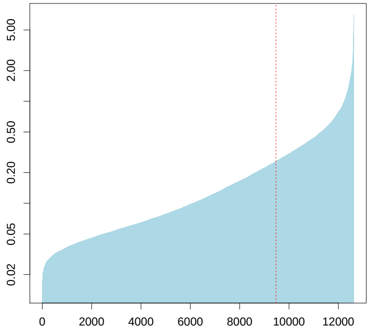
Predicted class

- Define the feature selection and the classifier estimation as a combined optimization process
- Include classifier optimization in the feature selection process
- More elegant/relevant but also more computing intensive than a filter

Non-specific filtering

	gene 1	gene 2	...	gene p	class
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$	+
sample 2	...	...	...	...	+
...	...	...	...	...	...
sample n-1	...	...	...	...	-
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	-

- genes with a small variance across all training samples are unlikely to be discriminating between classes
- keep only those genes (*e.g.* 25 %) with the larger variances
  - before normalization to unit variance!



Fold changes

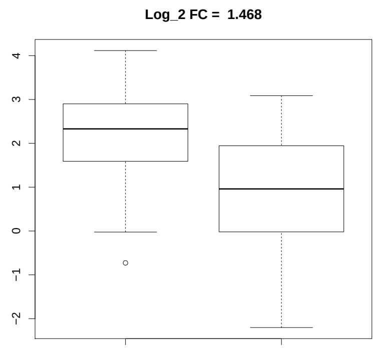
	gene 1	gene 2	...	gene p	class
sample 1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,p}$	+
sample 2	...	...	...	...	+
...	...	...	...	...	...
sample n-1	...	...	...	...	-
sample n	$x_{n,1}$	$x_{n,2}$	...	$x_{n,p}$	-

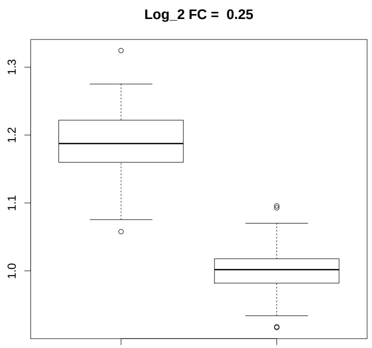
Select genes with the larger fold changes between both conditions

$$\frac{\bar{x}_1}{\bar{x}_2} \text{ or } \log \frac{\bar{x}_1}{\bar{x}_2} = \log \bar{x}_1 - \log \bar{x}_2 \text{ or } \bar{x}_1 - \bar{x}_2$$

Comments on fold changes

- whenever $\bar{x}_1 < \bar{x}_2$, one considers a small value as important
  - $\log_2 \frac{\bar{x}_1}{\bar{x}_2}$ should be ≥ 1 or ≤ -1
- is a two-fold change significant?
  - dependence on the measurement technology
  - dependence on the class conditional variance





t-Test relevance index

- A feature relevance $J(x)$ can be defined according to the distance between the average feature value in each class
- The larger the distance the better, relatively to standard deviations

t-Test statistic (actually Welch t-statistics)

$$J(x) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

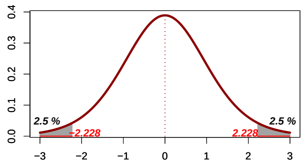
with n_1 (resp. n_2) the number of examples labeled as $+$ (resp. $-$) and the estimated variances in each class $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

Confidence measure

- The Welch statistics follows a t-distribution with a number of degrees of freedom equal to:

$$\frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)(n_1-1) + (S_2^2/n_2)(n_2-1)}$$

- p-values assess the significance of the difference between the two class means



- A feature is selected if its associated p-value is below a prescribed threshold (*e.g.* $5\% \Rightarrow |J(x)| \geq 2.228$ when d.f. $= 10$)

The R Project for Statistical Computing

An efficient way of computing *p*-values, and **many** other useful things. . .

http://www.r-project.org/

```
> t.test(x1,x2)
        Welch Two Sample t-test
data:  x1 and x2
t = 0.9183, df = 7.002, p-value = 0.389
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.908216  2.061549
sample estimates:
mean of x mean of y
 1.866667  1.290000
```

where `x1` (resp. `x2`) is the vector of expression values of a given gene from samples labeled as class 1 (resp. class 2)

$p-$value $> 0.05 \Rightarrow$ the difference between the 2 class means is not considered significant for this feature ⇒ discard the feature

Alternatives to a simple t-Test

- Mann-Whitney rank test is an alternative *non-parametric* test
- ANOVA offers a generalization of the t-Test in a multi-class (> 2) setting
- Pairwise t-tests between one class and the others is a common alternative
- Kruskal-Wallis is a generalization of Mann-Whitney to multi-class

The multiple test problem

A microarray experiment to distinguish between patients with a positive or a negative diagnosis

Among 50,000 gene expression values measured in each experiment, only those genes that are differently expressed, with a *p*-value $\leq 0.05 = \alpha$, are selected

The probability of type I error of a statistical test

Conclude that the mean expression values among the 2 classes are significantly different for a given gene while they are not ⇒ the feature is falsely selected with probability α

Test multiplicity

- The test will be performed for each gene ⇒ 50,000 times from the same experiment
- If $\alpha = 0.05$, we are expecting to select wrongly $50,000 \times .05 = 2,500$ genes !!

Multiple test correction

Bonferroni correction

Divide the critical value (e.g. $\alpha = .05$) by the number of tests n_t performed

Example: $\frac{\alpha}{n_t} = \frac{.05}{50,000} = 10^{-6}$

⇒ only genes with associated *p*-values $\leq 10^{-6}$ are selected

Very conservative, often leads to select no feature

False Discovery Rate correction

Benjamini-Hochberg correction

1. Select a confidence level α (*e.g.* 0.05)
2. Rank the p-values (one for each feature) in increasing order $p_1 \leq p_2 \leq \cdots \leq p_{n_t}$
3. Iterate over the n_t features
   ($n_t = p$ with data in $\mathbb{R}^p$, not to be confused with $p-$values)
   - Find the maximal index i such that $\frac{p_i \times n_t}{i} < \alpha$
4. Keep all features up to index i_{max}

Notes:

- If $p_{n_t} < 0.05$ FDR correction leads to select all features
- FDR correction is equivalent to Bonferonni correction whenever a single feature is selected: $p_1 \times n_t < \alpha \Leftrightarrow p_1 < \frac{\alpha}{n_t}$
- Those corrections do not change the relative ranking of features, just the selection threshold
- See R function `p.adjust`

Feature ranking with mutual information

$$I(X;Y) = -\sum_{ij} P(x_i, y_j) \log_2 \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$

$$= -\sum_{ij} P(x_i, y_j) \log_2 \frac{P(y_j|x_i)}{P(y_j)}$$

- A feature X is more relevant if its mutual information with the class value is higher
- If X tends to bring no information to predict Y then
  $P(y_j|x_i) \approx P(y_j)$ and $I(X;Y) \rightarrow 0$
- $I(X;Y) = 0$ **if and only if** X and Y are independent
- $I(X;Y)$ is invariant under rescaling of the variables X and Y (often rescaled to unit variance)

Univariate versus multivariate filters

- Correlation measures, t-Test (ANOVA), and $I(X;Y)$ are univariate filters
- Mutual information can be used to select several variables at a time $I(X_1, \ldots, X_k; Y)$ **but** MI depends on the distributions $P(X_1, \ldots, X_k, Y)$, $P(X_1, \ldots, X_k)$ and $P(Y)$, which need to be reliably estimated
  - replace the joint problem by an approximation with a greedy selection of features

Maximum relevance minimum redundancy

[Peng et al., 05]

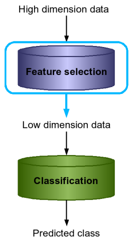
1. Select the feature with maximum mutual information with the response
   - $\hat{X} = \text{argmax}_X \, I(X;Y)$
   - $\Phi = \{\hat{X}\}$ // Initialize the set of selected features
   - $F = \{X_1, \ldots, X_p\} \setminus \{\hat{X}\}$ // The remaining set of features
2. **Repeat**

   $$\hat{X} = \text{argmax}_{X \in F} \left[\underbrace{I(X;Y)}_{\text{maximize relevance}} \underbrace{- \frac{1}{|\Phi|} \sum_{X_j \in \Phi} I(X;X_j)}_{\text{minimize redundancy}} \right]$$

   $\Phi \leftarrow \Phi \cup \{\hat{X}\}$; $F \leftarrow F \setminus \{\hat{X}\}$

   until an appropriate number of features are selected

Filters in a nutshell



High dimension data

Feature selection

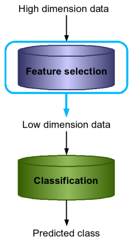
Low dimension data

Classification

Predicted class

- Use only the training data + class labels during selection
- Filters offer interesting baselines which are fast to compute
- Popular univariate filters are based on a t-Test (with multiple test correction) or mutual information
  - they ignore the interactions between genes!
- Maximum relevance minimum redundancy is a popular multivariate extension

Are filters independent from a predictive model?



High dimension data

Feature selection

Low dimension data

Classification

Predicted class

- The two step approach is sometimes considered as a benefit since the features are claimed to be selected independently from the subsequent classifier/regression model
  - Is it really better? (see embedded methods)
  - Is it really the case?

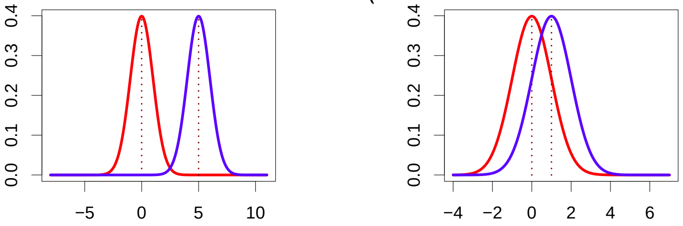
t-Test revisited

t-Test statistic

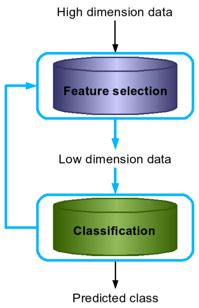
A feature x is selected whenever the difference between the class means is significant (after correction for multiplicity)

$$J(x) = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

Equivalently, one easily discriminates between the classes using 2 uni-dimensional Gaussian classifiers (with a common variance)



Wrapper principle



High dimension data

Feature selection

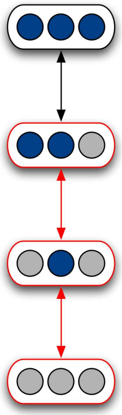
Low dimension data

Classification

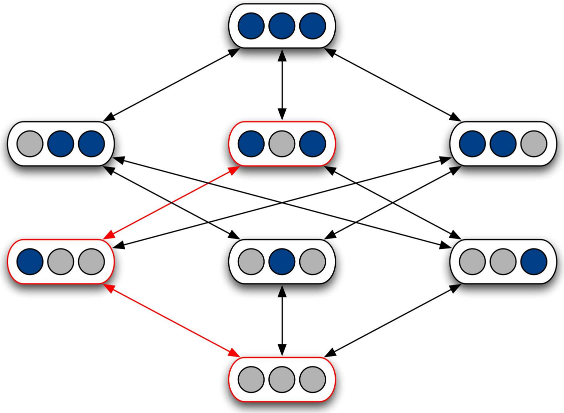
Predicted class

- Estimate a classifier from a given subset of all possible features
- Select the feature subset that optimizes the performance of the classifier (usually on an independent validation set)
  - Feature selection depends on the evaluation protocol of the classifier
  - There are $\mathcal{O}(2^p)$ possible subsets

Univariate feature ranking



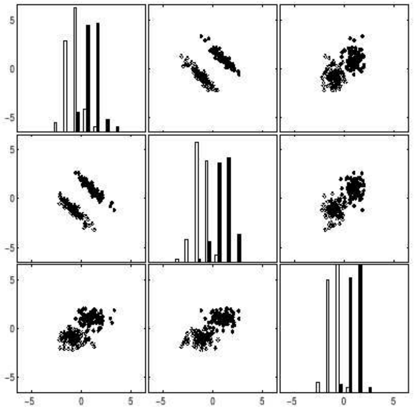
Multivariate Forward/Backward selection



- Forward selection goes bottom-up
- Backward selection goes top-down

Search order matters

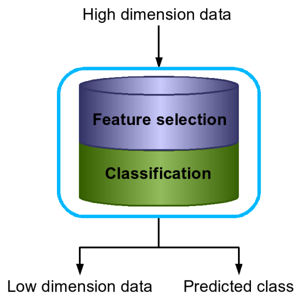
- x_3 alone is better than x_1 or x_2 alone, but x_1 together with x_2 offer the best discrimination
- **2 best features:**
  - Univariate feature ranking selects (x_3, x_2)
  - Forward selection selects (x_3, x_1)
  - Backward selection selects (x_1, x_2)
- **Single best feature:**
  - Forward selection or univariate feature ranking selects x_3
  - Backward selection selects x_2



Wrappers in a nutshell

- A wrapper with **univariate feature ranking** offers a **good baseline**
  - The t-Test statistics can be used to rank features only (no need for multiple test correction nor fixing a confidence measure)
  - Classifier performance is used to decide how many features to keep
  - This can outperform a pure filter approach while not increasing much the computational cost
- A **backward selection** may be preferable over a forward selection, but should not be used to select just a few features (what “a few” means depends on the data. . .)
- More sophisticated search strategies are possible (backward + forward, randomized search, . . .)
- If one can afford the computational cost of a multivariate selection, one should probably consider an embedded approach

Embedded Methods



High dimension data

Feature selection

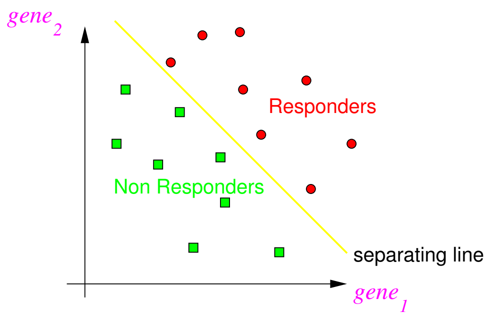
Classification

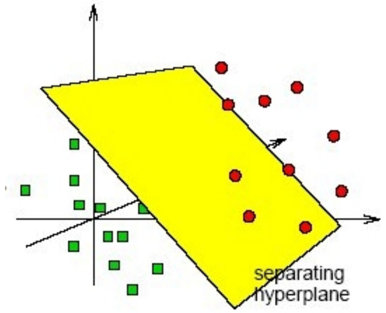
Low dimension data

Predicted class

- Define the feature selection and the classifier estimation as a combined optimization process
- The features are selected as a by-product of the estimated classifier and its parameters

Linear Discriminants

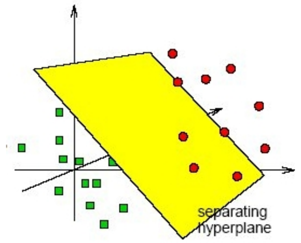




- The actual number of dimensions may be $\approx 10,000$ for microarray classification
- The linear discriminant is a hyperplane in $\mathbb{R}^{\geq 10,000}$
- Decision rule: $\text{sign}\ (\sum_{j=1}^{p} w_j x_j + w_0) = \text{sign}\ (\mathbf{w}^\top \mathbf{x})$ (with $x_0 \triangleq 1$)

  $\Rightarrow |w_j|$ is a measure of the importance of the j^{th} feature

Linear Separability

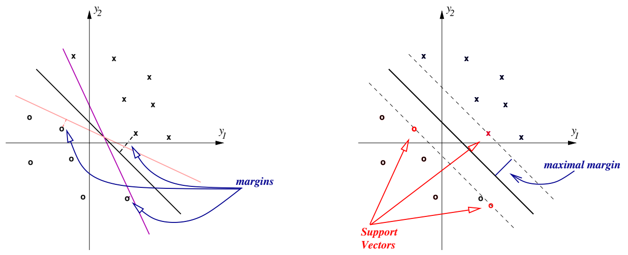


Facts

- The data is linearly separable if the two classes can be perfectly separated by a hyperplane
- A hyperplane in $\mathbb{R}^{10,000}$ can separate perfectly at least $10,001$ (unaligned) points, given any possible 2 class labeling
- There is no problem to find a perfect linear separator of less than 100 points in $\mathbb{R}^{\geq 10,000}$ (*e.g.* for microarray data)
- The problem is that there are many *apparently perfect* models

Linear Support Vector Machines in a nutshell

- When the data is linearly separable the separating hyperplane is not unique but the maximal margin hyperplane separates the data with the largest margin
- For each separating hyperplane, there is an associated set of support vectors



Recursive Feature Elimination [Guyon *et al.* , 02]

Embedded Backward Selection

1. Estimate a SVM on a given set of dimensions (initially p dimensions)
   - Decision rule: sign $(\sum_{j=1}^{p} w_j x_j + w_0)$
2. Consider $|w_j|$ as the relevance of the j^{th} dimension
3. Remove the least relevant dimension(s)
4. Iterate 1 to 3 on a reduced feature set

General Summary

- Feature selection aims at reducing the dimensionality of the data while preserving the interpretation of the original features
- **Filters** methods use only the data + class labels:
  - simple, fast, generally univariate (often an implicit use of a classifier)
- **Wrappers** take the performance of the classifier into account
  - Multivariate as soon as the classifier is multivariate
  - Often computing intensive
- **Embedded methods** take the structure of the classifier into account
  - More elegant and often faster than wrappers, not always better in terms of performance
  - A way to get an insight into a black-box classifier

Further information

- LINGI2262 Machine Learning: classification and evaluation (Semester 2)
- LELEC2870 Machine Learning: regression, dimensionality reduction and data visualization (Semester 1)
- LINGI2369 Artificial Intelligence and Machine Learning Seminar (Semester 1)

Further Reading I

Guyon, I., Gunn, S., Nikarvesh, M. and Zadeh, L.A. (editors) (2006).
Feature Extraction: Foundations and Applications.
Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009).
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
(2nd edition), Springer.

Abeel, T., Helleputte, T., Dupont, P. and Saeys, Y. (2010)
Robust biomarker identification for cancer diagnosis with ensemble feature selection methods
Bioinformatics, Vol. 26 (3), pp. 392-398.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. (2003)
A comparison of normalization methods for high density oligonucleotide array data based on variance and bias
Bioinformatics, Vol. 19 (2), pp. 185-193.

Further Reading II

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002).
Gene selection for cancer classification using support vector machines.
Machine learning, **46**, 389–422.

Peng, H., Long, F., and Ding, C. (2005).
Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.
IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, N° 8, pp. 1226-1238