Bioinformatics Applications and Feature Selection for SVMs

S. Mukherjee

Class 23, 2001

Outline

I. Basic Molecular biology
II. Some Bioinformatics problems
III. Microarray technology
    a. Purpose
    b. cDNA and Oligonucleotide arrays
    c. Yeast experiment
IV. Cancer classification using SVMs
V. Rejects and Confidence of classification
VI. Feature Selection for SVMs
    a. Leave-one-out bounds
    b. The algorithm
VII Results on several datasets

What is Bioinformatics

Pre 1995

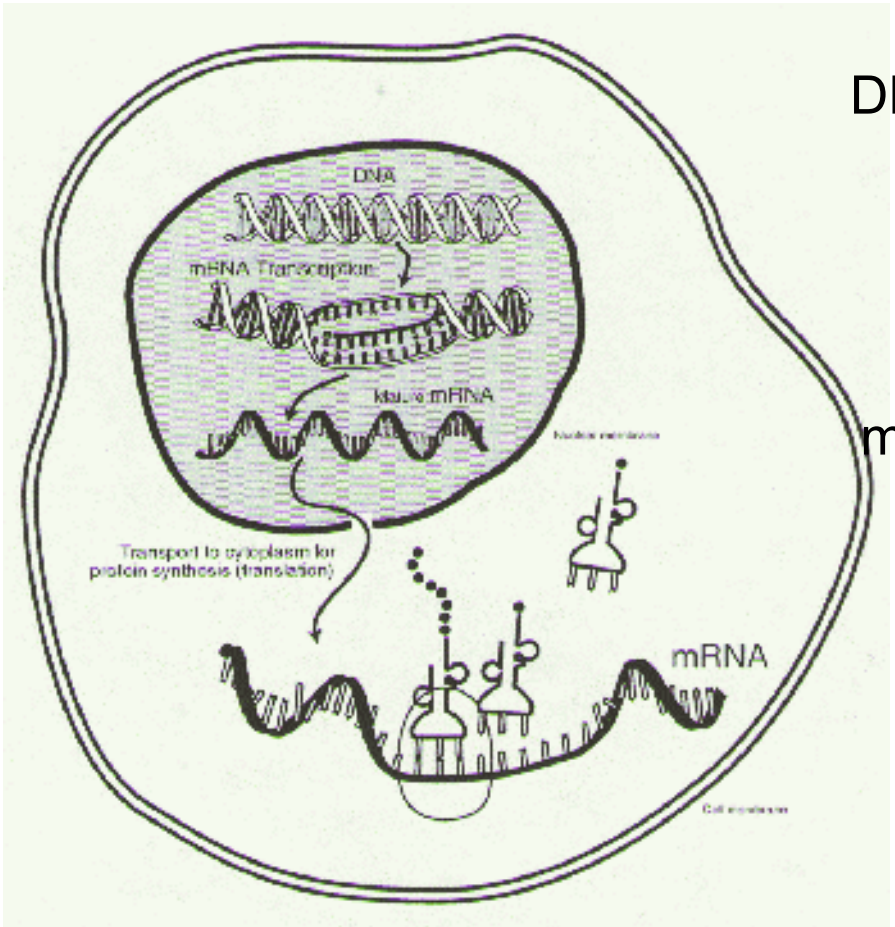
Application of computing technology to providing statistical and database solutions to problems in molecular biology.

Post 1995

Defining and addressing problems in molecular biology using methodologies from statistics and computer science.

The genome project, genome wide analysis/screening of disease, genetic regulatory networks, analysis of expression data.

Some Basic Molecular Biology



DNA: CGAACAAACCTCGAACCTGCT

Transcription

mRNA: GCU UGU UUA CGA

Translation

Polypeptide: Ala Cys Leu Arg

Examples of Problems

Gene sequence problems: Given a DNA sequence state which sections are coding or noncoding regions. Which sections are promoters etc...

Protein Structure problems: Given a DNA or amino acid sequence state what structure the resulting protein takes.

Gene expression problems: Given DNA/gene microarray expression data infer either clinical or biological class labels or genetic machinery that gives rise to the expression data.

Protein expression problems: Study expression of proteins and their function.

Microarray Technology

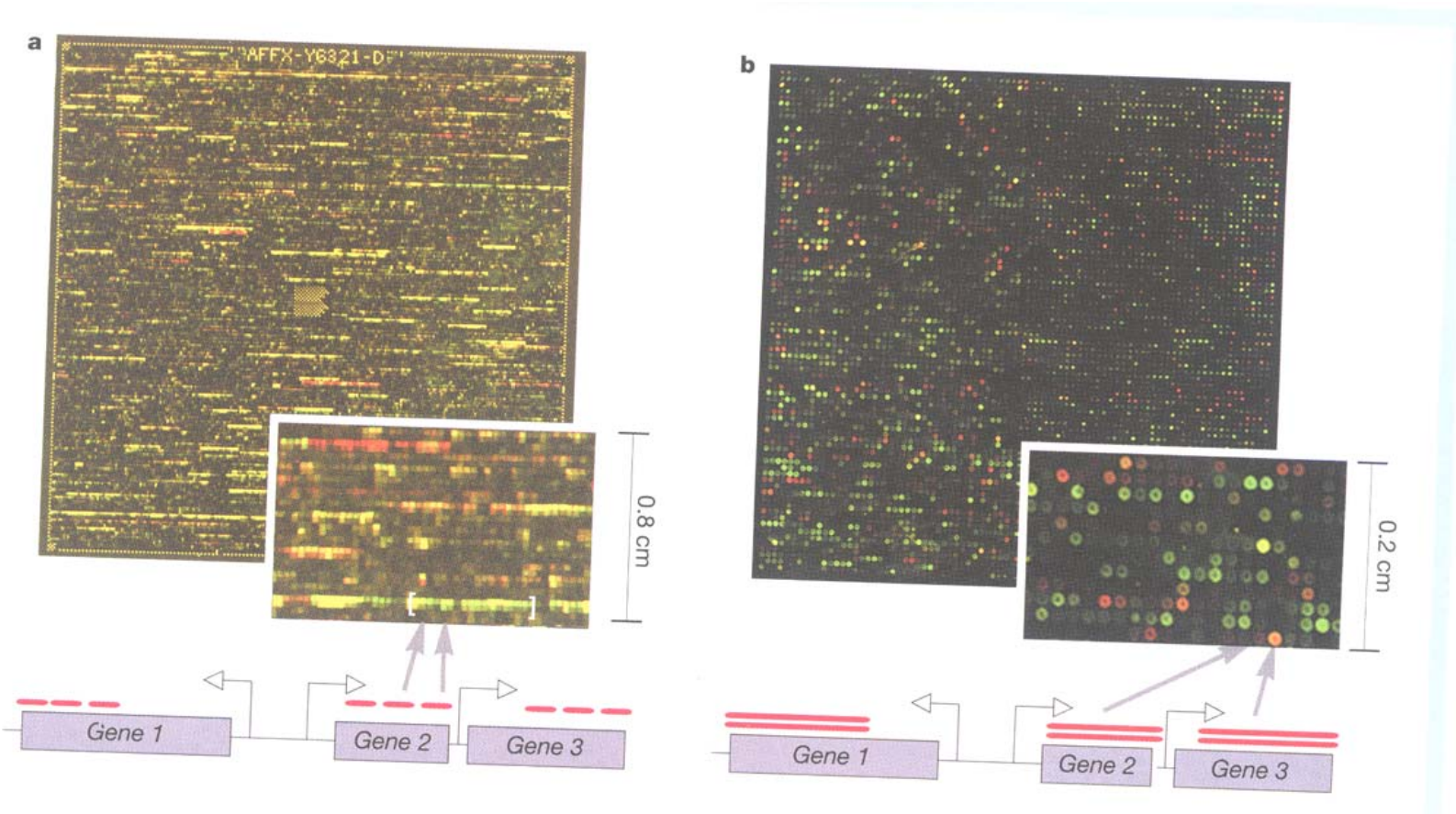
Basic idea:

The state of the cell is determined by proteins.
A gene codes for a protein which is assembled via mRNA.
Measuring amount particular mRNA gives measure of amount of corresponding protein.
Copies of mRNA is expression of a gene.

Microarray technology allows us to measure the expression of thousands of genes at once.

Measure the expression of thousands of genes under different experimental conditions and ask what is different and why.

Oligo vs cDNA arrays



Lockhart and Winzler 2000

A DNA Microarray Experiment

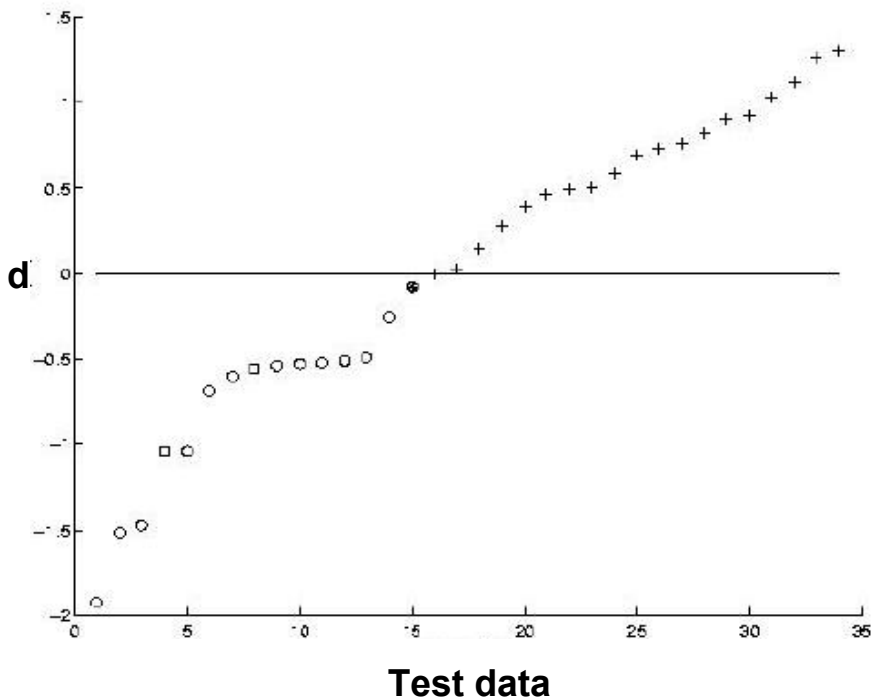
Cancer Classification

38 examples of Myeloid and Lymphoblastic leukemias
Affymetrix human 6800, (7128 genes including control genes)

34 examples to test classifier

Results: 33/34 correct

d perpendicular distance from hyperplane



Gene expression and Coregulation

This model considers the gene expressions and coregulation of Sonic Hedgehog and TrkC so $x = \{e_{sh}, e_{Trk}\}$.
$\phi(x) \rightarrow \{e_{sh}^2, e_{Trk}^2, e_{sh}e_{Trk}, e_{sh}, e_{Trk}, 1\}$.

Construct a linear classifier for $\phi(x)$:
$f(x) = sign(w \cdot \phi(x) + b)$.

A linear classifier is constructed in this high dimensional space called *feature space*.

We can now apply the same equations that we used for the linear SVM except we replace x with $\phi(x)$.

A linear classifier in the feature space is a quadratic classifier in the original space.

Nonlinear classifier

The feature space can be huge, consider previous model but with 7000 genes, $\phi(\boldsymbol{x})$ has about 24 million features.

The feature space never has to be explictly computed however. In the SVM only dot products need to be taken: $\phi(x) \cdot \phi(\boldsymbol{y})$.

For the above model

$$\phi(x) \cdot \phi(y) = (x \cdot y + 1)^2 = K(x, \boldsymbol{y}).$$

K(x,y) is called a kernel function. Examples

$K(x, \boldsymbol{y}) = (\boldsymbol{x} \cdot y + 1)^n$

$K(x, \boldsymbol{y}) = \exp(-||x - y||^2/2\sigma^2)$.

Nonlinear SVM

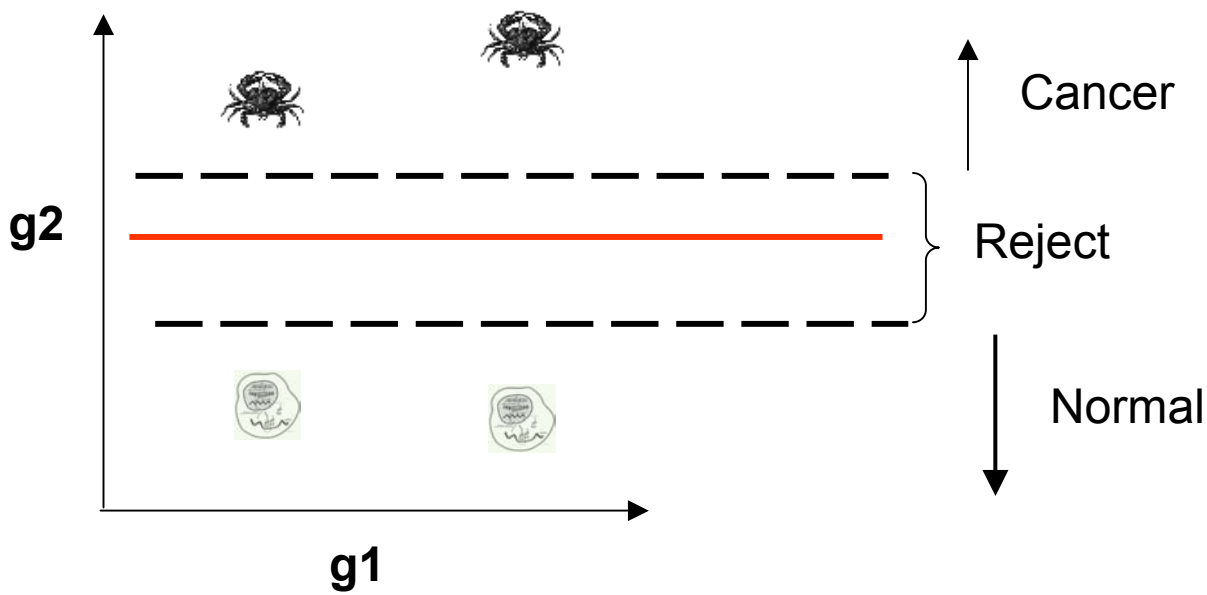
Nonlinear SVM does not help when using all genes but does help when removing top genes, ranked by Signal to Noise (Golub et al).

genes removed	1st order	2nd order	3rd order
10	2	1	1
20	3	2	1
30	3	3	2
40	3	3	2
50	3	2	2
100	3	3	2
200	3	3	3
300	3	4	4
400	4	4	4
500	4	4	4
600	4	5	5
700	3	3	3
800	3	3	3
900	3	4	7
1000	3	5	6
1100	4	6	6
1200	5	6	7
1300	7	8	8
1400	7	7	7
1500	7	7	8

Rejections

Golub et al classified 29 test points correctly, rejected 5 of which 2 were errors using 50 genes

Need to introduce concept of rejects to SVM



Rejections

To get confidences we need to get the conditional probability

$$P(c = \pm 1 | \mathbf{x}).$$

Assume that

$$P(c = 1 | \mathbf{x}) \approx P(c = 1 | \xi \geq d)$$

where d is the perpendicular distance from the optimal hyperplane, $d = f(\mathbf{x})$, and $P(c = 1 | \xi \geq d)$ is the probability that this distance is at least d.

From Bayes law we know that

$$P(c = 1 | d) \propto P(\xi \geq d | c = 1) P(c = 1)$$

assume $P(c = 1) = P(c = -1) = .5$ so we need to estimate $P(\xi \geq d | c = 1)$, this is the problem of estimating a CDF from data.

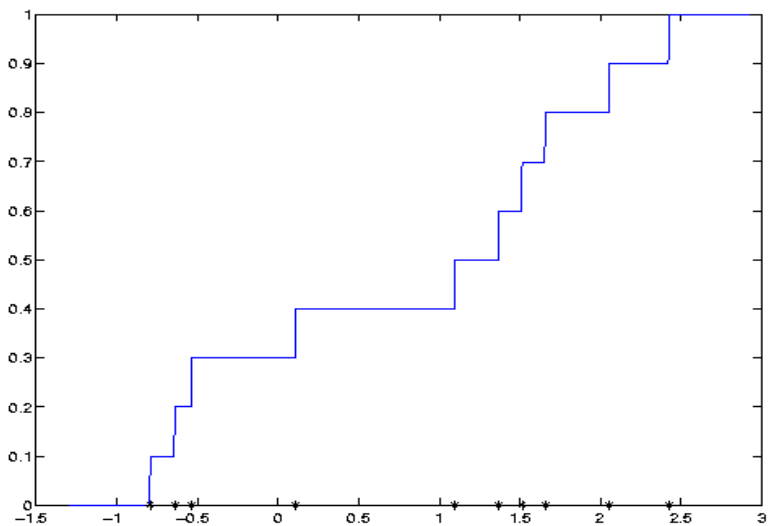
The problem of selecting a d for a cut-off is now basically a statistical test: given a particular confidence value find the d for which it is satisfied

$$C(c = 1 | d) = 1 - P(\xi \geq d | c = 1).$$

Estimating a CDF

Need to get from empirical CDF to estimate of true CDF

$$F_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell}\theta(x - x_i).$$



The Regularized Solution

Solving

$$\min_{p} \sum_{i=1}^{\ell} \left(\int_{-\infty}^{x_i} p(t)dt - F_\ell(x_i) \right)^2 + \gamma_\ell ||p||_K^2$$

results in the solutions

$$\hat{p}(t) = \frac{1}{\ell} \sum_{i=1}^{\ell} K_{\gamma_\ell}(||t - x_i||)$$

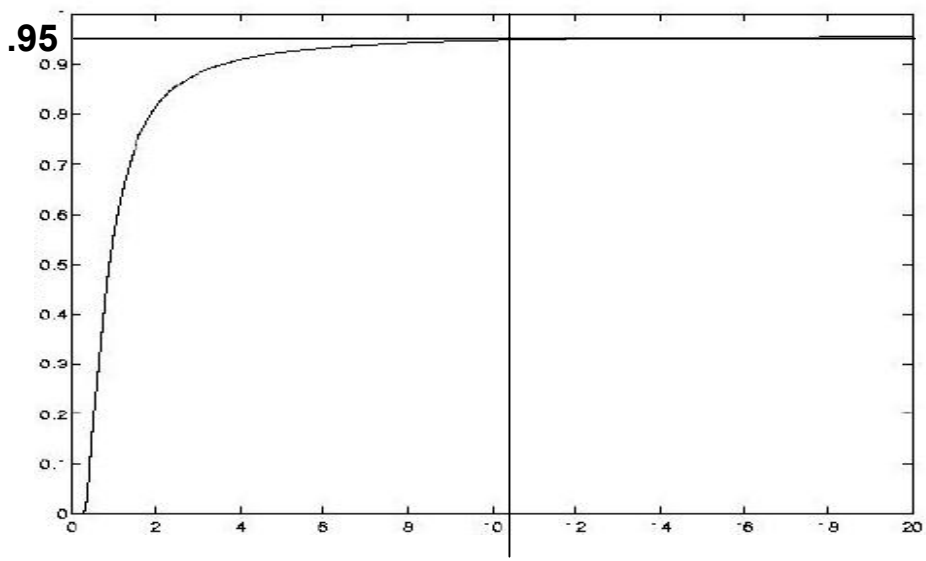
and

$$\hat{F}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \int_{-\infty}^{x} K_{\gamma_\ell}(||t - x_i||)dt.$$

Rejections for SVM

95% confidence or p = .05 d = .107

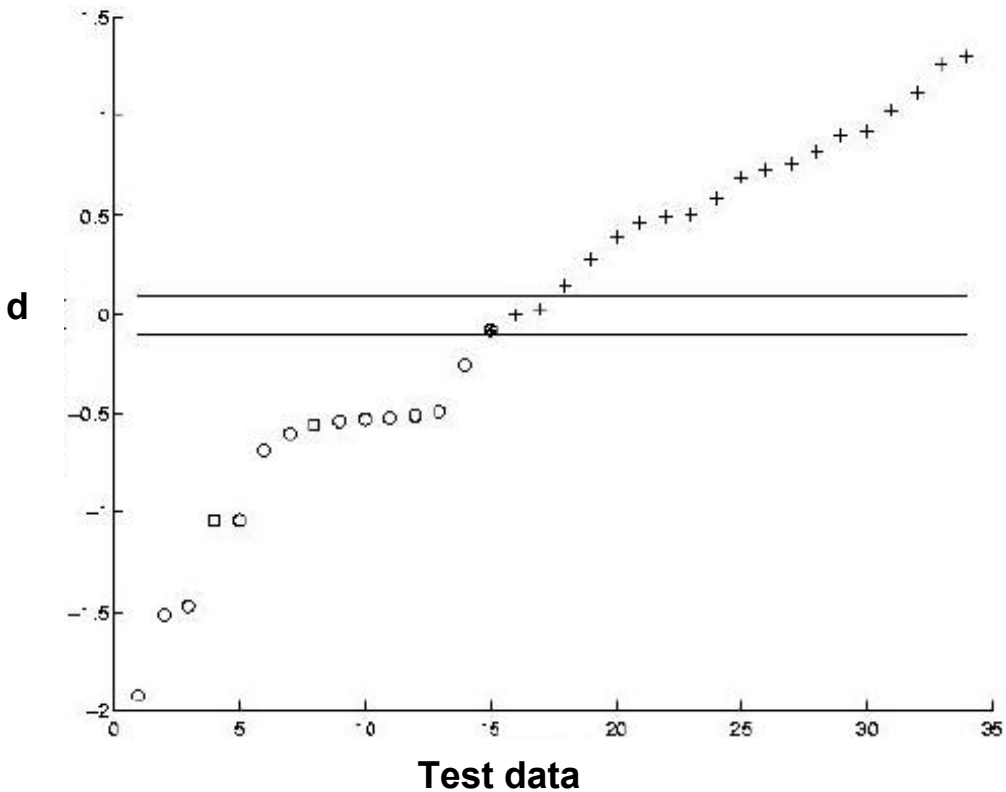
P(c=1 | d)



1/d

Results with rejections

Results: 31 correct, 3 rejected of which 1 is an error



Why Feature Selection

- SVMs as stated use all genes/features

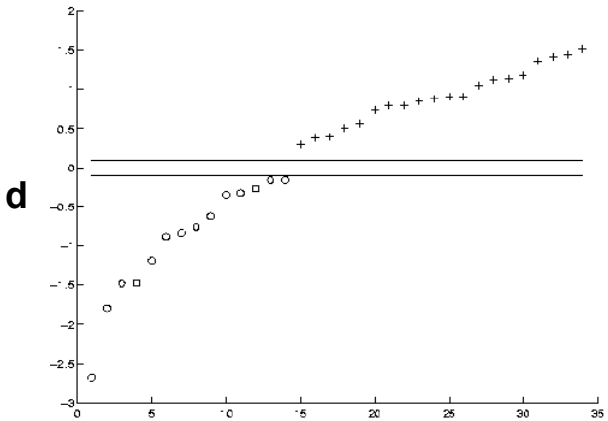
- Molecular biologists/oncologists seem to be conviced that only a small subset of genes are responsible for particular biological properties, so they want which genes are are most important in discriminating

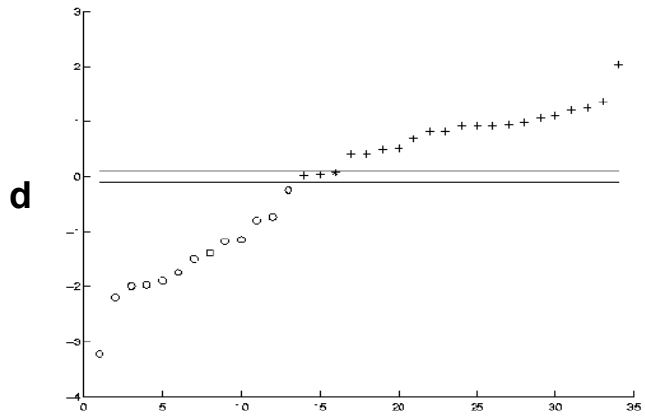
- Practical reasons, a clinical device with thousands of genes is not financially practical

- Possible performance improvement

Results with Gene Selection

AML vs ALL: 40 genes 34/34 correct, 0 rejects.
5 genes 31/31 correct, 3 rejects of which 1 is an error.





B vs T cells for AML: 10 genes 33/33 correct, 0 rejects.

Leave-one-out Procedure

The leave-one-out procedure:
remove one point from the training set, train on the remaining points, and test on the point left out.

This procedure is repeated on all points and

$$\mathcal{L}(D_\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell}\Theta(-y_i f_i(\mathbf{x}_i)).$$

The leave-one-out estimator is almost unbiased:

$$E_{D_\ell}[E_{\mathbf{x},y}[\Theta(-y f_{D_\ell}(\mathbf{x}))]] = E_{D_{\ell+1}}[\mathcal{L}(D_{\ell+1})].$$

The Basic Idea

Use leave-one-out (LOO) bounds for SVMs as a criterion to select features by searching over all possible subsets of n features for the ones that minimizes the bound.

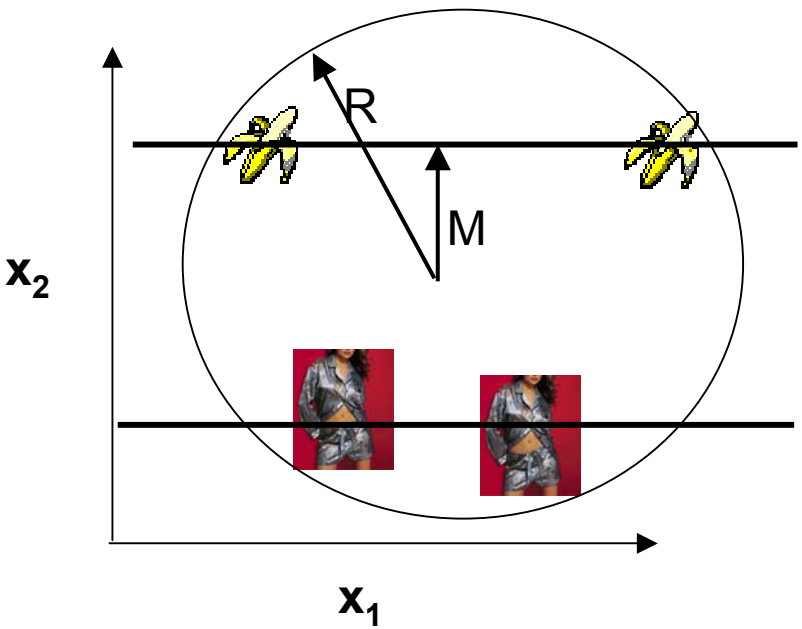
When such a search is impossible because of combinatorial explosion, scale each feature by a real value variable and compute this scaling via gradient descent on the leave-one-out bound. One can then keep the features corresponding to the largest scaling variables.

The rescaling can be done in the input space or in a “Principal Components” space.

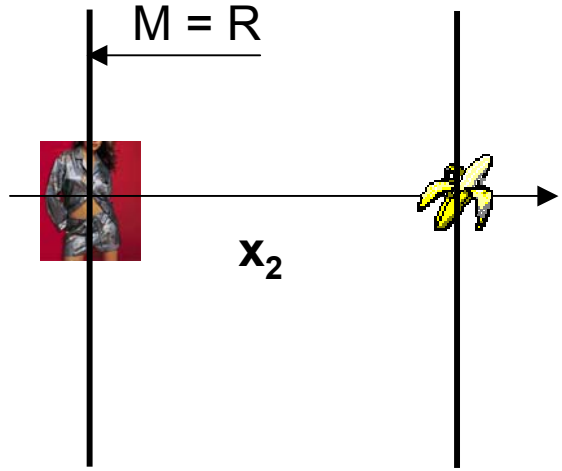
Pictorial Demonstration

Rescale features to minimize the LOO bound R^2/M^2

$R^2/M^2 > 1$



$R^2/M^2 = 1$



SVM Functional

To the SVM classifier we add an extra scaling parameters for feature selection:

$$f(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i K_\sigma(\mathbf{x}_i, \mathbf{x}) + b$$

where $\sigma \in \mathbf{R}^n$ and $K_\sigma(\mathbf{x}, \mathbf{y}) = K(\sigma * \mathbf{x}, \sigma * \mathbf{x})$, where $\mathbf{a} * \mathbf{b}$ indicates element-wise multiplication.

where the parameters α, b are computed by maximizing the the following functional, which is equivalent to maximizing the margin:

$$W^2(\alpha, \sigma) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K_\sigma(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{\ell} y_i \alpha_i = 0$.

Radius Margin Bound

Theorem 1 *if the images of training data D_ℓ belong to a sphere of size R and have margin M, the following bound holds:*

$$\mathcal{L}(D_\ell) \leq T = \frac{1}{\ell}\frac{R_{D_\ell}^2}{M_{D_\ell}^2} = \frac{1}{\ell}R^2(\beta,\sigma)W^2(\alpha,\sigma).$$

This bound is derived as an application of Novikoff's theorem.

R_{D_ℓ} can be computed by maximizing the following

$$R(\beta,\sigma) = \sum_{i=1}^{\ell}\beta_i K_\sigma(\mathbf{x}_i,\mathbf{x}_i) - \sum_{i,j=1}^{\ell}\beta_i\beta_j K_\sigma(\mathbf{x}_i,\mathbf{x}_j)$$

subject to

$$\sum_{i=1}^{\ell}\beta_i = 1 \qquad 0 \leq \beta.$$

Jaakkola-Haussler Bound

Theorem 2 *For kernel machines without the bias term:*

$$\mathcal{L}(D_\ell) \leq T = \frac{1}{\ell} \sum_{i=1}^{\ell} \Theta \left(\alpha_i K(\mathbf{x}_i, \mathbf{x}_i) - 1 \right).$$

This bound is based upon the following inequality

$$y_i(f(\mathbf{x}_i) - f_i(\mathbf{x}_i)) \leq \alpha_i K(\mathbf{x}_i, \mathbf{x}_i).$$

Span Bound

Theorem 3 *Under the assumption the set of support vectors does not change when removing example i, the following bound holds:*

$$\mathcal{L}(D_\ell) \leq T = \frac{1}{\ell} \sum_{i=1}^{\ell} \Theta\left(\alpha_i S_i^2 - 1\right),$$

where S_i is the distance between point $\phi(\mathbf{x}_i)$ and the set Λ_i where

$$\Lambda_i = \left\{ \sum_{j/i, \alpha_j > 0} \lambda_j \phi(\mathbf{x}_j), \quad \sum_{j/i} \lambda_j = 1 \right\}.$$

This bound is derived in a very similar fashion as the Jaakkola-Haussler bound except the bias term is encorporated. It is based upon the following equality

$$y_i(f(\mathbf{x}_i) - f_i(\mathbf{x}_i)) = \alpha_i S_i^2.$$

The Algorithm

The following steps are used to compute α and σ

1. Initialize $\sigma = \{1, ..., 1\}$.

2. Solve the standard SVM algorithm

$$\alpha(\sigma) = \arg\max_{\alpha} W(\alpha, \sigma).$$

3. Minimize the estimate of error T with respect to σ with a gradient step.

5. If local minima of T is not reached goto step 3.

6. Discard dimensions corresponding to small elements in σ and return to step 2.

Computing Gradients

The gradient with respect to the margin

$$\frac{\partial W^2(\alpha)}{\partial \sigma_i} = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j \frac{\partial K(\mathbf{x}_i, \mathbf{x}_j)}{\partial \sigma_i}.$$

The gradient with respect to the radius

$$\frac{\partial R^2(\beta)}{\partial \sigma_i} = \sum_{i=1}^{\ell} \beta_i \frac{\partial K(\mathbf{x}_i, \mathbf{x}_i)}{\partial \sigma_i} - \sum_{i,j=1}^{\ell} \beta_i \beta_j \frac{\partial K(\mathbf{x}_i, \mathbf{x}_j)}{\partial \sigma_i}.$$

The gradient with respect to span

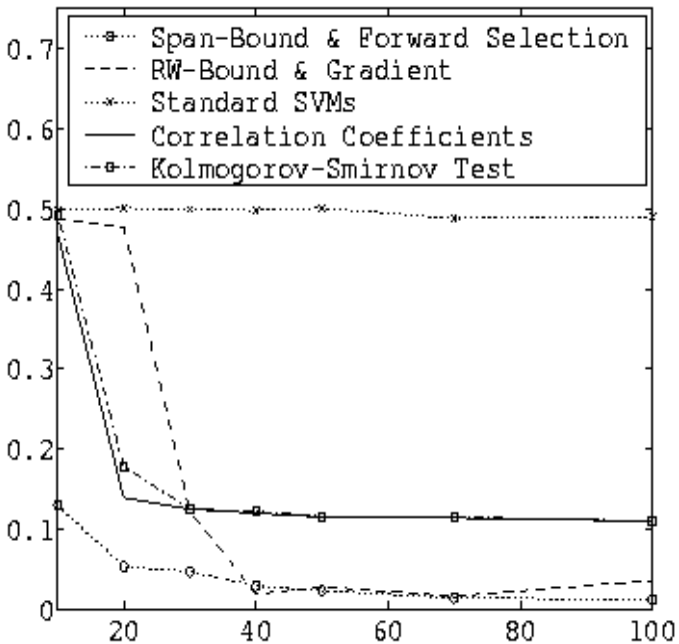
$$\frac{\partial S_i^2(\beta)}{\partial \sigma_i} = S_i^4 \left(\tilde{\mathbf{K}}_{SV}^{-1} \frac{\partial \tilde{\mathbf{K}}_{SV}}{\partial \sigma_i} \tilde{\mathbf{K}}_{SV}^{-1} \right),$$

where

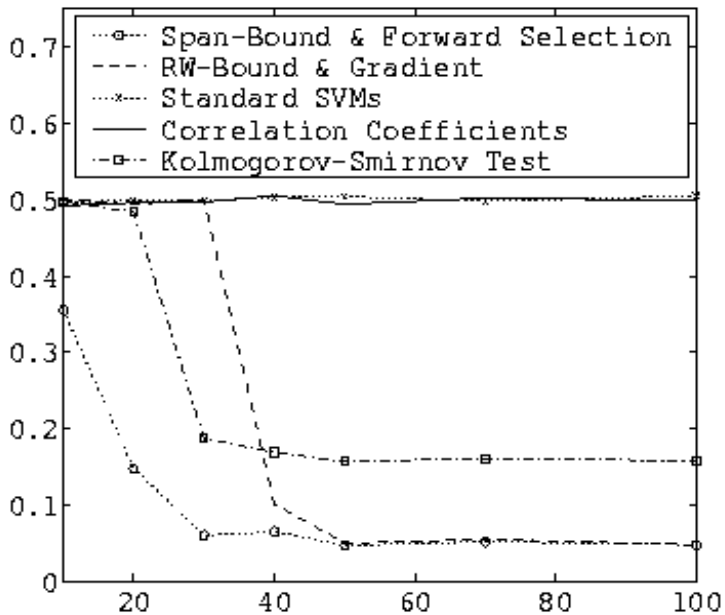
$$\tilde{\mathbf{K}}_{SV} = \begin{pmatrix} \mathbf{K} & 1 \\ 1^T & 0 \end{pmatrix}.$$

Toy Data

Linear problem with 6 relevant dimensions of 202

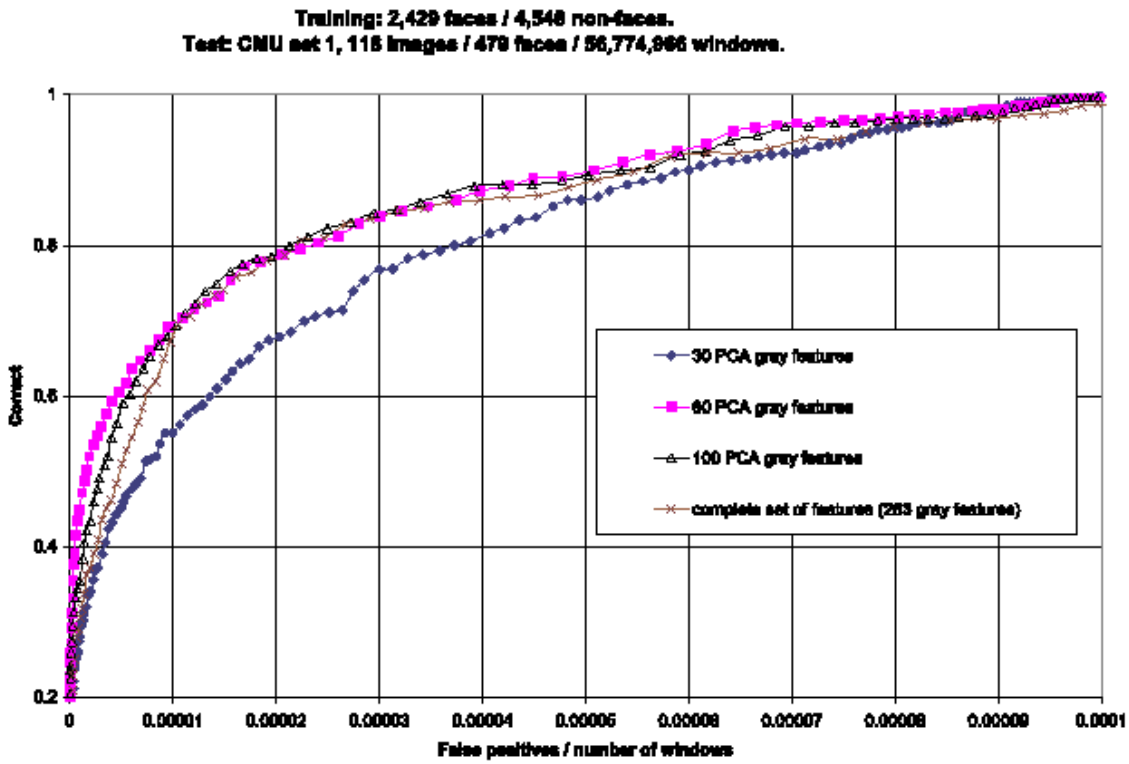


Nonlinear problem with 2 relevant dimensions of 52



Face Detection

On the CMU testset consisting of 479 faces and 57,000,000 non-faces we compare ROC curves obtained for different number of selected features. We see that using more than 60 features does not help.



Class 23, 2001

Molecular Classification of Cancer

Dataset	Total Samples	Class 0	Class 1
Leukemia Morphology (train)	38	27 ALL	11 AML
Leukemia Morpholgy (test)	34	20 ALL	14 AML
Leukemia Lineage (ALL)	23	15 B-Cell	8 T-Cell
Lymphoma Outcome (AML)	15	8 Low risk	7 High risk

Dataset	Total Samples	Class 0	Class 1
Lymphoma Morphology	77	19 FSC	58 DLCL
Lymphoma Outcome	58	20 Low risk	14 High risk
Brain Morphology	41	14 Glioma	27 MD
Brain Outcome	50	38 Low risk	12 High risk

Class 23, 2001

Morphology Classification

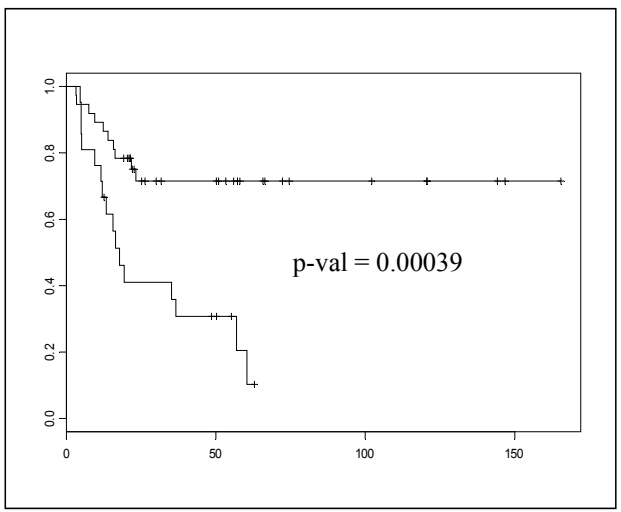
Dataset	Algorithm	Total Samples	Total errors	Class 1 errors	Class 0 errors	Number Genes
Leukemia Morphology (trest) AML vs ALL	SVM	35	0/35	0/21	0/14	40
	WV	35	2/35	1/21	1/14	50
	k-NN	35	3/35	1/21	2/14	10
Leukemia Lineage (ALL) B vs T	SVM	23	0/23	0/15	0/8	10
	WV	23	0/23	0/15	0/8	9
	k-NN	23	0/23	0/15	0/8	10
Lymphoma FS vs DLCL	SVM	77	4/77	2/32	2/35	200
	WV	77	6/77	1/32	5/35	30
	k-NN	77	3/77	1/32	2/35	250
Brain MD vs Glioma	SVM	41	1/41	1/27	0/14	100
	WV	41	1/41	1/27	0/14	3
	k-NN	41	0/41	0/27	0/14	5

Outcome Classification

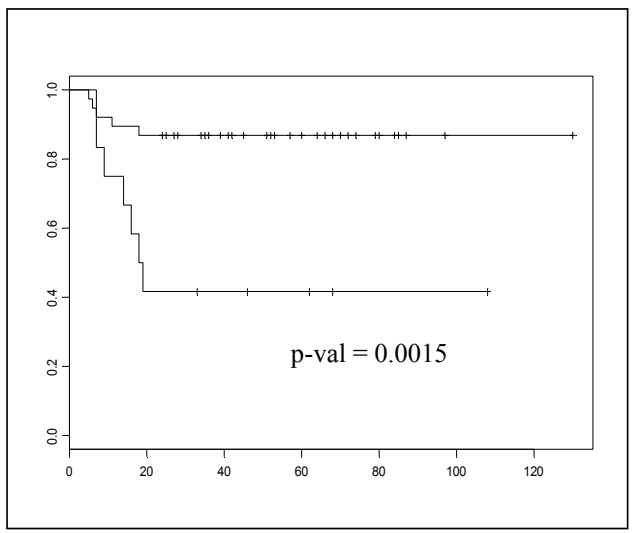
Dataset	Algorithm	Total Samples	Total errors	Class 1 errors	Class 0 errors	Number Genes
Lymphoma LBC treatment outcome	SVM	58	13/58	3/32	10/26	100
	WV	58	15/58	5/32	10/26	12
	k-NN	58	15/58	8/32	7/26	15
Brain MD treatment outcome	SVM	50	7/50	6/12	1/38	50
	WV	50	13/50	6/12	7/38	6
	k-NN	50	10/50	6/12	4/38	5

Outcome Classification

Error rates ignore temporal information such as when a patient dies. Survival analysis takes temporal information into account. The Kaplan-Meier survival plots and statistics for the above predictions show significance.



Lymphoma



Medulloblastoma