Optimization in Data Mining

Olvi L. Mangasarian
with
G. M. Fung, J. W. Shavlik, Y.-J. Lee, E.W. Wild
& Collaborators at ExonHit – Paris

University of Wisconsin – Madison
&
University of California- San Diego

Occam's Razor

A Widely Held "Axiom" in Machine Learning & Data Mining

"Entities are not to be multiplied beyond necessity"

William of Ockham (English Philosopher & Theologian)
1287 Surrey - 1347 Munich

"Everything should be made as simple as possible, but not simpler"

Albert Einstein
1879 Munich- 1955 Princeton

"Simplest is Best"

What is Data Mining?

- Data mining is the process of analyzing data in order to extract useful knowledge such as:
  - Clustering of unlabeled data
    - Unsupervised learning
  - Classifying labeled data
    - Supervised learning
  - Feature selection
    - Suppression of irrelevant or redundant features
- Optimization plays a fundamental role in data mining via:
  - Support vector machines or kernel methods
    - State-of-the-art tool for data mining and machine learning

What is a Support Vector Machine?

- An optimally defined surface
- Linear or nonlinear in the input space
- Linear in a higher dimensional feature space
- Feature space defined by a linear or nonlinear kernel

$$K(A, X) \to Y,$$

$$A \in R^{m \times n}, \quad X \in R^{n \times k}, \quad \text{and} \quad Y \in R^{m \times k}$$

Principal Topics

- Data clustering as a concave minimization problem
  - K-median clustering and feature reduction
  - Identify class of patients that benefit from chemotherapy
- Linear and nonlinear support vector machines (SVMs)
  - Feature and kernel function reduction
- Enhanced knowledge-based classification
  - LP with implication constraints
- Generalized Newton method for nonlinear classification
  - Finite termination with or without stepsize
- Drug discovery based on gene macroarray expression
  - Identify class of patients likely to respond to new drug
- Multisurface proximal classification
  - Nonparallel classifiers via generalized eigenvalue problem

Clustering in Data Mining

General Objective

- Given: A dataset of m points in n-dimensional real space
- Problem: Extract hidden distinct properties by clustering the dataset into k clusters

Concave Minimization Formulation
1-Norm Clustering: k-Median Algorithm

- Given: Set $\mathcal{A}$ of *m* points in R^n represented by the matrix $A \in R^{m \times n}$, and a number k of desired clusters
- Find: Cluster centers $C_1, \ldots, C_k \in R^n$ that minimize the sum of 1-norm distances of each point: $A_1, A_2, \ldots, A_m,$ to its closest cluster center.
- Objective Function: Sum of *m* minima of k linear functions, hence it is piecewise-linear concave
- Difficulty: Minimizing a general piecewise-linear concave function over a polyhedral set is NP-hard

Clustering via Finite Concave Minimization

- Minimize the sum of 1-norm distances between each data point A_i and the closest cluster center C_ℓ :

$$\min_{C_\ell \in R^n,\, D_{i\ell} \in R^n} \sum_{i=1}^{m} \min_{\ell=1\ldots,k} \{e' D_{i\ell}\}$$

$$\text{s.t.} \quad -D_{i\ell} \le A_i' - C_\ell \le D_{i\ell},$$

$$i = 1,\ldots,m,\ \ell = 1,\ldots,k,$$

where e is a column vector of ones.

K-Median Clustering Algorithm

Finite Termination at Local Solution
Based on a Bilinear Reformulation

Step 0 (Initialization): Pick *k* initial cluster centers

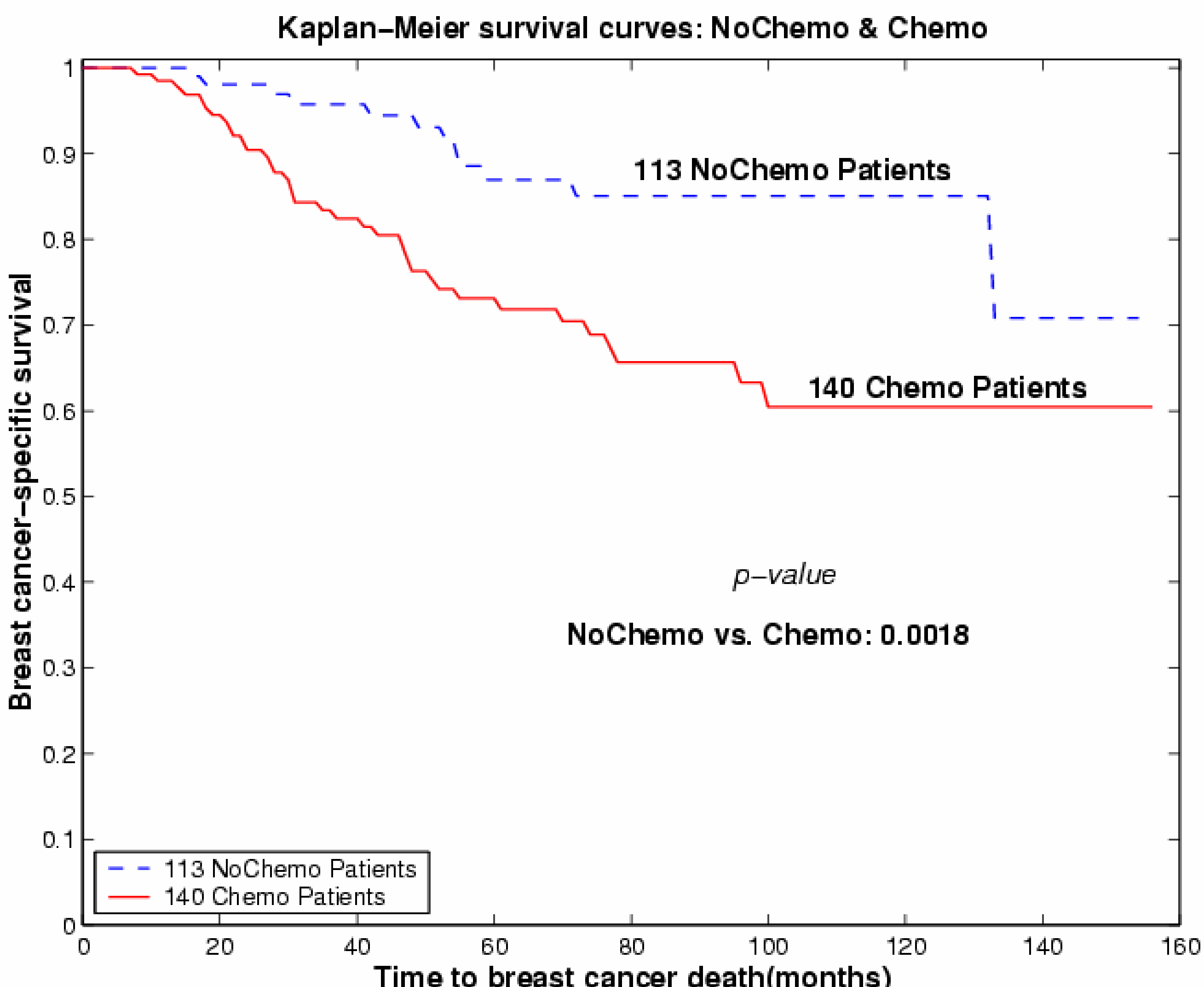
Step 1 (Cluster Assignment): Assign points to the cluster with the nearest cluster center in 1-norm

Step 2 (Center Update) Recompute location of center for each cluster as the cluster median (closest point to all cluster points in 1-norm)

Step3 (Stopping Criterion) Stop if the cluster centers are unchanged, else go to Step 1

Algorithm terminates in a finite number of steps, at a local solution

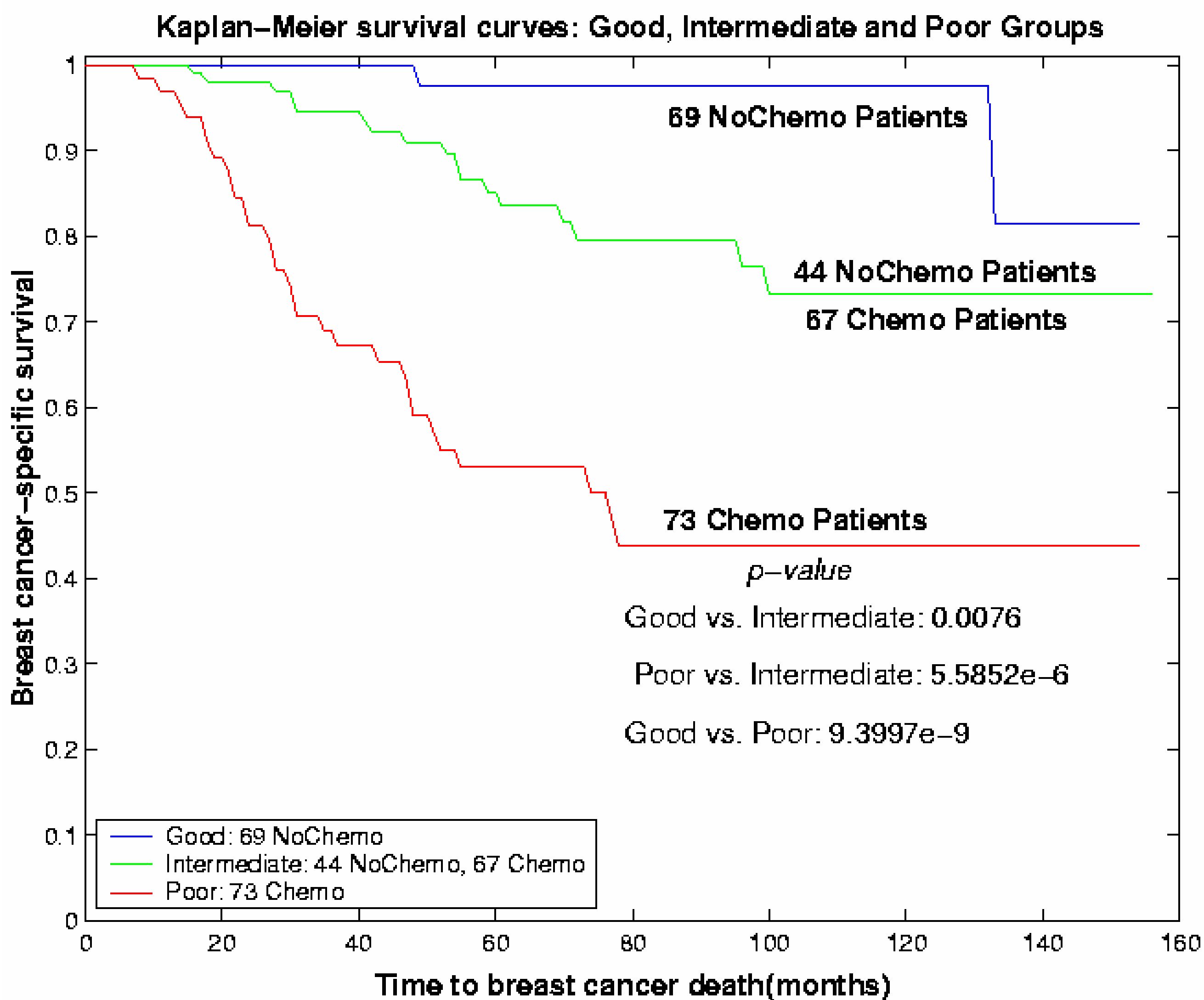
Breast Cancer Patient Survival Curves With & Without Chemotherapy



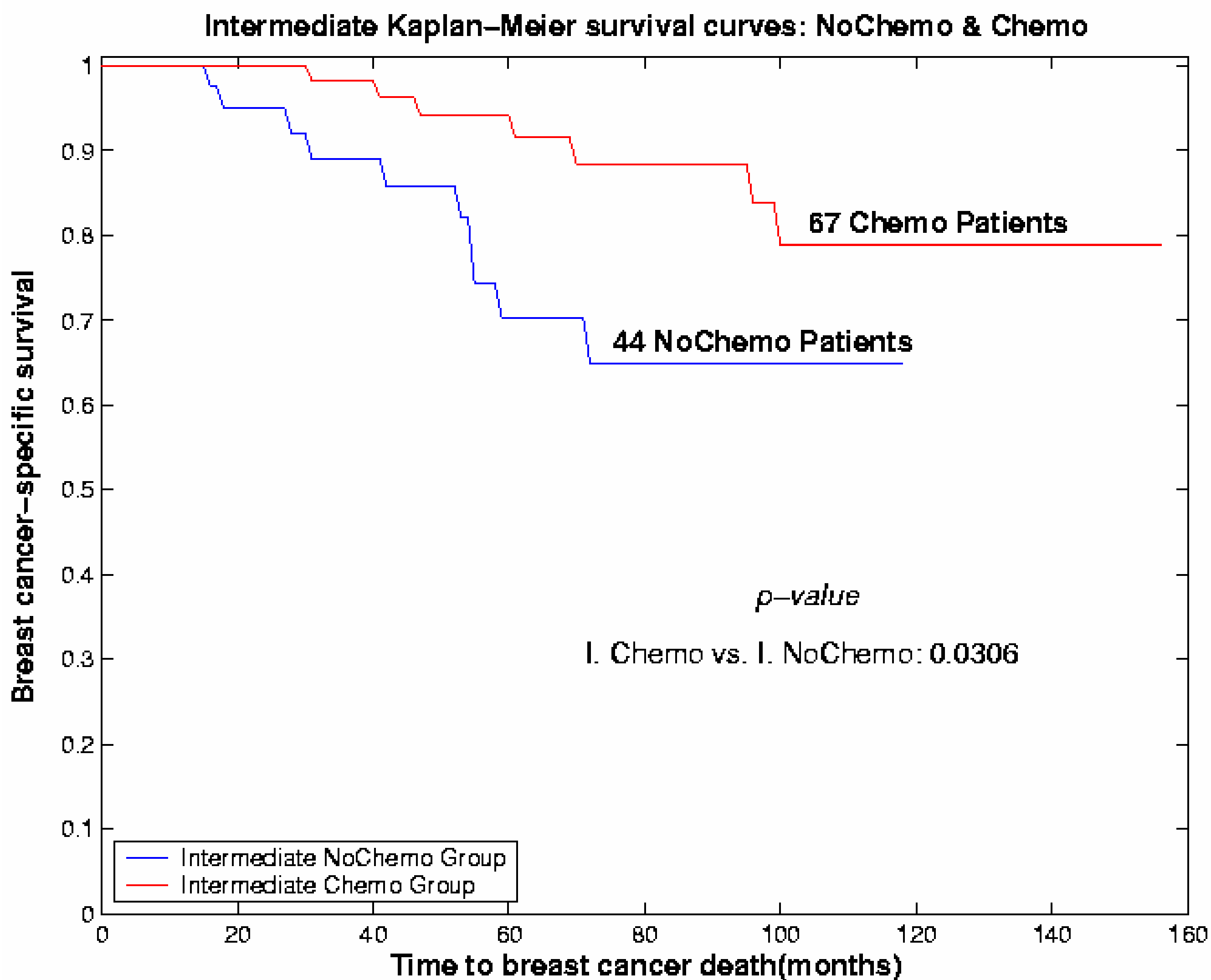
Survival Curves for 3 Groups:

Good, Intermediate & Poor Groups

(Generated Using k-Median Clustering)



Survival Curves for Intermediate Group: Split by Chemo & NoChemo



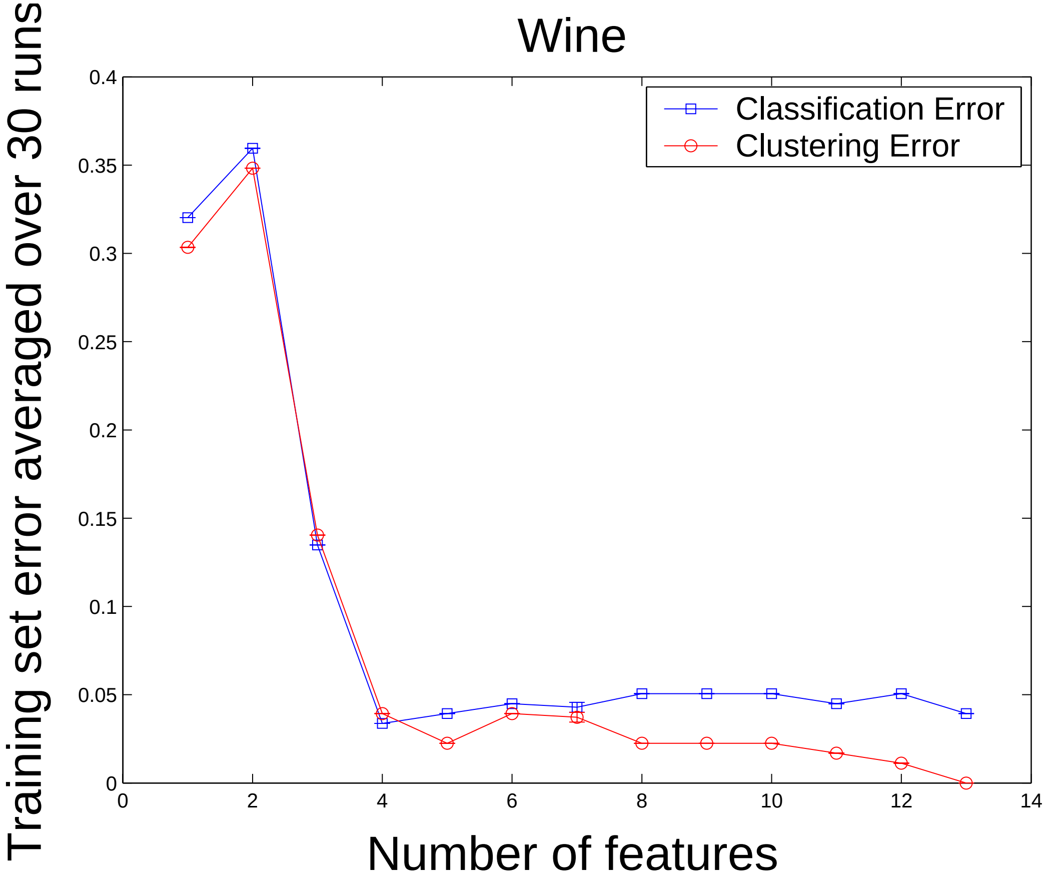
Feature Selection in k-Median Clustering

- Find a reduced number of input space features such that clustering in the reduced space closely replicates the clustering in the full dimensional space

Basic Idea

- Based on nondifferentiable optimization theory, make a simple but fundamental modification in the second step of the k-median algorithm
- In each cluster, find a point closest in the 1-norm to all points in that cluster and to the zero median of ALL data points
- Based on increasing weight given to the zero data median, more features are deleted from problem
- Proposed approach can lead to a feature reduction as high as 69%, with clustering comparable to within 4% to that with the original set of features

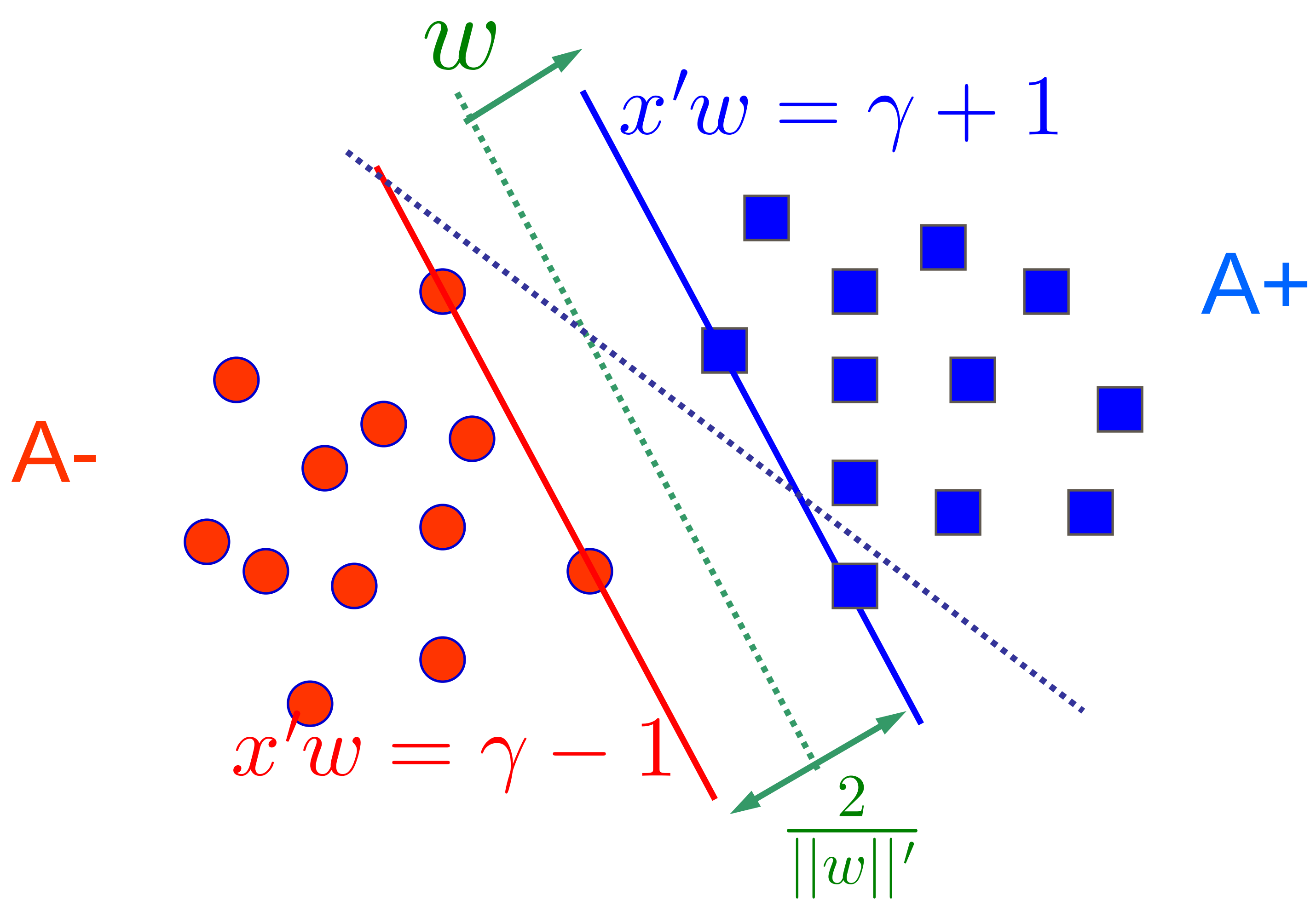
3-Class Wine Dataset
178 Points in 13-dimensional Space



Support Vector Machines

- Linear & nonlinear classifiers using kernel functions

Support Vector Machines
Maximize the Margin between Bounding Planes



Support Vector Machine

Algebra of 2-Category Linearly Separable Case

- Given m points in n dimensional space
- Represented by an m-by-n matrix A
- Membership of each A_i in class +1 or –1 specified by:
  - An m-by-m diagonal matrix D with +1 & -1 entries
- Separate by two bounding planes, $x'w = \gamma \pm 1$:

$$A_i w \geqq \gamma + 1, \quad \text{for} \quad D_{ii} = +1,$$
$$A_i w \leqq \gamma - 1, \quad \text{for} \quad D_{ii} = -1.$$

- More succinctly:

$$D(Aw - e\gamma) \geqq e,$$

where e is a vector of ones.

Feature-Selecting 1-Norm Linear SVM

- 1-norm SVM:

$$\min_{y \geqslant 0, w, \gamma} \quad \nu e'y + \|w\|_1$$

$$\text{s. t.} \quad D(Aw - e\gamma) + y \geqslant e \quad ,$$

where $D_{ii} = \pm 1$ are elements of the diagonal matrix D denoting the class of each point A_i of the dataset matrix A

- Very effective in feature suppression

➢ For example, 5 out of 30 cytological features are selected by the 1-norm SVM for breast cancer diagnosis with over 97% correctness.

➢ In contrast, 2-norm and ∞-norm SVMs suppress no features.

1- Norm Nonlinear SVM

- Linear SVM: (Linear separating surface: $x'w = \gamma$)

$$\min_{y \geqslant 0, w, \gamma} \quad \nu e'y + \|w\|_1 \qquad \text{(LP)}$$
$$\text{s.t.} \quad D(Aw - e\gamma) + y \geqslant e$$

Change of variable $w = A'Du$ and maximizing the margin

in the "dual space", gives:

$$\min_{y \geqslant 0, u, \gamma} \quad \nu e'y + \|u\|_1$$
$$\text{s.t.} \quad D(AA'Du - e\gamma) + y \geqslant e$$

- Replace AA' by a nonlinear kernel $K(A, A')$:

$$\min_{y \geqslant 0, u, \gamma} \quad \nu e'y + \|u\|_1$$
$$\text{s.t.} \quad D(K(A, A')Du - e\gamma) + y \geqslant e$$

2- Norm Nonlinear SVM

$$\min_{y \geqslant 0, u, \gamma} \frac{\nu}{2}\|y\|_2^2 + \frac{1}{2}\|u, \gamma\|_2^2$$

$$\text{s.t. } D(K(A, A')Du - e\gamma) + y \geqslant e$$

Equivalently:

$$\min_{u, \gamma} \frac{\nu}{2}\|(e - D(KA, A')Du - e\gamma))_+\|_2^2 + \frac{1}{2}\|u, \gamma\|_2^2$$

The Nonlinear Classifier

- The nonlinear classifier:

$$K(x', A')Du = \gamma$$

$$K(A, A') : R^{m\times n} \times R^{n\times m} \longmapsto R^{m\times m}$$

- K is a nonlinear kernel, e.g.:
- Gaussian (Radial Basis) Kernel：

$$K(A, A')_{ij} = \varepsilon^{-\mu\|A_i - A_j\|_2^2}, \ \ i, j = 1, \ldots, m$$

- The ij-entry of $K(A, A')$ represents "similarity" between the data points A_i and A_j (Nearest Neighbor)
- Can generate highly nonlinear classifiers

Data Reduction in Data Mining

- RSVM:Reduced Support Vector Machines

Difficulties with Nonlinear SVM for Large Problems

- The nonlinear kernel $K(A, A') \in R^{m \times m}$ is fully dense
  - Long CPU time to compute $m \times m$ elements of nonlinear kernel K(A,A′)
  - Runs out of memory while storing $m \times m$ elements of K(A,A′)
- Computational complexity depends on m
  - Complexity of nonlinear SSVM $\sim O((m+1)^3)$
- Separating surface depends on almost entire dataset
  - Need to store the entire dataset after solving the problem

Overcoming Computational & Storage Difficulties

Use a "Thin" Rectangular Kernel

- Choose a small random sample $\bar{A} \in R^{\bar{m} \times n}$ of A
  - The small random sample $\bar{A}$ is a representative sample of the entire dataset
  - Typically $\bar{A}$ is 1% to 10% of the rows of A
- Replace $K(A, A')$ by $K(A, \bar{A}') \in R^{m \times \bar{m}}$ with corresponding $\bar{D} \subset D$ in nonlinear SSVM
  - Only need to compute and store $m \times \bar{m}$ numbers for the rectangular kernel
- Computational complexity reduces to $O((\bar{m}+1)^3)$
- The nonlinear separator only depends on $\bar{A}$

Using $K(\bar{A}, \bar{A}')$ gives lousy results!

Reduced Support Vector Machine Algorithm

Nonlinear Separating Surface: $K(x', \bar{A}')\bar{D}\bar{u} = \gamma$

(i) Choose a random subset matrix $\overline{A} \in R^{\overline{m} \times n}$ of entire data matrix $A \in R^{m \times n}$

(ii) Solve the following problem by a generalized Newton method with corresponding $\overline{D} \subset D$:

$$\min_{(\overline{u},\gamma) \in R^{\overline{m}+1}} \frac{\nu}{2}\|(e - D(K(A, \overline{A}')\bar{D}\bar{u} - e\gamma))_+\|_2^2 + \frac{1}{2}\|\bar{u}, \gamma\|_2^2$$

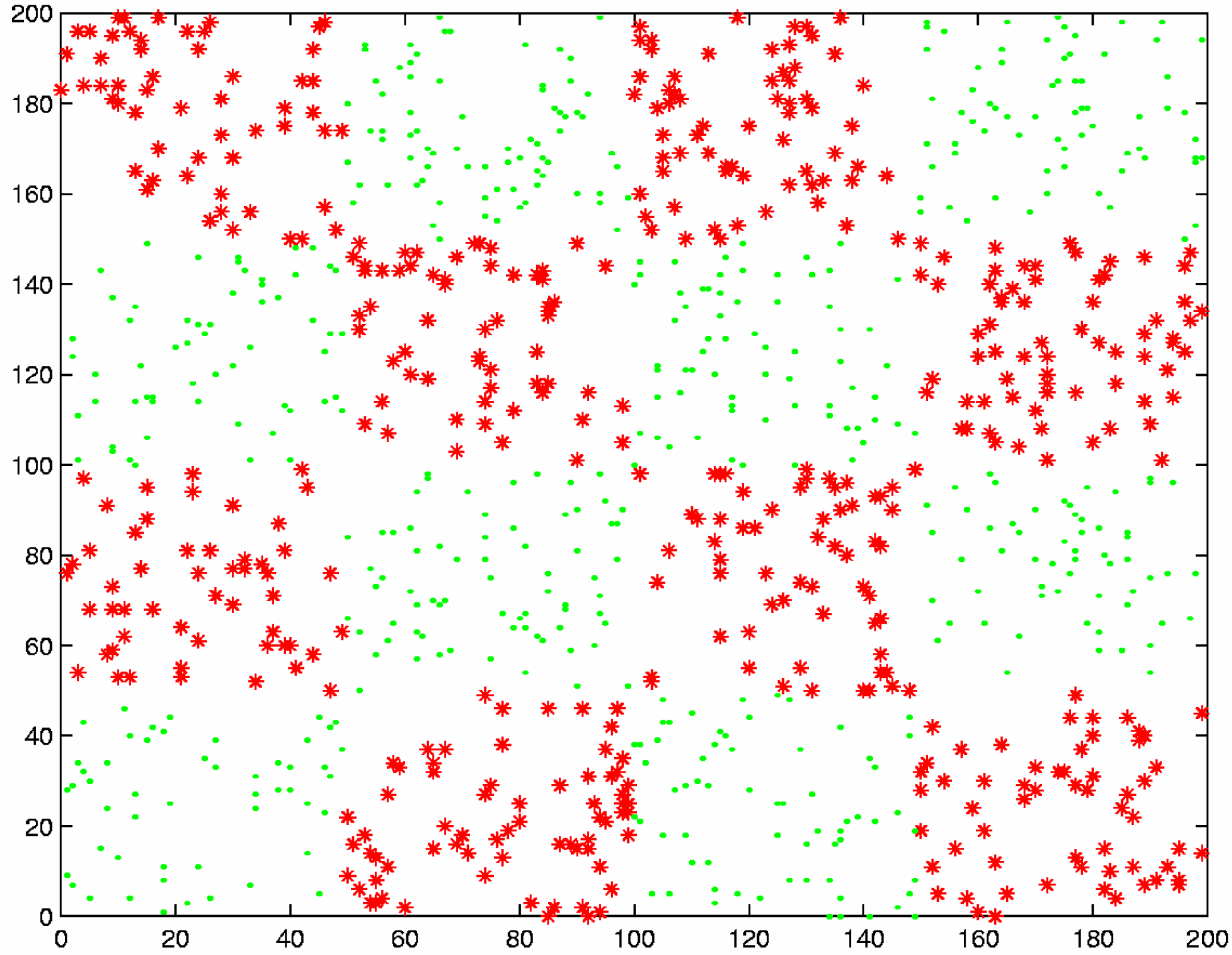
(iii) The separating surface is defined by the optimal solution $(\overline{u}, \gamma)$ in step (ii):

$$K(x', \bar{A}')\bar{D}\bar{u} = \gamma$$

A Nonlinear Kernel Application

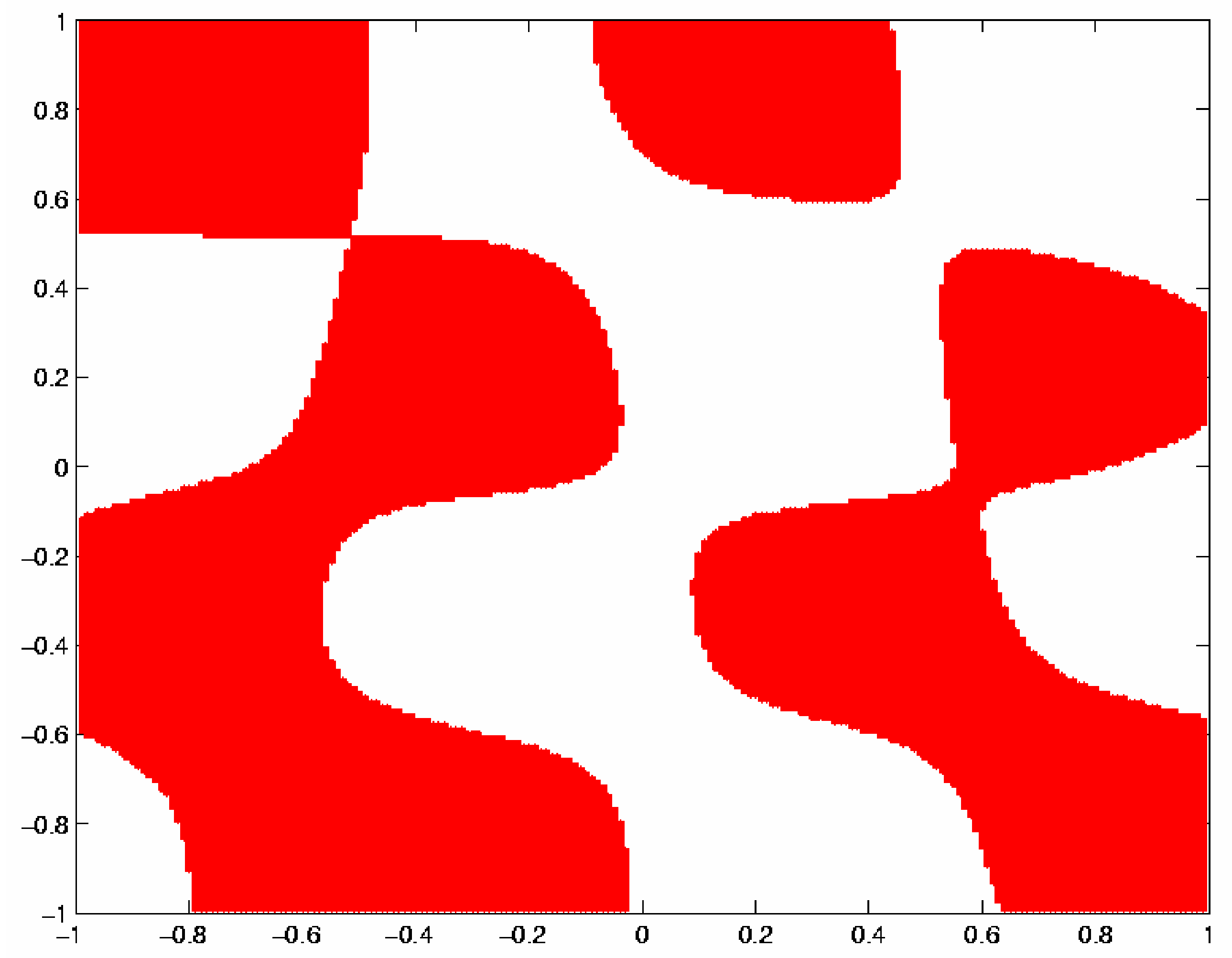
Checkerboard Training Set: 1000 Points in R^2

Separate 486 Asterisks from 514 Dots



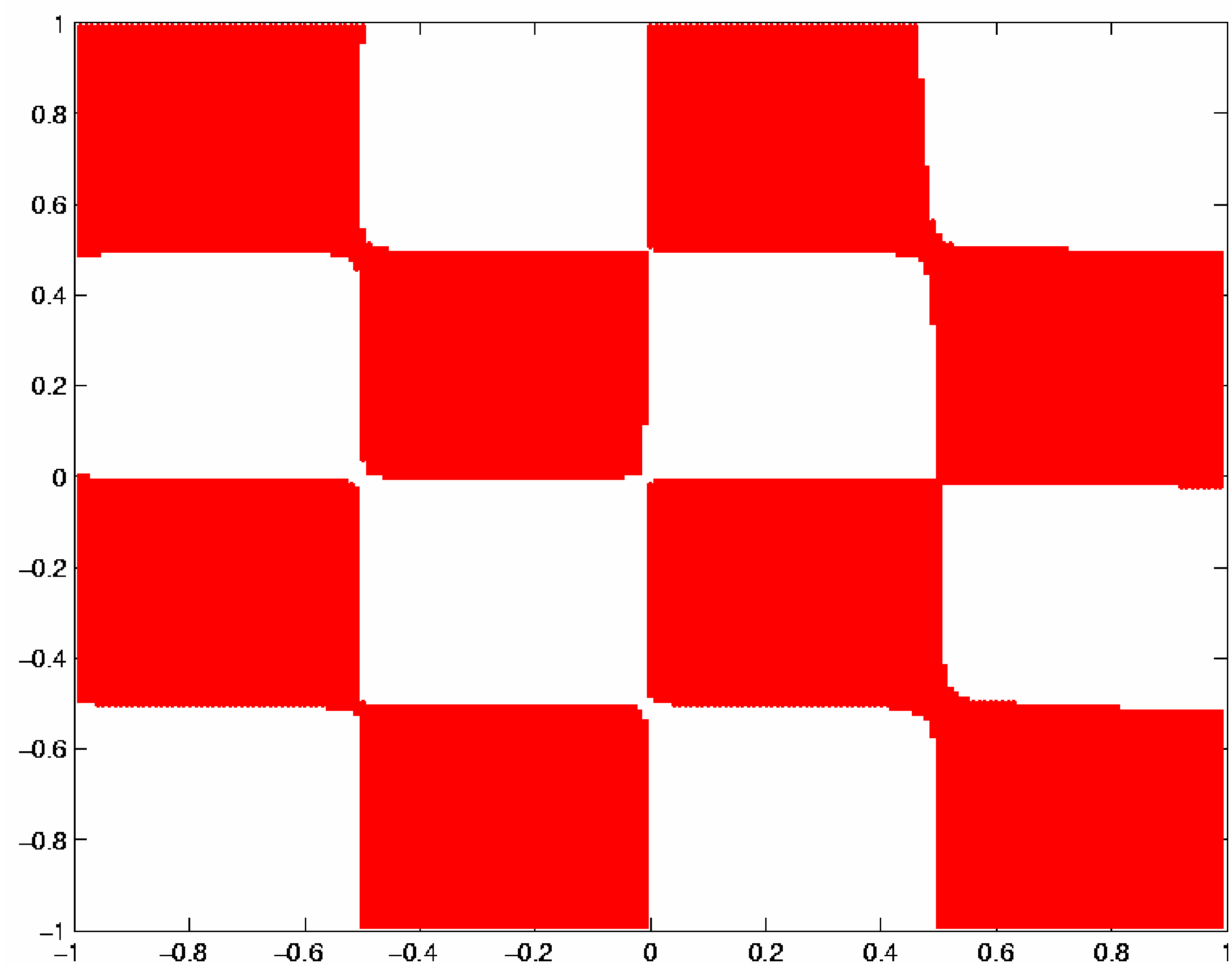
Conventional SVM Result on Checkerboard Using 50 Randomly Selected Points Out of 1000

$$K(\overline{A}, \overline{A}') \in R^{50 \times 50}$$



RSVM Result on Checkerboard
Using SAME 50 Random Points Out of 1000

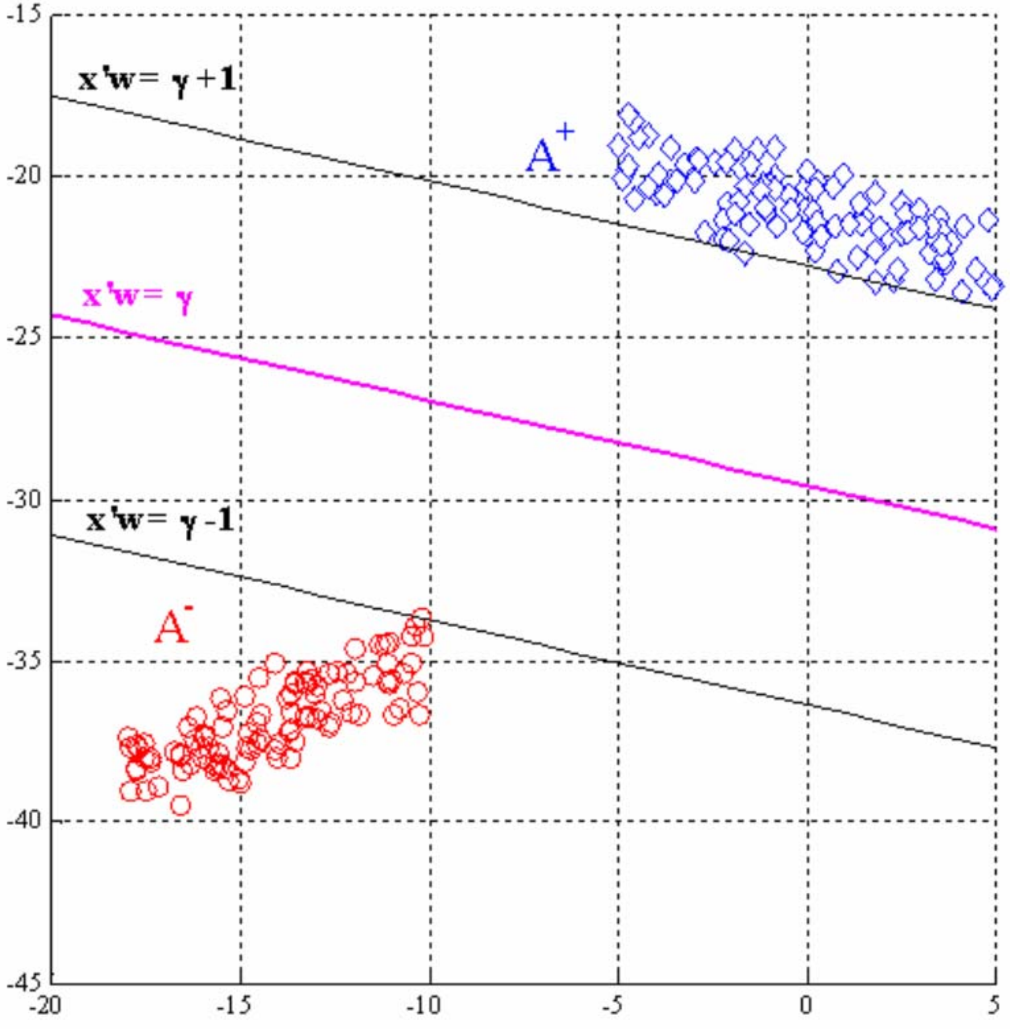
$$K(A, \overline{A}') \in R^{1000 \times 50}$$



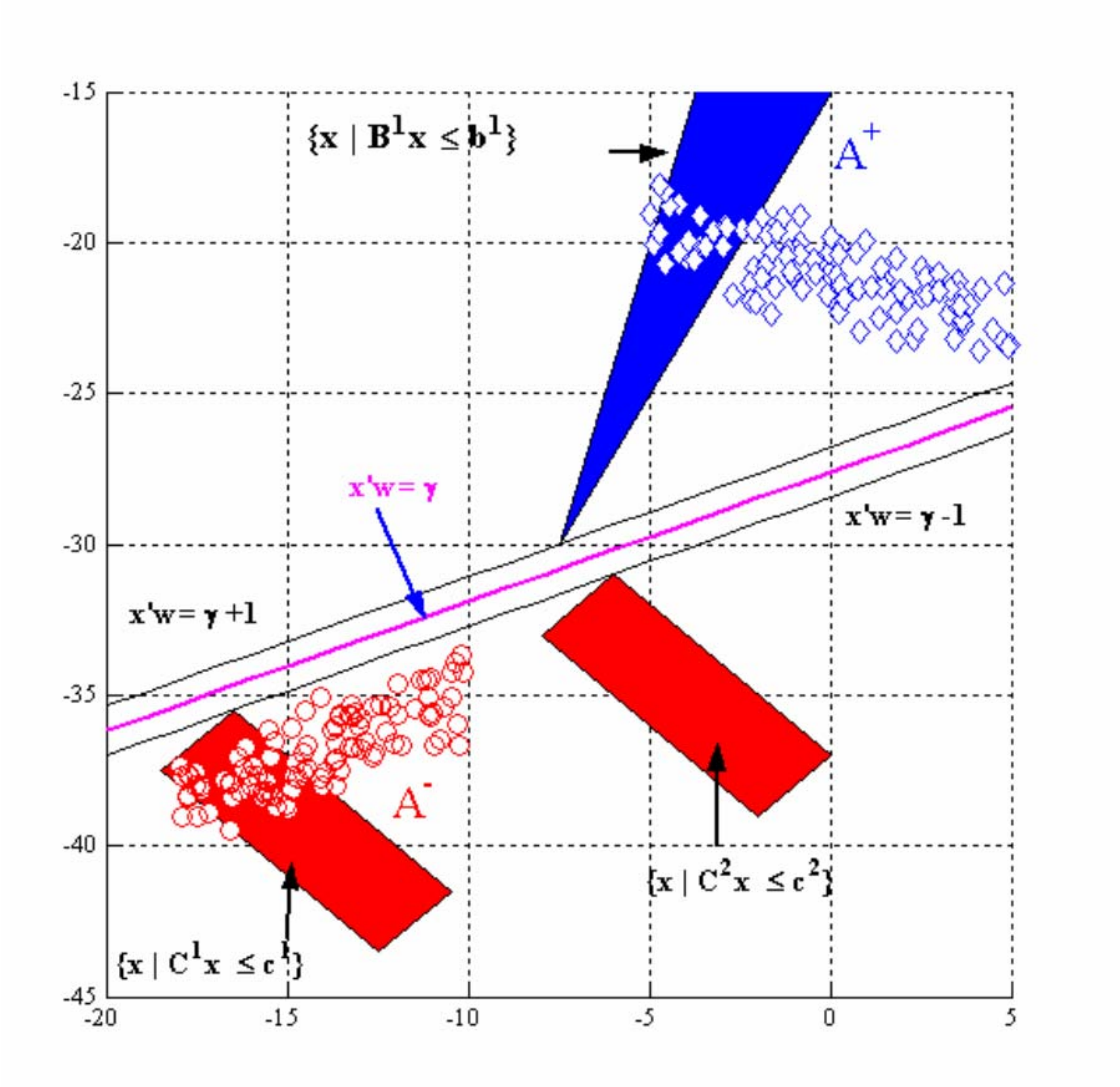
Knowledge-Based Classification

- Use prior knowledge to improve classifier correctness

Conventional Data-Based SVM



Knowledge-Based SVM via Polyhedral Knowledge Sets



Incoporating Knowledge Sets Into an SVM Classifier

- Suppose that the knowledge set: $\{x \mid Bx \leqslant b\}$ belongs to the class A+. Hence it must lie in the halfspace :

$$\{x|x'w \geqslant \gamma + 1\}$$

- We therefore have the implication:

$$Bx \leqslant b \quad \Rightarrow \quad x'w \geqslant \gamma + 1$$

- This implication is equivalent to a set of constraints that can be imposed on the classification problem.

Knowledge Set Equivalence Theorem

$$Bx \leqslant b \implies x'w \geqslant \gamma + 1,$$

or, for a fixed (w, γ) :

$$Bx \leqslant b, \quad x'w < \gamma + 1, \text{ has no solution } x$$

$$\Updownarrow \{x \mid Bx \leqslant b\} \neq \emptyset$$

$$\exists u: \ B'u + w = 0, \quad b'u + \gamma + 1 \leqslant 0, \ u \geqslant 0$$

Knowledge-Based SVM Classification

- Adding one set of constraints for each knowledge set to the 1-norm SVM LP, we have:

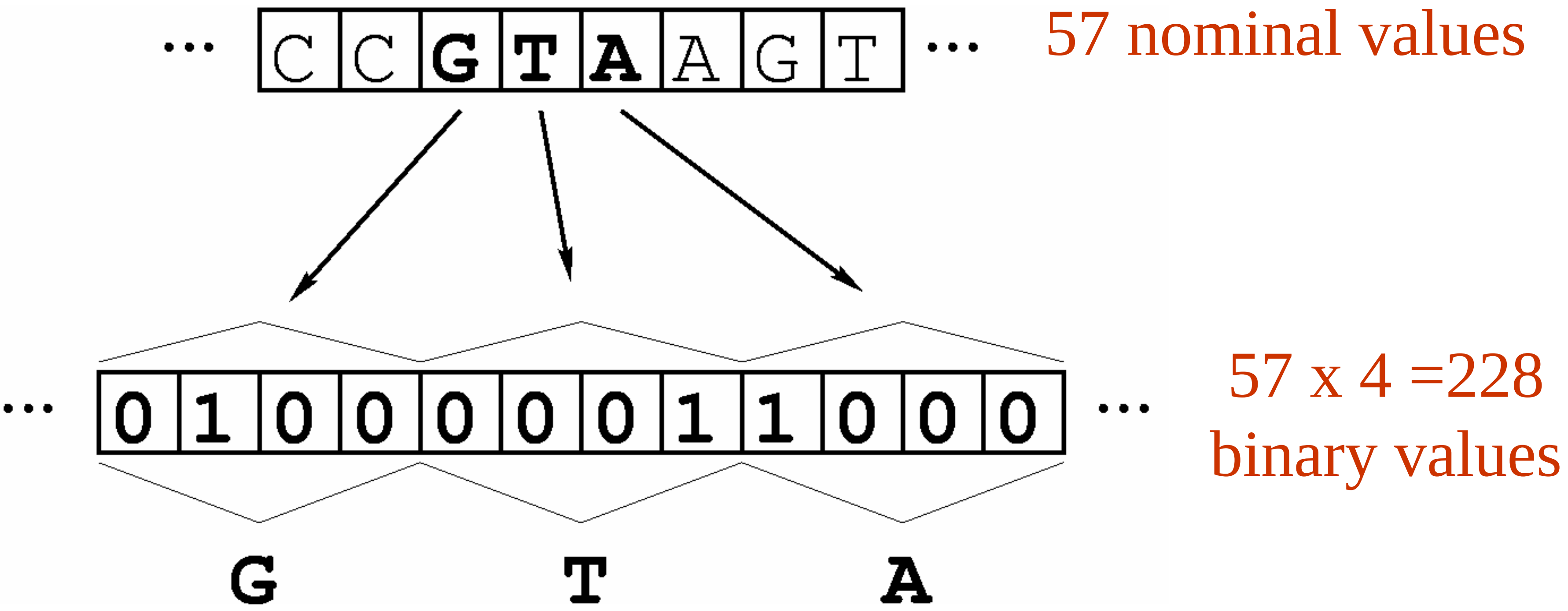
$$\begin{array}{rrcl}
\displaystyle\min_{w,\gamma,y,u^i,v^j} & \nu e'y + \|w\|_1 & & \\
\text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\
& y & \geq & 0 \\
& & & \\
& {B^i}'u^i + w & = & 0 \\
& {b^i}'u^i + \gamma + 1 & \leq & 0 \\
& u^i & \geq & 0, \; i = 1, \ldots, k \\
& {C^j}'v^j - w & = & 0 \\
& {c^j}'v^j - \gamma + 1 & \leq & 0 \\
& v^j & \geq & 0, \; j = 1, \ldots, \ell
\end{array}$$

Numerical Testing
DNA Promoter Recognition Dataset

- Promoter: Short DNA sequence that precedes a gene sequence.
- A promoter consists of 57 consecutive DNA nucleotides belonging to {A,G,C,T} .
- Important to distinguish between promoters and nonpromoters
- This distinction identifies starting locations of genes in long uncharacterized DNA sequences.

The Promoter Recognition Dataset Numerical Representation

- Input space mapped from 57-dimensional nominal space to a real valued 57 x 4=228 dimensional space.



Promoter Recognition Dataset Prior Knowledge Rules as Implication Constraints

- Prior knowledge consist of the following 64 rules:

$$\begin{bmatrix} R1 \\ or \\ R2 \\ or \\ R3 \\ or \\ R4 \end{bmatrix} \wedge \begin{bmatrix} R5 \\ or \\ R6 \\ or \\ R7 \\ or \\ R8 \end{bmatrix} \wedge \begin{bmatrix} R9 \\ or \\ R10 \\ or \\ R11 \\ or \\ R12 \end{bmatrix} \Longrightarrow PROMOTER$$

Promoter Recognition Dataset
Sample Rules

$$R4: \quad (p_{-36} = T) \wedge (p_{-35} = T) \wedge (p_{-34} = G) \wedge (p_{-33} = A) \wedge (p_{-32} = C),$$

$$R8: \quad (p_{-12} = T) \wedge (p_{-11} = A) \wedge (p_{-07} = T),$$

$$R10: \quad (p_{-45} = A) \wedge (p_{-44} = A) \wedge (p_{-41} = A).$$

A sample rule is:

$$R4 \wedge R8 \wedge R10 \Longrightarrow PROMOTER$$

The Promoter Recognition Dataset Comparative Algorithms

- KBANN Knowledge-based artificial neural network [Shavlik et al]
- BP: Standard back propagation for neural networks [Rumelhart et al]
- O'Neill's Method Empirical method suggested by biologist O'Neill [O'Neill]
- NN: Nearest neighbor with k=3 [Cost et al]
- ID3: Quinlan's decision tree builder[Quinlan]
- SVM1: Standard 1-norm SVM [Bradley et al]

The Promoter Recognition Dataset Comparative Test Results with Linear KSVM

Total leave-one-out error

KSVM & other classification algorithms

106-point Promoter Dataset: 53 Promoters, 53 Nonpromoters

Method	Number of Errors (out of 106)
KBANN	4
KSVM	5
BP	8
SVM_1	9
O'Neill	12
NN	13
ID3	19

Finite Newton Classifier

- Newton for SVM as an unconstrained optimization problem

Fast Newton Algorithm for SVM Classification

Standard quadratic programming (QP) formulation of SVM:

$$\begin{array}{lrcl} \min\limits_{w,\gamma,y} & \frac{\nu}{2}\|y\|_2^2 + \frac{1}{2}\|w,\gamma\|_2^2 & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0, \end{array}$$

At solution of QP:

$$y = (e - D(Aw - e\gamma))_+,$$

where $(\cdot)_+ = \max\{\cdot, 0\}$. Hence QP is equivalent to the nonsmooth SVM:

$$\min_{w,\gamma} \frac{\nu}{2}\|(e - D(Aw - e\gamma))_+\|_2^2 + \frac{1}{2}\|w,\gamma\|_2^2$$

Once, but not twice differentiable. However Generlized Hessian exists!

Generalized Newton Algorithm

$$f(z) = \frac{\nu}{2}\left\|(Cz-h)_+\right\|^2 + \frac{1}{2}\left\|z\right\|^2$$

$$z^{i+1} = z^i - \partial^2 f(z^i)^{-1}\nabla f(z^i)$$

$$\nabla f(z) = \nu C'(Cz-h)_+ + z$$

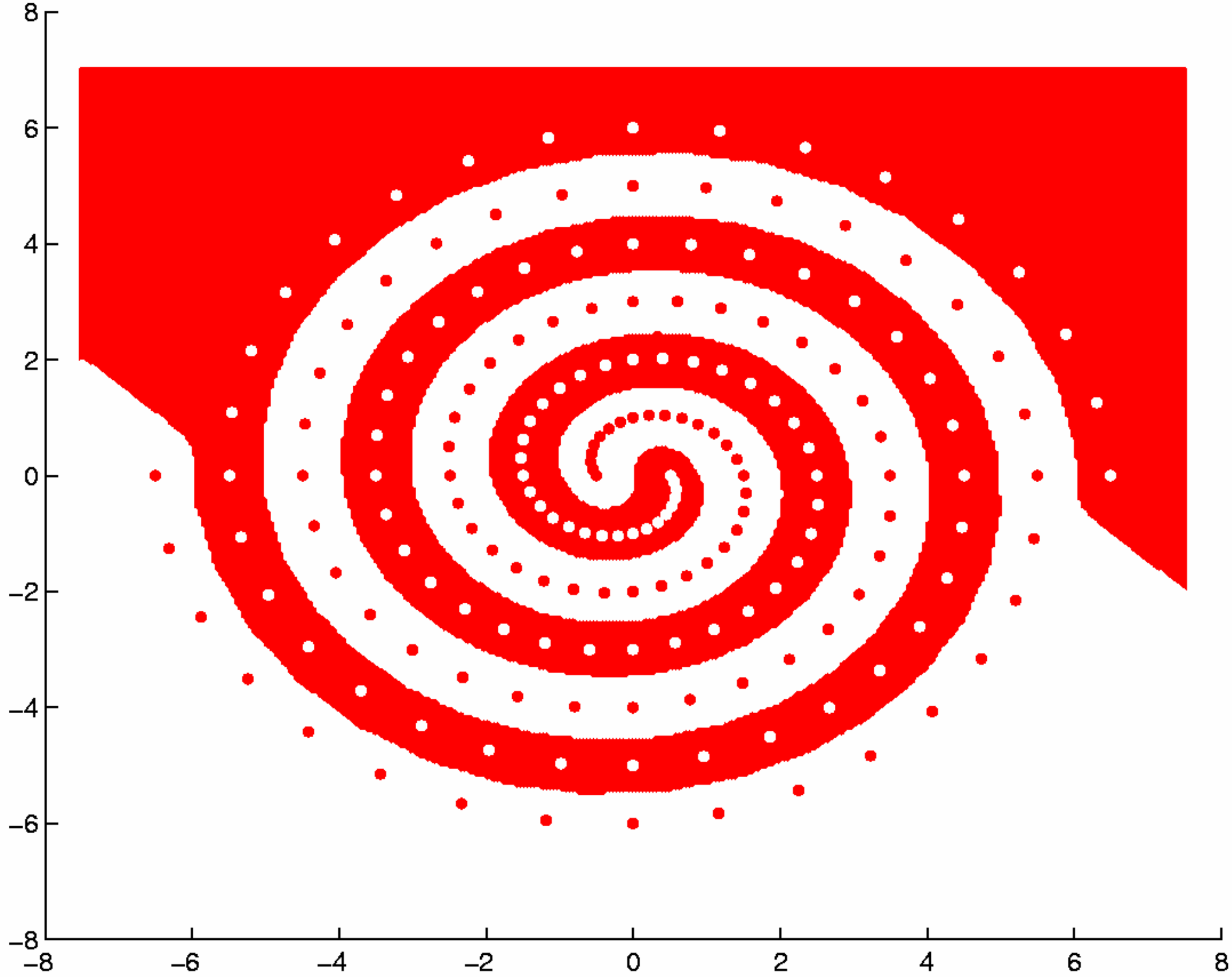
$$\partial^2 f(z) = \nu C' diag(Cz-h)_* C + I$$

$$\text{where } (Cz-h)_* = 0 \text{ if } (Cz-h) \le 0, \text{ else } (Cz-h)_* = 1.$$

- Newton algorithm terminates in a finite number of steps
  - With an Armijo stepsize (unnecessary computationally)
- Termination at global minimum
- Error rate decreases linearly
- Can generate complex nonlinear classifiers
  - By using nonlinear kernels: K(x,y)

Nonlinear Spiral Dataset

94 Red Dots & 94 White Dots



SVM Application to Drug Discovery

- Drug discovery based on gene expression

Breast Cancer Drug Discovery Based on Gene Expression

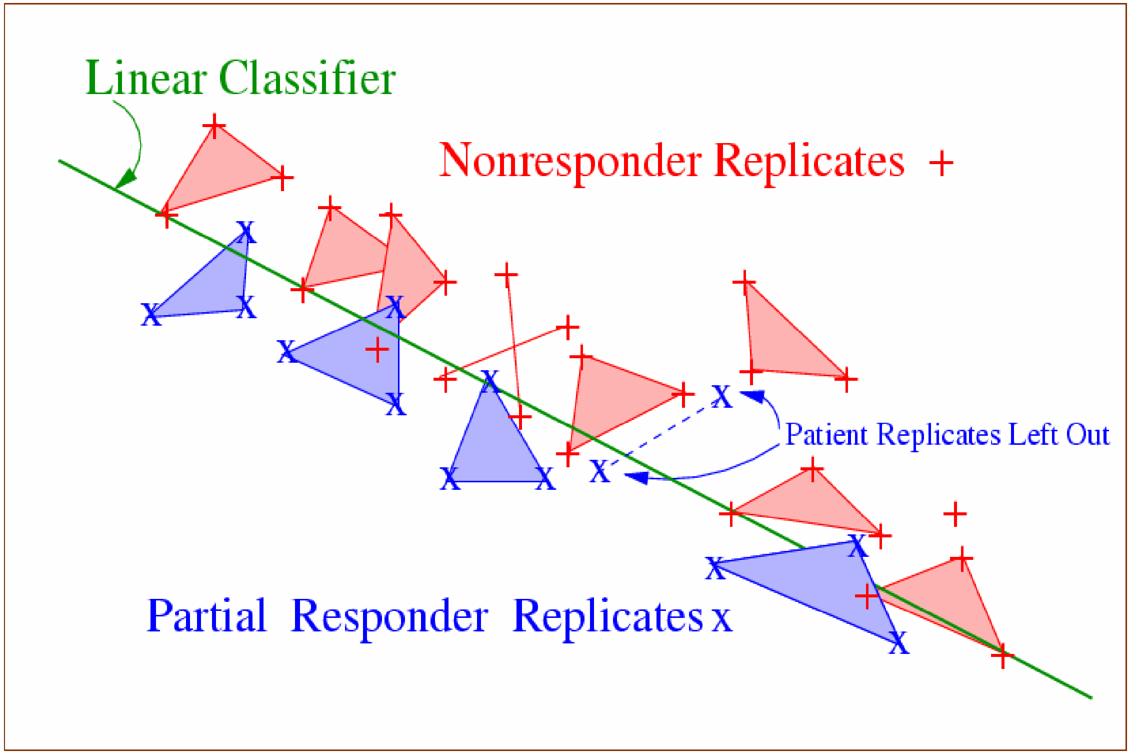
Joint with ExonHit - Paris (Curie Dataset)

- 35 patients treated by a drug cocktail
- 9 partial responders; 26 nonresponders
- 25 gene expressions out of 692 selected by ExonHit
- 1-Norm SVM and greedy combinatorial approach selected 5 genes out of 25
- Most patients had 3 distinct replicate measurements
- Distinguishing aspects of this classification approach:
  - Separate convex hulls of replicates
  - Test on mean of replicates

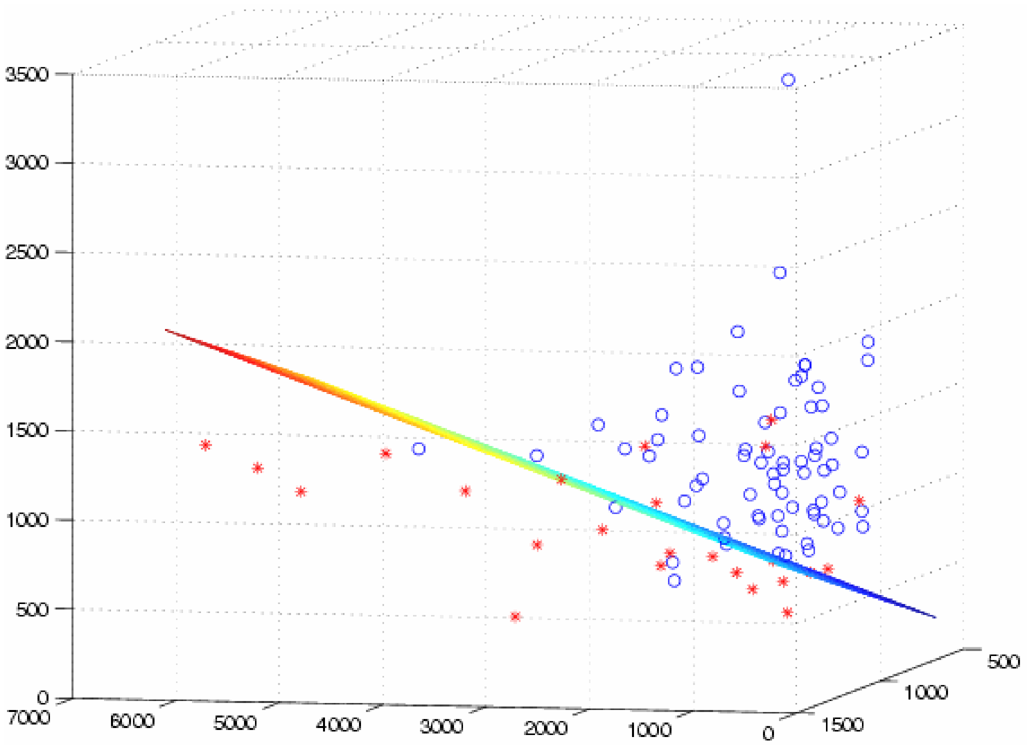
Separation of Convex Hulls of Replicates

10 Synthetic Nonresponders: 26 Replicates (Points)

5 Synthetic Partial Responders: 14 Replicates (Points)



Linear Classifier in 3-Gene Space
35 Patients with 93 Replicates
26 Nonresponders & 9 Partial Responders



In 5-gene space, leave-one-out correctness was 33 out of 35, or 94.2%

Generalized Eigenvalue Classification

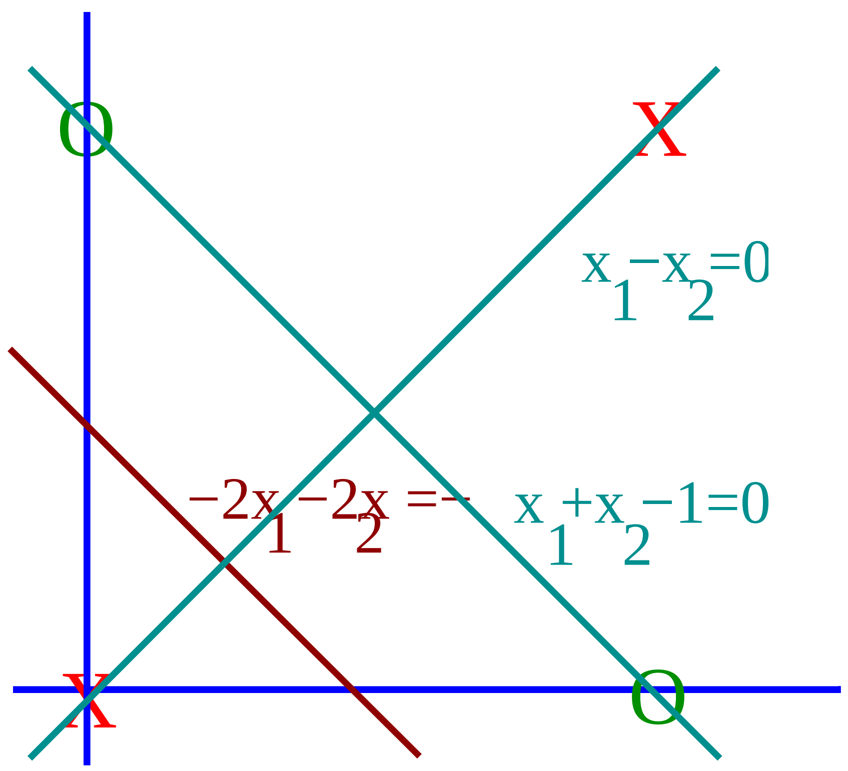
- Multisurface proximal classification via generalized eigenvalues

Multisurface Proximal Classification

- Two distinguishing features:
  - Replace halfspaces containing datasets A and B by planes proximal to A and B
  - Allow nonparallel proximal planes
- First proximal plane: $x' w^1 - \gamma^1 = 0$
  - As close as possible to dataset A
  - As far as possible from dataset B
- Second proximal plane: $x' w^2 - \gamma^2 = 0$
  - As close as possible to dataset B
  - As far as possible from dataset A

Classical Exclusive "Or" (XOR) Example

$$A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Multisurface Proximal Classifier

As a Generalized Eigenvalue Problem

$$\min_{(w,\gamma)\neq 0} \frac{\|Aw - e\gamma\|^2 / \|[\begin{smallmatrix} w \\ \gamma \end{smallmatrix}]\|^2}{\|Bw - e\gamma\|^2 / \|[\begin{smallmatrix} w \\ \gamma \end{smallmatrix}]\|^2}.$$

- Simplifying and adding regularization terms gives:

$$\min_{(w,\gamma)\neq 0} \frac{\|Aw - e\gamma\|^2 + \delta\|[\begin{smallmatrix} w \\ \gamma \end{smallmatrix}]\|^2}{\|Bw - e\gamma\|^2 + \delta\|[\begin{smallmatrix} w \\ \gamma \end{smallmatrix}]\|^2}$$

- Define:

$$\begin{aligned} G &:= [A \quad -e]'[A \quad -e] + \delta I, \\ H &:= [B \quad -e]'[B \quad -e] + \delta I, \\ z &:= \begin{bmatrix} w \\ \gamma \end{bmatrix}, \end{aligned}$$

Generalized Eigenvalue Problem

The optimization problem reduces to minimizing the Rayleigh quotient:

$$\min_{z \neq 0} r(z) := \frac{z'Gz}{z'Hz}.$$

- The Rayleigh quotient ranges over the interval $[\lambda_1, \lambda_{n+1}]$ for $\|z\|_2 = 1$.
- λ_1 and λ_{n+1} are the minimum and maximum eigenvalues of the generalized eigenvalue problem:

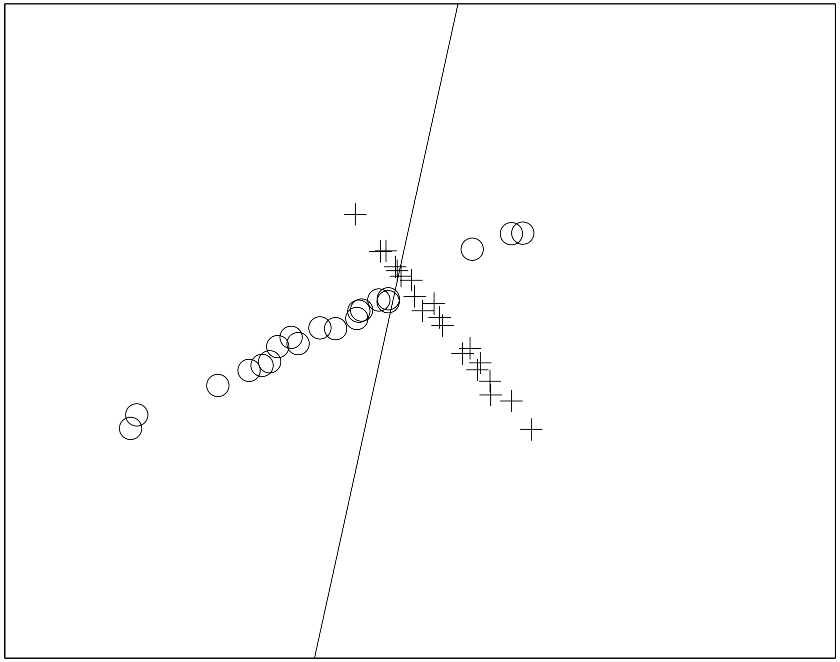
$$Gz = \lambda H z, \quad z \neq 0.$$

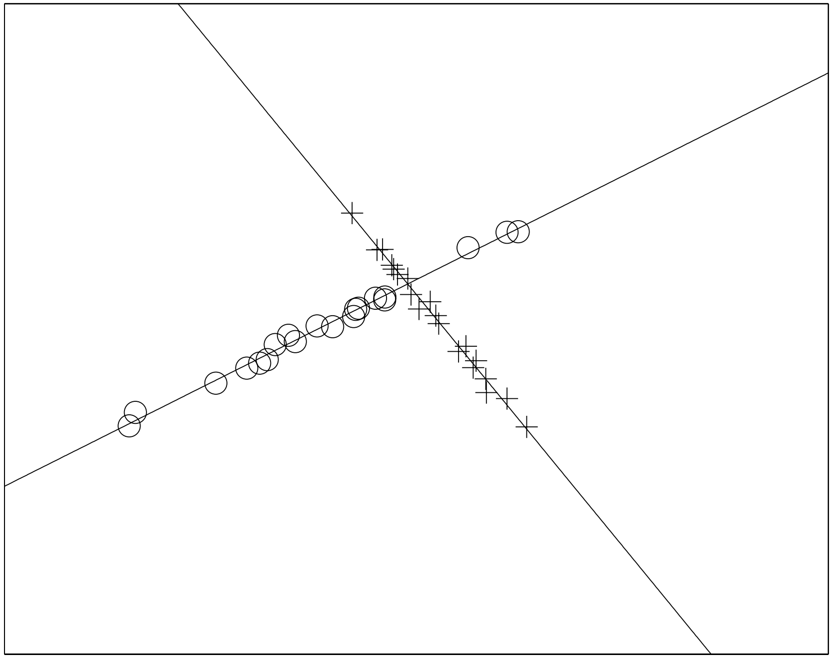
The eigenvectors z^1 corresponding to the smallest eigenvalue λ_1 and z^{n+1} corresponding to the largest eigenvalue λ_{n+1} determine the two nonparallel proximal planes. $eig(G, H)$

A Simple Example

Linear Classifier

Generalized Eigenvalue Classifier





80% Correctness

100% Correctness

Also applied successfully to real world test problems

Conclusion

- Variety of optimization-based approaches to data mining
  - Feature selection in both clustering & classification
  - Enhanced knowledge-based classification
  - Finite Newton method for nonlinear classification
  - Drug discovery based on gene macroarrays
  - Proximal classifaction via generalized eigenvalues

- Optimization is a powerful and effective tool for data mining, especially for implementing Occam's Razor
  - "Simplest is best"