Gene expression signature for cancer classification

Junior Barrera

BIOINFO / IME-USP

Team

Hugo A. Armelin

Junior Barrera

Helena Brentaini

Marcel Brun

Y. Chen

Edward R. Dougherty

Roberto M. Cesar Jr.

Daniel Dantas

Gustavo Esteves

Marco D. Gubitoso

Nina S. T. Hirata

Roberto Hirata Jr

Luiz F. Reis

Paulo S. Silva

Sandro de Souza

Walter Trepode

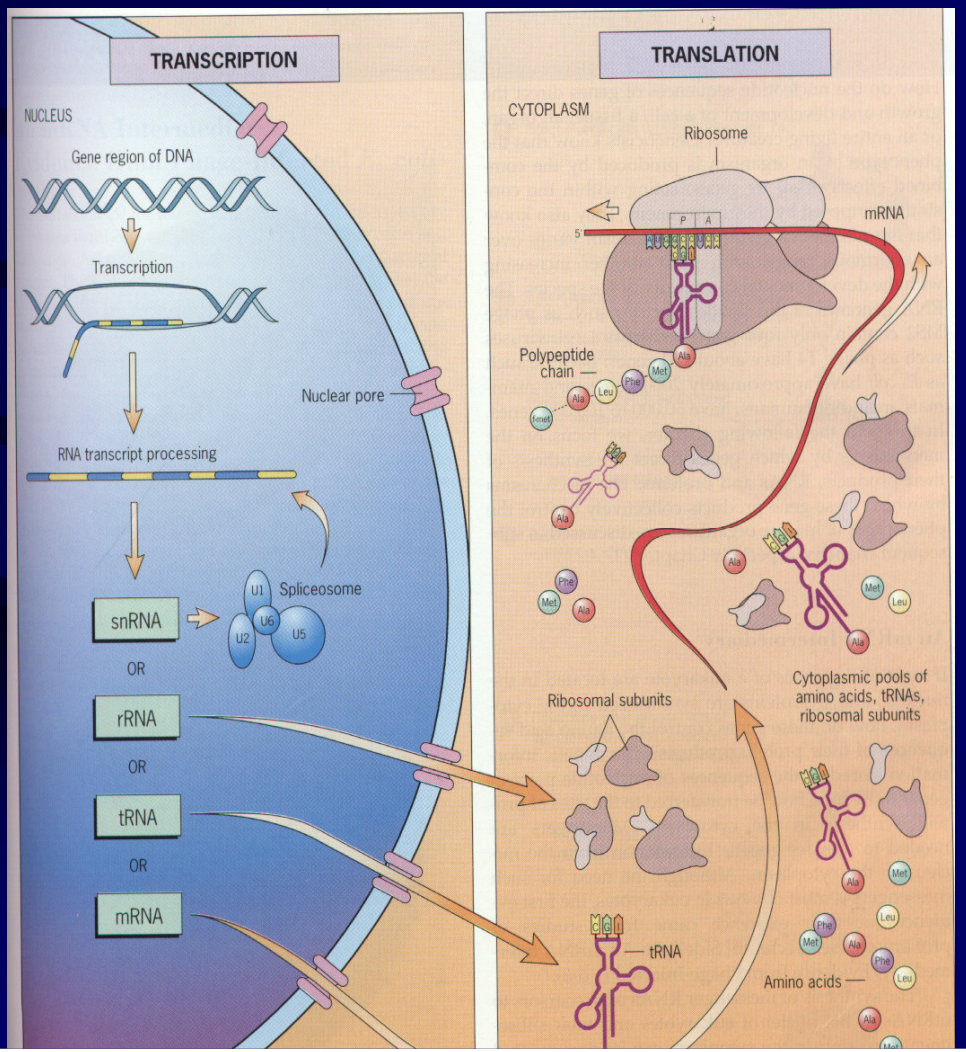
....

Layout

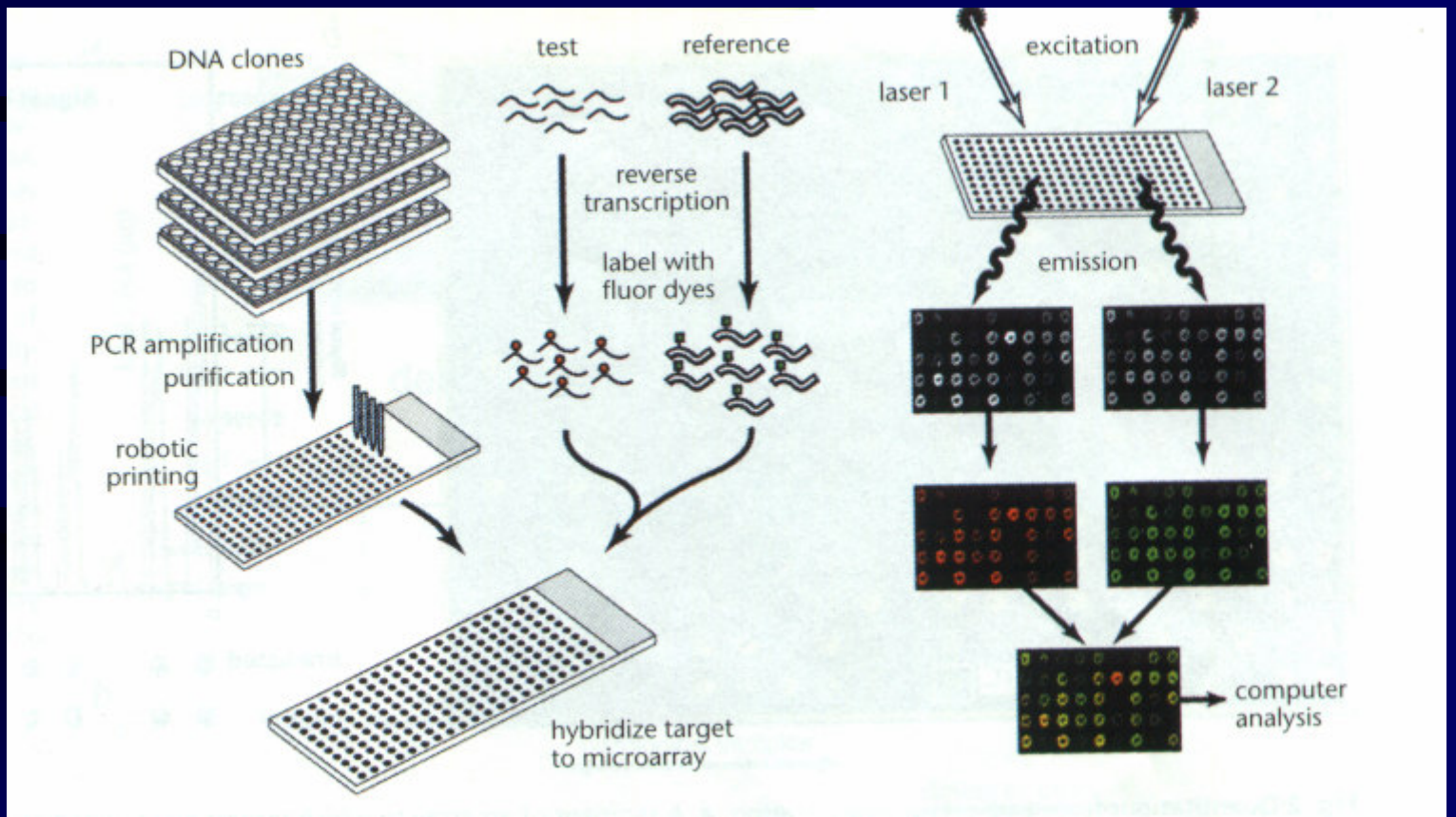
- Introduction
- Classifiers: concept and design
- Dimensionality reduction
- Mean conditional entropy
- Strong features : concept and algorithms
- Validation

Introduction

Gene Expression



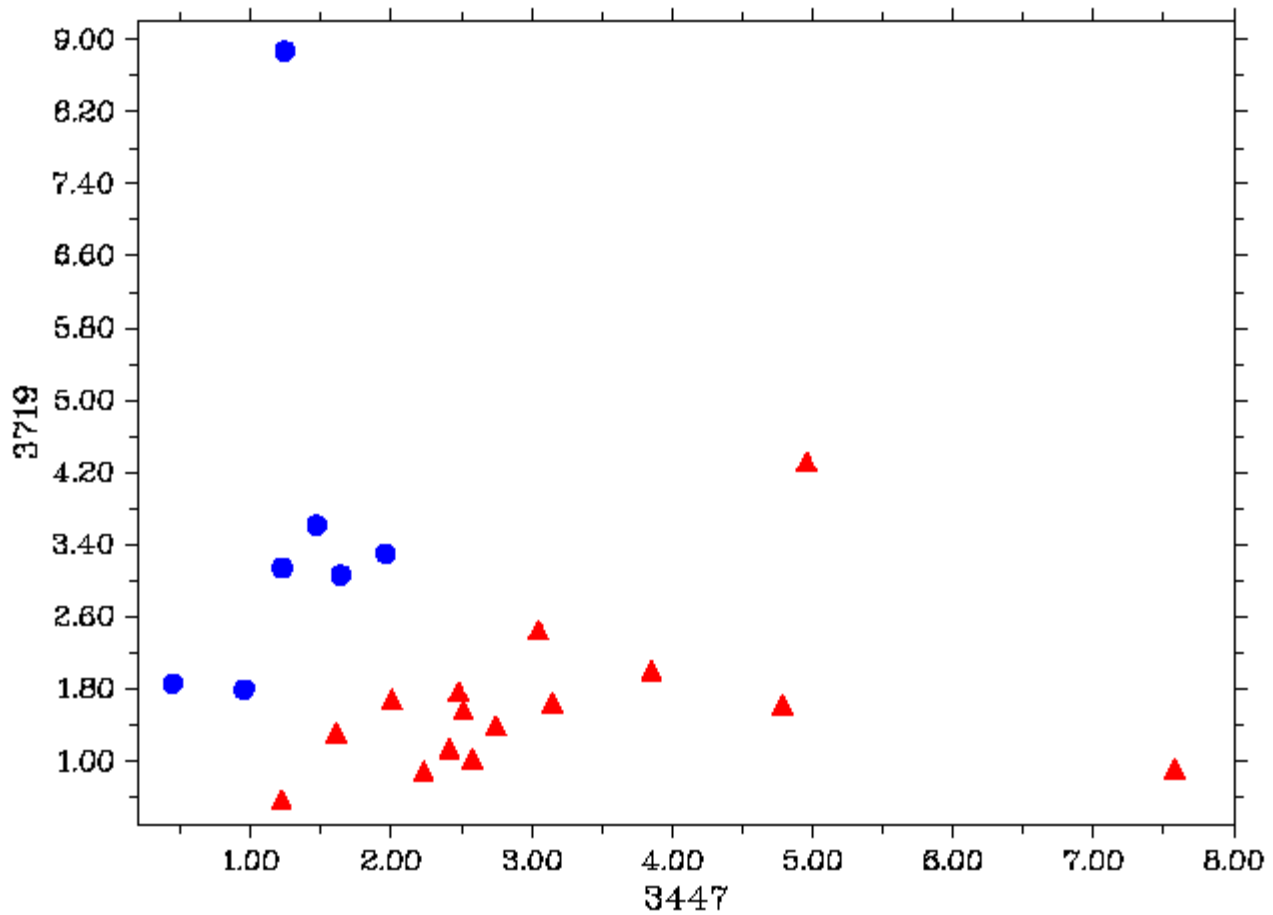
Expression measure



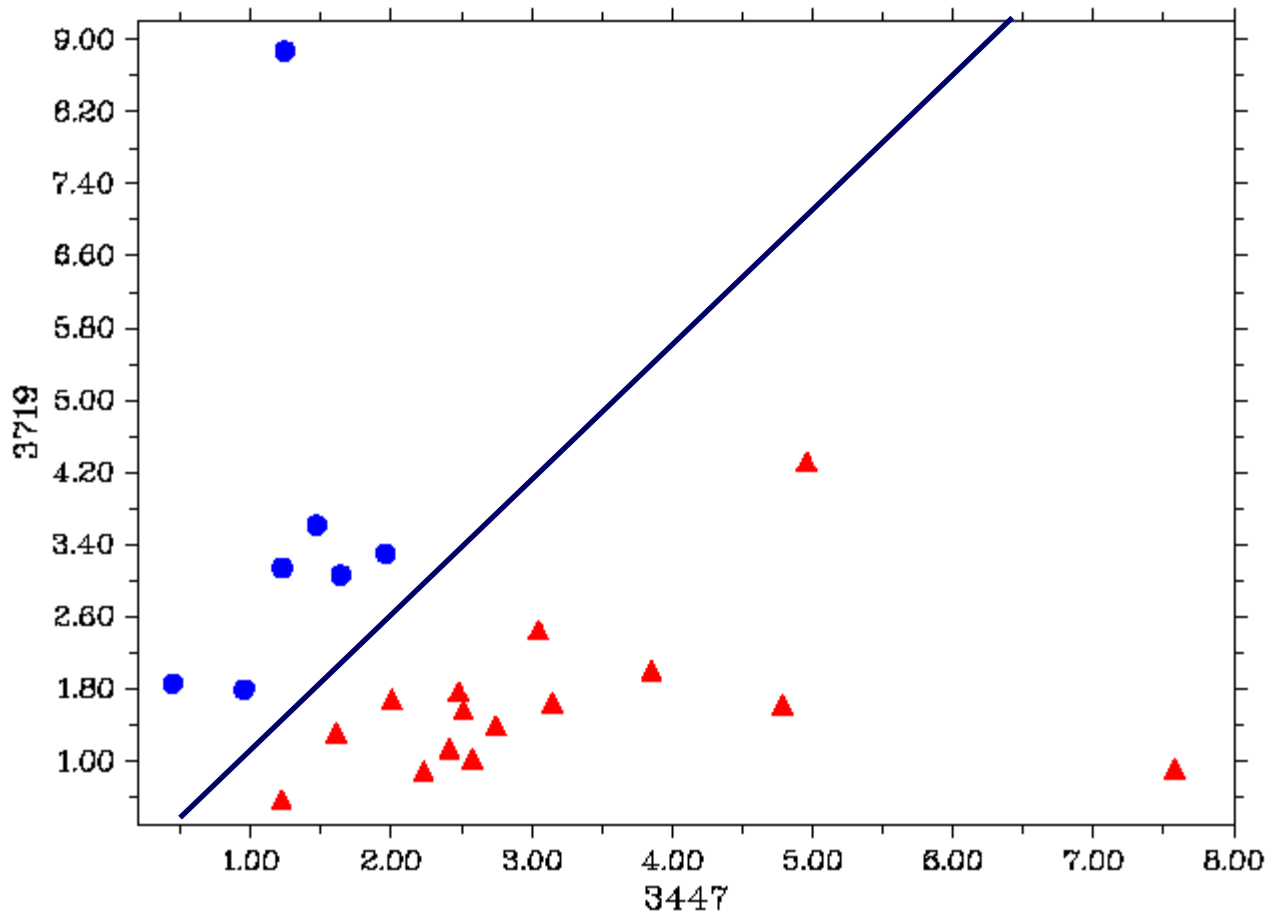
Goal: find a small subset of genes whose expression values are enough to recognize two or more cancer types.

Classifiers: concept and design

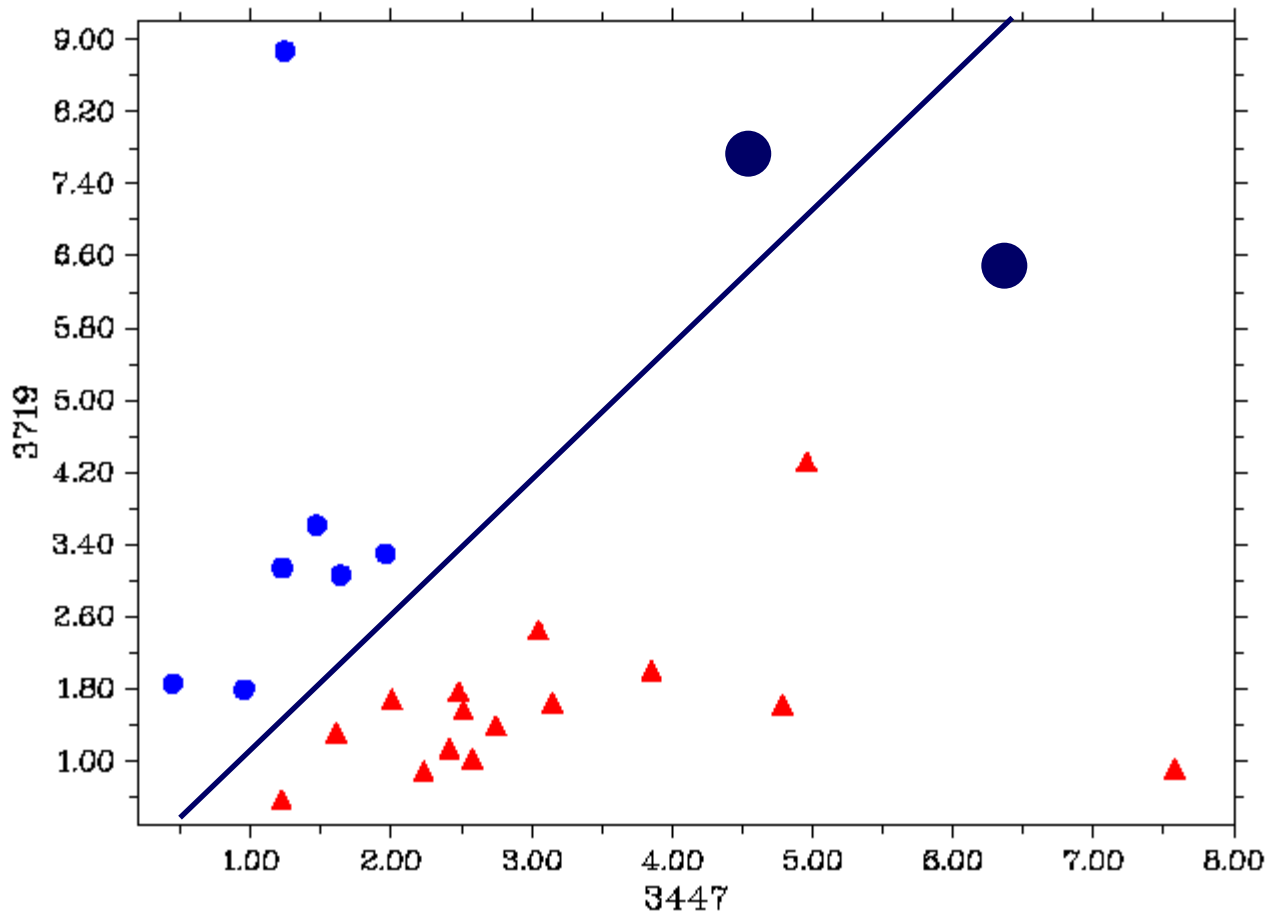
Classifiers



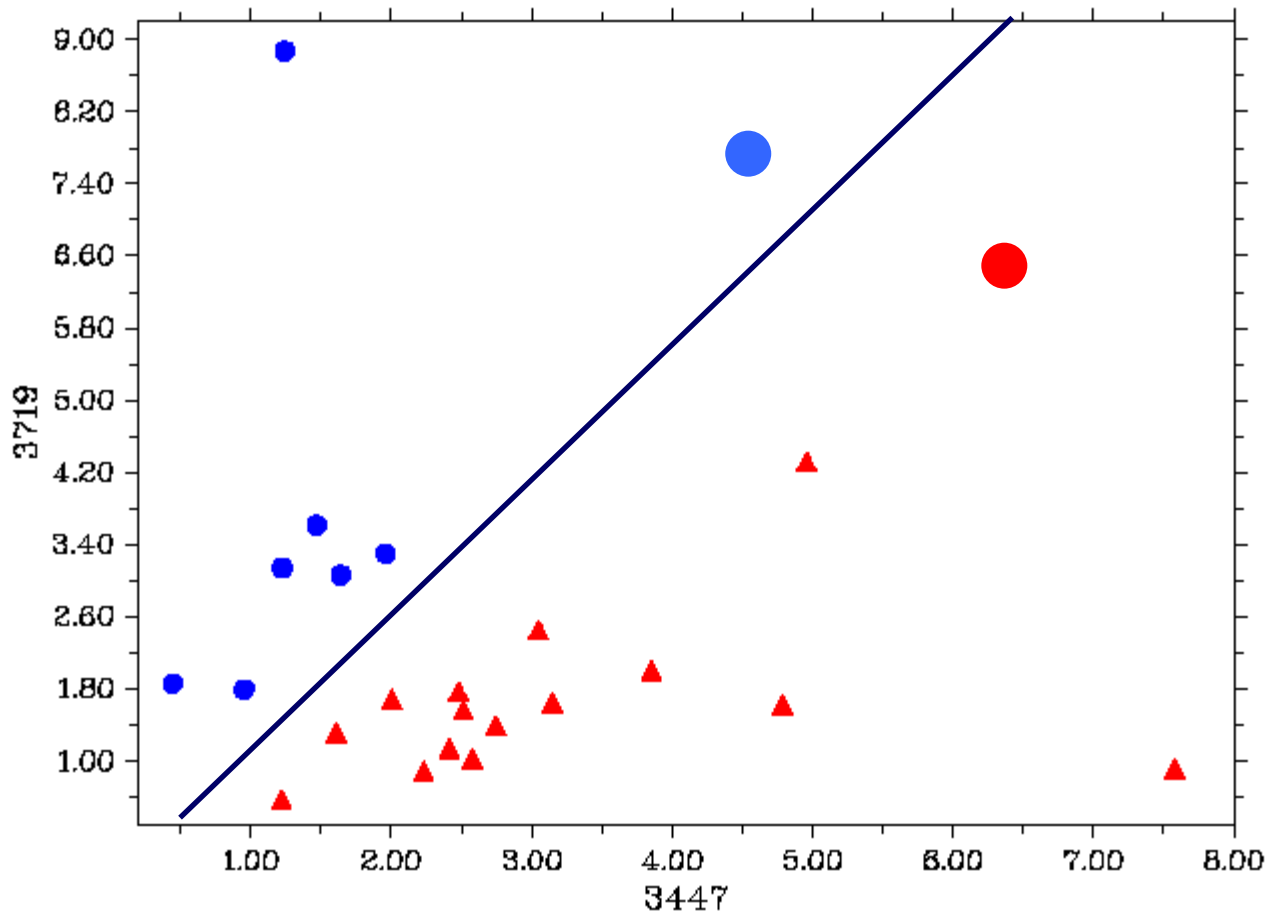
Classifiers



Classifiers



Classifiers



Classifier Design

- Design goal is to find a function with minimum risk.
- Risk (expected loss) of a function :

$$R(\psi) = E[l(\psi(X), Y)]$$

X is a random set
Y is a binary random variable

- Loss function

$$l : \{0,1\} \times \{0,1\} \rightarrow R^{+}$$

MAE example

- *Example* : MAE loss function

$$l_{MAE}(a,b) = |a-b| \qquad a,b \in \{0,1\}$$

$$MAE\langle\Psi\rangle = E[|\psi(X) - Y|]$$

- Optimal MAE function

$$\psi(X) = \begin{cases} 1 & p(1,X) > p(0,X) \\ 0 & p(1,X) \leq p(0,X) \end{cases}$$

PAC learning

L is Probably Approximately Correct (PAC)

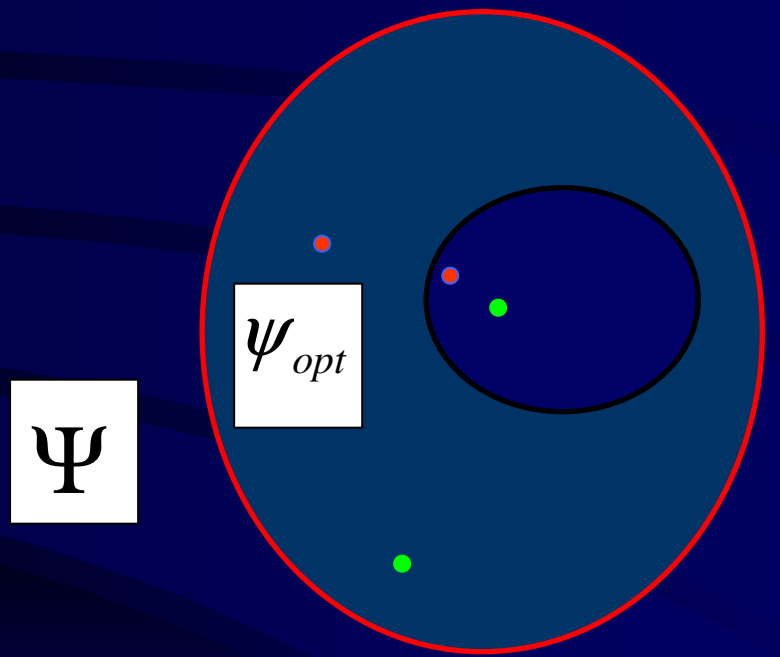
For $m > m(\varepsilon, \delta)$ examples

$$\Pr(|R(\psi) - R(\psi_{opt})| < \varepsilon) > 1 - \delta$$

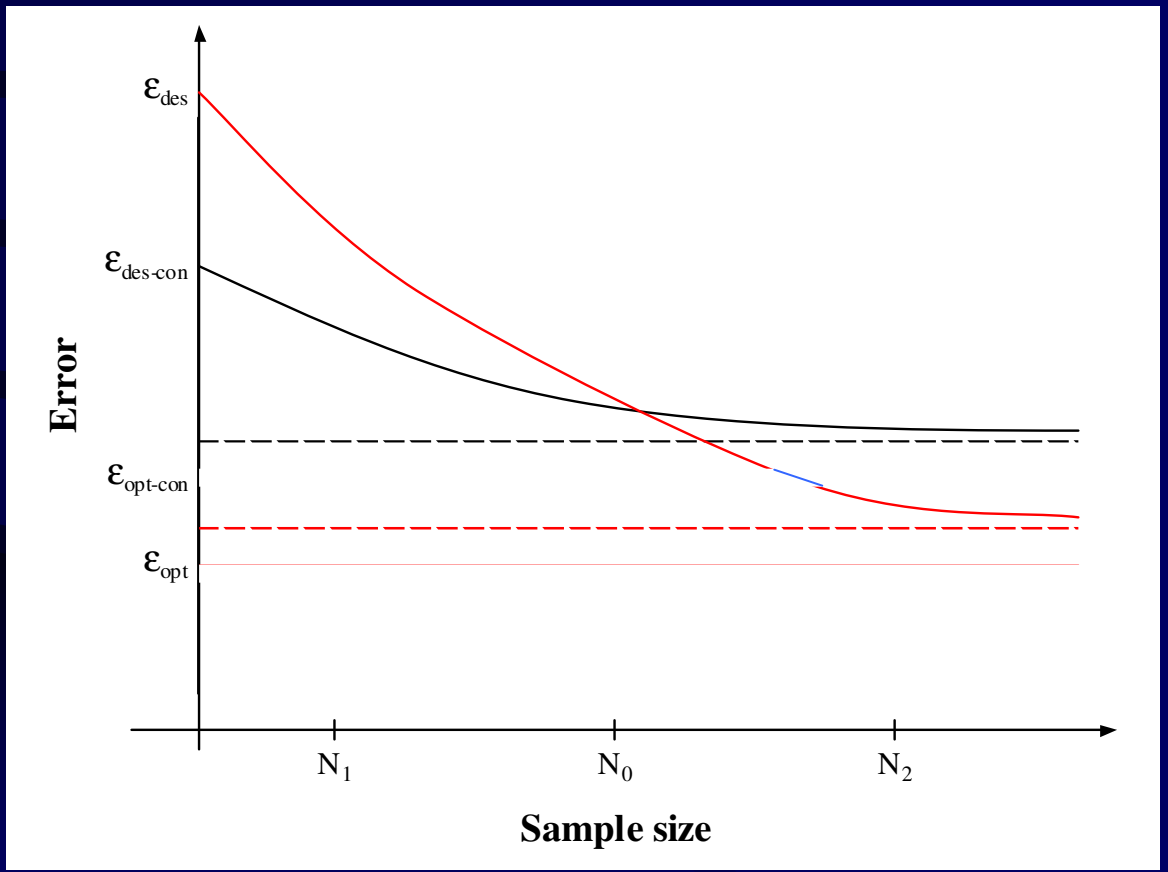
$$\varepsilon, \delta \in (0,1)$$

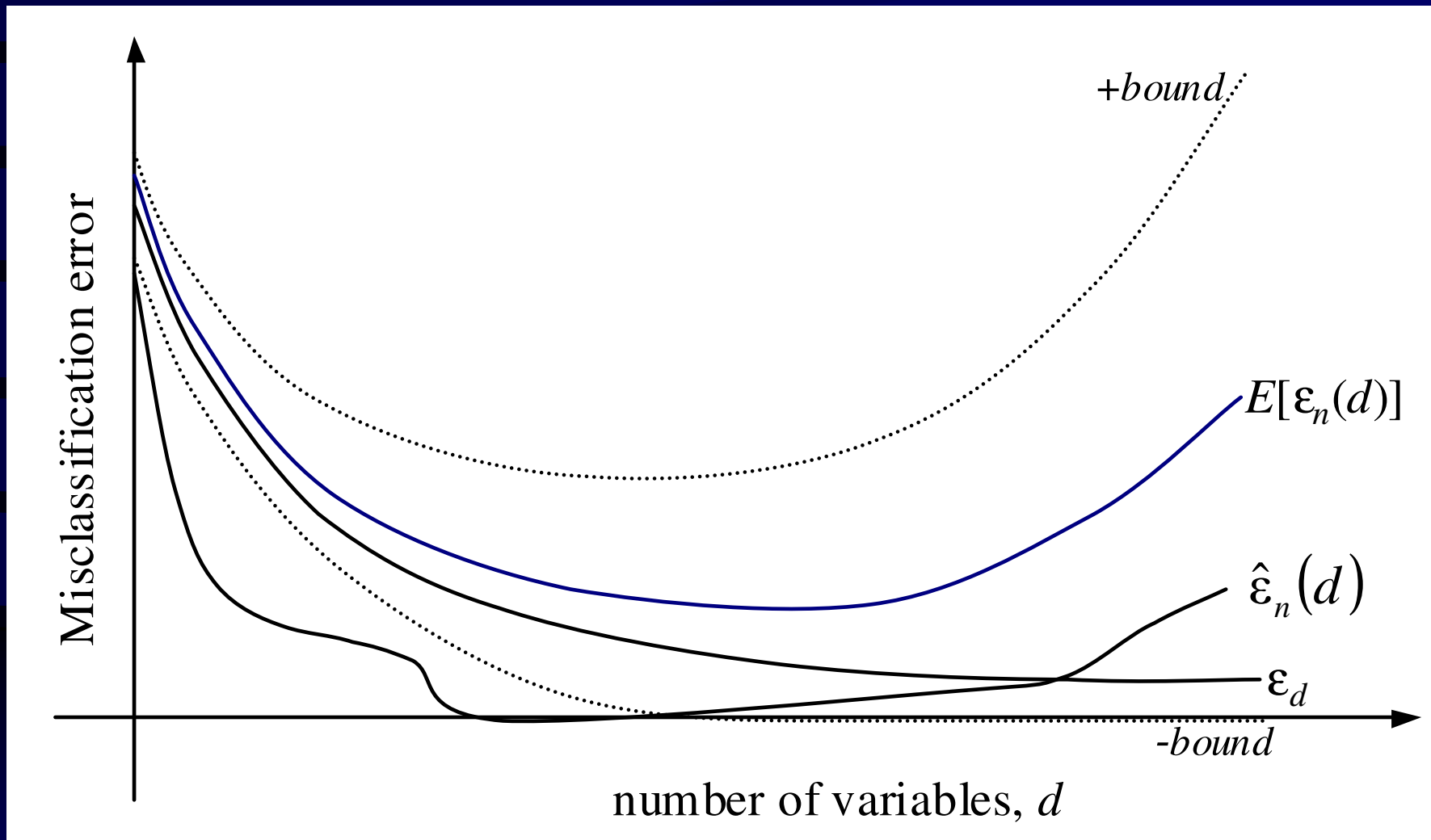
Dimensionality Reduction

Constraints



Constraints




+bound
Misclassification error
$E[\varepsilon_n(d)]$
$\hat{\varepsilon}_n(d)$
ε_d
-bound
number of variables, d

Mean conditional entropy

Entropy

- Distribution measure
- $H(X) = - \sum p(x) \log p(x)$
- decreases when the probability mass is more concentrated.
- Maximum for uniform distribution
- Invariant to redistribution of the probability mass, keeping the same proportion

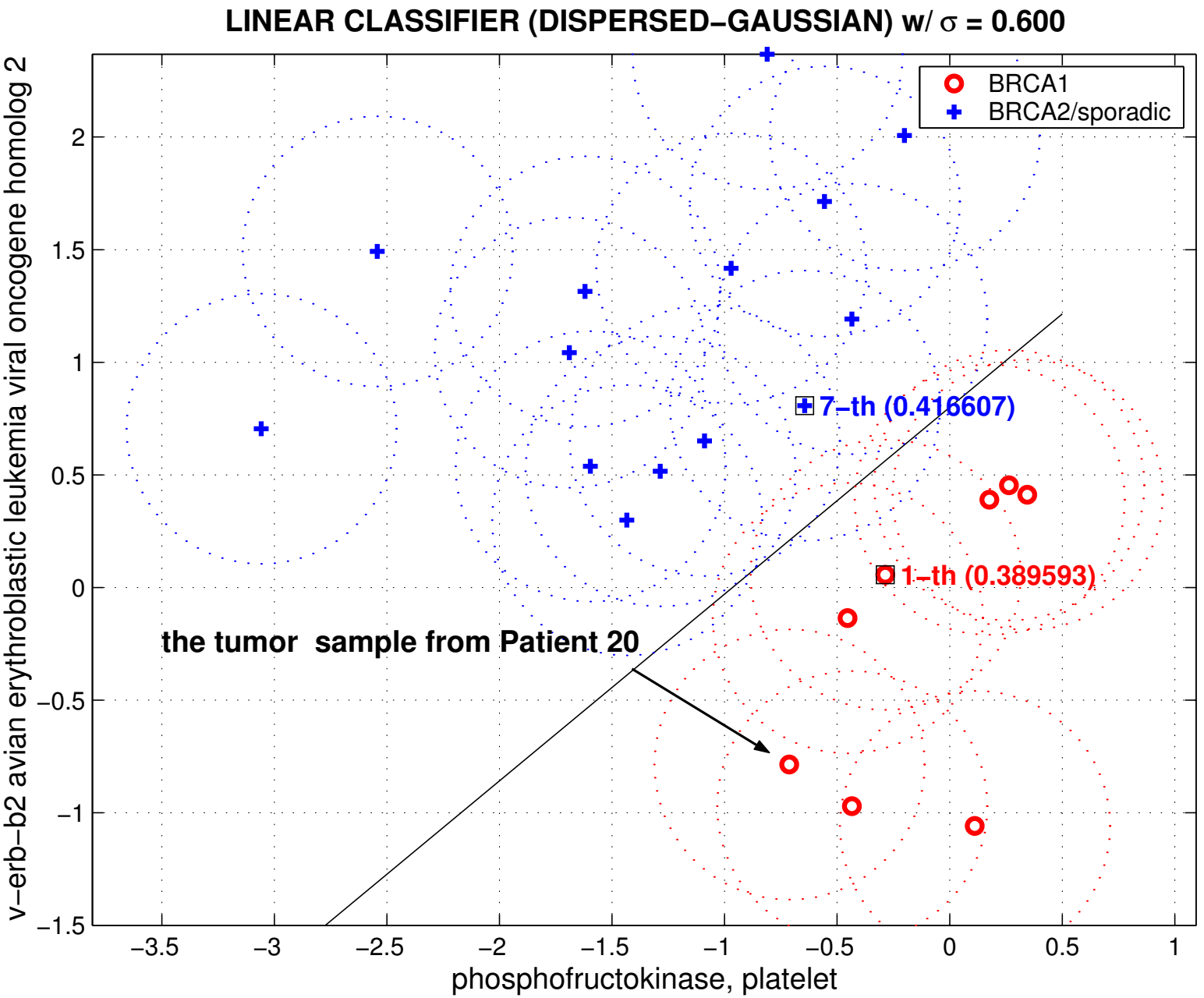
Expected Conditional Entropy

- $E[H(Y/X)] = \sum - p(x) \sum p(Y/x) \log p(Y/x)$
- When $E[H(Y/X)]$ is smaller, the pattern recognition problem is simpler

Finding the best features

- For each subset of features, p(x) and p(y/x) may be estimated from data. The best set of features has the smallest estimated conditional entropy.
- The feature space forms a Boolean lattice, that can be explored exhaustively or partially.

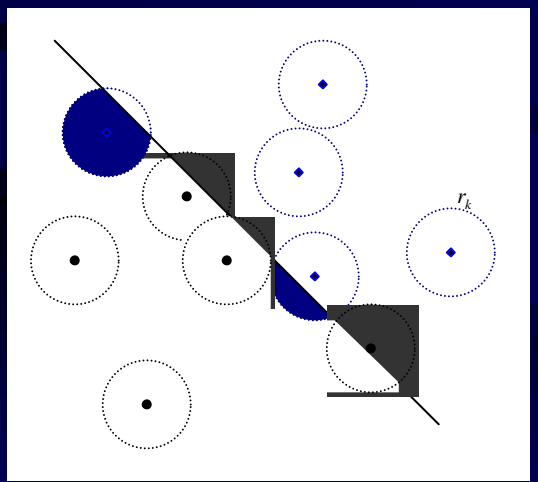
Strong features: concept and algorithms

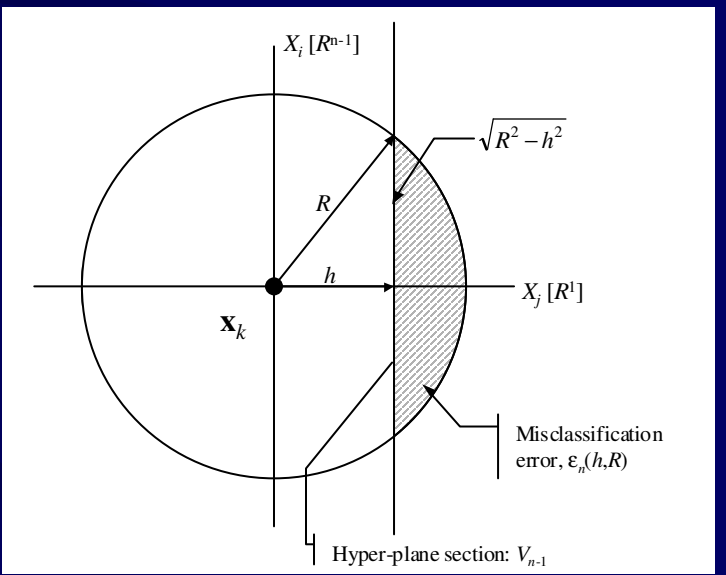

LINEAR CLASSIFIER (DISPERSED–GAUSSIAN) w/ σ = 0.600
BRCA1
BRCA2/sporadic
7–th (0.416607)
1–th (0.389593)
the tumor sample from Patient 20
v–erb–b2 avian erythroblastic leukemia viral oncogene homolog 2
phosphofructokinase, platelet
2
1.5
1
0.5
0
–0.5
–1
–1.5
–3.5
–3
–2.5
–2
–1.5
–1
–0.5
0
0.5
1

Approach

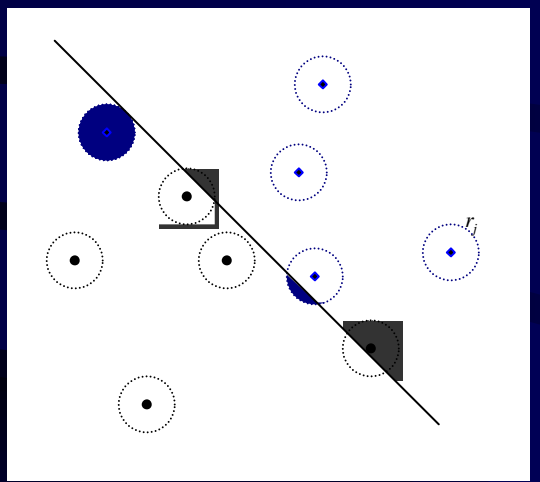
randomize data
compute classifier using genes subsets
measure error for different dispersions
choose the subset that balance small error and high dispersion.
A supercomputer is required.

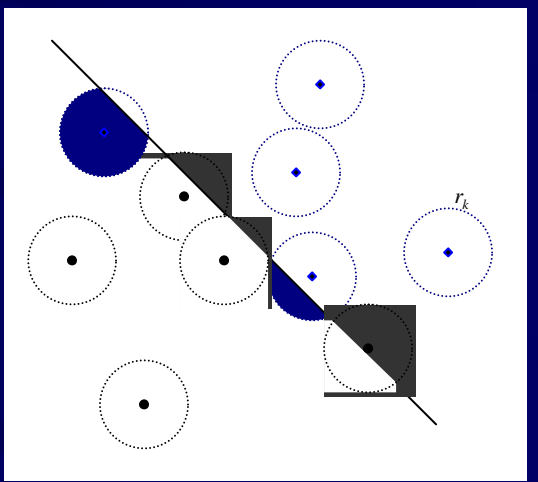
Linear classifier
Dispersion centered in the sample
Flat round dispersion model
Error computed analytically (faster)

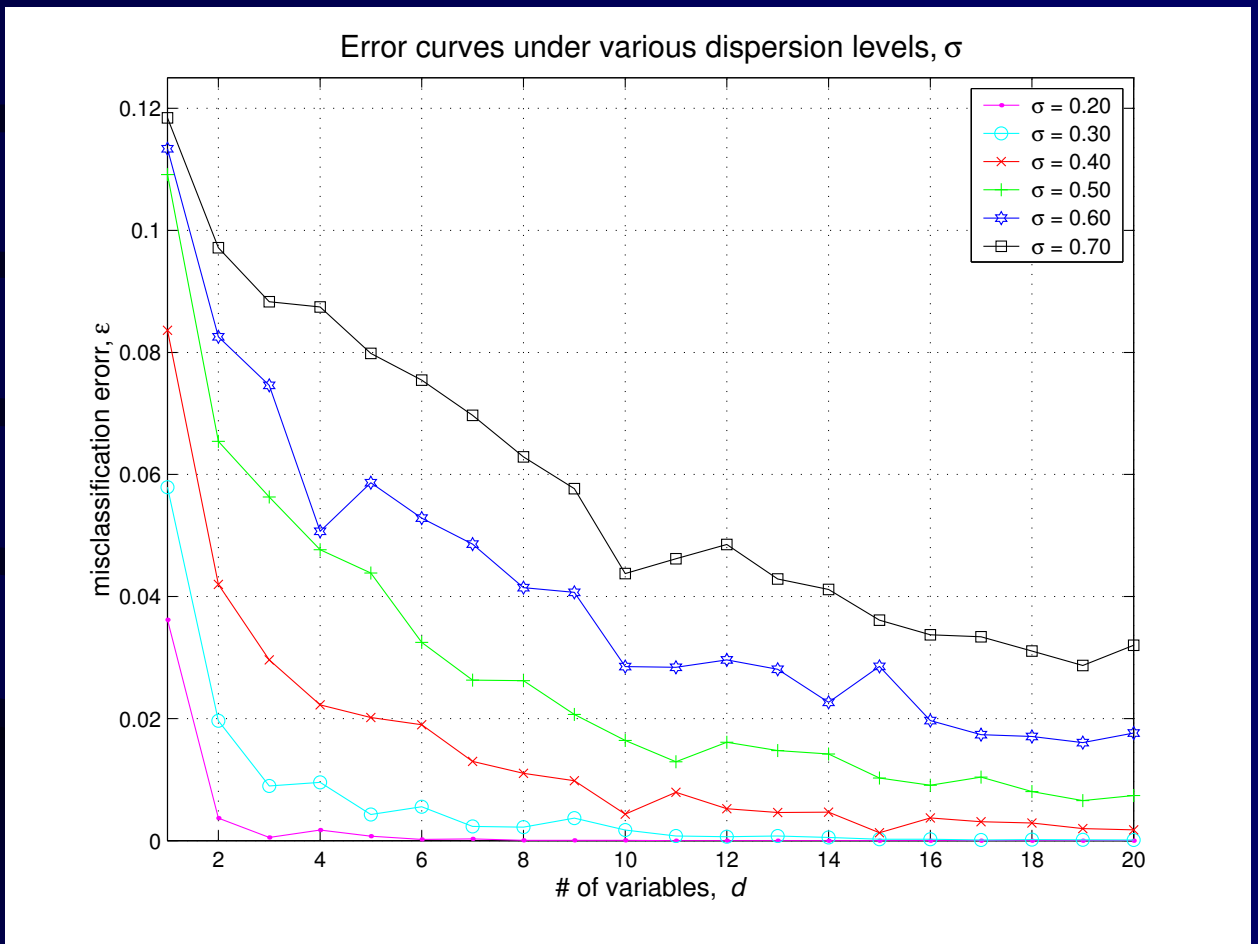




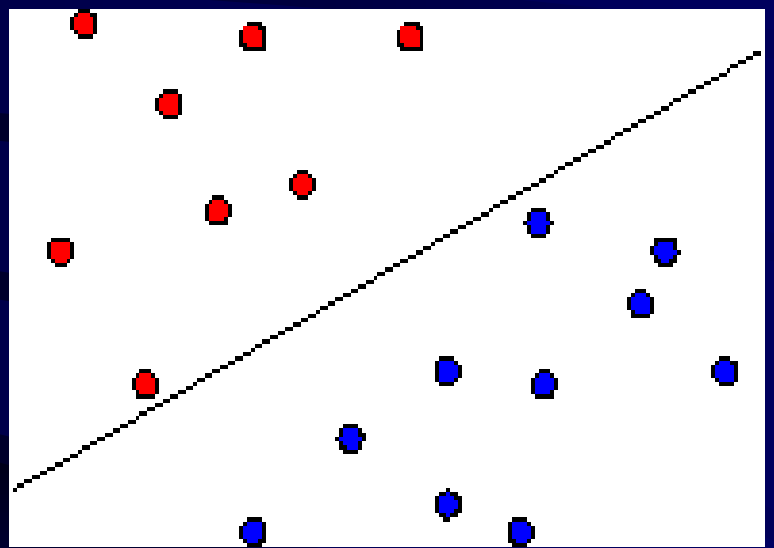
Robustness analysis

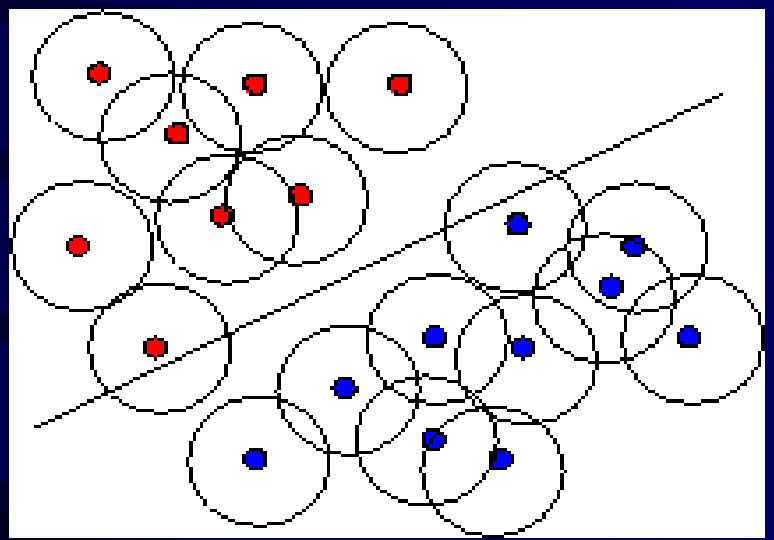


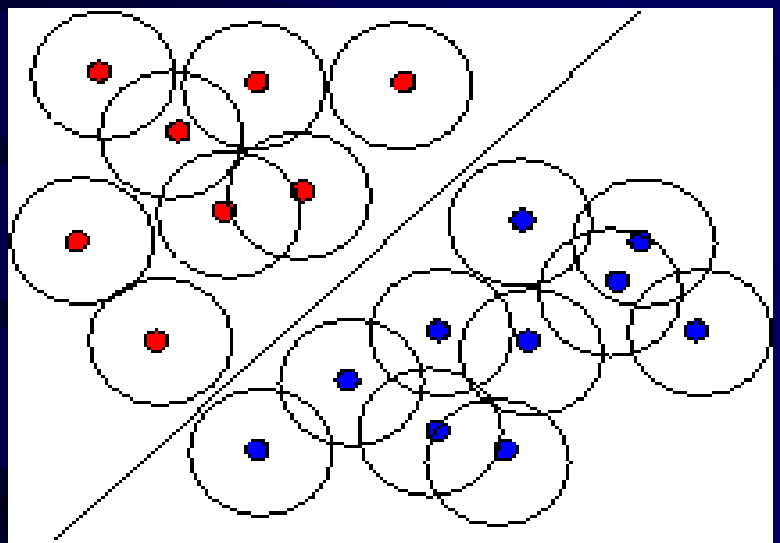



Error curves under various dispersion levels, σ
misclassification erorr, ε
# of variables, d
σ = 0.20
σ = 0.30
σ = 0.40
σ = 0.50
σ = 0.60
σ = 0.70
0
0.02
0.04
0.06
0.08
0.1
0.12
2
4
6
8
10
12
14
16
18
20

Algorithm based on linear programming







Steps

- The best linear classifier uses about 20-25 genes
- Genes used are eliminated and the best linear classifier is computed, more 20-25 genes are separated
- The procedure is repeated till having about 100 genes
- The full search is applied in the selected subset of genes

Validation

- Expression of chosen subsets of genes are measured several times in low cost experiments
- If the experiments reveal compact clusters the subset of genes chosen should be correct.

Classifiers