Multicategory Support Vector Machines

Yoonkyung Lee[a]

www.stat.ohio-state.edu/∼yklee

Department of Statistics
The Ohio State University

---

[a]Research partly supported by NSF and NIH Grants to Grace Wahba. Joint work with Yi Lin and Grace Wahba, UW-Madison.

Classification Problems

- Training data set :

  Covariate (attribute) $\mathbf{x}_i \in R^d$, and Class label $y_i \in \{1, \cdots, k\}$, $i = 1, \cdots, n$.

  k is the number of classes.

- Task :

  Learn a classification rule matching $\mathbf{x}_i$ with y_i,

  the mapping $\phi(\mathbf{x}) : R^d \to \{1, \cdots, k\}$

Example: Classification Problem

- Fisher (1936): Iris data

  Iris Setosa, Iris Versicolour and Iris Virginica

- Golub et al.(*Science*, 1999): Acute leukemia data

  ALL (acute lymphoblastic leukemia) with subtype B-cell and T-cell lineage, and AML (acute myeloid leukemia)

Training data set: Acute leukemia gene expression profiles

	$gene_1$	$gene_2$	$\cdots$	$gene_{7129}$	$class$
$patient_1$	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,7129}$	ALL − B
$patient_2$	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,7129}$	ALL − B
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{20}$	$x_{20,1}$	$x_{20,2}$	$\cdots$	$x_{20,7129}$	ALL − T
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{28}$	$x_{28,1}$	$x_{28,2}$	$\cdots$	$x_{28,7129}$	AML
$\vdots$	$\vdots$			$\vdots$	$\vdots$
$patient_{38}$	$x_{38,1}$	$x_{38,2}$	$\cdots$	$x_{38,7129}$	AML

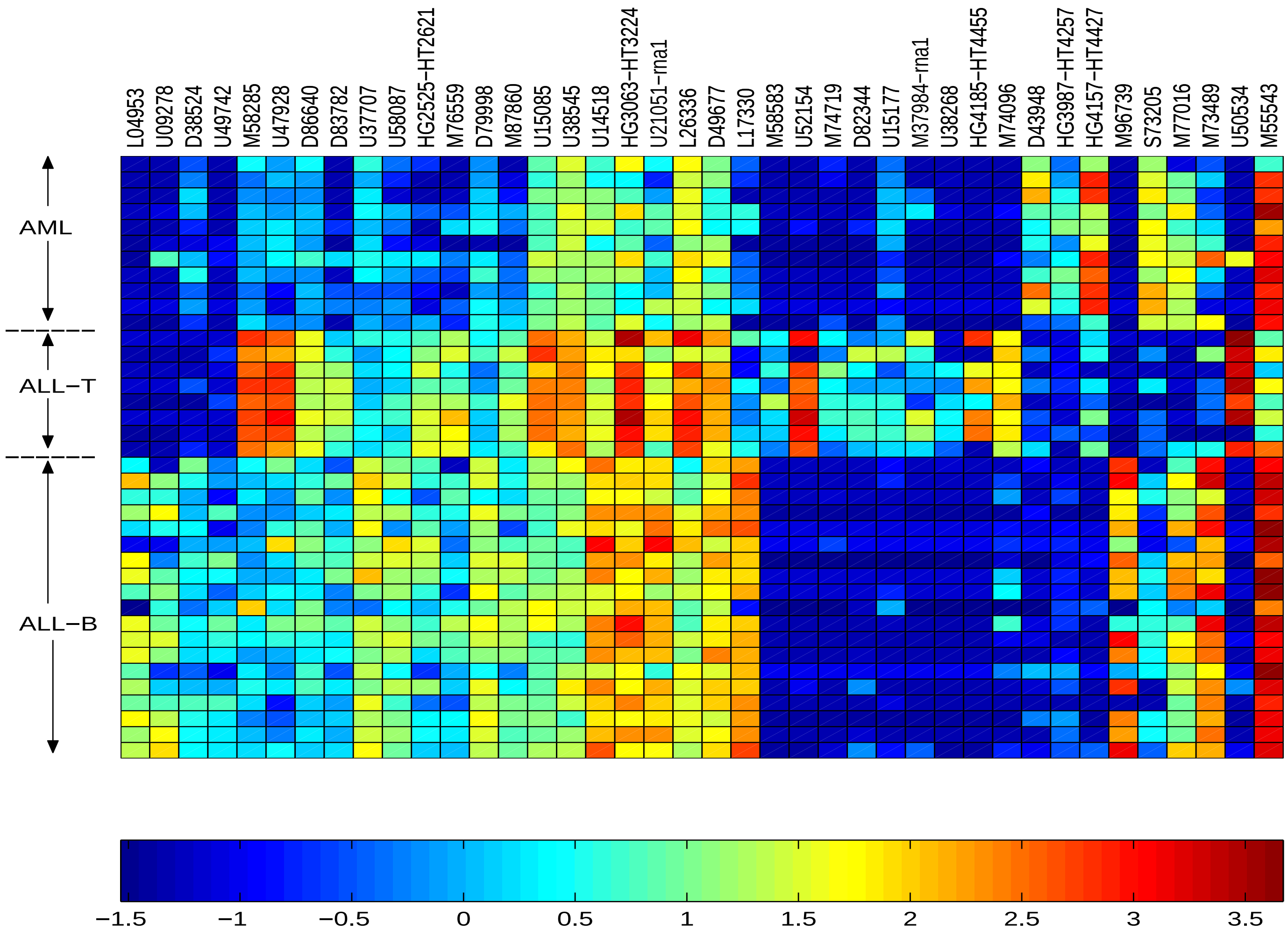


Figure 1: The heat map shows the expression levels of 40 most important genes for the training samples when they are appropriately standardized. Each row corresponds to a sample, which is grouped into the three classes, and the columns represent genes. The 40 genes are clustered in a way the similarity within each class and the dissimilarity between classes are easily recognized.

The Bayes Classification Rule

- Misclassification Loss function :

$$l(y, \phi(\mathbf{x})) = I(y \neq \phi(\mathbf{x}))$$

- (X, Y) : a random sample from $P(\mathbf{x}, y)$, and $p_j(\mathbf{x}) = P(Y = j | X = \mathbf{x})$

- The Bayes rule minimizes the expected misclassification rate $E[l(Y, \phi(\mathbf{X}))] = P(Y \neq \phi(\mathbf{X}))$

$$\phi_B(\mathbf{x}) = arg \max_{j=1,\cdots,k} p_j(\mathbf{x})$$

Classification Tools

Linear Discriminant Analysis	Quadratic Discriminant Analysis
Logistic Regression	Density Estimation
Classification Tree	Nearest Neighbor
Neural Networks	Support Vector Machine

Support Vector Machine for Binary Case

Boser, Guyon, & Vapnik (1992), Vapnik (1995), Burges (1998), Cristianini & Shawe-Taylor (2000), http://www.kernel-machines.org

- Training data set :
  $\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in R^d, y_i \in \{-1, 1\},$ for $i = 1, \cdots, n\}$
- Linear SVM: $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ where $\mathbf{w} \in R^d$ and $b \in R$.
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$

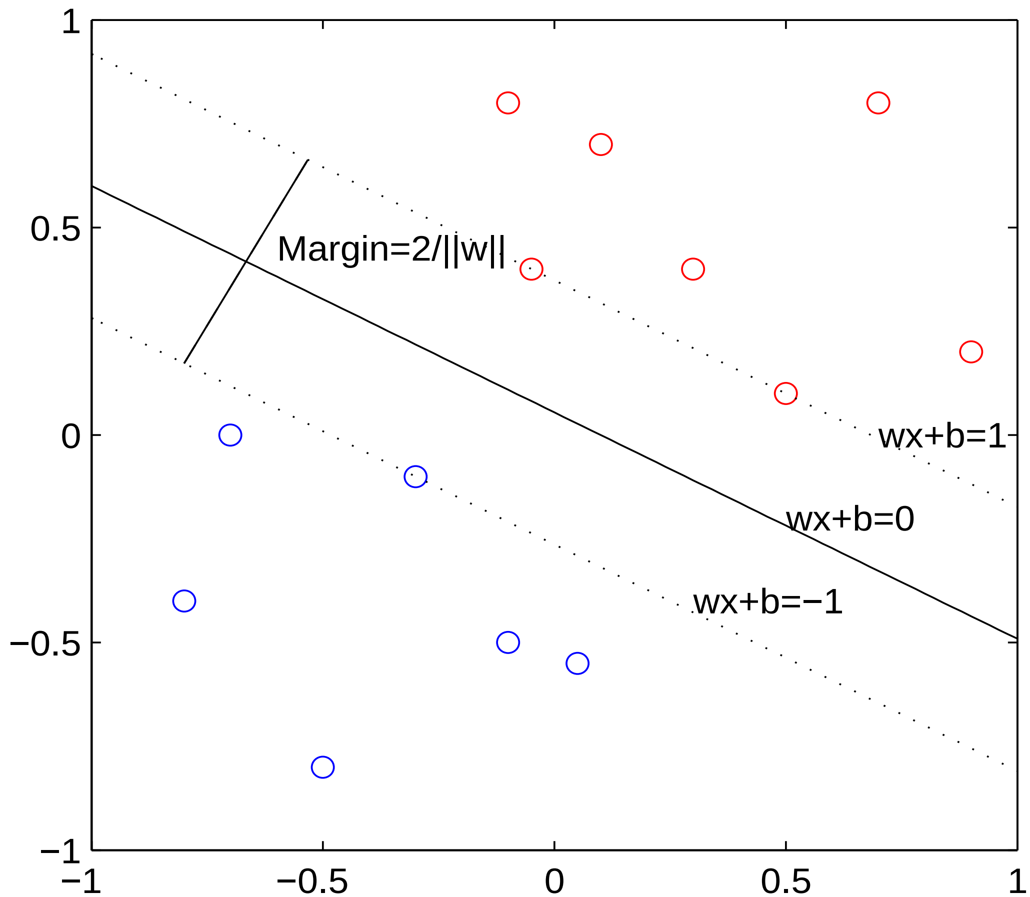


Figure 2: Linear Support Vector Machine example in separable case. Red and blue circles indicate positive and negative examples, respectively. The solid line is the separating hyperplane with the maximum margin in this example.

- Separable case :

  Find $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ minimizing $\|\mathbf{w}\|^2$

$$\text{subject to} \quad \begin{array}{ll} f(\mathbf{x}_i) \geq 1 & \text{if } y_i = 1 \\ f(\mathbf{x}_i) \leq -1 & \text{if } y_i = -1. \end{array}$$

  Or more succinctly,

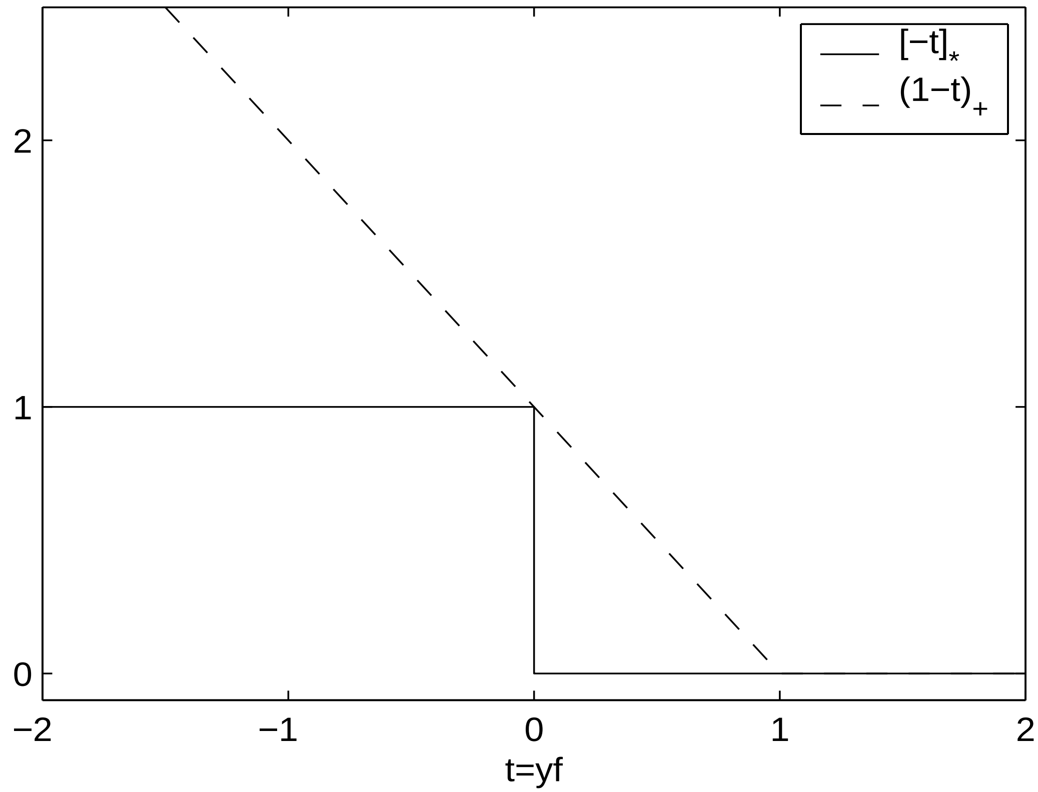
$$y_i f(\mathbf{x}_i) \geq 1 \text{ for } i = 1, \cdots, n.$$

- Nonseparable case :

  Find $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(\mathbf{x}_i))_+ + \lambda \|\mathbf{w}\|^2$$

  where $(x)_+ = \max(x, 0)$.



Loss function $(1 - y_i f(\mathbf{x}_i))_+$ is an upper bound of the misclassification loss function $I(y_i \neq \phi(\mathbf{x}_i)) = [-y_i f(\mathbf{x}_i)]_* \leq (1 - y_i f(\mathbf{x}_i))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

SVM in Regularization Framework

Wahba (1998)

- $f(\mathbf{x}) = h(\mathbf{x}) + b$ with $h \in \mathcal{H}_K$ a Reproducing Kernel Hilbert space (RKHS) with reproducing kernel K.
- Classification rule : $\phi(\mathbf{x}) = sign(f(\mathbf{x}))$
- SVM scheme : to find f minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\mathbf{x}_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K}$$

- Linear SVM: $\mathcal{H}_K = \{h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} | \mathbf{w} \in R^d\}$
  with $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2$ and $\|h\|^2_{\mathcal{H}_K} = \|\mathbf{w} \cdot \mathbf{x}\|^2_{\mathcal{H}_K} = \|\mathbf{w}\|^2$
- Nonlinear SVM: $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1 \cdot \mathbf{x}_2)^d$, $\exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$

Review of Reproducing Kernel Hilbert Space

Aronszajn (1950), Wahba (1990)

- An RKHS $\mathcal{H}_K$ is a Hilbert space of real valued functions defined on an index set $\mathcal{T}$ for which all the evaluation functionals are bounded linear functionals. Then there is a positive definite function called Reproducing Kernel K unique to the RKHS such that

  (i) $K(t, \cdot) \in \mathcal{H}_K$ for each $t \in \mathcal{T}$

  (ii) $f(t) = (K(t, \cdot), f(\cdot))_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$

  Consequently, $K(t, s) = (K(t, \cdot), K(s, \cdot))_{\mathcal{H}_K}$

- (Moore-Aronszajn theorem) To every positive definite function K on $\mathcal{T} \times \mathcal{T}$, there corresponds a unique RKHS $\mathcal{H}_K$ of real valued functions on $\mathcal{T}$ and vice versa.

- By Representer Theorem (Kimeldorf, & Wahba 1971),

$$f(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b$$

  So, $\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j)$

- Solve a constrained optimization problem through its Primal-Dual relation. Introduce a Lagrange multiplier for each data point.

- Dual problem in terms of Lagrange multipliers: Quadratic Programming problem

- If the Lagrange multiplier for the i th example is zero, then $c_i = 0$. Removing such example $(\mathbf{x}_i, y_i)$ would not affect the solution at all. *Support Vectors*: Examples with nonzero coefficient.

SVM and the Bayes Rule for Binary case

- Under the label of $Y = 1$ or -1,
  Let $p^+(\mathbf{x}) = P(Y = 1|X = \mathbf{x})$.
- the Bayes rule : $\phi_B(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$

Lemma 1 (Lin 1999). *The minimizer of* $E[(1-Yf(X))_+]$ *is* $f(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$.

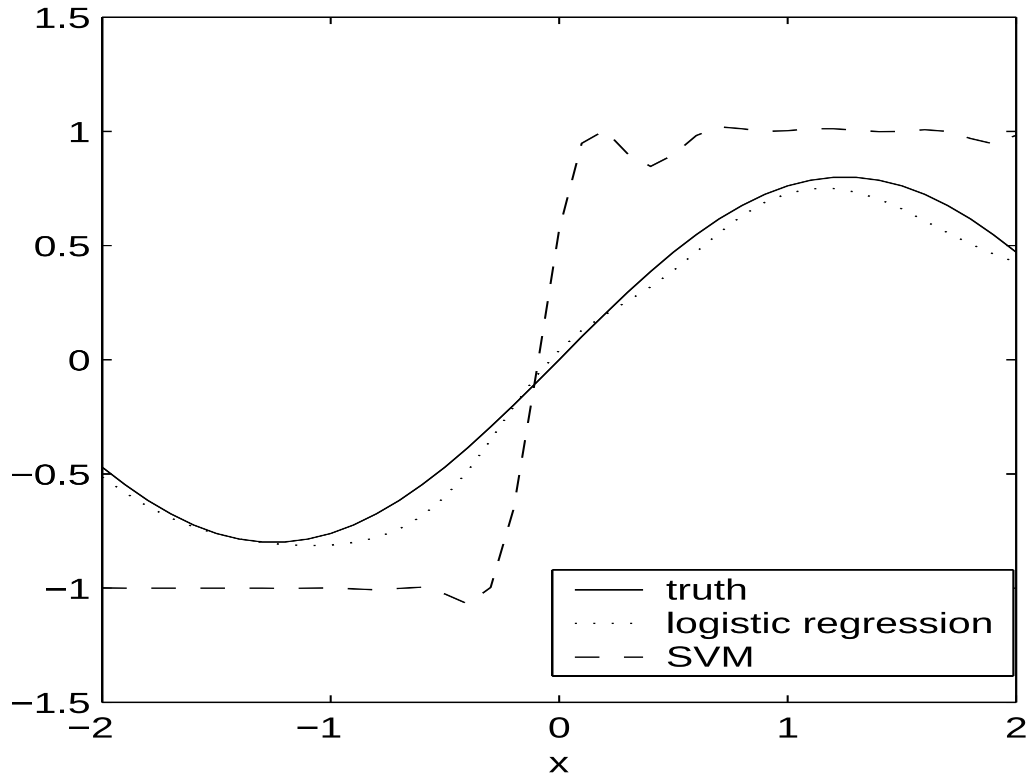


Figure 3: The solid line: two times the true conditional probability - 1, the dashed line: the SVM solution, and the dotted line: two times the conditional probability estimate via the logistic regression - 1. The estimates are based on 300 data points with y generated from the true conditional probability, given x from the standard normal distribution. Note that SVM outputs are lumped around $\pm$ 1.

Characteristics of Support Vector Machines

- Classification accuracy
- Flexibility
- Handle high dimensional data
- Sparsity - Quadratic programming problem
- No probability estimates

Fisher Consistency or Admissibility

- Two-class :

  The population minimizer of the loss function has the same sign as the Bayes decision rule, $\phi_B(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$.

  Lin (2001), Zhang (2001), Steinwart (2002), Bartlett, Jordan, and McAuliffe (2003).

- Multiclass :

  The population minimizer of the loss function predicts the same class as the Bayes decision rule, $\phi_B(\mathbf{x}) = arg\max_{j=1,\cdots,k} p_j(\mathbf{x})$ where $p_j(\mathbf{x}) = P(Y = j|X = \mathbf{x})$.

  Lee, Lin, and Wahba (2002), Zhang (2003).

Approaches to Multiclass by Binary SVM

- Combine binary classifiers to solve multiclass problems.

  Dietterich and Bakiri (1995), Allwein, Schapire and Singer (2000).

- One-vs-rest approach : very common for SVMs.

  (i) Solve k binary problems.

  (ii) Assign a test sample the class giving the largest $f_j(\mathbf{x})$, where $f_j(\mathbf{x})$ is the SVM solution from training class j versus the rest.

  One-vs-rest approach is not Fisher consistent!

- Some multiclass extensions:

  Weston & Watkins(1998), Vapnik(1998), Crammer & Singer(2000), Guermeur (2000).

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 $h_j \in \mathcal{H}_K$.

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

MSVM properties

- Binary SVM is a special case of Multicategory SVM when $k = 2$.
- Handle equal or different misclassification costs in a unified way.
- Multicategory SVM is Fisher consistent.

Lemma 2 (Lee, Lin & Wahba 2001). *The minimizer of* $E[L(Y) \cdot (\mathbf{f}(X) - Y)_+]$ *under sum-to-zero constraint is* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_k(\mathbf{x}))$ *with*

$$f_j(\mathbf{x}) = \begin{cases} 1 & \text{if } j = arg\max_{\ell=1,\cdots,k} p_\ell(\mathbf{x}) \\ -\frac{1}{k-1} & \text{otherwise} \end{cases}$$

Other multiclass extensions

Weston & Watkins(1998), Vapnik(1998), Guermeur (2000).

$$l(y_i, \mathbf{f}(\mathbf{x}_i)) = \sum_{j=1, j \neq y_i}^{k} (f_j(\mathbf{x}_i) - f_{y_i}(\mathbf{x}_i) + 2)_+$$

where the induced classifier is $\phi(\mathbf{x}) = arg\, \max_j f_j(\mathbf{x})$.

Lemma 3. *Consider the case of $k = 3$ with $p_1 < 1/3 < p_2 < p_3 < 1/2$ at a given point $\mathbf{x}$. To insure uniqueness, without loss of generality we can fix $f_1(\mathbf{x}) = -1$. Then the unique minimizer of $E\left[l(Y, \mathbf{f}(X))\right]$ is $(f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})) = (-1, 1, 1)$.*

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*

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

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

- Solve a constrained optimization problem through its dual problem in terms of Lagrange multipliers.
- Dual problem : Quadratic Programming problem
- If the Lagrange multipliers for the i th example are zero vector, then $(c_{i1}, \cdots, c_{ik}) = 0$.
  Support Vectors: Examples with $(c_{i1}, \cdots, c_{ik}) \neq 0$ for $i = 1, \cdots, n$.
- Multicategory SVM retains the sparsity of solution in the same way as binary SVM!

Numerical Study

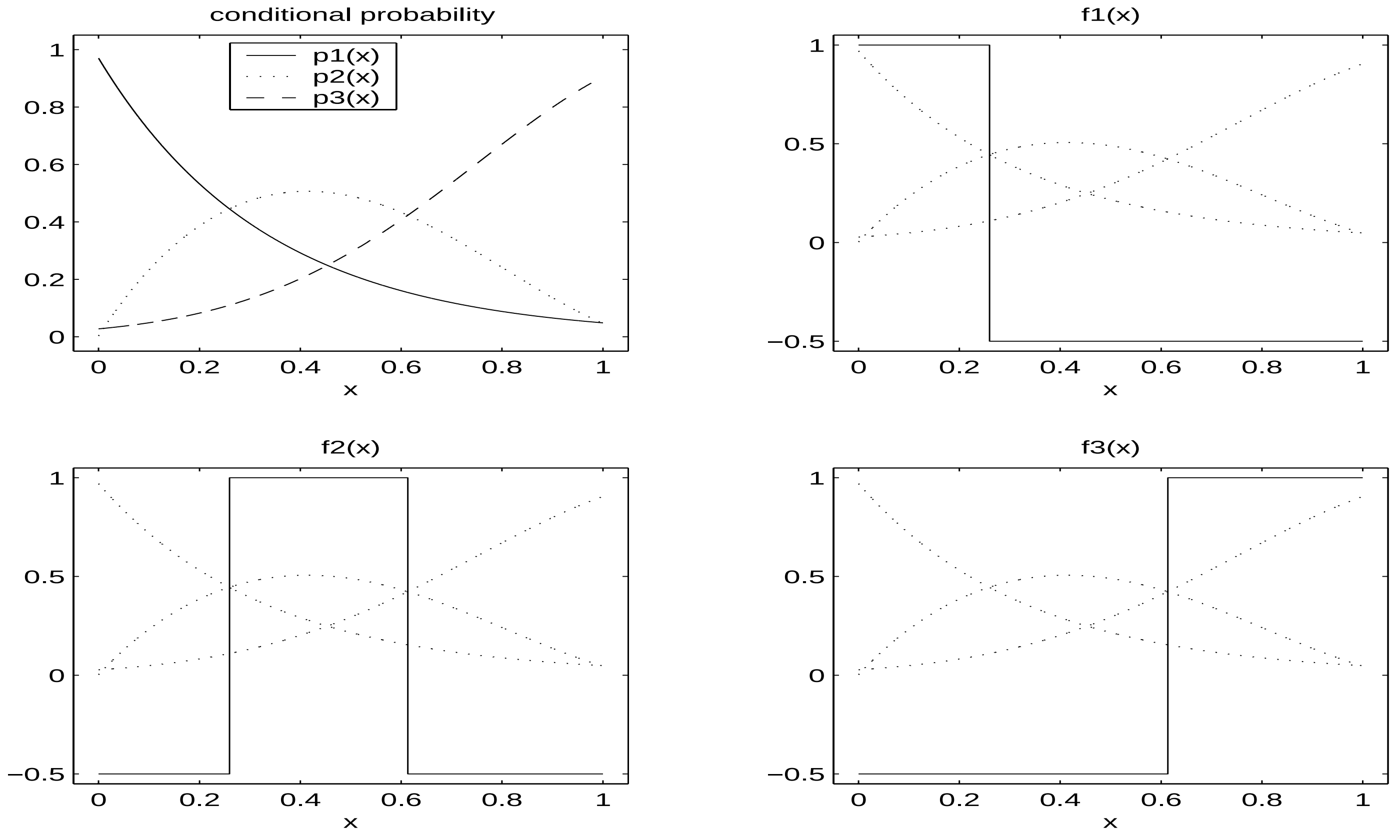


Figure 4: Conditional probabilities and multicategory SVM target functions for three-class example.

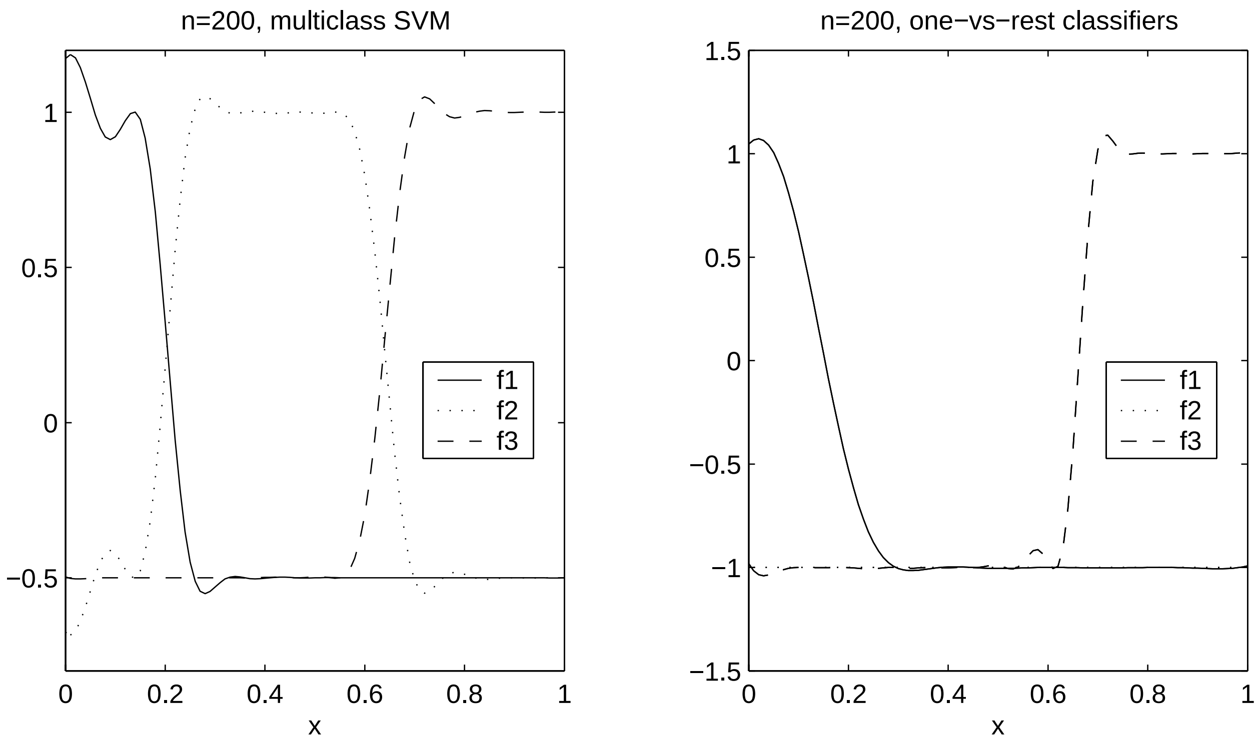


Figure 5: Comparison between multicategory SVM and one-versus-rest method. The Gaussian kernel function is used, and the tuning parameters λ, and σ are simultaneously chosen by $GCKL = E_{true} \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{Y}_i) \cdot (\mathbf{f}_\lambda(x_i) - \mathbf{Y}_i)_+$. Over 10000 newly generated examples, the multicategory SVM has a misclassification rate 0.3890, while that of the one-vs-rest approach is 0.4243

Comparisons

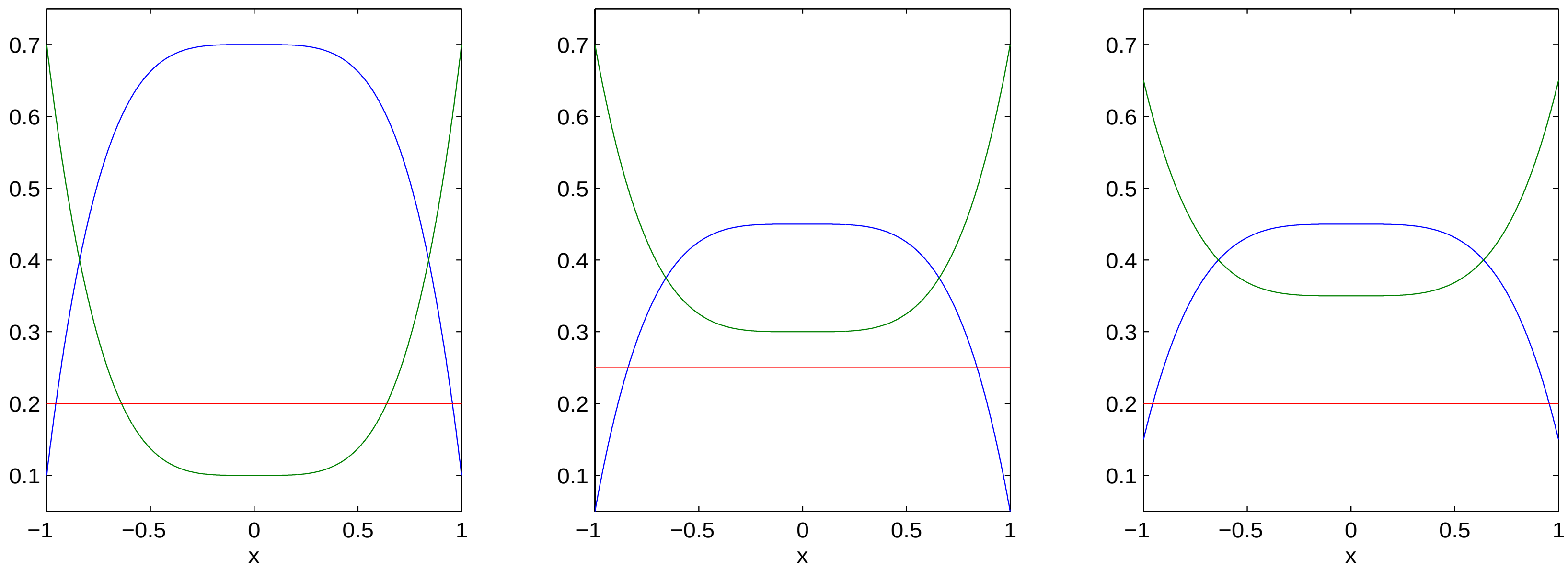


Figure 6: Underlying true conditional probabilities in three situations. Scenario 1: dominant class (left), Scenario 2: the lowest two classes compete (center), and Scenario 3: the highest two classes compete (right).

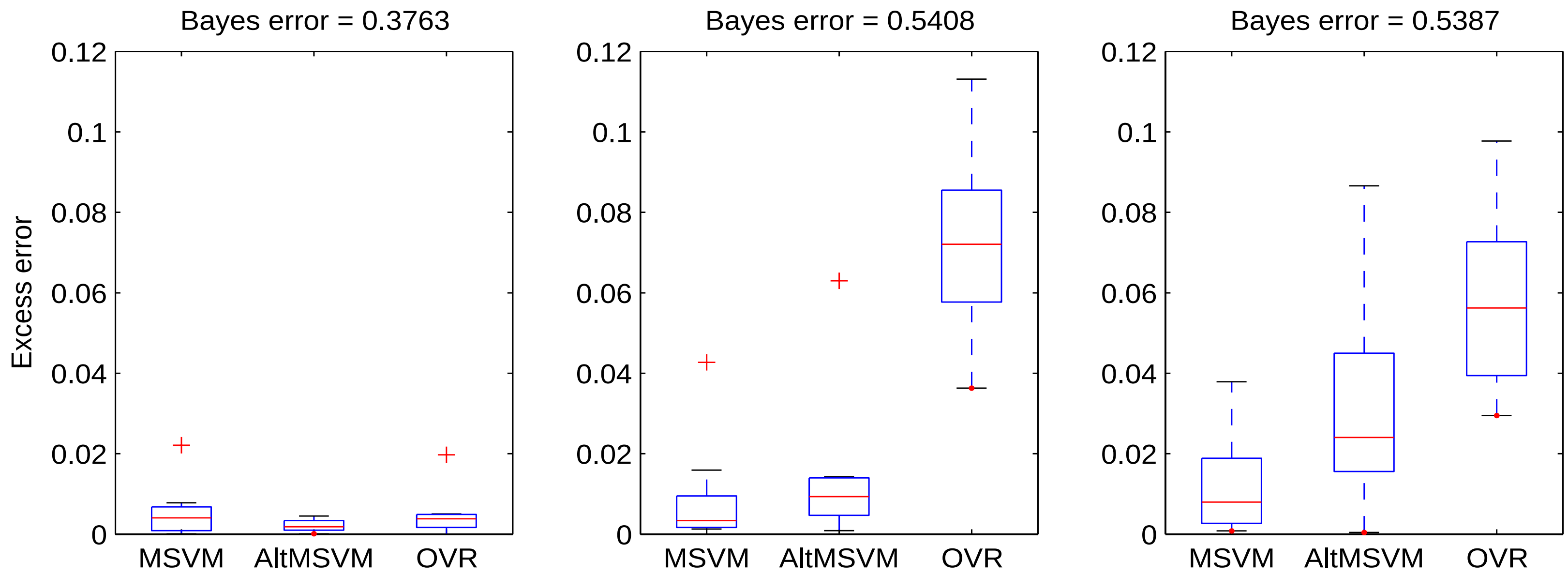


Figure 7: Boxplots of approximated classification error rates from 10 replicates of sample size 200. Scenario 1: left, Scenario 2: center, and Scenario 3: right.

What happened in Scenario 3?

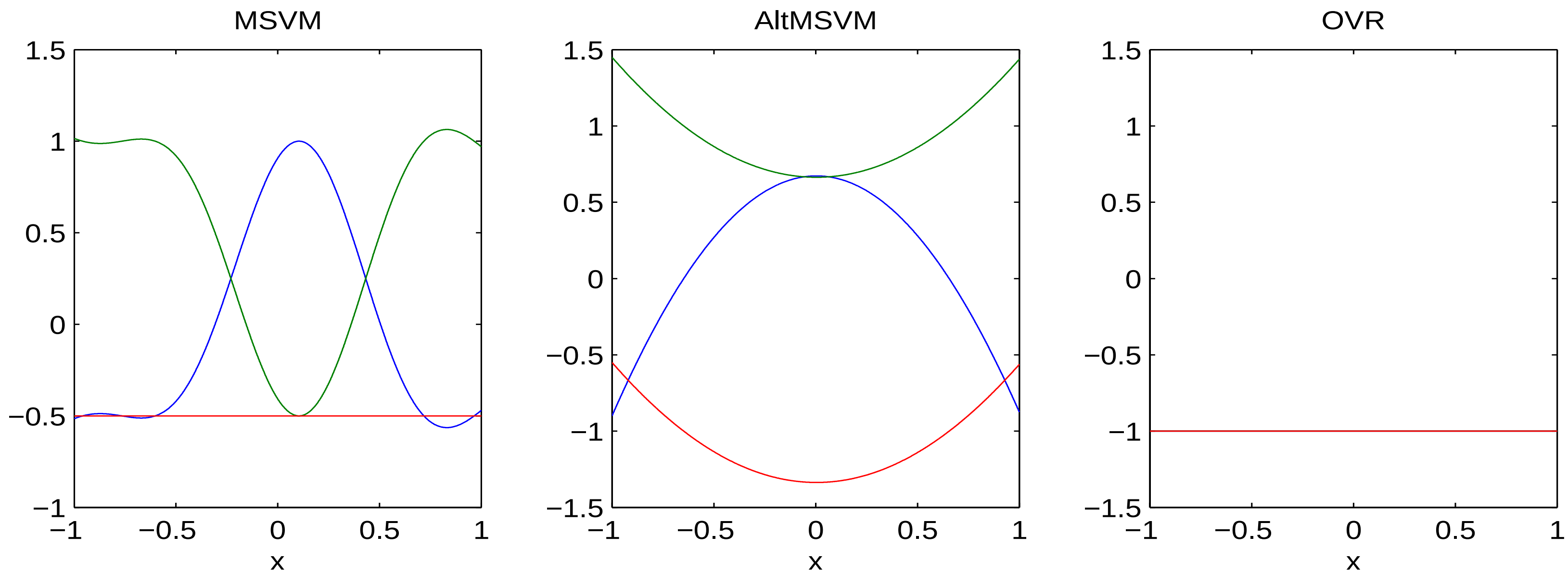


Figure 8: Estimated functions (f_1, f_2, f_3) by the three methods under Scenario 3.

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.

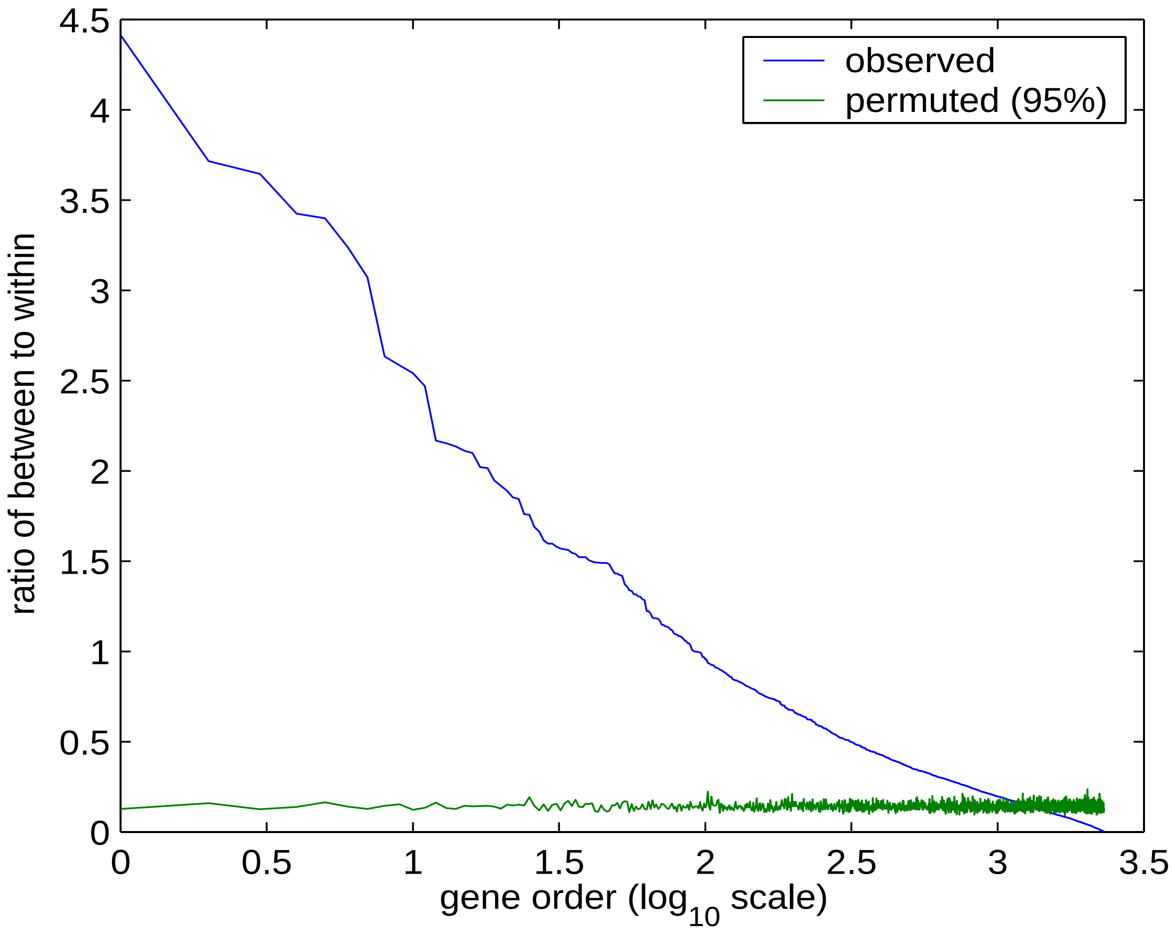


Figure 9: Observed ratios of between-class SS to within-class SS and the 95 percentiles of the corresponding ratios for expression levels with randomly permuted class labels.

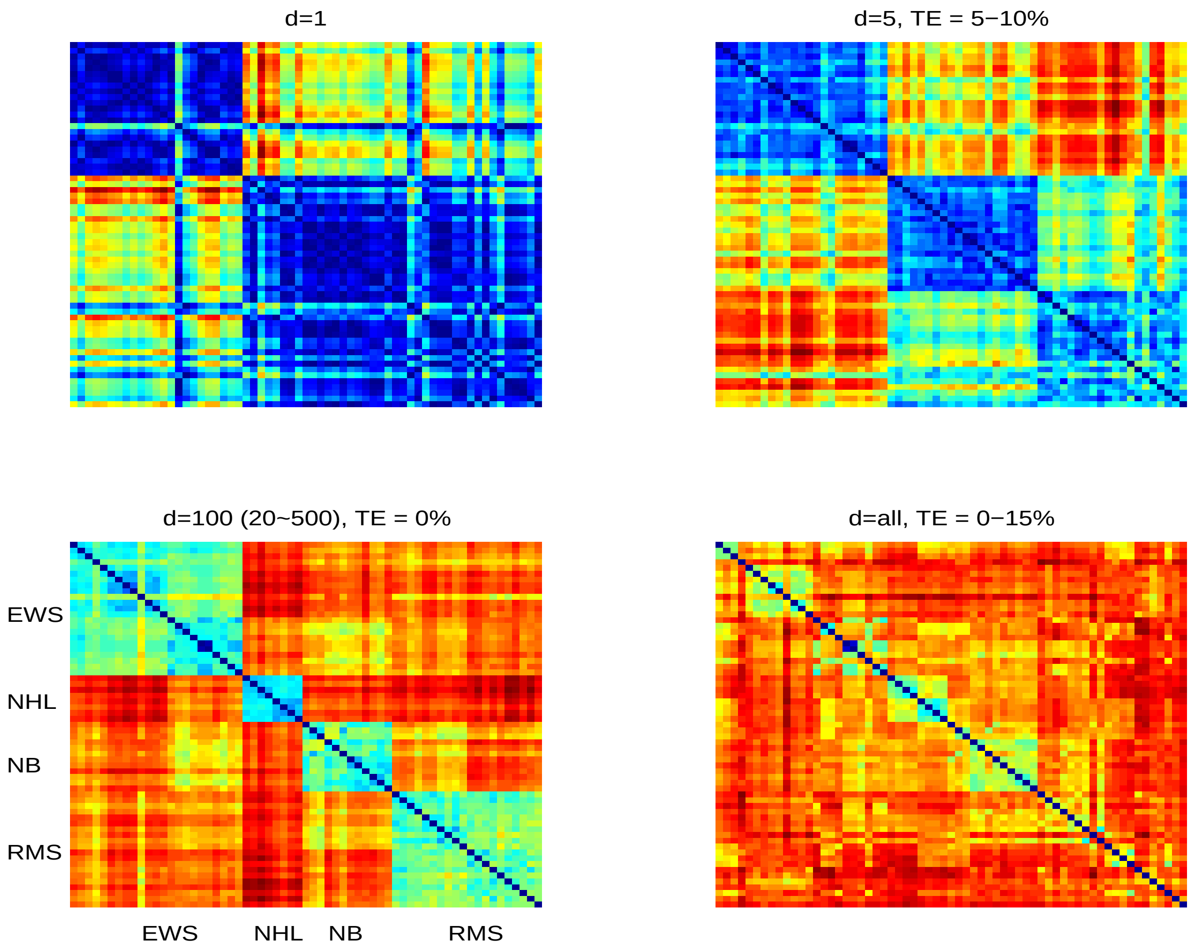


Figure 10: Pairwise distance matrices for the training data as the numbers of genes included change, and test error rates of MSVM with Gaussian kernel.

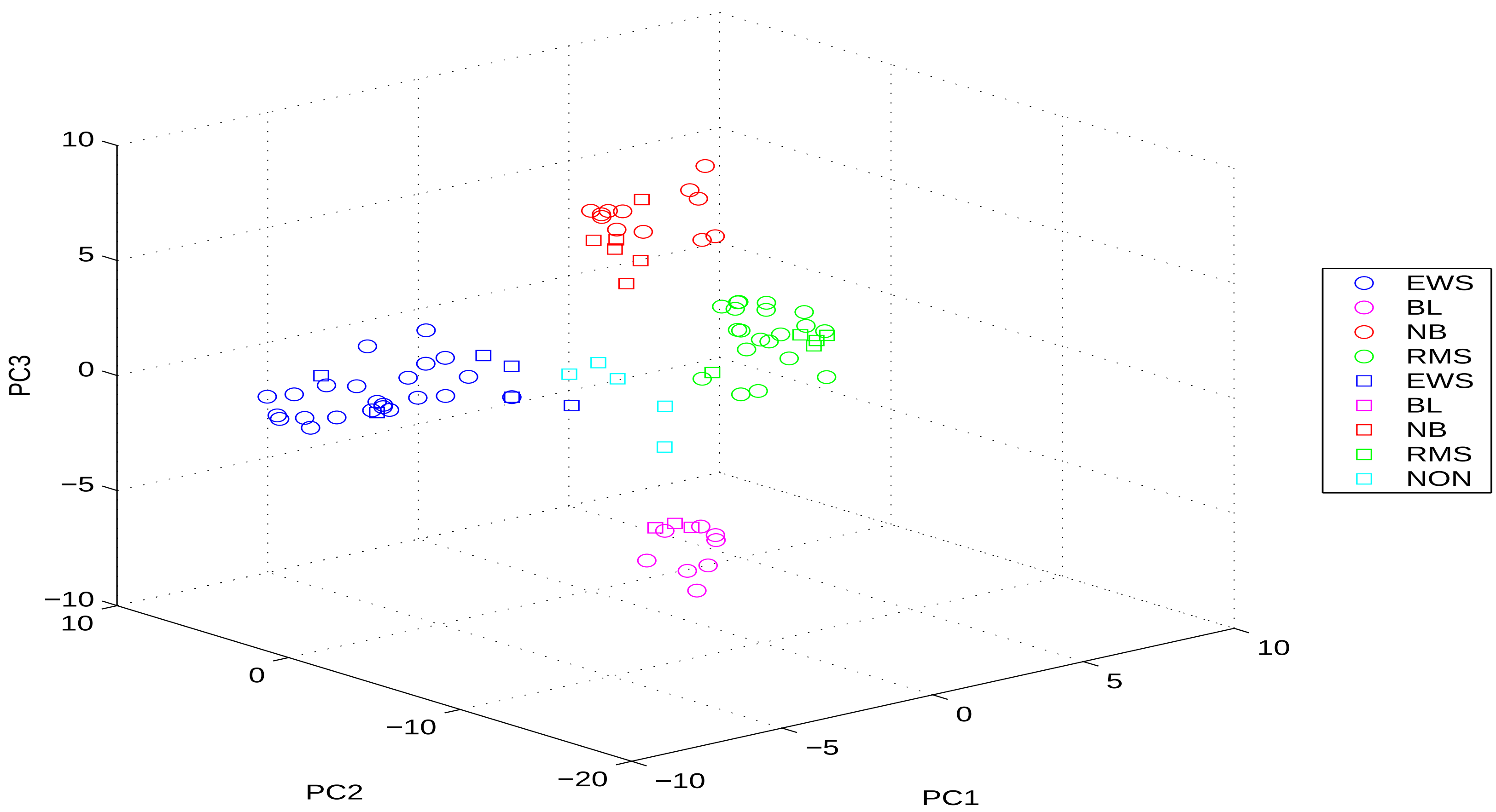


Figure 11: Three principal components of 100 gene expression levels (circles: training samples, squares: test samples including non SRBCT samples). The tumor types are distinguished by colors.

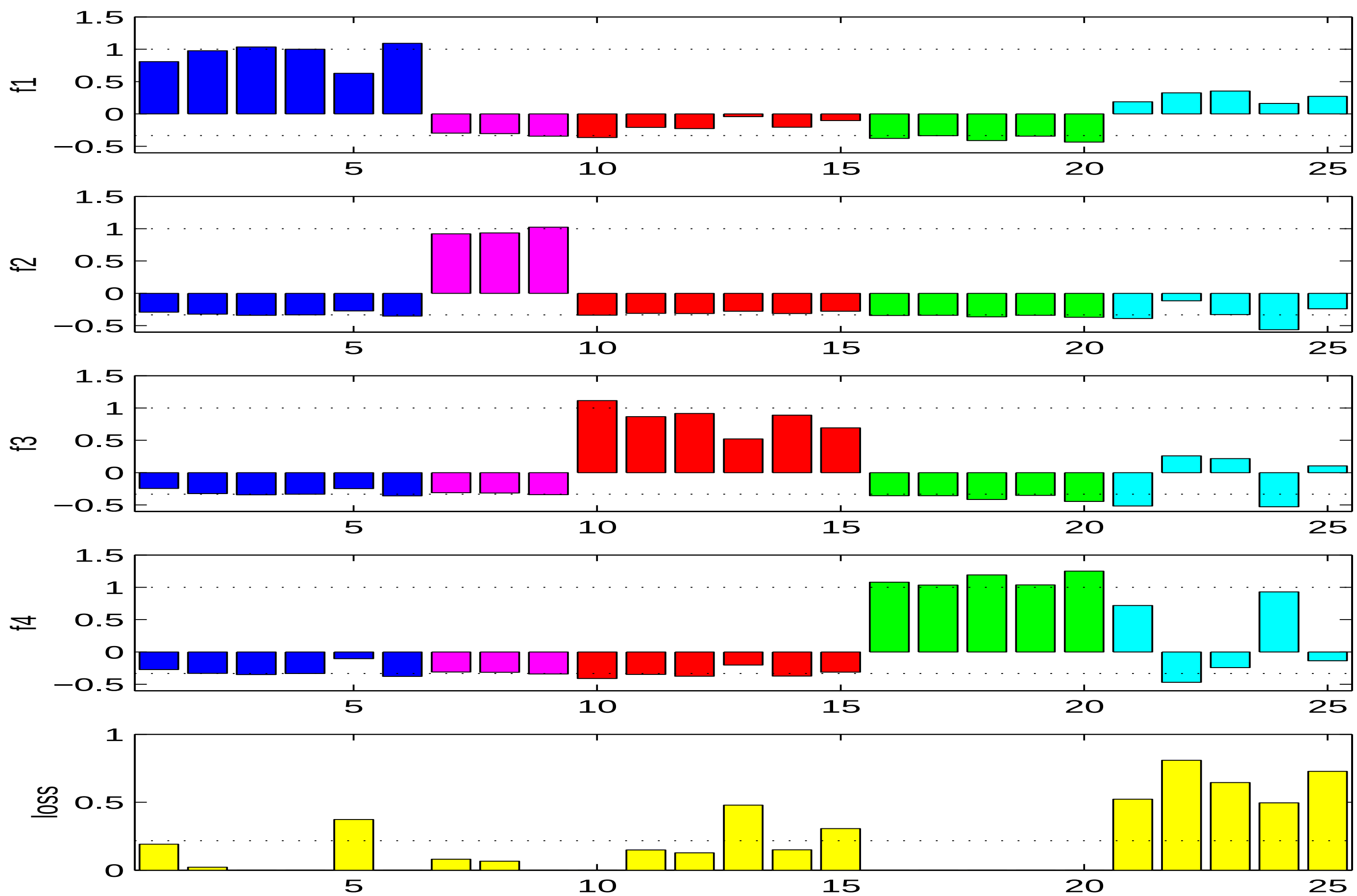


Figure 12: Predicted decision vectors (f_1, f_2, f_3, f_4) at the test samples. EWS: $(1, -1/3, -1/3, -1/3)$, BL: $(-1/3, 1, -1/3, -1/3)$, NB: $(-1/3, -1/3, 1, -1/3)$, and RMS: $(-1/3, -1/3, -1/3, 1)$. The colors indicate the true class identities of the test samples.

Concluding remarks

- Extended SVMs for the binary case to the multicategory case which encompass the binary SVM as a special case.
- Discussed the properties of the MSVM with emphasis on Fisher consistency.
- Multicategory SVM has a great potential for real applications.
- Variable selection and model selection (tuning) for further improvement.
- Generalization error bound analysis (Desyatnikov and Meir 2003).
- Establish theoretical properties.
- Off-the-shelf package.

Slides of this talk and papers are available from

http://www.stat.ohio-state.edu/$\sim$yklee

References

[1] Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proc. 17th International Conf. on Machine Learning*, pages 9–16. Morgan Kaufmann, San Francisco, CA, 2000.

[2] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

[3] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. In *Computational Learing Theory*, pages 35–46, 2000.

[4] Ilya Desyatnikov and Ron Meir. Data-dependent bounds for multi-category classification based on convex losses. In *Proceedings of Fifteenth Annual Conference on Computational Learning Theory 2003, (COLT-03)*, 2003.

[5] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[6] Y. Guermeur. Combining discriminant models with new multi-class SVMs. Technical Report NC-TR-2000-086, NeuroCOLT2, LORIA Campus Scientifique, 2000.

[7] Y. Lee, Y. Lin, and G. Wahba. 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, 2002.

[8] Yoonkyung Lee and Cheol-Koo Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9):1132–1139, 2003.

[9] Y. Lin. A note on margin-based loss functions in classification. Technical Report 1044, Department of Statistics, University of Wisconsin, Madison WI, 2001.

[10] Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

[11] Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers, 2002.

[12] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[13] J. Weston and C. Watkins. Support vector machines for multiclass pattern recognition. In *Proceedings of the Seventh European Symposium On Artificial Neural Networks*, 4 1999.

[14] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. Technical Report rc22155, IBM Research, Yorktown Heights, NY, 2001.

[15] Tong Zhang. An infinity-sample theory for multicategory large margin classification, 2003.