Optimization in Support Vector Machines, and Application to Microarray Data

Yoonkyung Lee

www.stat.ohio-state.edu/∼yklee

Department of Statistics
The Ohio State University

Support Vector Machines

- Classification accuracy.
- Flexibility - implicit embedding through kernel.
- Handle high dimensional data - some myth.
- Sparsity - quadratic programming problem.
- No probability estimates.

References

- Lee, Y., Lin, Y., and Wahba, G. (2002)
  Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data, Technical Report 1064. Department of Statistics, University of Wisconsin-Madison.

- Lee, Y. and Lee, C.-K. (2003)
  Classification of Multiple Cancer Types by Multicategory Support Vector Machines Using Gene Expression Data, Bioinformatics, vol. 19, 1132-1139, 2003.

Multicategory SVM

- Class codes :

  $\mathbf{y}_i = (y_{i1}, \cdots, y_{ik})$ with $y_{ij} = 1$ and $-\frac{1}{k-1}$ elsewhere if example i falls into class j.

$$\text{When } k = 3, \qquad \mathbf{y}_i = \begin{cases} (1, -\frac{1}{2}, -\frac{1}{2}) & \text{for class 1} \\ (-\frac{1}{2}, 1, -\frac{1}{2}) & \text{for class 2} \\ (-\frac{1}{2}, -\frac{1}{2}, 1) & \text{for class 3} \end{cases}$$

- Class separating functions :

  $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ with $\sum_{j=1}^{k} f_j(\mathbf{x}) = 0$ for any $\mathbf{x} \in R^d$, and $f_j(\mathbf{x}) = h_j(\mathbf{x}) + b_j$ with $h_j \in \mathcal{H}_K$.

  e.g. $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^t \mathbf{x}_2$, or $(1 + \mathbf{x}_1^t \mathbf{x}_2)^2$, or $\exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$

- Classification rule : $\phi(\mathbf{x}) = arg\max_j f_j(\mathbf{x})$.

- Multicategory SVM formulation :

  Find $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$, with sum-to-zero constraint, minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} L_{cat(i)j}(f_j(\mathbf{x}_i) - y_{ij})_+ + \frac{\lambda}{2}\sum_{j=1}^{k}\|h_j\|^2_{\mathcal{H}_K}$$

  where

  $cat(i)$: the category of $\mathbf{y}_i$ and $L_{jj'}$: the cost of misclassifying j as j'.

  When $L_{jj'} = I(j \neq j')$,

  $\sum_{j=1}^{k} L_{cat(i)j}(f_j(\mathbf{x}_i) - y_{ij})_+ = \sum_{j\neq cat(i)}(f_j(\mathbf{x}_i) + \frac{1}{k-1})_+$.

Representer theorem for Multicategory SVM

Theorem 1. *To find* $(f_1(\mathbf{x}), \cdots, f_k(\mathbf{x})) \in \prod_1^k(\{1\} + \mathcal{H}_K)$, *with the sum-to-zero constraint, minimizing the MSVM objective function is equivalent to find* $(f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ *of the form*

$$f_j(\mathbf{x}) = b_j + \sum_{i=1}^{n} c_{ij} K(\mathbf{x}_i, \mathbf{x}) \quad \text{for } j = 1, \cdots, k$$

with the sum-to-zero constraint only at $\mathbf{x}_i$ *for* $i = 1, \cdots, n$, *minimizing the objective function.*

How to deal with non-differentiable function $(x)_+$?

Convince yourself that

$$(x)_+ = \left\{ \begin{array}{rl} & \min\ \xi \\ \text{subject to} & x \leq \xi \\ & \xi \geq 0 \end{array} \right.$$

- $\boxed{\text{Primal problem}}$: Minimize

$$L_P(\mathbf{c}, \mathbf{b}, \xi) = \frac{1}{n} \sum_{j=1}^{k} L_j^t \xi_{\cdot j} + \frac{\lambda}{2} \sum_{j=1}^{k} \mathbf{c}_{\cdot j}^t K \mathbf{c}_{\cdot j}$$

subject to

$$b_j \mathbf{e} + K\mathbf{c}_{\cdot j} - \mathbf{y}_{\cdot j} \le \xi_{\cdot j} \quad \text{for } j = 1, \cdots, k$$

$$\xi_{\cdot j} \ge 0 \quad \text{for } j = 1, \cdots, k$$

$$\textstyle\sum_{j=1}^{k} (b_j \mathbf{e} + K\mathbf{c}_{\cdot j}) = 0$$

where

$\mathbf{c}_{\cdot j} = (c_{1j}, \cdots, c_{nj})^t$, $K = (K(\mathbf{x}_i, \mathbf{x}_j))$,
$L_j = (L_{cat(1)j}, \cdots, L_{cat(n)j})^t$, $\mathbf{y}_{\cdot j} = (y_{1j}, \cdots, y_{nj})^t$, and
$\xi_{\cdot j} = (\xi_{1j}, \cdots, \xi_{nj})^t$.

Introducing Lagrange multipliers α_j for $b_j\mathbf{e} + K\mathbf{c}_{\cdot j} - \mathbf{y}_{\cdot j} \le \xi_{\cdot j}$,

- $\boxed{\text{Dual problem}}$: Maximize

$$L_D(\alpha) = -\frac{1}{2n}\sum_{j=1}^{k}(\alpha_j - \bar{\alpha})^t K(\alpha_j - \bar{\alpha}) - \lambda\sum_{j=1}^{k}\alpha_j^t\mathbf{y}_{\cdot j}$$

subject to

$$0 \le \alpha_j \le L_j \quad \text{for } j = 1, \cdots, k$$

$$(\alpha_j - \bar{\alpha})^t\mathbf{e} = 0 \quad \text{for } j = 1, \cdots, k$$

Quadratic programming problem

How to determine $\mathbf{c}_{\cdot j}$ and b_j from α?

- $\mathbf{c}_{\cdot j} = -\frac{1}{n\lambda}(\alpha_j - \bar{\alpha})$ for $j = 1, \cdots, k$.

  By Karush-Kuhn-Tucker complementarity conditions, the solution should satisfy

$$
\begin{aligned}
\alpha_j \perp (b_j \mathbf{e} + K\mathbf{c}_{\cdot j} - \mathbf{y}_{\cdot j} - \xi_{\cdot j}) \quad & \text{for } j = 1, \cdots, k \\
\gamma_j = (L_j - \alpha_j) \perp \xi_{\cdot j} \quad & \text{for } j = 1, \cdots, k
\end{aligned}
$$

- If $(\alpha_{i1}, \cdots, \alpha_{ik}) = \mathbf{0}$, then $(c_{i1}, \cdots, c_{ik}) = \mathbf{0}$.

  Support Vectors: data points with $(c_{i1}, \cdots, c_{ik}) \neq \mathbf{0}$.

Small Round Blue Cell Tumors of Childhood

- Khan et al. (2001) in *Nature Medicine*
- Tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS)
- Number of genes : 2308
- Class distribution of data set

Data set	EWS	BL(NHL)	NB	RMS	total
Training set	23	8	12	20	63
Test set	6	3	6	5	20
Total	29	11	18	25	83

- Gene selection : Dudoit et al. (2000)

  For gene ℓ, the ratio of between classes sum of squares to within class sum of squares is defined as

$$\frac{BSS(\ell)}{WSS(\ell)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j)(\bar{x}_{\cdot\ell}^{(j)} - \bar{x}_{\cdot\ell})^2}{\sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j)(x_{i\ell} - \bar{x}_{\cdot\ell}^{(j)})^2}$$

  Pick genes with the largest ratios.

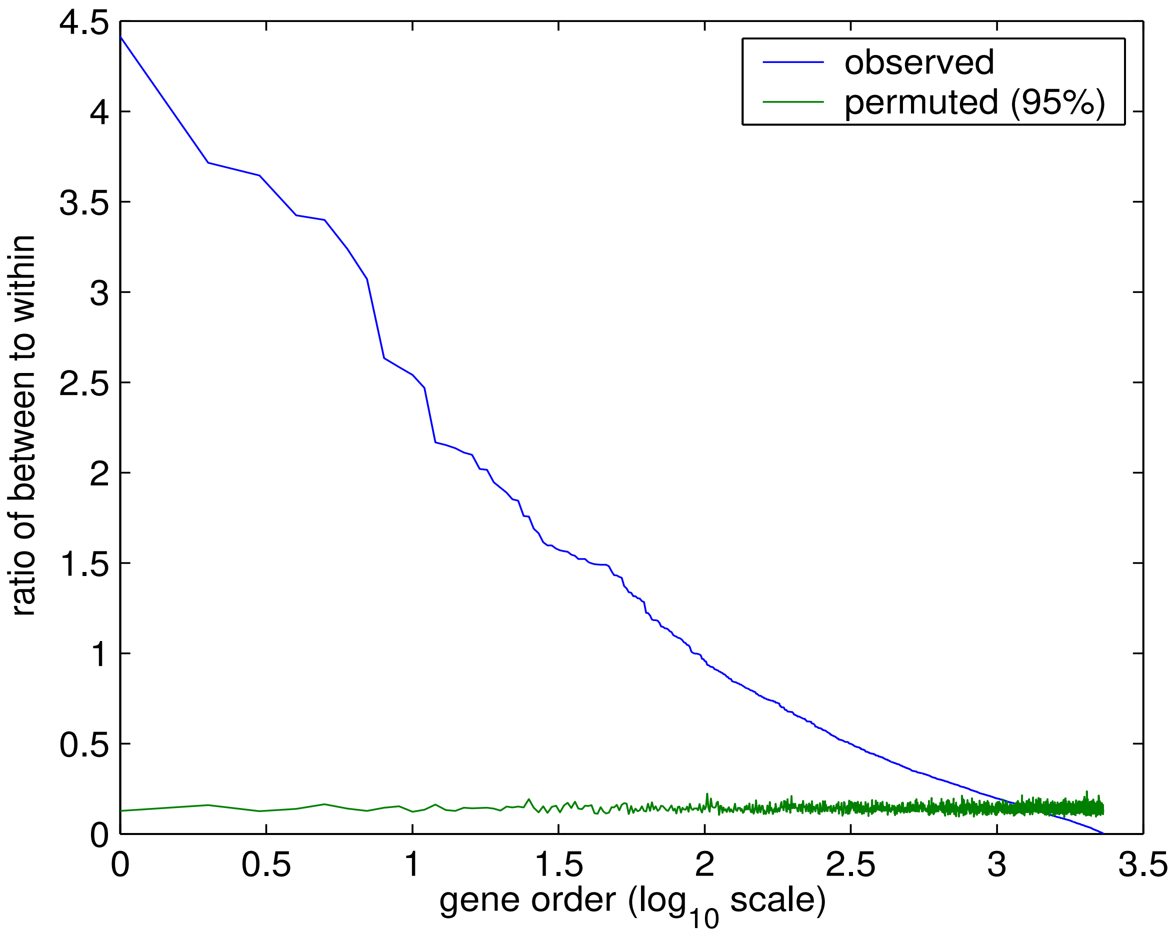

Figure 1: Observed ratios of between-class SS to within-class SS and the 95 percentiles of the corresponding ratios for expression levels with randomly permuted class labels.

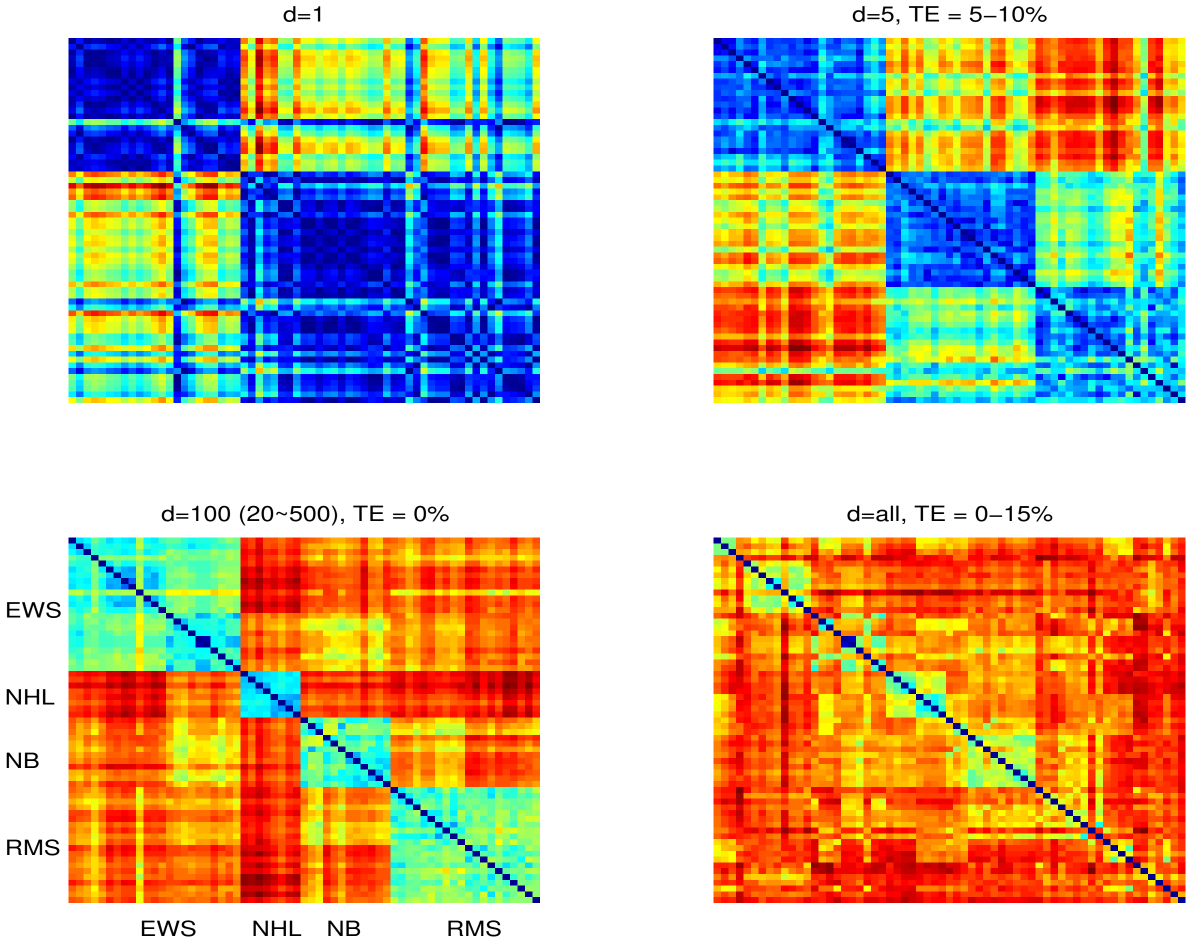

Figure 2: Pairwise distance matrices for the training data as the numbers of genes included change, and test error rates of MSVM with Gaussian kernel.

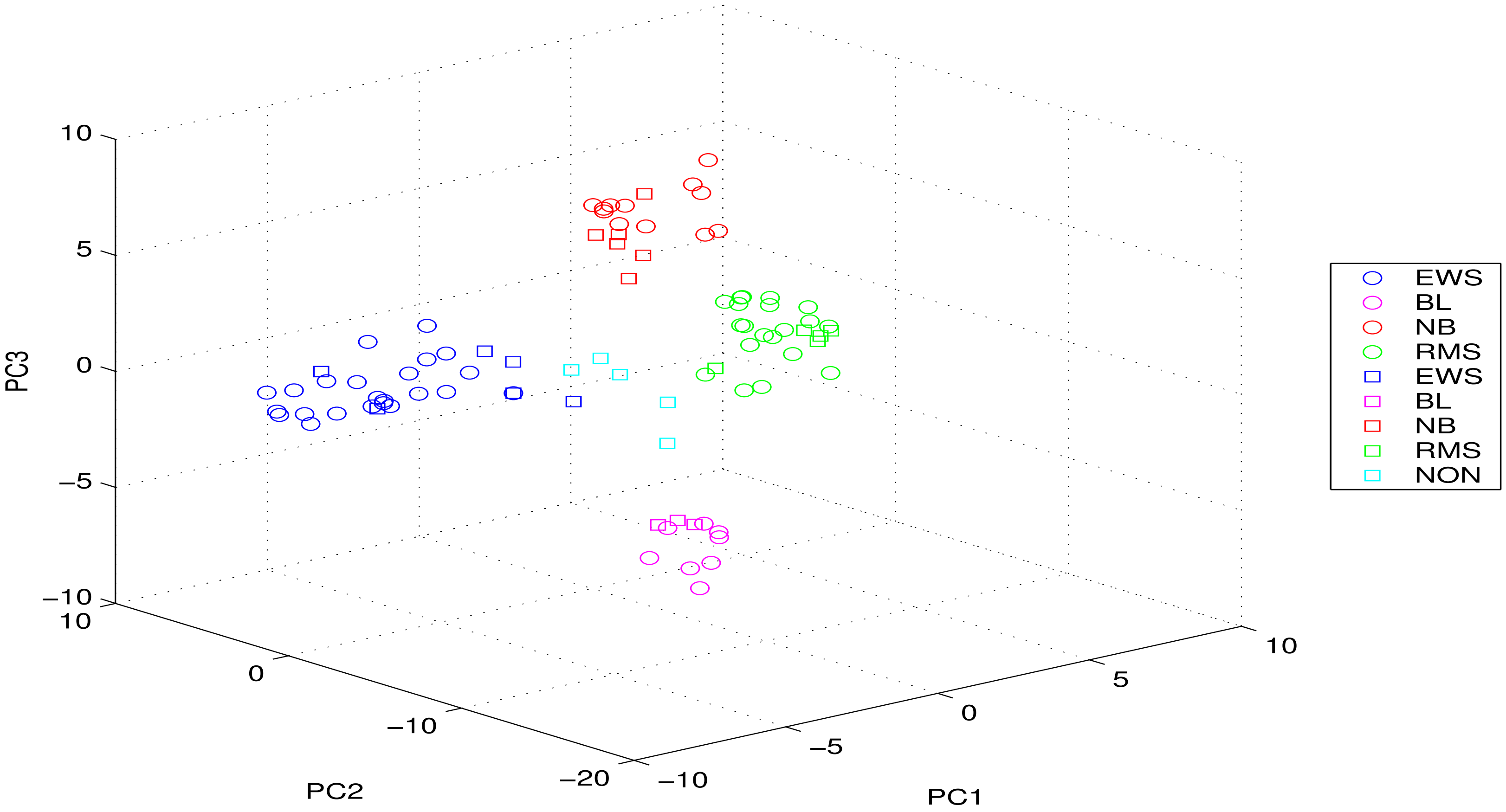

Figure 3: Three principal components of 100 gene expression levels (circles: training samples, squares: test samples including non SRBCT samples). The tumor types are distinguished by colors.

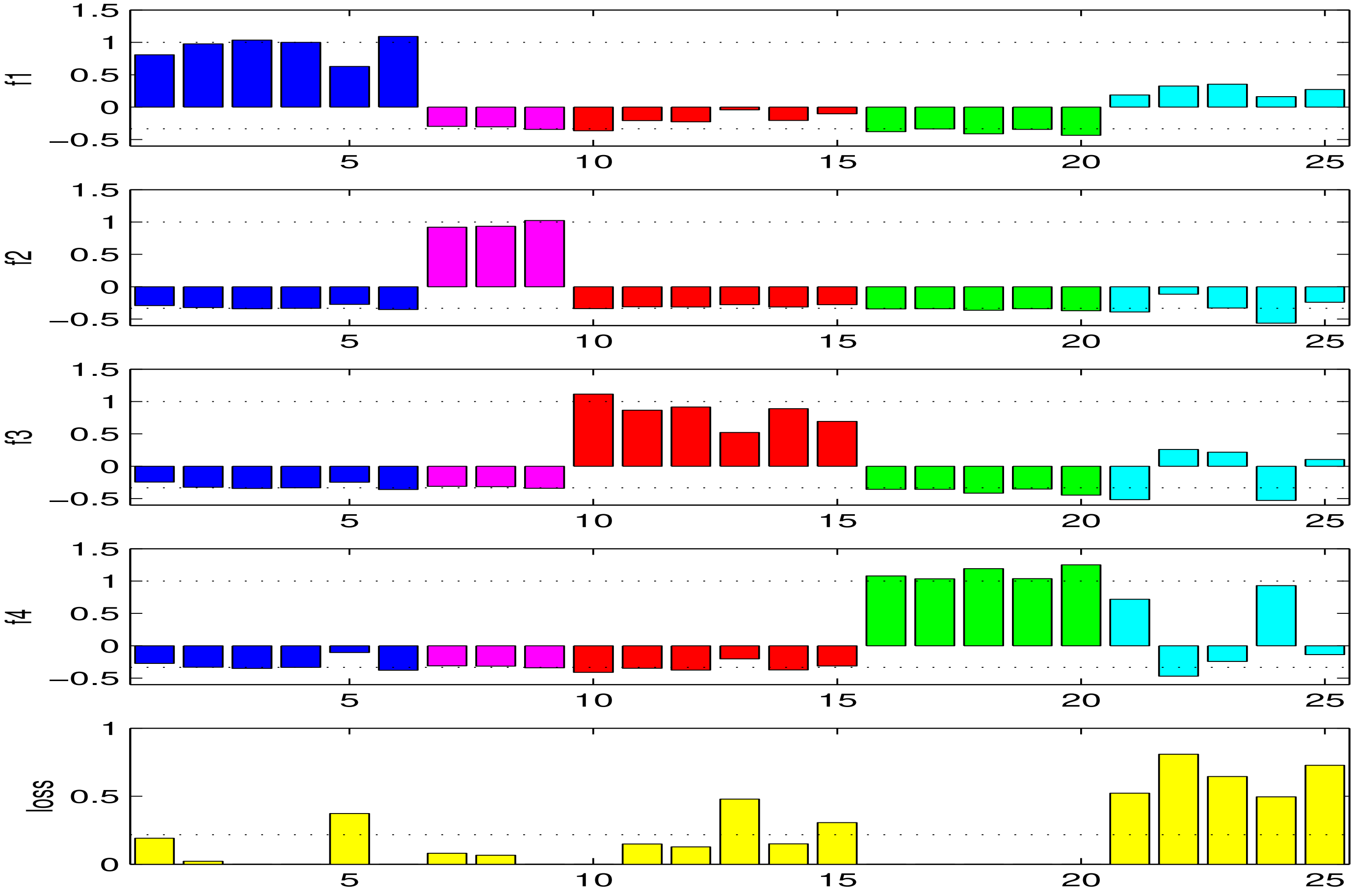

Figure 4: Predicted decision vectors (f_1, f_2, f_3, f_4) at the test samples. EWS: $(1, -1/3, -1/3, -1/3)$, BL: $(-1/3, 1, -1/3, -1/3)$, NB: $(-1/3, -1/3, 1, -1/3)$, and RMS: $(-1/3, -1/3, -1/3, 1)$. The colors indicate the true class identities of the test samples.

Concluding remarks

- Optimization problem in SVM is a quadratic programming problem.
- Covariates appear in SVM formulation only through kernel evaluations.
- Selective choice of variables would improve accuracy. Integrate variable selection with learning classification rule.
- The effect of the number of variables on classification accuracy depends on data at hand. (gene expression, text, image data)