Knowledge-Based Breast Cancer Prognosis

Computation and Informatics in Biology and Medicine
Training Program Annual Retreat
October 13, 2006

Olvi Mangasarian
UW Madison & UCSD La Jolla
Edward Wild
UW Madison

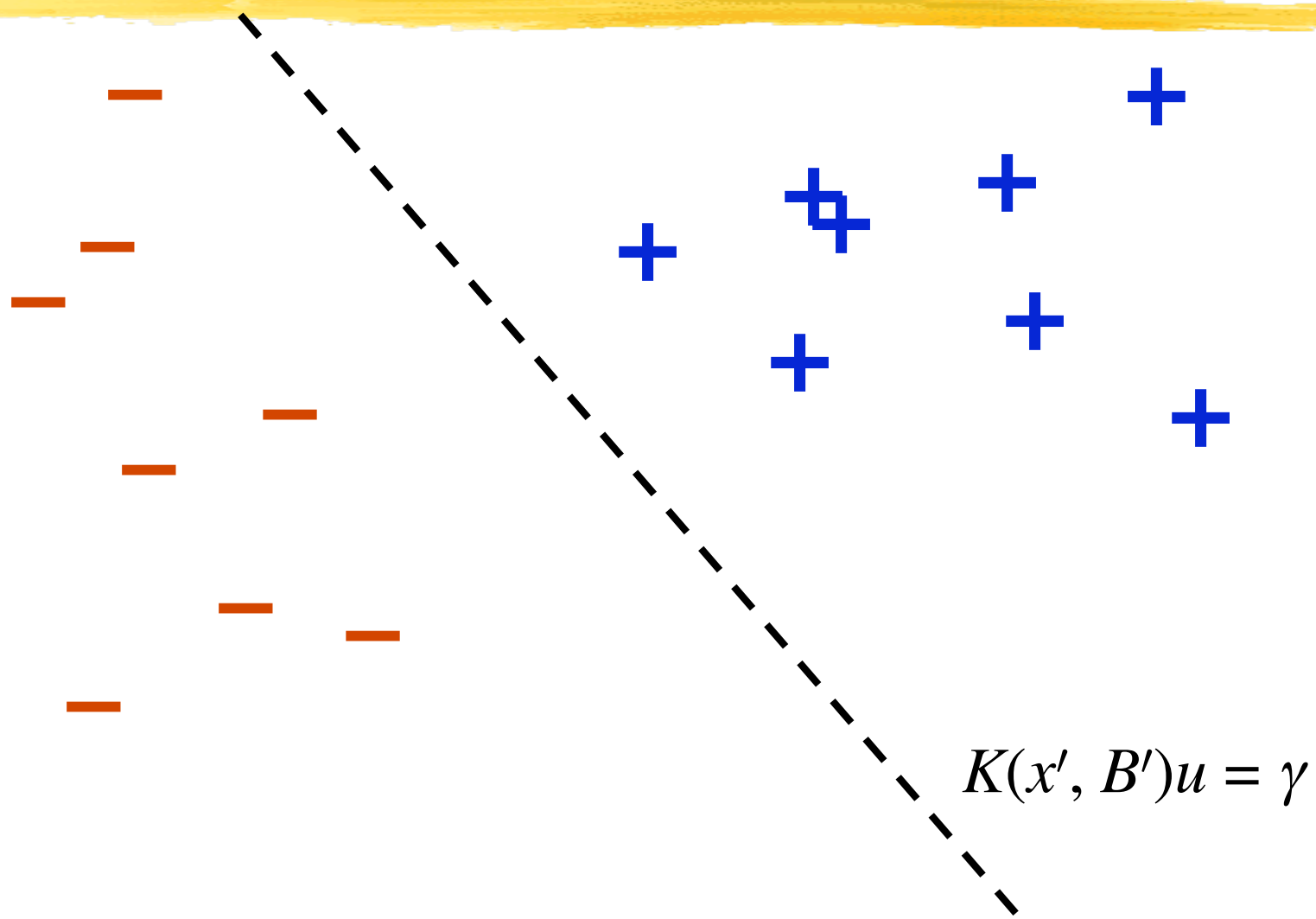
Objectives

- Primary objective: Incorporate prior knowledge over *completely arbitrary sets* into:
  - function approximation, and
  - classification
  - without transforming (kernelizing) the knowledge
- Secondary objective: Achieve transparency of the prior knowledge for practical applications
- Use prior knowledge to improve accuracy on two difficult breast cancer prognosis problems

Classification and Function Approximation

- Given a set of m points in n-dimensional real space R^n with corresponding labels
  - Labels in $\{+1, -1\}$ for classification problems
  - Labels in R for approximation problems
  - Points are represented by rows of a matrix $A \in R^{m \times n}$
  - Corresponding labels or function values are given by a vector y
  - Classification: $y \in \{+1, -1\}^m$
  - Approximation: $y \in R^m$
- Find a function $f(A_i) = y_i$ based on the given data points A_i
  - $f : R^n \rightarrow \{+1, -1\}$ for classification
  - $f : R^n \rightarrow R$ for approximation

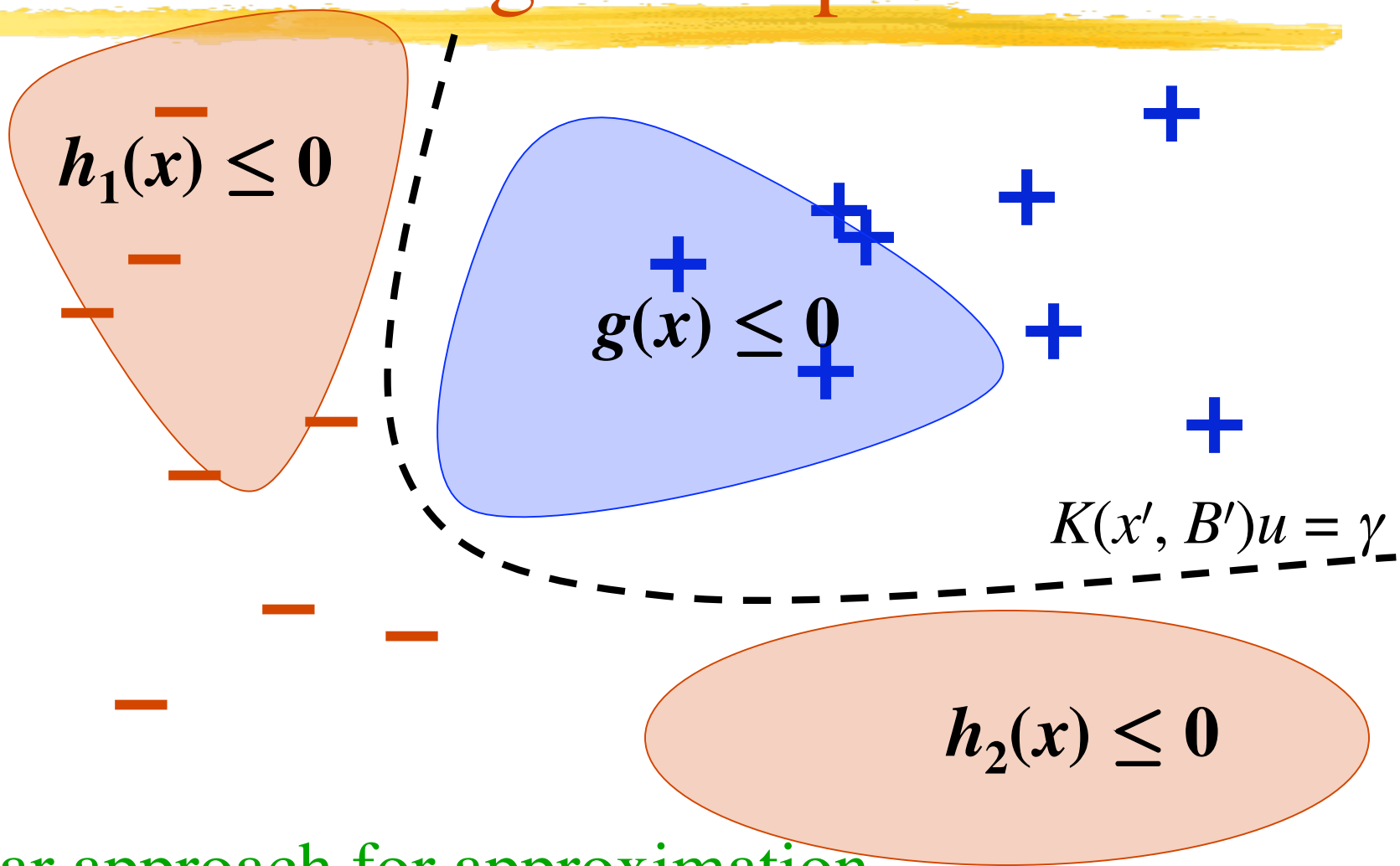
Graphical Example *with no* Prior Knowledge Incorporated



Classification and Function Approximation

- Problem: utilizing only given data may result in a poor classifier or approximation
  - Points may be noisy
  - Sampling may be costly
- Solution: use prior knowledge to improve the classifier or approximation

Graphical Example *with* Prior Knowledge Incorporated



Similar approach for approximation

Kernel Machines

- Approximate f by a nonlinear kernel function K using parameters $u \in R^k$ and γ in R
- A kernel function is a nonlinear generalization of scalar product
- $f(x) \approx K(x', B')u - \gamma, \;\; x \in R^n, K{:}R^n \times R^{n\times k} \rightarrow R^k$
- $B \in R^{k\times n}$ is a *basis matrix*
  - Usually, $B = A \in R^{m\times n}$ = Input data matrix
  - In Reduced Support Vector Machines, B is a small subset of the rows of A
  - B may be any matrix with n columns

Kernel Machines

- Introduce *slack variable* s to measure error in classification or approximation
- Error s in kernel approximation of given data:
  - $-s \leq K(A, B')u - \gamma e - y \leq s$, e is a vector of ones in R^m
  - Function approximation: $f(x) \approx K(x', B')u - \gamma$
- Error s in kernel classification of given data
  - $K(A^+, B')u - \gamma e + s^+ \geq e$, $s^+ \geq 0$
  - $K(A^-, B')u - \gamma e - s^- \leq -e$, $s^- \geq 0$
- More succinctly, let: $D = \mathrm{diag}(y)$, the $m \times m$ matrix with diagonal y of ± 1's, then:
  - $D(K(A, B')u - \gamma e) + s \geq e$, $s \geq 0$
  - Classifier: $f(x) \approx \mathrm{sign}(K(x', B')u - \gamma)$

Kernel Machines in Approximation OR Classification

$$\min_{(u,\gamma,s,a)} \quad e'a + \nu e's$$

$$\text{s.t.} \quad -a \leq u \leq a$$

$$-s \leq K(A, B')u - \gamma e - y \leq s$$

OR

$$D(K(A, B')u - \gamma e) - s \geq e, \; s \geq 0$$

Positive parameter ν controls trade off between

- solution complexity: $e'a = \|u\|_1$ at solution
- data fitting: $e's = \|s\|_1$ at solution

Nonlinear Prior Knowledge in Function Approximation

- Start with arbitrary *nonlinear* knowledge implication
- g, h are arbitrary functions on Γ
- $g{:}\Gamma \rightarrow R^k$, $h{:}\Gamma \rightarrow R$
- $g(x) \leq 0 \Rightarrow K(x', B')u - \gamma \geq h(x), \ \forall x \in \Gamma \subset R^n$

$$\Uparrow$$

$$\exists v \geq 0{:}\ v'g(x) + K(x', B')u - \gamma - h(x) \geq 0 \ \ \forall x \in \Gamma$$

- Linear in v, u, γ

Theorem of the Alternative for Convex Functions

- Assume that $g(x)$, $K(x', B')u - \gamma$, $-h(x)$ are *convex* functions of x, that Γ is convex and $\exists\, x \in \Gamma : g(x) < 0$. Then either:
  - I. $g(x) \leq 0$, $K(x', B')u - \gamma - h(x) < 0$ has *a* solution $x \in \Gamma$, ***or***
  - II. $\exists v \in R^k$, $v \geq 0$: $K(x', B')u - \gamma - h(x) + v'g(x) \geq 0\ \forall x \in \Gamma$
  - But *never both*.
- If we can find $v \geq 0$: $K(x', B')u - \gamma - h(x) + v'g(x) \geq 0$ $\forall x \in \Gamma$, then by above theorem
  - $g(x) \leq 0$, $K(x', B')u - \gamma - h(x) < 0$ has *no* solution $x \in \Gamma$ or equivalently:
  - $g(x) \leq 0 \Rightarrow K(x', B')u - \gamma \geq h(x),\ \forall x \in \Gamma$

Incorporating Prior Knowledge

$$\begin{aligned} \min_{(u,\gamma,s,a,v)} \quad & e'a + \nu e's \\ \text{s.t.} \quad & -s \le K(A,B')u - \gamma e - y \le s, \\ & -a \le u \le a, \\ & K(x',B')u - \gamma - h(x) + v'g(x) \ge 0, x \in \Gamma \\ & v \ge 0. \end{aligned}$$

Linear semi-infinite program: infinite number of constraints

Discretize to obtain a finite linear program

Add term in objective to drive prior knowledge error to zero

$$\begin{aligned} \min_{(u,\gamma,s,a,v,z^i,\ldots,z^k)} \quad & e'a + \nu e's + \sigma \sum_{i=1}^{k} z_i \\ \text{s.t.} \quad & -s \le K(A,B')u - \gamma e - y \le s, \\ & -a \le u \le a, \\ & K(x^{i\prime},B')u - \gamma - h(x^i) + v'g(x^i) + z_i \ge 0, \\ & v \ge 0,\ z_i \ge 0,\ i = 1,\ldots,k. \end{aligned}$$

Slacks z_i allow knowledge to be satisfied inexactly at the point x^i

$g(x^i) \le 0 \Rightarrow K(x^{i\prime}, B')u - \gamma \ge h(x^i),\ i = 1, \ldots, k$

Incorporating Prior Knowledge in Classification (Very Similar)

- Implication for positive region
- $g(x) \leq 0 \Rightarrow K(x', B')u - \gamma \geq \alpha, \ \forall x \in \Gamma \subset R^n$

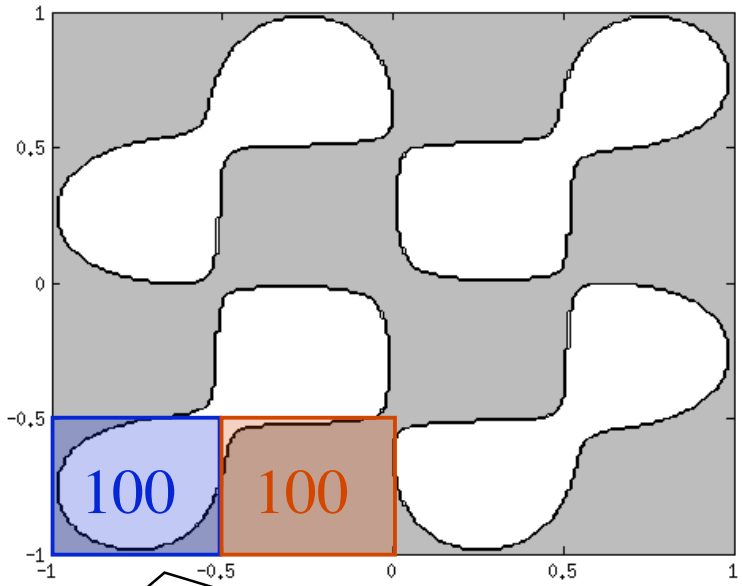
$$\Uparrow$$

- $\exists v \geq 0, K(x', B')u - \gamma - \alpha + v'g(x) \geq 0, \ \forall x \in \Gamma$
- Similar implication for negative regions
- Add discretized constraints to linear program

Incorporating Prior Knowledge in Classification

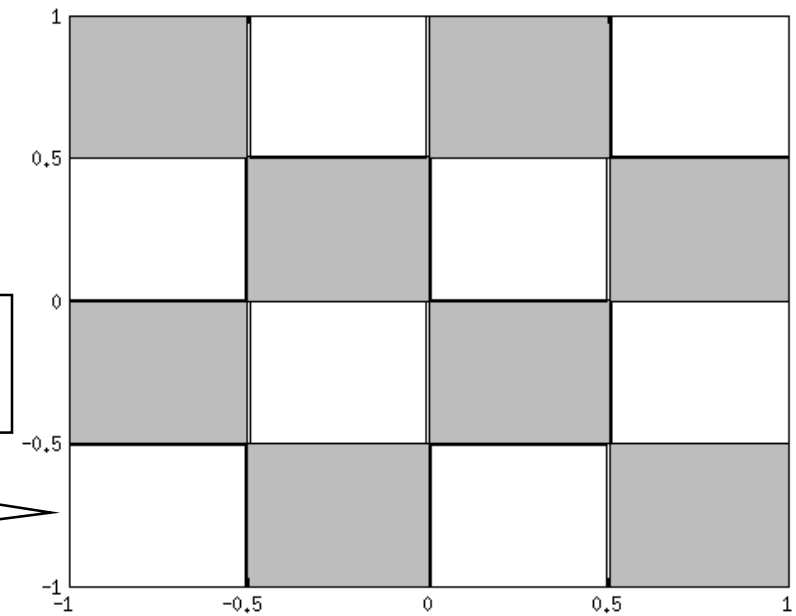
$$
\begin{array}{rl}
\min\limits_{(u,\gamma,s,a,v,p,z_1,\ldots,z_k,q_1,\ldots,q_t)} & e'a + \nu e's + \sigma(\sum\limits_{i=1}^{k} z_i + \sum\limits_{j=1}^{t} q_j) \\
\text{s.t.} & D(K(A,B')u - \gamma e) + s \geq e, \\
& -a \leq u \leq a, \\
& s \geq 0, \\
& K(x^{i'},B')u - \gamma - \alpha + v'g(x^i) + z_i \geq 0, \\
& v \geq 0,\ z_i \geq 0,\ i = 1,\ldots,k, \\
& -K(x^{j'},B')u + \gamma - \alpha + p'h(x^j) + q_j \geq 0, \\
& p \geq 0,\ q_j \geq 0,\ j = 1,\ldots,t.
\end{array}
$$

Checkerboard Dataset: Black and White Points in R^2



Classifier based on the 16 points at the center of each square *and no* prior knowledge

Prior knowledge given at 100 points in the two left-most squares of the bottom row



Perfect classifier based on the same 16 points *and* the prior knowledge

Predicting Lymph Node Metastasis as a Function of Tumor Size

- Number of metastasized lymph nodes is an important prognostic indicator for breast cancer recurrence
  - Determined by surgery *in addition* to the removal of the tumor
  - Optional procedure especially if tumor size is small
- Wisconsin Prognostic Breast Cancer (WPBC) data
  - Lymph node metastasis and tumor size for 194 patients
- Task: predict the number of metastasized lymph nodes given tumor size alone

Predicting Lymph Node Metastasis

- Split data into two portions
  - Past data: 20% used to find prior knowledge
  - Present data: 80% used to evaluate performance
- Simulates acquiring prior knowledge from an expert

Prior Knowledge for Lymph Node Metastasis as a Function of Tumor Size

- Generate prior knowledge by fitting past data:
  - $h(x) := K(x', B')u - \gamma$
  - B is the matrix of the past data points
- Use density estimation to decide where to enforce knowledge
  - $p(x)$ is the empirical density of the past data
- Prior knowledge utilized on approximating function $f(x)$:
  - Number of metastasized lymph nodes is greater than the predicted value on past data, with tolerance of 1%
  - $p(x) \geq 0.1 \Rightarrow f(x) \geq h(x) - 0.01$

Predicting Lymph Node Metastasis: Results

Approximation	Error
Prior knowledge $h(x)$ based on past data 20%	6.12 RMSE
$f(x)$ without knowledge based on present data 80%	5.92 LOO
$f(x)$ with knowledge based on present data 80%	**5.04** LOO

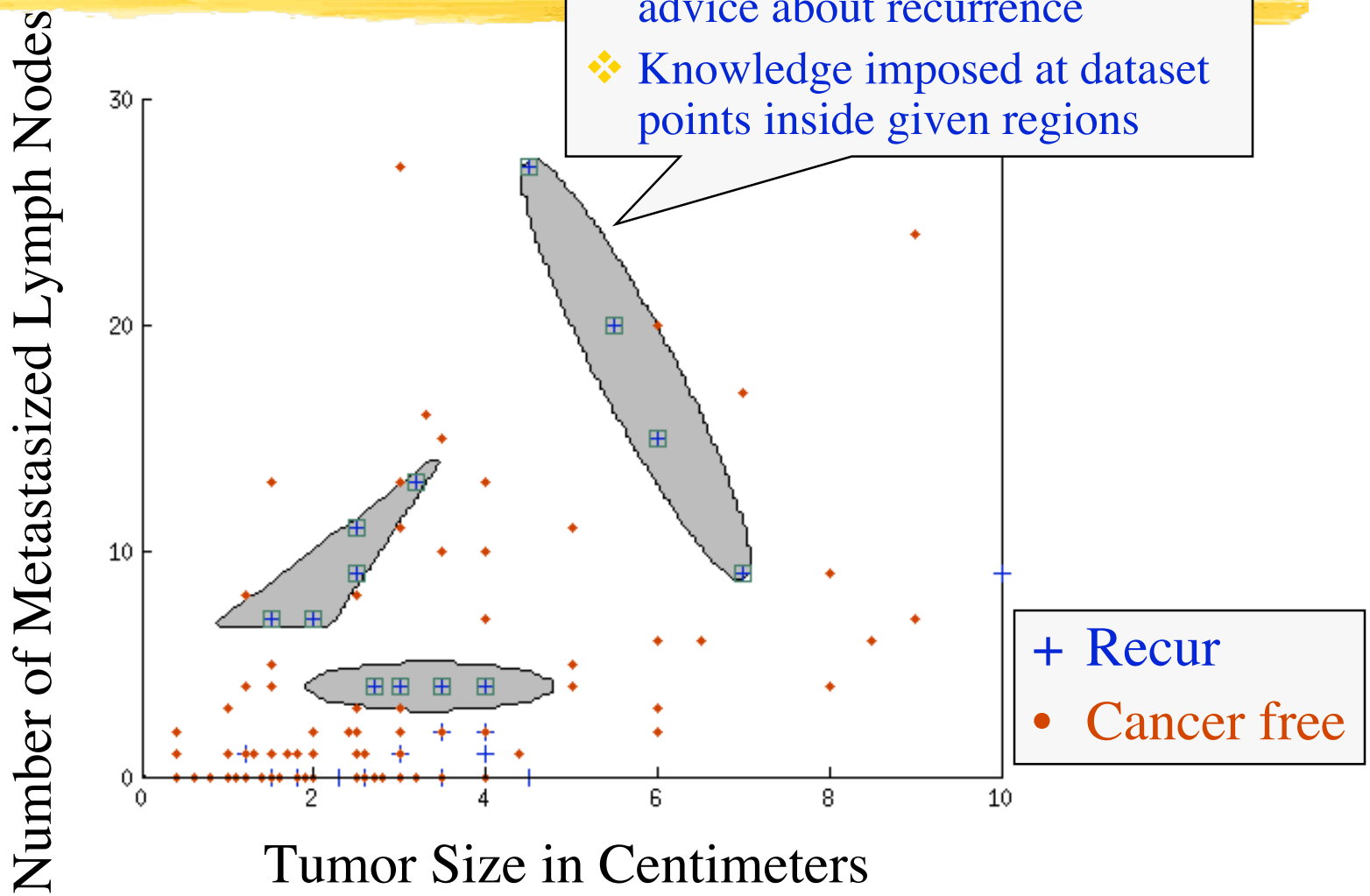
- RMSE: root-mean-squared-error
- LOO: leave-one-out error
- Improvement due to knowledge: 14.9%

Predicting Breast Cancer Recurrence Within 24 Months

- Wisconsin Prognostic Breast Cancer (WPBC) dataset
  - 155 patients monitored for recurrence within 24 months
  - 30 cytological features
  - 2 histological features: number of metastasized lymph nodes and tumor size
- Predict whether or not a patient remains cancer free after 24 months
- 82% of patients remain disease free
- 86% accuracy (Bennett, 1992) best previously attained
- Prior knowledge allows us to incorporate additional information to improve accuracy

Generating WPBC Prior Knowledge

- Gray regions indicate areas where $g(x) \leq 0$
- Simulate oncological surgeon's advice about recurrence
- Knowledge imposed at dataset points inside given regions



WPBC Results

Classifier	Misclassification Rate
Without Knowledge	18.1%
With Knowledge	**9.0%**

49.7 % improvement due to knowledge

35.7 % improvement over best previous predictor

Conclusion

- General nonlinear prior knowledge incorporated into kernel classification and approximation
  - Implemented as linear inequalities in a linear programming problem
  - Knowledge appears transparently
- Demonstrated effectiveness of nonlinear prior knowledge on two real world problems from breast cancer prognosis
- Future work
  - Prior knowledge with more general implications
  - User-friendly interface for knowledge specification
- More information
  - http://www.cs.wisc.edu/~olvi/
  - http://www.cs.wisc.edu/~wildt/