Smooth Support Vector Machines

Yuh-Jye Lee

University of Wisconsin – Madison

www.cs.wisc.edu/˜yuh-jye

Overall Outline

Part I: SSVM Theory and Algorithm

Part II: Application to Breast Cancer Prognosis

Part I Outline
SSVM Theory and Algorithm

- The linear (kernel) support vector machine (SVM)
  - Quadratic programming formulation
  - Unconstrained smooth minimization formulation (SSVM)
- The generalized (nonlinear kernel) support vector machine (GSVM)
  - QP and smooth unconstrained minimization formulations
- Newton-Armijo algorithm for SSVM
  - Globally & quadratically convergent
- Numerical tests
  - Highest test set accuracy on 6 public datasets
  - Fastest on 5 out of 6 datasets
  - Capable of accurate & sharp nonlinear separation
- Summary and future work

What is a Support Vector Machine (SVM)?

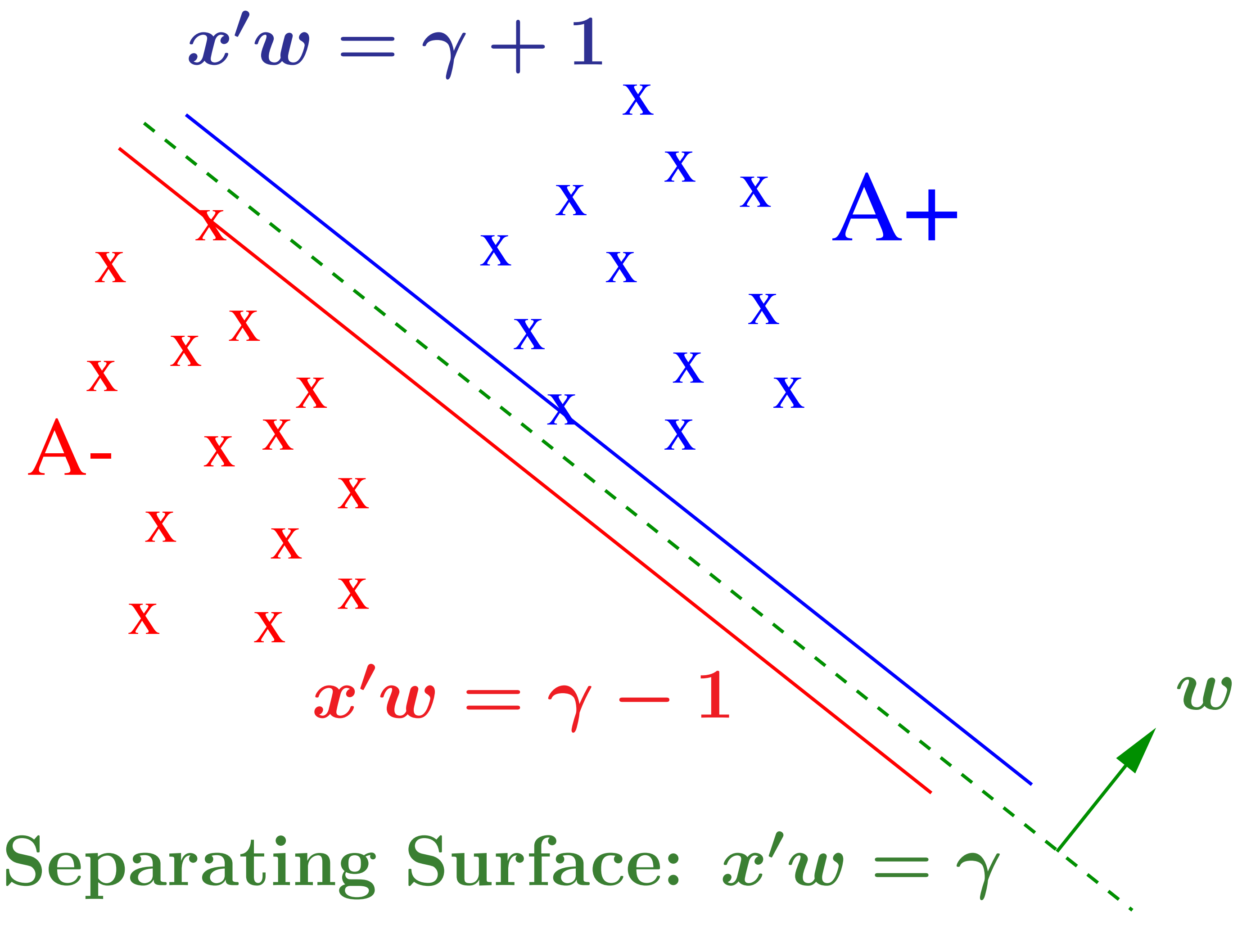
- **An optimally defined surface**
- **Typically nonlinear in the input space**
- **Linear in a higher dimensional "feature" space**
- **Implicitly defined by a kernel function**

What are Support Vector Machines Used For?

- **Classification**
- **Regression & Data Fitting**
- **Supervised & Unsupervised Learning**

Simplest SVM is a linear one, which is an explicitly defined plane in the input space!

Geometry of the Classification Problem
The Fundamental 2-Category Strictly Linearly Separable Case



Algebra of the Classification Problem
The Fundamental 2-Category Strictly Linearly Separable Case

- Given m points in the n dimensional real space R^n
- Represented by an $m \times n$ matrix A
- Membership of each point A_i in the classes 1 or -1 is specified by:
  - An $m \times m$ diagonal matrix D with ± 1 along its diagonal
- Separate by two bounding planes: $x'w = \gamma \pm 1$ such that:

$$\begin{aligned} A_i w &\geq \gamma + 1, \text{ for } D_{ii} = +1, \\ A_i w &\leq \gamma - 1, \text{ for } D_{ii} = -1. \end{aligned}$$

- More succinctly:

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

where e is a vector of ones.

Preliminary Approach to the (Linear) Support Vector Machine
Robust Linear Programming (RLP)
"Margin" of Separation not Maximized

Solve the following mathematical program:

$$\begin{array}{rrcl} \min\limits_{w,\gamma,y} & e'y & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e \\ & y & \geq & 0, \end{array} \qquad (RLP)$$

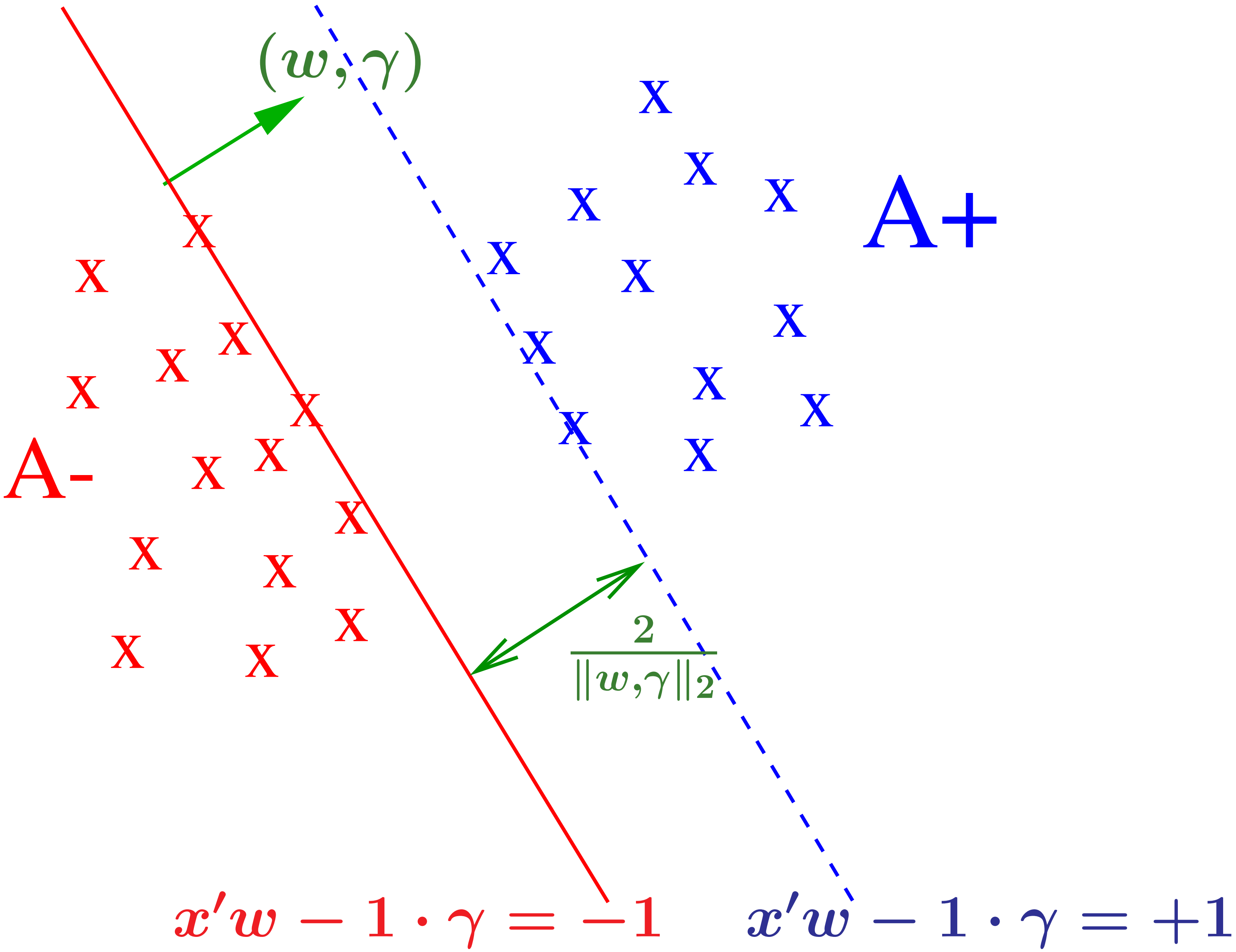
where:

$$y \; := \; \text{nonnegative error (slack) vector}$$

Note: $y = 0$ iff convex hulls of $A+$ and $A-$ do not intersect.

The (Linear) Support Vector Machine

Maximize Margin between Bounding Planes



The (Linear) Support Vector Machine Formulation
Quadratic Programming Formulation

Solve the following mathematical program for some $\nu > 0$:

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

Other norms and functions lead to other formulations:
$SVM_{\|\cdot\|_1}$, $SVM_{\|\cdot\|_2^2}$ & FSV.

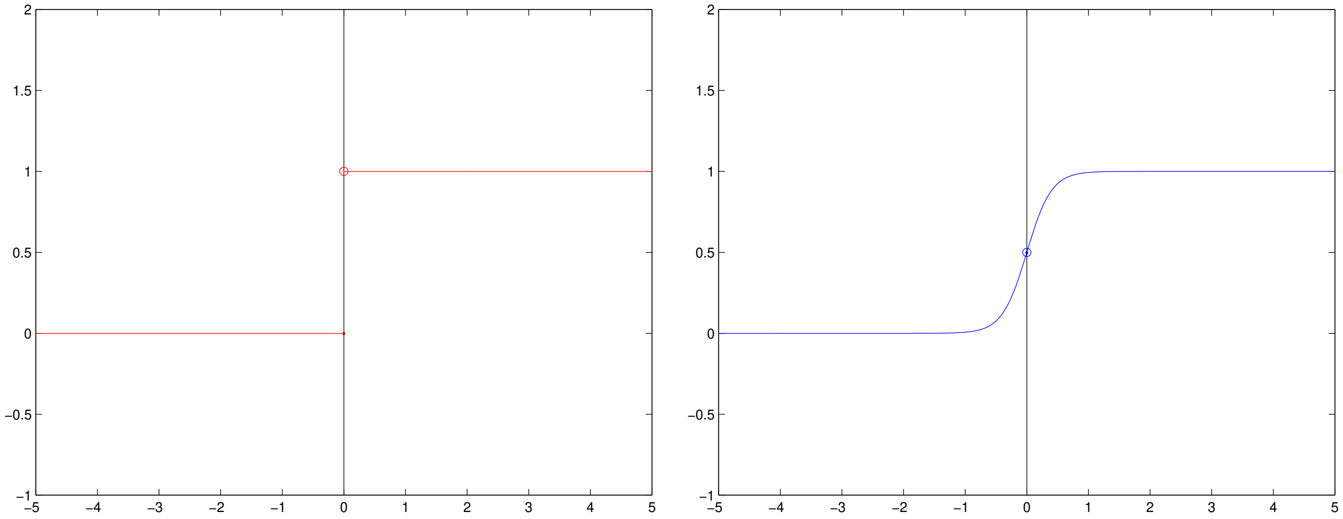
At solution of QP:

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

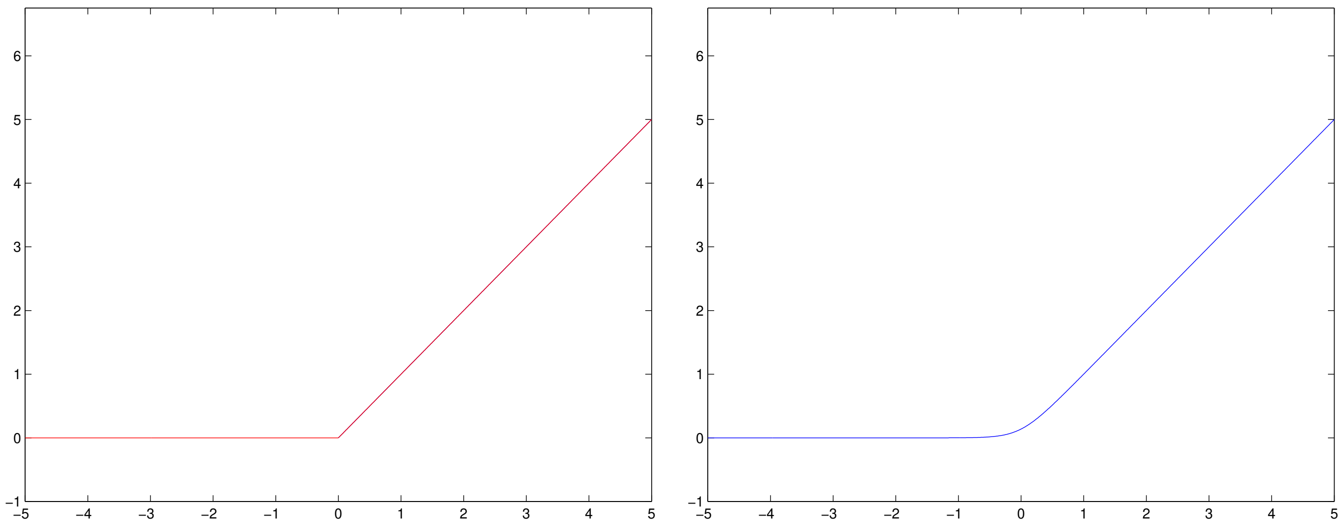
where $(\cdot)_+ = \max\{\cdot, 0\}$**. Hence QP is equivalent to the nonsmooth SVM:**

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

The Step Function $(x)_*$ and the Sigmoid-Function $\frac{1}{1+\varepsilon^{-\alpha x}}$



The Plus Function $(x)_+$ and the p-Function $p(x,5)$



SSVM: The Smooth Support Vector Machine
Smoothing the Plus Function

Integrating a standard smooth approximation of the step function, the sigmoid function of neural networks:

$$s(x,\alpha) = \frac{1}{1+\varepsilon^{-\alpha x}},$$

gives a smooth and excellent approximation to the plus function:

$$p(x,\alpha) = x + \frac{1}{\alpha}\log(1+\varepsilon^{-\alpha x}),\ \alpha > 0.$$

Replacing the plus function $(\cdot)_+$ in the nonsmooth SVM by the smooth $p(\cdot,\alpha)$, gives our SSVM:

$$\min_{(w,\gamma)\in R^{n+1}} \Phi_\alpha(w,\gamma) \;:=\; \min_{(w,\gamma)\in R^{n+1}} \frac{\nu}{2}\|p(e-D(Aw-e\gamma),\alpha)\|_2^2 + \frac{1}{2}\|w,\gamma\|_2^2.$$

Relating the Nonsmooth and Smooth SVMs

Theorem Let $A \in R^{m\times n}$ and $b \in R^{m\times 1}$. Define the real valued functions $f(x)$ and $g(x,\alpha)$ for $x \in R^n$ and $\alpha > 0$:

$$f(x) = \frac{1}{2}\|(Ax-b)_+\|_2^2 + \frac{1}{2}\|x\|_2^2$$

and

$$g(x,\alpha) = \frac{1}{2}\|p((Ax-b),\alpha)\|_2^2 + \frac{1}{2}\|x\|_2^2.$$

Then:

(i) There exists a unique solution $\bar{x}$ of $\min\limits_{x\in R^n} f(x)$ and a unique solution $\bar{x}_\alpha$ of $\min\limits_{x\in R^n} g(x,\alpha)$.

(ii) For all $\alpha > 0$, we have the following error bound:

$$\|\bar{x}_\alpha - \bar{x}\|_2^2 \le \frac{m}{2}((\frac{\log 2}{\alpha})^2 + 2\xi\frac{\log 2}{\alpha}), \qquad (*)$$

where the constant ξ is defined as:

$$\xi = \max_{1\le i\le m} |(A\bar{x}-b)_i|.$$

Thus, $\bar{x}_\alpha$ converges to $\bar{x}$ as α goes to infinity with an error bound given by (*).

A Newton-Armijo Algorithm for the SSVM: $\min_{(w,\gamma)\in R^{n+1}} \Phi_\alpha(w,\gamma)$

Start with any $(w^0,\gamma^0) \in R^{n+1}$. Having (w^i,γ^i), stop if the gradient of the objective function is zero. Else compute (w^{i+1},γ^{i+1}) as follows:

(i) Newton Direction: Determine direction $d^i \in R^{n+1}$ by setting equal to zero the linearization of $\nabla\Phi_\alpha(w,\gamma)$ around (w^i,γ^i) which gives $n+1$ linear equations in $n+1$ variables:

$$\nabla^2\Phi_\alpha(w^i,\gamma^i)d^i = -\nabla\Phi_\alpha(w^i,\gamma^i)'.$$

(ii) Armijo Stepsize: Choose a stepsize $\lambda_i \in R$ such that:

$$(w^{i+1},\gamma^{i+1}) = (w^i,\gamma^i) + \lambda_i d^i$$

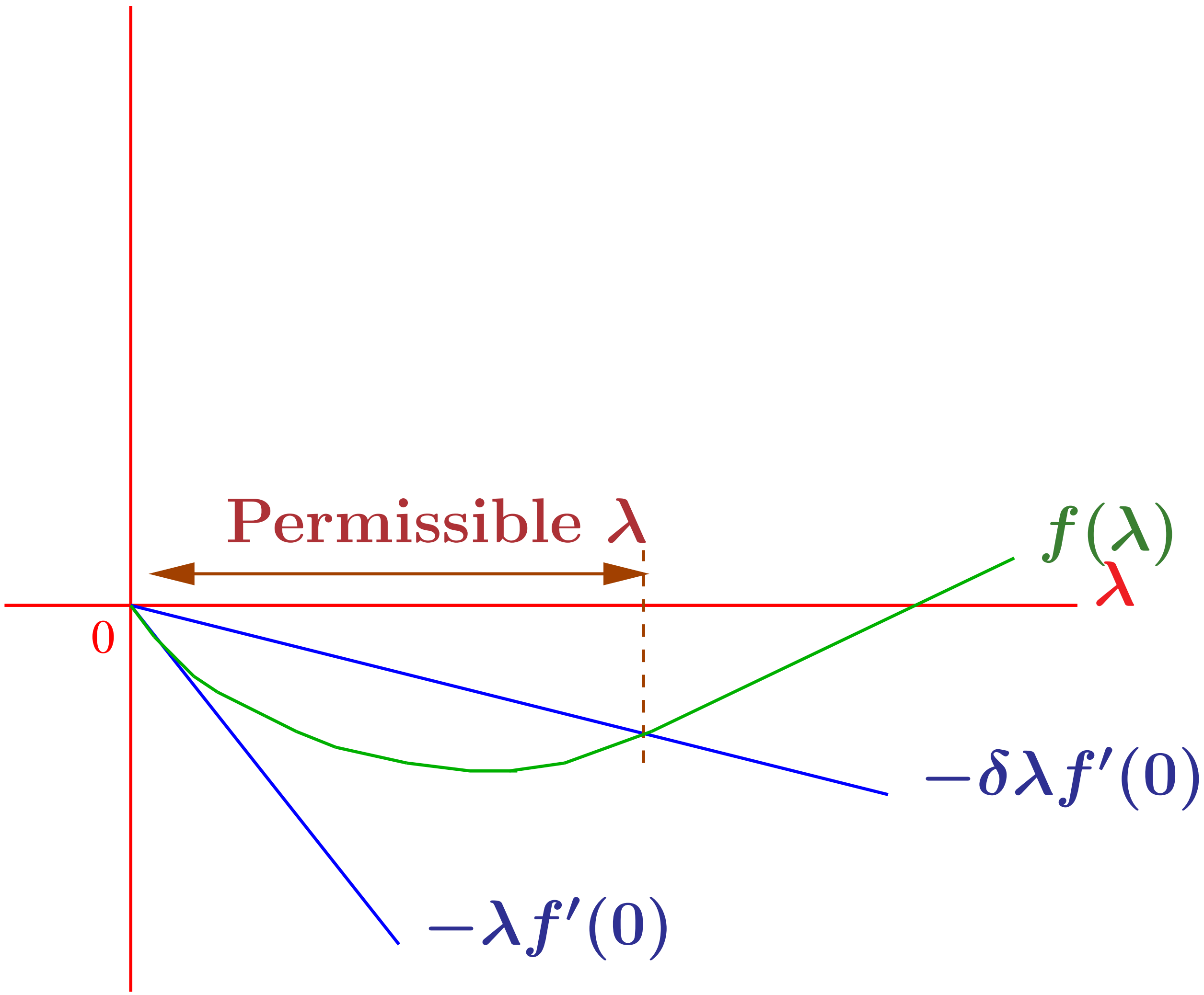
where $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \ldots\}$ such that:

$$\Phi_\alpha(w^i,\gamma^i) - \Phi_\alpha((w^i,\gamma^i) + \lambda_i d^i) \geq -\delta\lambda_i\nabla\Phi_\alpha(w^i,\gamma^i)d^i,$$

where $\delta \in (0, \frac{1}{2})$.

Armijo Stepsize

$$f(\lambda) := \Phi_\alpha(w^i, \gamma^i) - \Phi_\alpha((w^i, \gamma^i) + \lambda d^i)$$
$$f'(0) = -\nabla\Phi_\alpha(w^i, \gamma^i) d^i$$



Global Quadratic Convergence of the Newton-Armijo Algorithm

Theorem Let $\{(w^i, \gamma^i)\}$ be a sequence generated by the Newton-Armijo Algorithm and $(\bar{w}, \bar{\gamma})$ be the unique solution of the smooth support vector machine SSVM problem.

(i) The sequence $\{(w^i, \gamma^i)\}$ converges to the unique solution $(\bar{w}, \bar{\gamma})$ from any initial point (w^0, γ^0) in R^{n+1}.

(ii) For any initial point (w^0, γ^0), there exists an integer $\bar{i}$ such that the stepsize λ_i of the Algorithm equals 1 for $i \geq \bar{i}$ and the sequence $\{(w^i, \gamma^i)\}$ converges to $(\bar{w}, \bar{\gamma})$ Q-quadratically.

GSVM: The Generalized Support Vector Machine
Motivation & Quadratic Programming Formulation

Linear Support Vector Machine: (Linear separating surface: $x'w = \gamma$)

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

By "duality", set $w = A'Du$: (Linear separating surface: $x'A'Du = \gamma$)

$$\begin{array}{rl} \min\limits_{y \geq 0, u, \gamma} & \frac{\nu}{2}\|y\|_2^2 + \frac{1}{2}\|A'Du, \gamma\|_2^2 \\ \text{s.t.} & D(AA'Du - e\gamma) + y \;\geq\; e. \end{array}$$

Nonlinear Support Vector Machine: (Replace AA' by a nonlinear kernel $K(A, A') : R^{m \times n} \times R^{n \times m} \longrightarrow R^{m \times m}$, and $\|A'Du, \gamma\|_2^2$ by a convex function of (u, γ), typically a norm. Nonlinear separating surface: $K(x', A')Du = \gamma$)

$$\begin{array}{rl} \min\limits_{y \geq 0, u, \gamma} & \frac{\nu}{2}\|y\|_2^2 + \frac{1}{2}\|u, \gamma\|_2^2 \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y \;\geq\; e. \end{array} \qquad (GSVM)$$

SSVM with a Nonlinear Kernel
Nonlinear Separating Surface

Thus by using a nonlinear kernel function: $K(X,Y) : R^{m\times n} \times R^{k\times \ell} \longrightarrow R^{m\times \ell}$ we generate a nonlinear separating surface:

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

determined by solving the quadratic program:

$$\begin{array}{rl} \min\limits_{y\geq 0,\ u,\gamma} & \frac{\nu}{2}\|y\|_2^2 + \frac{1}{2}\|u,\ \gamma\|_2^2 \\ \text{s.t.} & D(K(A,A')Du - e\gamma) + y \ \geq \ e \end{array}$$

Since at a solution of the above QP, $y = (e - (D(K(A,A')Du - e\gamma))_+$, hence the QP is equivalent to the unconstrained problem:

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

Replacing the plus function by its smooth approximation the p function, gives our SSVM with a nonlinear kernel as follows:

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

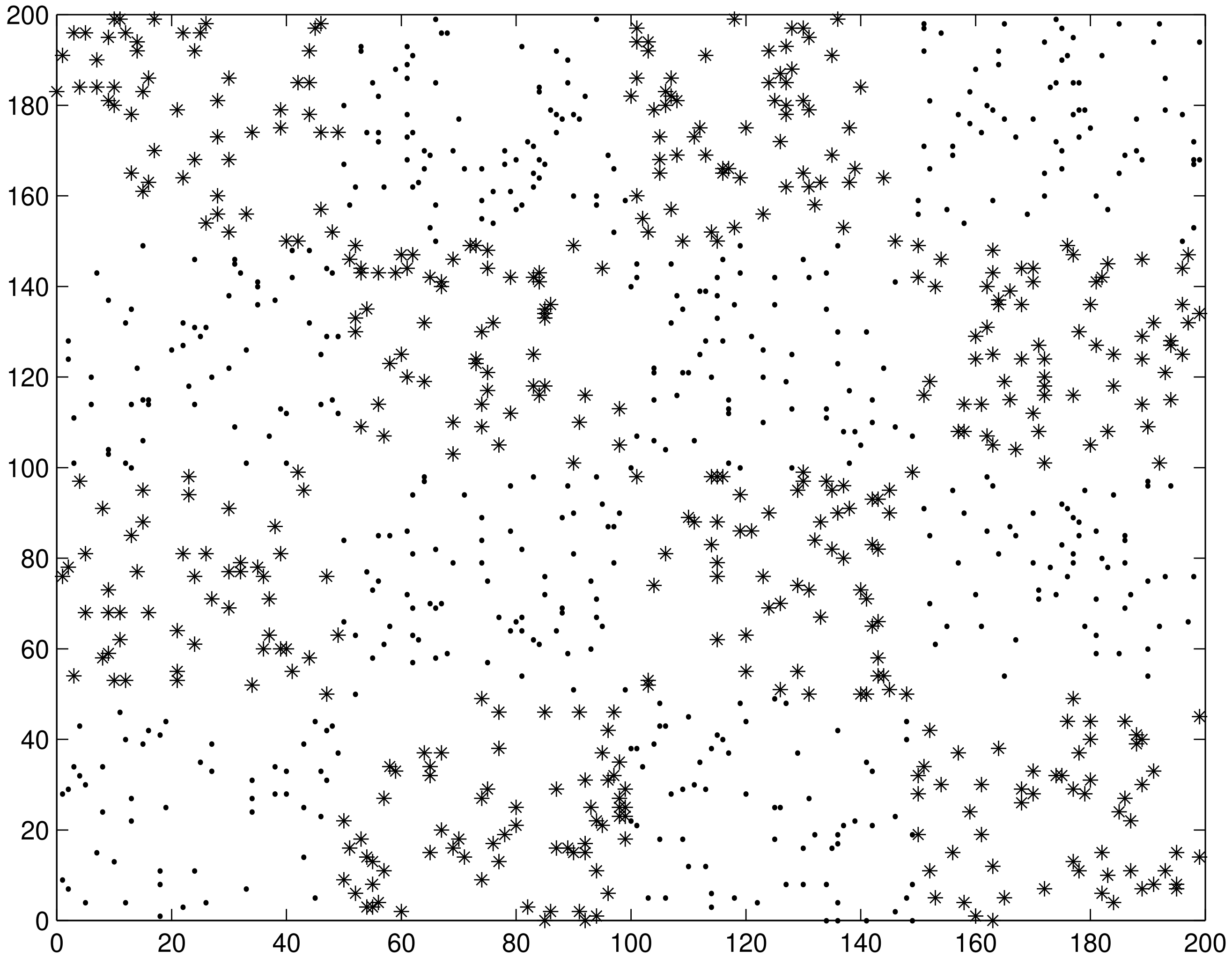
Examples of Kernels

- Examples ($A \in R^{m \times n}$, $a \in R^m$, $\mu \in R$ and d integer):
  - Polynomial Kernel: $(AA' + \mu aa')^d_{\bullet}$,
    where $[\cdot]^d_{\bullet}$ denotes component-wise exponentiation as in MATLAB.
  - Radial Basis (Gaussian) Kernel: $\varepsilon^{-\mu\|A_i - A_j\|^2}$, $i, j = 1, \ldots, m$.
  - Neural Network Kernel: $(AA' + \mu aa')_{\bullet *}$,
    where $[\cdot]_{\bullet *}$ denotes the step function: $R \longrightarrow \{0, 1\}$ component-wise.

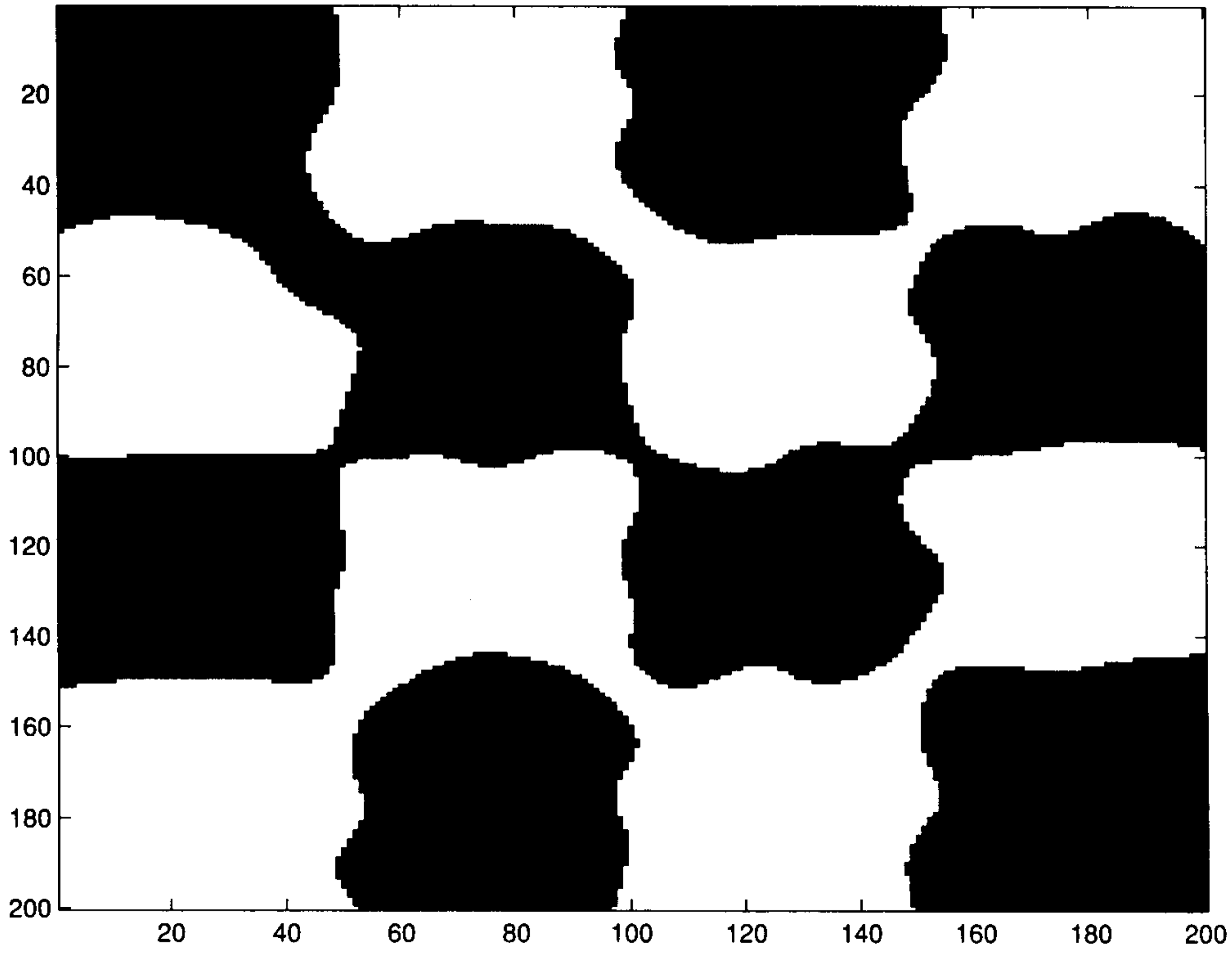
A Nonlinear Kernel Application

Checkerboard Training Set: 1000 Points in R^2

Separate 486 Asterisks from 514 Dots

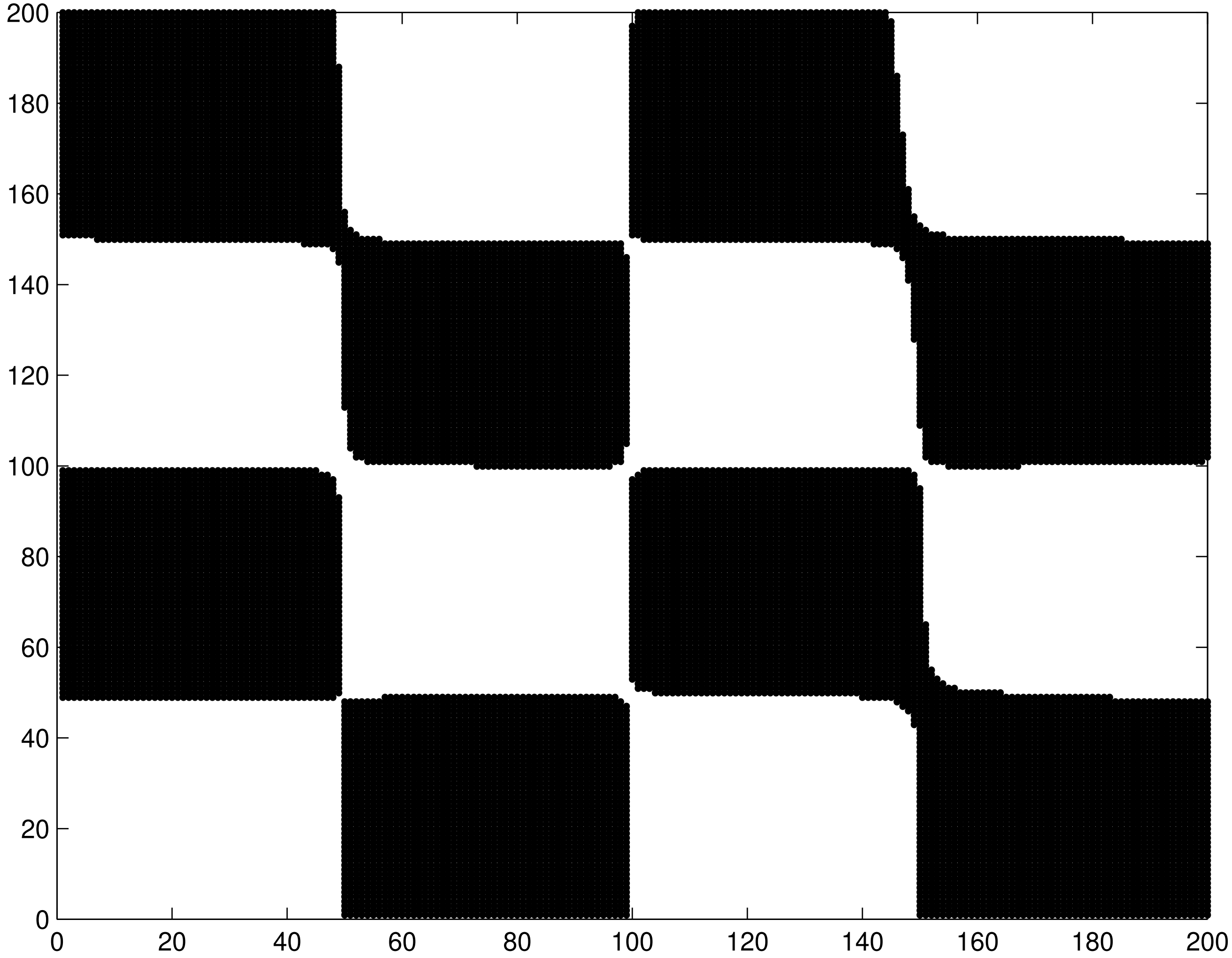


Checkerboard Gaussian Kernel Separation
Best Previously Available Result [Kaufman 1998]



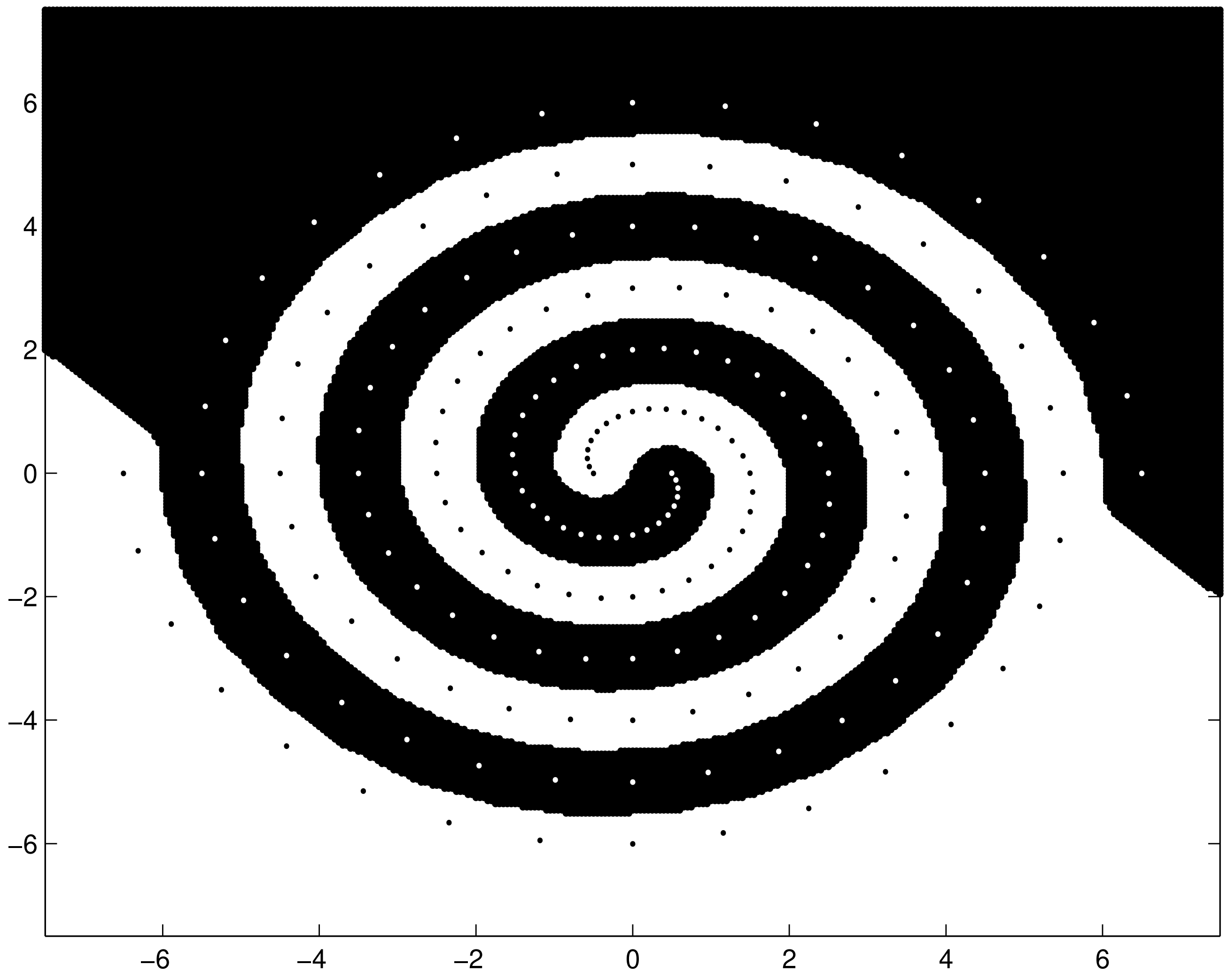
Checkerboard Gaussian Kernel Separation

Gaussian Kernel $K(A, A') = \varepsilon^{-2\|A_i - A_j\|_2^2}, \ i, j = 1, 2, \ldots, 1000.$



Two-spiral Gaussian Kernel Separation

Gaussian Kernel $K(A, A') = \varepsilon^{-4\|A_i - A_j\|_2^2}, \;\; i, j = 1, 2, \ldots, 194.$



Numerical Testing
Comparison of SSVM with Other SVMs

DATA SETS:

- Four sets from the UC Irvine Repository
- Two breast cancer prognosis sets from UW Datasets
- All publicly available

OTHER SVMs

- RLP: Robust linear programming
- $SVM_{\|\cdot\|_1}$: SVM with $\|\cdot\|_\infty$ margin
- $SVM_{\|\cdot\|_2^2}$: SVM with $\|\cdot\|_2$ margin
- FSV: SVM Conca<u>v</u>e – Support Vector Suppression

MACHINES:

- Cluster of 4 Sun Enterprise E6000 machines each consisting of:
  - 16 UltraSPARC II 250 MHz Processors with 2 Gig RAM,

  Total: 64 Processors, 8 Gig RAM

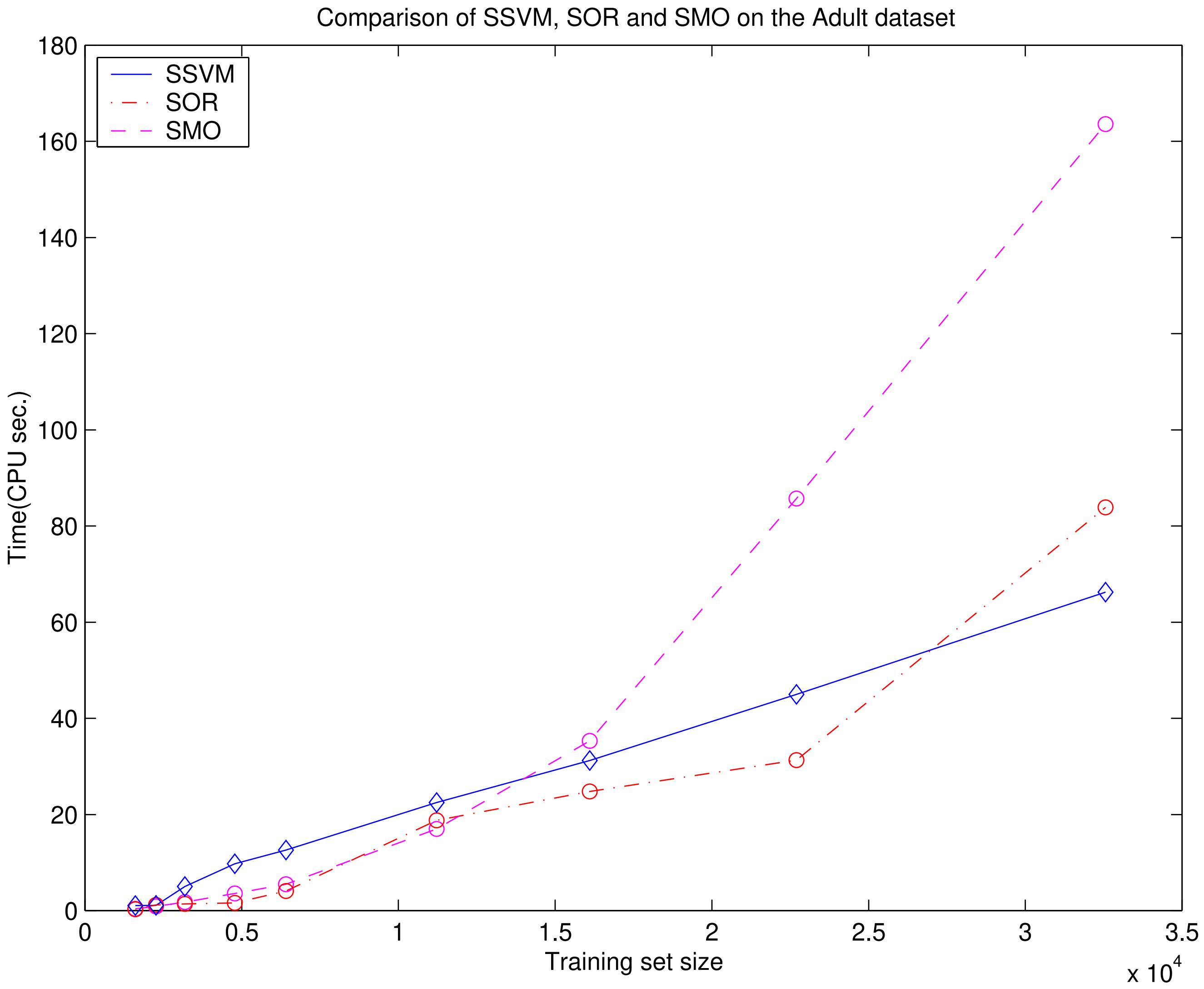
Comparison of SSVM with Other SVMs

Ten-Fold Cross Validation Testing Correctness % Red=Best Ten-Fold Cross Validation Running Time *Sec* Red=Best					
Dataset	Method				
$m \times n$	$SSVM$	RLP	$SVM_{\|\cdot\|_1}$	$SVM_{\|\cdot\|_2^2}$	FSV
WPBC(24 months) 155×32	83.47 4.05	67.12 6.47	71.08 6.25	75.46 -	66.42 -
WPBC(60 months) 110×32	68.18 3.15	63.50 2.81	66.23 3.72	66.21 -	67.05 -
Ionosphere 351×34	89.63 8.84	86.04 10.38	86.10 42.41	85.75* -	84.07 -
Cleveland 297×13	86.13 4.61	83.87 12.74	84.55 18.71	75.86* -	80.94 -
Pima Indians 768×8	78.12 9.93	76.16 195.21	74.47 286.59	* -	74.60 -
BUPA Liver 345×6	70.33 4.19	64.34 17.26	64.03 19.94	60.95* -	65.20 -

Comparison of SSVM with Other Methods on Adult Dataset

Testing Correctness % Red=Best Running Time *Sec* Red=Best					
Training Set	Method				
Test Set	*SSVM*	*SOR*	*SMO*	SVM^{light}	*RLP*
1605×123 30957×123	**84.27** 1.0	84.06 **0.3**	- 0.4	84.25 5.4	78.68 9.9
3185×123 29377×123	**84.63** 5.0	84.23 **1.4**	- 1.8	84.40 21.0	77.83 80.1
6414×123 26148×123	**84.59** 12.6	84.30 **4.1**	- 5.5	84.43 87.6	71.85 218.8
16101×123 16461×123	**84.97** 31.2	84.62 **24.8**	- 35.3	84.83 667.2	72.52 632.6
22697×123 9865×123	**85.35** 44.9	85.06 **31.3**	- 85.7	85.17 1425.6	77.43 991.9
32562×123 16282×123	85.01 **66.3**	84.96 83.9	- 163.6	**85.05** 2184.0	83.25 1561.1

Comparison of SSVM, SOR and SMO on Adult Dataset



Summary and Future Work

- SSVM: A new formulation of support vector machines as a smooth unconstrained minimization problem
- A very fast Newton-Armijo algorithm is faster than standard quadratic programming formulations
- SSVM is equally efficient on linear and nonlinear kernels
- SSVM generalizes better on all test problems than other SVMs
- Future work:
  - Support vector regression via SSVMs
  - Massive classification and regression via parallel SSVMs
  - Column and row chunking of SSVMs for massive problems
  - Support vector selection

Part II Outline

Application to Breast Cancer Prognosis

- **Statistical tools of survival analysis**
  - **Kaplan-Meier survival curve**
  - **Logrank test**
- **Breast cancer prognosis**
  - **Lymph node status as surrogate for chemotherapy prescription**
  - **Feature selection via $SVM_{\|\cdot\|_1}$**
    - **Extract 6 critical features out of 31 features for prognosis**
  - **Classification via GSVM with Gaussian kernel**
    - **Classify 253 patients into three prognosis groups :**
      Good, Intermediate and Poor
- **Summary and future work**

Survival Curve Estimation

S_{t_j} = Kaplan-Meier estimate of survival probability at j^{th} observed time

$$S_{t_j} = \begin{cases} 1, & j = 0 \\ S_{t_{j-1}}(1 - \text{risk}_{t_j}), & j = 1, 2, \ldots \end{cases}$$

$$\begin{aligned} \text{risk}_{t_j} &= \text{instantaneous risk of death} \\ &\approx \frac{\text{no. of new patients died at time } t_j}{\text{no. of patients remaining in patient pool at time } t_j} \end{aligned}$$

Kaplan-Meier survival curve

- Plotted as a step function with initial survival probability of $S_{t_0} = 1$
- Step down to the next survival probability as we move to the next observed time

Logrank Statistic for Two Groups

Provides an *overall* comparison of different Kaplan-Meier survival curves

- **Notation**
  - n_{ij} : the number of patients at j^{th} observed time of group $i, i = 1, 2$
  - m_{ij} : the number of deaths at j^{th} observed time of group $i, i = 1, 2$
  - e_{ij} : the *expected* number of deaths at j^{th} observed time of group i.
    For example: for $i = 1$

$$e_{1j} = \frac{n_{1j}}{n_{1j} + n_{2j}} \times (m_{1j} + m_{2j})$$

  - $O_i - E_i$: the total difference between *observed* and *expected* number

$$O_i - E_i = \sum_j (m_{ij} - e_{ij}), \quad i = 1, 2$$

- **Logrank statistic :**

$$\frac{(O_1 - E_1)^2}{\mathrm{Var}(O_1 - E_1)} \sim \chi^2_1$$

Pairwise Survival Curve Comparison

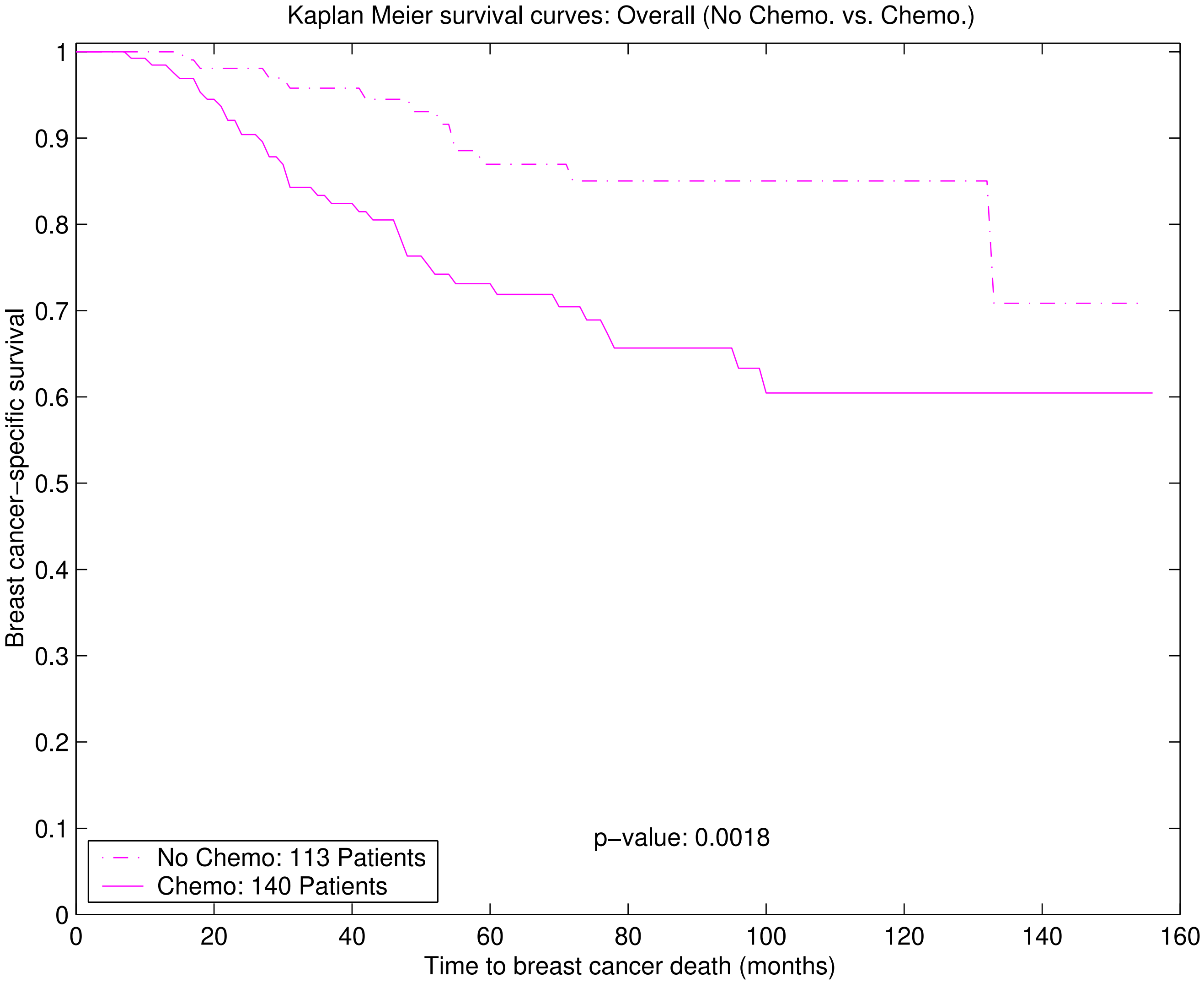
- **Hypothesis**
  - H_0 : No difference between survival curves
- **Testing statistic**
  - Logrank statistic $\sim \chi^2$ with 1 df under H_0
- ***p*-value**
  - Probability that two survival curves are the same based on the logrank statistic
  - Small p-value $\Longleftrightarrow$ Large logrank statistic
  - Typically reject H_0 if p-value ≤ 0.05 ($\Longleftrightarrow$ logrank statistic ≥ 7.89)

Overall Survival Curves for 253 Breast Cancer Patients

Two Groups Classified by Therapy:

Dashes: No Chemotherapy (113) - Solid: Chemotherapy (140)

p-value = 0.0018



Breast Cancer Prognosis Analysis & Chemotherapy

- Feature selection
  - Use surgically-obtained lymph node status (a surrogate for chemotherapy treatment decision) as a classification criterion
  - Linear $SVM_{\|\cdot\|_1}$ automatically extracts six features:
    - Five out of thirty nuclear features obtained during diagnosis
    - Tumor size obtained from surgery
- Classification
  - Apply nonlinear (Gaussian) kernel on these six selected features
  - Use surgically-obtained lymph node status (0, 1 - 4, 5 -30) as a classification criterion
  - Using GSVM (nonlinear $SVM_{\|\cdot\|_1}$) classify patients into three prognostic groups: good, intermediate and poor
- Prognosis
  - Construct Kaplan-Meier survival curves for:
    - Each of the three classes
    - Within each class construct survival curves for patients taking and not taking chemotherapy

Feature Selection by a Linear Support Vector Machine $SVM_{\|\cdot\|_1}$
Classification Criterion: "Lymph Node Positive or Negative"

Linear Support Vector Machine:

$$\begin{array}{rrcl} \min\limits_{y\geq 0, w, \gamma} & \nu e'y + \|w\|_1 & & \\ \text{s.t.} & D(Aw - e\gamma) + y & \geq & e. \end{array} \qquad (SVM_{\|\cdot\|_1})$$

$*$ Linear separating surface: $x'w = \gamma$

Features Selected by $SVM_{\|\cdot\|_1}$

- The following 6 features out of 31 were selected:
  1. Mean Area
  2. Standard Error of Area
  3. Worst Area
  4. Worst Texture
  5. Worst Perimeter
  6. Tumor Size (from surgery)

Classification by a Nonlinear GSVM with a Gaussian Kernel
(“Nearest Neighbor” Flavor)

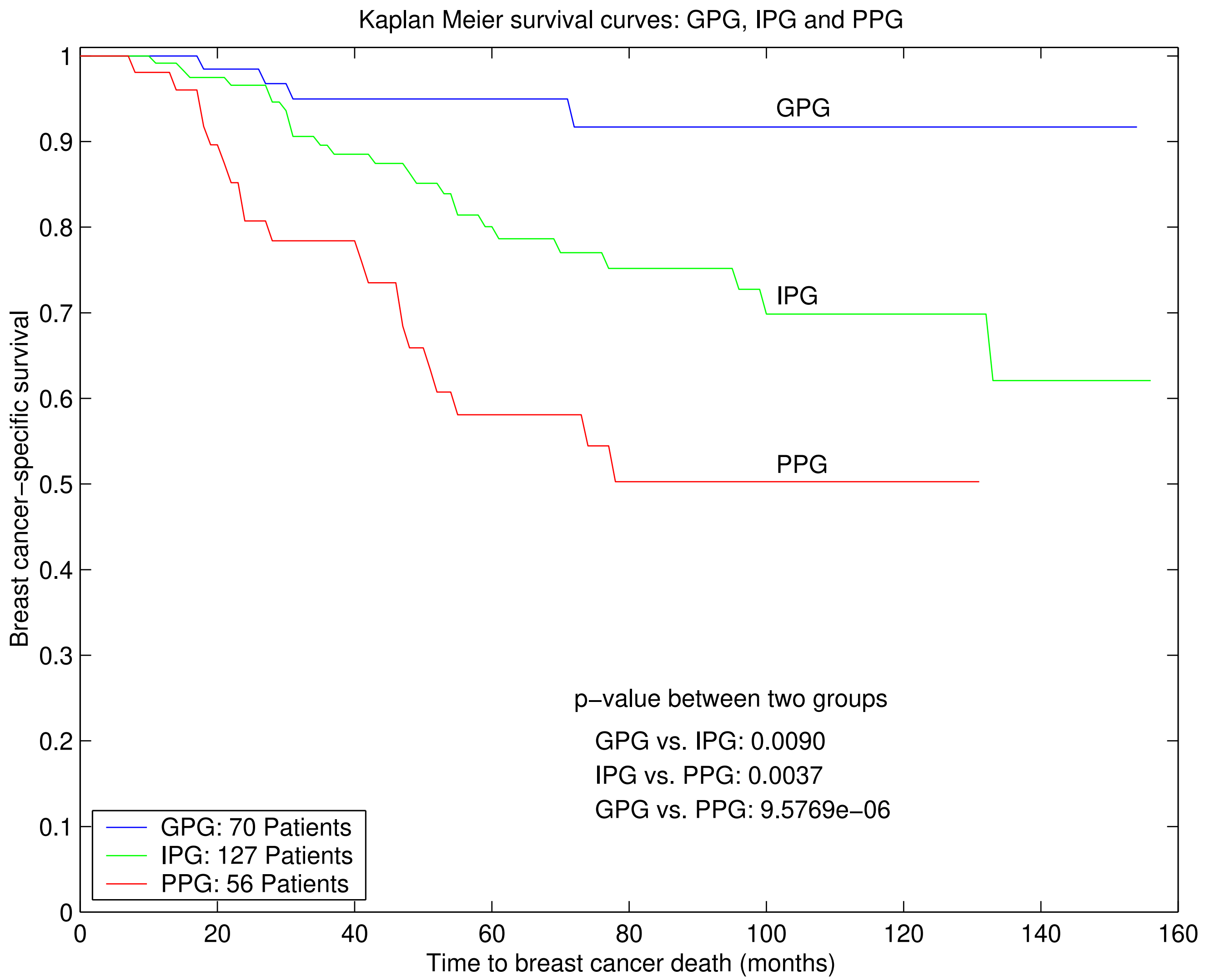
$SVM_{\|\cdot\|_1}$ **with a Gaussian Kernel:**

$$\begin{array}{rrcl} \displaystyle\min_{y\geq 0,u,\gamma} & \nu e'y + \|u\|_1 & & \\ \text{s.t.} & D(K(A,A')Du - e\gamma) + y & \geq & e. \end{array} \qquad (GSVM)$$

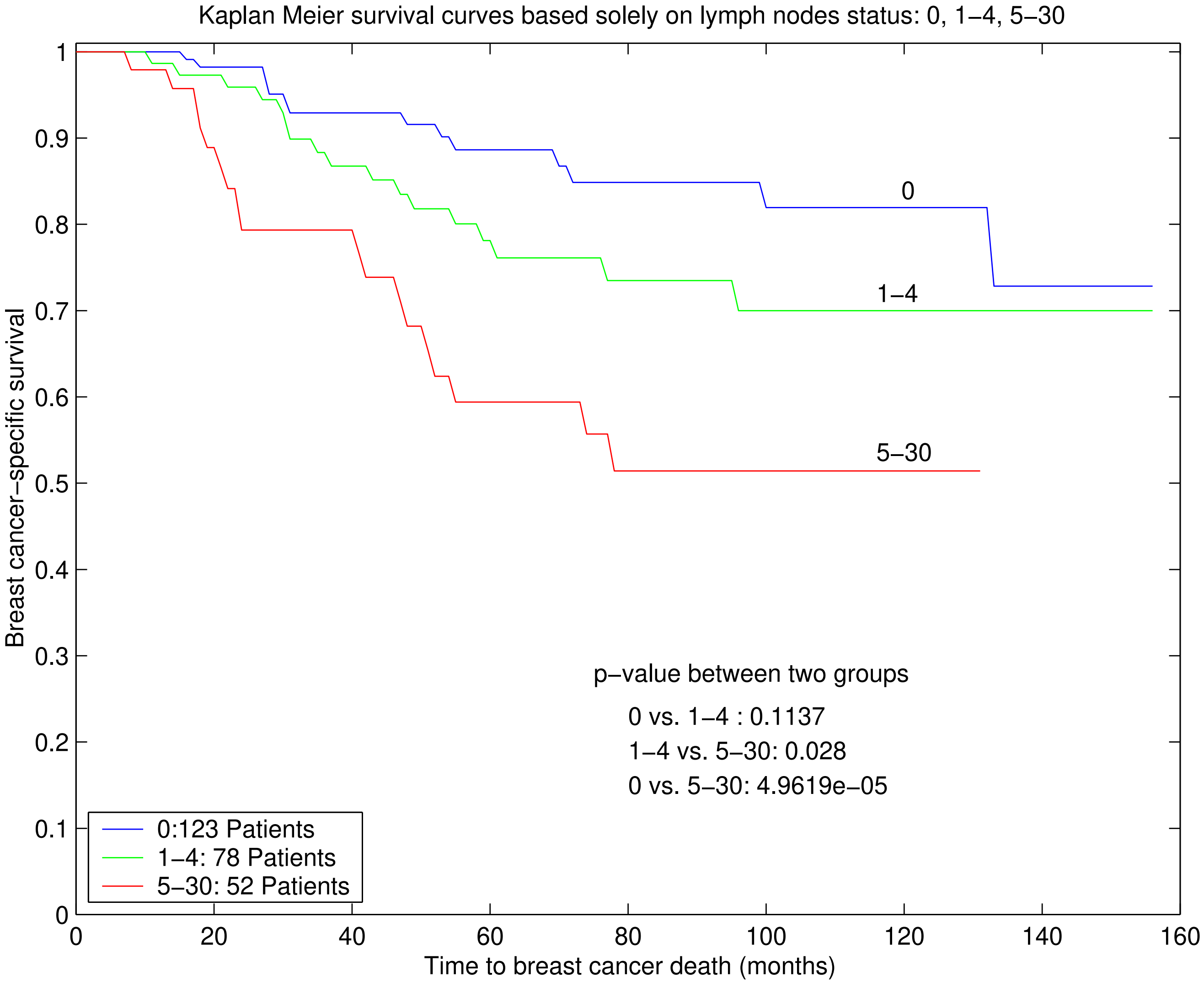
$*$ **Nonlinear separating surface:** $K(x', A')Du = \gamma$

- **Using the six features selected by $SVM_{\|\cdot\|_1}$, the GSVM achieved the following separation:**
  1. **Good (blue) prognosis group (0 lymph node metastasis) was separated from the rest for which:**
  2. **Intermediate (green) prognosis group (1-4 lymph node metastasis) was separated from**
  3. **Poor (red) prognosis group (lymph node metastasis > 4).**
- **Tenfold cross validation correctness: 73%**

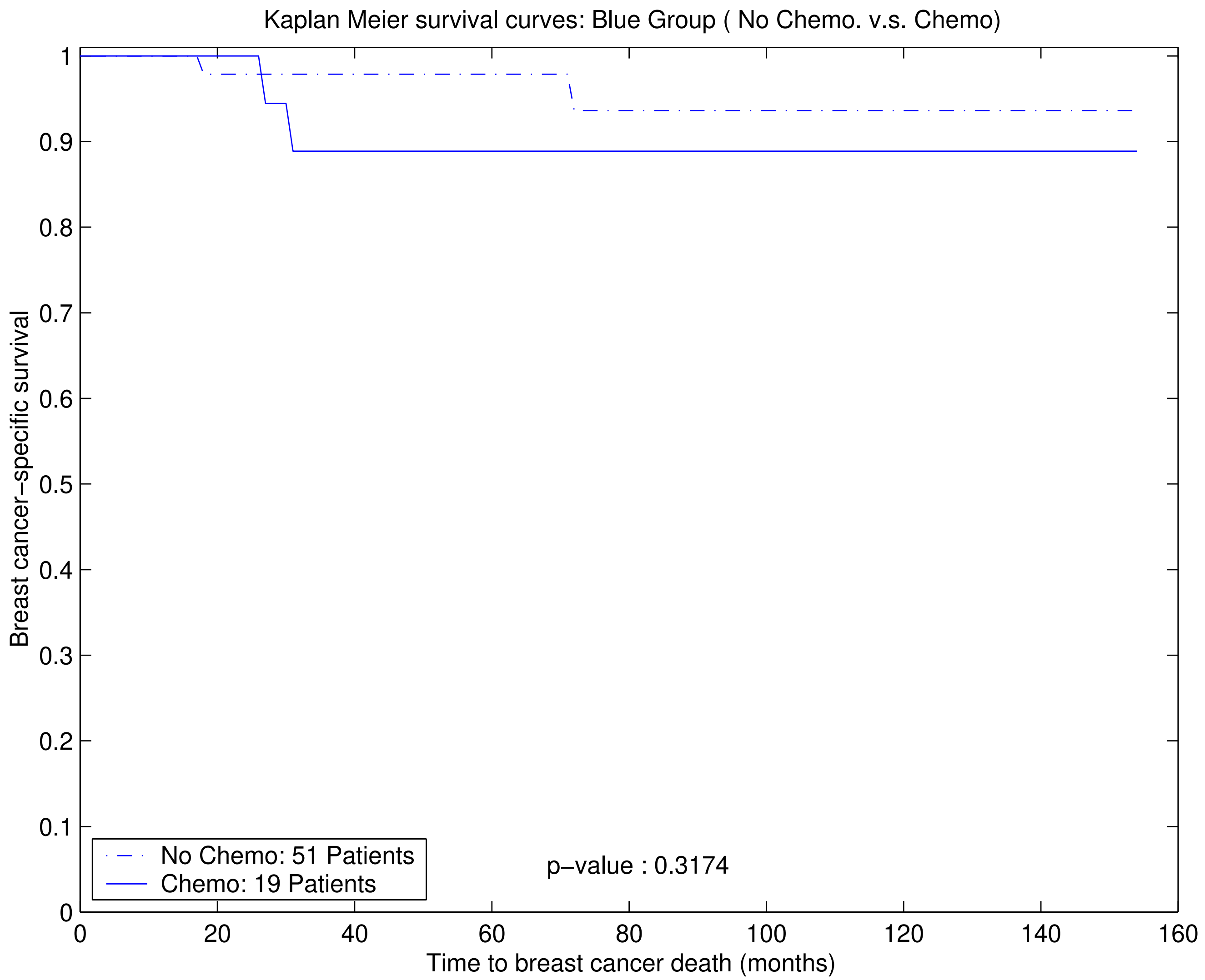
Survival Curves for 253 Breast Cancer Patients
Three Groups Classified by GSVM Using Six Predictive Features:
Good (70), Intermediate (127) & Poor (56)
p-values: BG=0.0090, RG =0.0037, BR=9.5769e-06



Survival Curves for 253 Breast Cancer Patients
Three Groups Classified by Lymph Node Status Solely:
0 nodes (123), 1-4 nodes (78) & 5-30 nodes (52)
p-values: 0 vs. 1-4 =0.1137, 5-30 vs.1-4 =0.028, 0 vs. 5-30=4.9619e-05



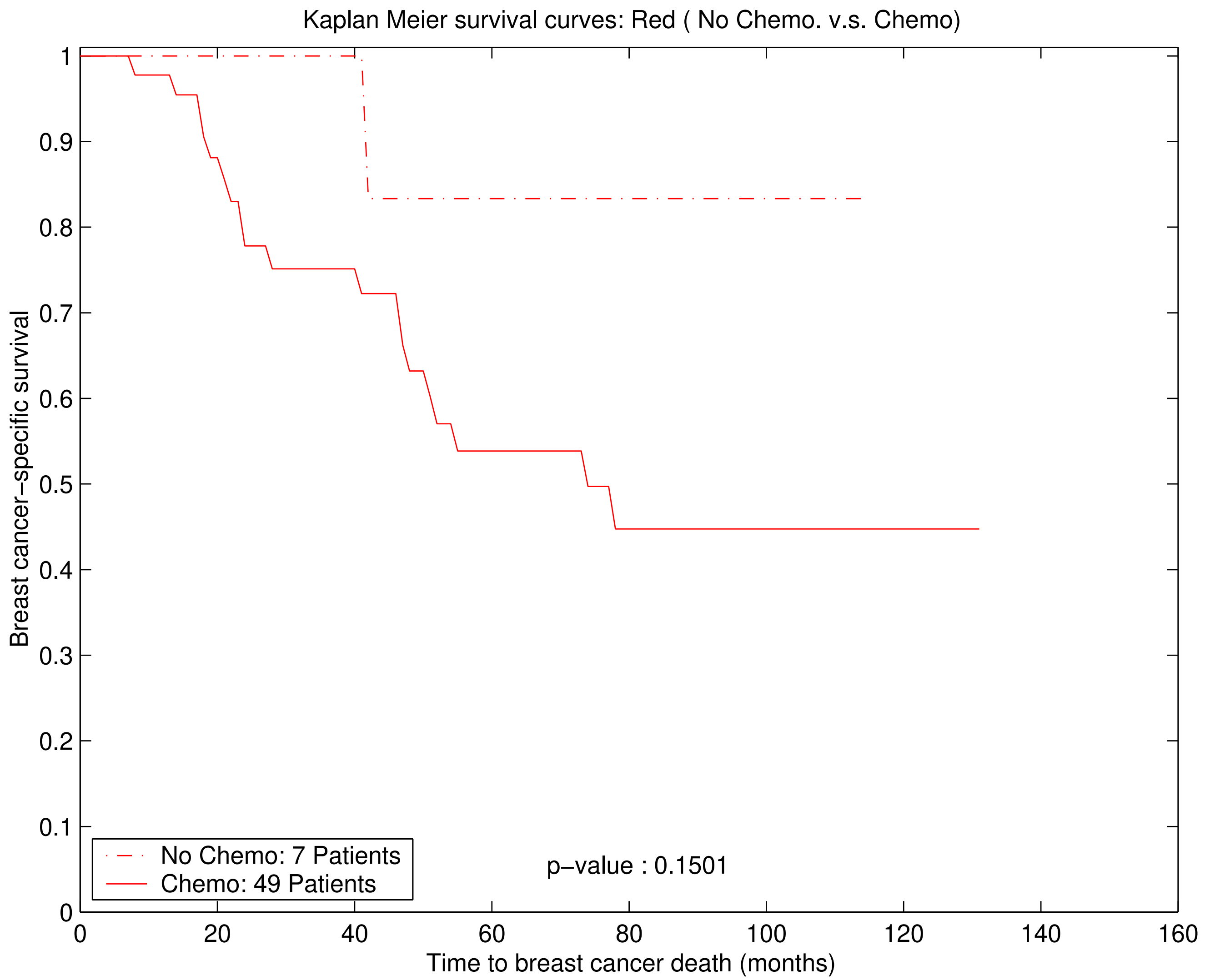
Survival Curves for Breast Cancer Patients:
Good Prognosis Group
Dashes: No Chemotherapy (51) - Solid: Chemotherapy (19)
p-value=0.3174



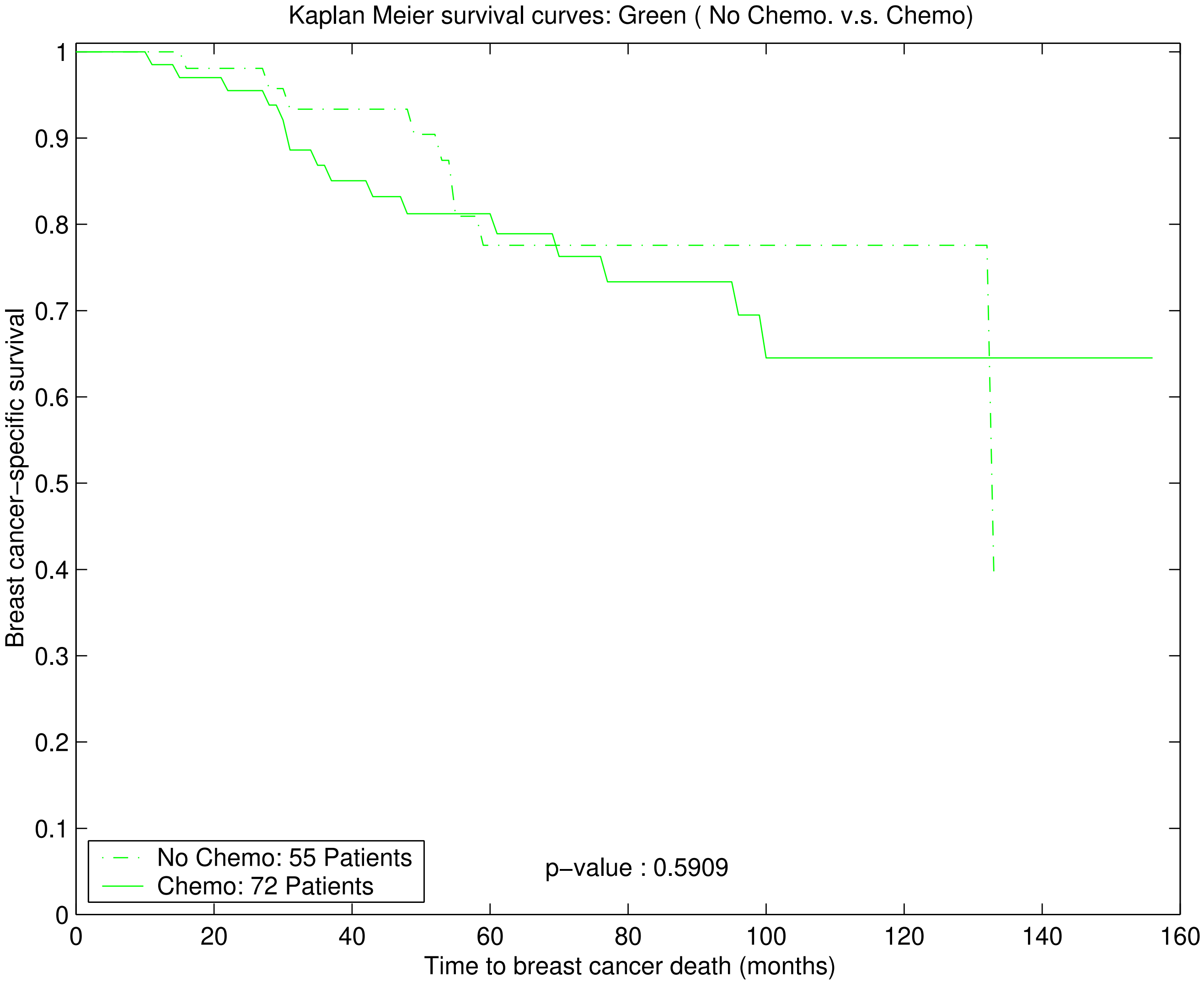
Survival Curves for Breast Cancer Patients: Poor Prognosis Group

Dashes: No Chemotherapy (7) - Solid: Chemotherapy (49)

p-value=0.1501



Survival Curves for Breast Cancer Patients:
Intermediate Prognosis Group
Dashes: No Chemotherapy (55) - Solid: Chemotherapy (72)
p-value=0.5909



Summary and Future Work

- Support vector machines can be used to:
  - Select features that are important for prognosis
  - Classify breast cancer patients into three groups with well separated survival curves
- Most (72.8%) patients in good group received no chemotherapy
- Most (87.5%) patients in poor group received chemotherapy
- Just over half the patients (56.7%) in intermediate group received chemotherapy
- Survival curves and possible chemotherapy decision for new patients can be determined by assigning them to one of the three groups based on six features without the need for lymph node status
- Future work:
  - Prognosis depending on other classification criteria
    * Whether patients had chemotherapy or not
    * The type of chemotherapy
  - Include other features such as age