Multicategory Support Vector Machines:

A population view of different approaches

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.

Outline

- Overview of Binary SVM
- Fisher consistency
- Some approaches to multiclass problems
- Multicategory SVM
- Comparisons
- Concluding remarks

Classification Problems

- Training data set :

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

  $i = 1, \cdots, n$, and k is the number of classes.

  (X, Y) : a random sample from $P(\mathbf{x}, y)$.

- Task :

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

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

Support Vector Machine for Binary Case

Vapnik (1995), The Nature of Statistical Learning Theory

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 $i = 1, \cdots, n\}$
- $f(\mathbf{x}) = h(\mathbf{x}) + b$ 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: $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ where $\mathbf{w} \in R^d$ and $b \in R$.
  $\mathcal{H}_K = \{h(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} | \mathbf{w} \in R^d\}$ 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: $f(\mathbf{x}) = \sum_{i=1}^n c_i K(\mathbf{x}_i, \mathbf{x}) + b$
  where K is a positive definite function.
  Popular choices of $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1 \cdot \mathbf{x}_2)^d$, $\exp(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|^2}{2\sigma^2})$

SVM and the Bayes Rule for Binary case

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

> **Lemma 1 (Lin 2001).** *The minimizer of* $E[(1-Yf(X))_+]$ *is* $f(\mathbf{x}) = sign(p^+(\mathbf{x}) - 1/2)$.

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* $(f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})) = (-1, 1, 1)$.

Comparisons

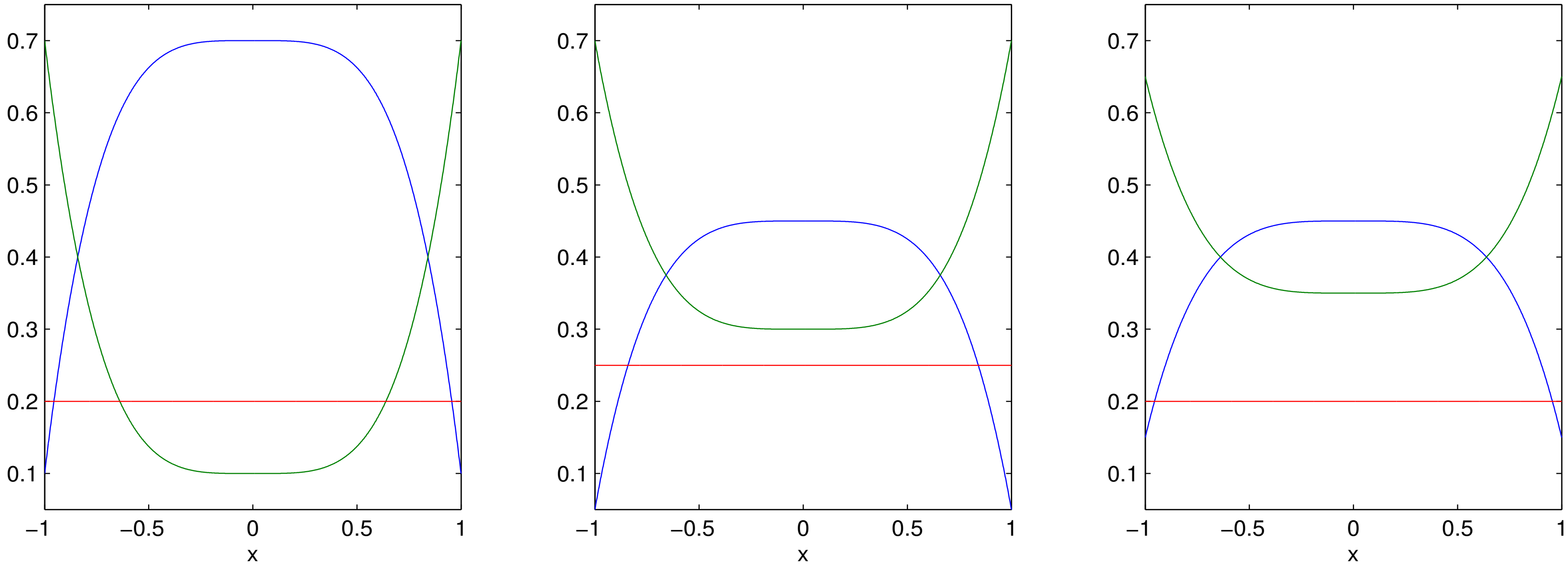

Figure 1: Underlying true conditional probabilities in three situations. Scenario 1: dominant class (left), Scenario 2: the lowest two classes compete (center), and Scenario 3: the highest two classes compete (right).

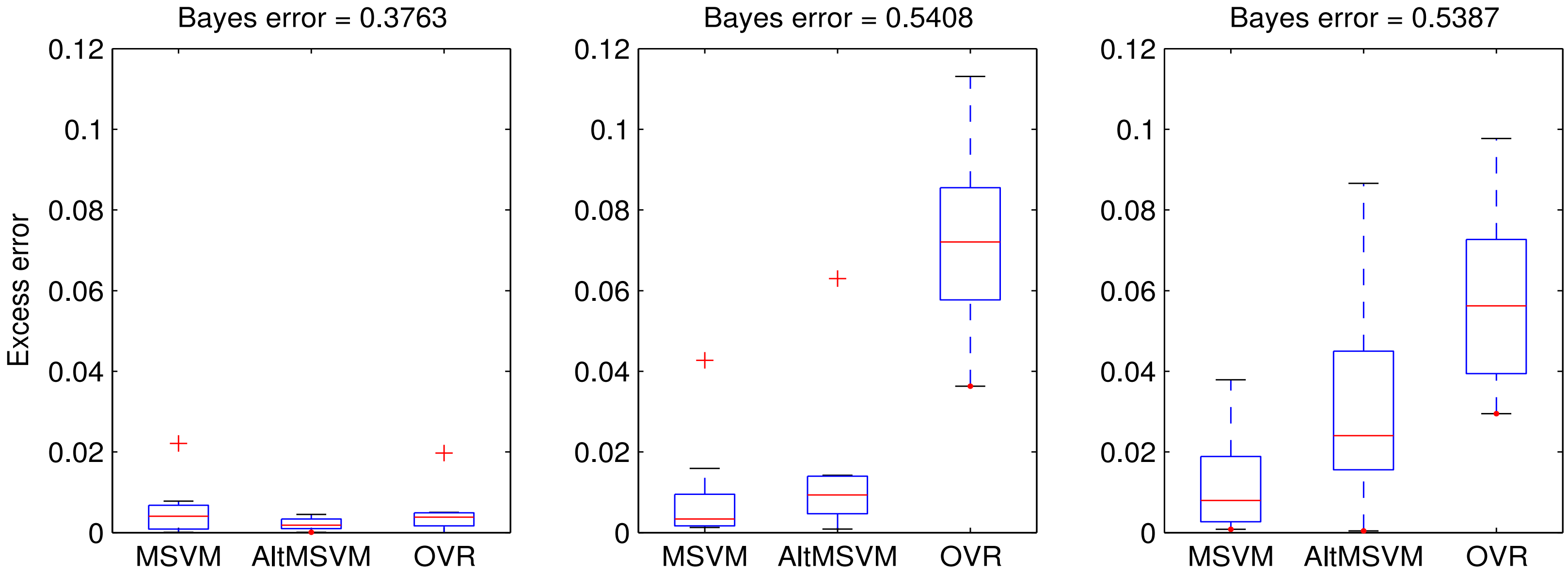

Figure 2: Boxplots of approximated classification error rates from 10 replicates of sample size 200. Scenario 1: left, Scenario 2: center, and Scenario 3: right.

What happened in Scenario 3?

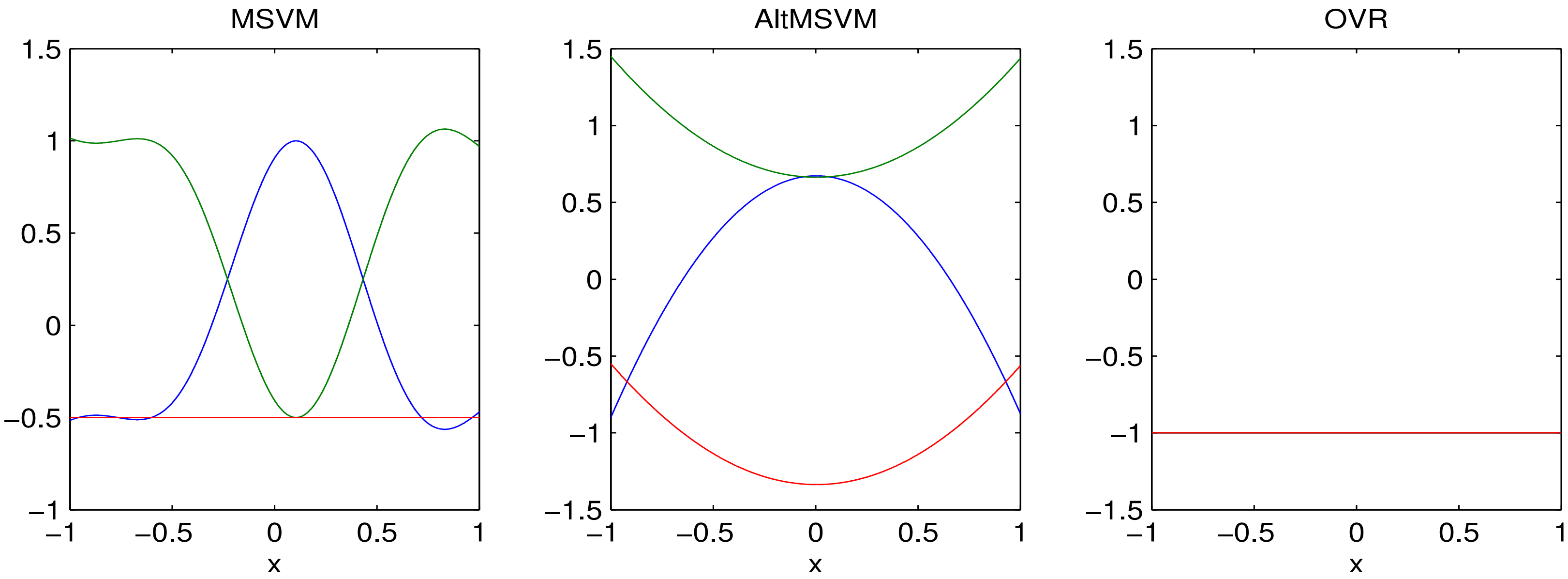

Figure 3: Estimated functions (f_1, f_2, f_3) by the three methods under Scenario 3.

Concluding remarks

- Discussed the properties of the MSVM with emphasis on Fisher consistency.
- Generalization error bound analysis (Desyatnikov and Meir 2003).
- Variable selection and model selection (tuning) for further improvement.
- Establish theoretical properties.

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

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

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

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

[12] 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.

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

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