Extension of Support Vector Machines

Statistical Inference with Reproducing Kernel Hilbert Space

Kenji Fukumizu

Institute of Statistical Mathematics, ROIS
Department of Statistical Science, Graduate University for Advanced Studies

May 30, 2008 / Statistical Learning Theory II

Outline

Multiclass classification - overview - I

- Multiclass classification:
  Classify x in one of L classes $\{1, 2, \ldots, L\}$.
  $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: explanatory variable
  - $Y_i \in \{1, \ldots, L\}$: labels for L classes.

  Make a classifier: $h : \mathcal{X} \to \{1, 2, \ldots, L\}$.
- The original SVM is applicable only to binary classification problems.
- There are some approaches to extending SVM to multiclass classification.
  - Direct construction of a multiclass classifier.
  - Combination of binary classifiers.

Multiclass classification - overview - II

An incomplete list of multiclass extension of SVM and related methods.

- Direct approach:
  - Multiclass SVM ([CS01],[WW98], [BB99], [LLW] etc.)
  - Kernel logistic regression ([ZH02], K.Tanabe, [KDSP05])
  - and others
- Combination approach:
  - How to divide the problem
    - one-vs-rest (one-vs-all)
      i-th class vs the other classes (L binary classification problems)
    - one-vs-one
      i-th vs j-th class ($L(L-1)/2$ binary classification problems)
    - Error correcting output code (ECOC) [DB95]
  - How to combine the binary classifiers
    - Hamming decoding
    - Bradly-Terry model ([HT98], [HWL06])
    - Learning of combiner (stacking [Shi08])

Multiclass SVM I

Multiclass SVM (Crammer & Singer 2001)

- Large margin criterion is generalized to multiclass cases.
- Efficient optimization.
- Implemented in SVM^{light}.

- Linear classifier for L-class classification
  - Data: $(X_1, Y_i), \ldots, (X_N, Y_N), \quad X_i \in \mathbb{R}^m, Y_i \in \{1, \ldots, L\}$.
  - Classifier:
$$h(x) = \arg \max_{\ell=1,\ldots,L} w_\ell^T x.$$
    L linear classifiers are used.
    (The bias term b_ℓ is omitted for simplicity.)
  - $w_\ell^T x$ ($\ell = 1, \ldots, L$) is the similarity score for the class ℓ. The class of the largest similarity is the answer of the classifier.

Multiclass SVM II

- Margin for multiclass problem:

$$Margin_i = w_{Y_i}^T X_i - \max_{\ell \neq Y_i} w_\ell^T X_i.$$

  - $W = (w_1, \ldots, w_L)$ correctly classifies the data (X_i, Y_i), if and only if $Margin_i \geq 0$.
  - The scale of the margin must be fixed.

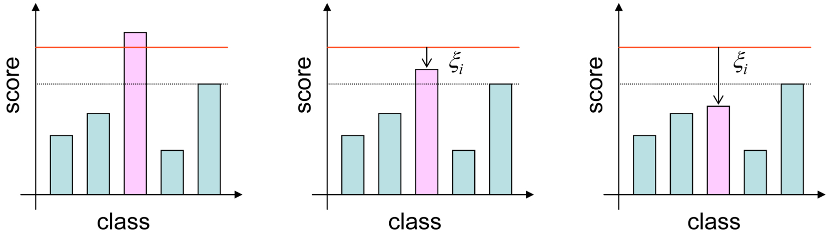
- Large margin classifier (hard margin)

$$\min_W \frac{1}{2} \|W\|^2 \qquad \text{subj. to} \quad w_{Y_i}^T X_i + \delta_{\ell Y_i} - w_\ell^T X_i \geq 1 \quad (\forall \ell, i).$$

  - If $\ell = Y_i$, the constraints are redundant.
  - If $\ell \neq Y_i$, the score must be at least 1 smaller than the score of the true class.

Multiclass SVM III

Meaning of margin



Multiclass SVM IV

Multiclass SVM (soft margin)

- Introducing slack variables $\xi_i \geq 0$ ($i = 1, \ldots, N$)

$$\max_{\ell}\left(w_\ell^T X_i + 1 - \delta_{\ell Y_i}\right) - w_{Y_i}^T X_i = \xi_i \qquad (\forall i).$$

ξ_i represents the break of the separability.

- Primal problem of multiclass SVM:

$$\min_{W,\xi} \frac{\beta}{2}\|W\|^2 + \sum_{i=1}^{N} \xi_i \qquad \text{subj. to} \quad w_{Y_i}^T X_i + \delta_{\ell Y_i} - w_\ell^T X_i \geq 1 - \xi_i \quad (\forall \ell, i).$$

Note: for $\ell = Y_i$, the inequality constraints become $\xi_i \geq 0$.

Dual of multiclass SVM I

- Lagrangian:

$$L(W,\xi,\eta) = \frac{\beta}{2}\|W\|^2 + \sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\sum_{\ell=1}^{L}\eta_{i\ell}\big((w_\ell - w_{Y_i})^T X_i - \delta_{\ell Y_i} + 1 - \xi_i\big).$$

$(\eta_{i\ell} \geq 0, \forall \ell, i)$

- Dual function:

$$\begin{aligned} \nabla_{\xi_i} L = 0 \quad &\Longrightarrow \quad \textstyle\sum_\ell \eta_{i\ell} = 1, \\ \nabla_{w_\ell} L = 0 \quad &\Longrightarrow \quad w_\ell = \beta^{-1}\textstyle\sum_i (\delta_{Y_i \ell} - \eta_{i\ell}) X_i \end{aligned}$$

- X_i is a support pattern if and only if $\eta_{i\ell}$ is *not* concentrated on the true label Y_i.
  (Note: $\eta_{i\ell} \geq 0$ and $\sum_\ell \eta_{i\ell} = 1$.)

Dual of multiclass SVM

- Let

$$\tau_i = e_{Y_i} - \eta_i, \qquad \text{where } e_r = (0, \ldots, 0, 1, 0, \ldots, 0).$$

- Dual problem:

$$\min_{\tau} : \quad g(\tau) = -\frac{1}{2}\sum_{i,j=1}^{N} \left(X_i^T X_j\right) \tau_i^T \tau_j + \beta \sum_{i=1}^{N} \tau_i^T e_{Y_i},$$
$$\text{subject to} \quad \tau_i \leq e_{Y_i} \quad (\forall i) \qquad \sum_{\ell=1}^{L} \tau_{i\ell} = 1.$$

- Classifier:

$$h(x) = \arg\max_{\ell=1,\ldots,L} \left(\sum_{i=1}^{N} \tau_{i\ell}^{*}(X_i^T x)\right).$$

- Kernelization: Just replace $(X_i^T X_j)$ and $(X_i^T x)$ by $k(X_i, X_j)$ and $k(X_i, x)$.

Efficient computation I

- The dual problem is QP with $L \times N$ variables. Direct application of a QP solver may be difficult.
- Efficient computation 1: Decomposition into N subproblems
  - Select an example $p \in \{1, \ldots, N\}$ one by one.
  - Solve a subproblem over τ_p.

$$(*) \qquad \min_{\tau_p} \frac{1}{2} a_p \tau_p^T \tau_p + b_p^T \tau_p,$$

$$\text{subj. to } \tau_p \le e_{Y_p}, \quad \sum_{\ell=1}^{L} \tau_{p\ell} = 1,$$

  where $a_p = k(X_p, X_p)$ and $b_p = \sum_{i \neq p} k(X_i, X_p)\tau_p - \beta e_{Y_p}$.
  - The example p is chosen by the degree of breaking KKT condition.

Efficient computation II

- Efficient computation 2: Optimization by fixed point algorithm
  - The subproblem over τ_p has a special form: the coefficient of quadratic term is a scaler matrix.
  - The solution of the dual of the subproblem (*) is reduced to a fixed point problem:

$$\theta^* = \frac{1}{L}\sum_{\ell=1}^{L} \max\{\theta^*, d_\ell\} - \frac{1}{L},$$

    where θ is a Lagrange multiplier and d_ℓ is a constant.
  - Use iteration

$$\theta^{new} = \frac{1}{L}\sum_{\ell=1}^{L} \max\{\theta^{old}, d_\ell\} - \frac{1}{L}.$$

1. Multiclass classification with SVM
2. Combination of binary classifiers
3. Structured output
4. Others

Combination of binary classifiers

- Base classifiers: make use of strong binary classifiers, and combine their outputs. *e.g.* SVM, AdaBoost, etc.
- Decomposition of a multiclass classification into binary classifications
  - 1-vs-rest
    i-class vs the other classes – L problems
  - 1-vs-1
    i-class vs j-class ($\forall i, j \in \{1, \ldots, L\}$) – $L(L-1)/2$ problems
  - More general approach = Error correcting output code (ECOC). ECOC attributes a code for each class.

class	f_1	f_2	f_3	f_4	f_5	f_6
C_1	-1	-1	-1	1	1	1
C_2	-1	1	1	-1	-1	1
C_3	1	-1	1	-1	1	-1
C_4	1	1	-1	-1	1	1

Combining base classifiers

- Hamming decoding for ECOC:
  Let $W_{\ell a}$ be the code of ECOC for the class ℓ and classifier f_a $(1 \leq \ell \leq L, 1 \leq a \leq M)$.

$$h(x) = \arg\min_{\ell} \|w_\ell - f(x)\|_{Hamming},$$

  where $f(x) = (f_1(x), \ldots, f_M(x)) \in \{\pm 1\}^M$.
  This is equivalent to

$$h(x) = \arg\max_{\ell} \sum_{a=1}^{M} W_{\ell a} f_a(x).$$

- In the case of one-vs-one, Hamming decoding coincides with majority vote, which returns the class with the most "votes".
- Bradly-Terry model:
  A probabilistic model for paired comparison. It can be applied when the output of $f_i(x)$ is continuous.

Learning combiner

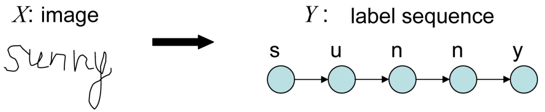
- Given base classifiers $\{f_i(x)\}_{a=1}^{M}$, consider a linear combination function
$$h(x) = \arg\max_{\ell} \textstyle\sum_{a=1}^{M} v_{\ell a} f_a(x).$$
- It is reasonable to expect that adapting v by the data increases the classification accuracy.
- A better combination is possible, if we avoid overfitting caused by reusing the data for both of base classifiers and combiner.

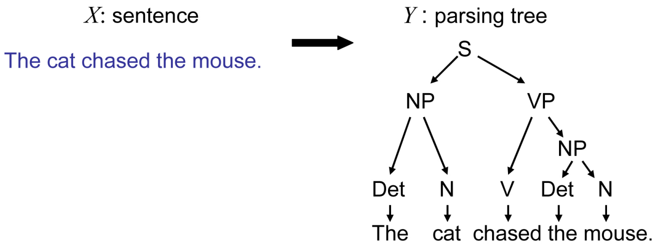
  Stacking via cross-validation ([Shi08]):

$$\min_{v} \sum_{i=1}^{N} \left\| Y_i - \textstyle\sum_{a=1}^{M} v_a f_a^{[-i]}(X_i) \right\|^2 + \lambda \|v\|^2.$$

Structured output

- The output of prediction may be structured object, such as label sequence (strings), trees, and graphs.





Large margin approach to structured output I

References

- Application to natural language processing [Col02].
- Max-Margin Markov Network ($\mathrm{M}^3\mathrm{N}$) [TGK04].
- Hidden Markov support vector machine [ATH03].

Approach

- Assign for x a structured object $y \in \mathcal{Y}$.
- $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: input variable,
  - $Y_i \in \mathcal{Y}$: structured object.
- Feature vector

$$F(x, y) = (f_1(x, y), \ldots, f_M(x, y))$$

Make a classifier: $h : \mathcal{X} \to \mathcal{Y}$

$$h(x) = \arg\max_{y \in \mathcal{Y}} w^T F(x, y).$$

Large margin approach to structured output II

Formulate the problem as a multiclass classification.
Each $y \in \mathcal{Y}$ is regarded as a *class*.

- Multiclass SVM gives

$$\min_{W,\xi} \frac{\beta}{2}\|w\|^2 + \textstyle\sum_{i=1}^{N}\xi_i$$

subj. to $\quad w^T F(X_i, Y_i) + \delta_{yY_i} - w^T F(X_i, y) \geq 1 - \xi_i \quad (\forall i, y \in \mathcal{Y}).$

- Problem:
  # constrains (= # dual variables) = $|\mathcal{Y}|$.
  This is prohibitive in many cases!
  e.g. for label sequence

$$|\mathcal{Y}| = |\text{Alphabet}|^{\text{length}}.$$

Large margin approach to structured output III

- The computational cost must be reduced by some methods.
  - Reducing the dual variables according to the graph structure [TGK04].
    The variables correspond to the nodes and edges.
  - Cutting plane method (selecting variables) [ATH03].

1. Multiclass classification with SVM
2. Combination of binary classifiers
3. Structured output
4. Others

Other topics

- Support vector regression. [MM00]
- ν-SVM: Another formulation of soft margin. [SSWB00]
  ν = an upper bound on the fraction of margin errors.
  ν = the lower bound on the fraction of support vectors.
- one-class SVM: (similar to estimating a level set of density function.)
- Large margin approach to ranking.

References I

Y. Altun, I. Tsochantaridis, and T. Hofmann.
Hidden markov support vector machines.
In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

Erin J. Bredensteiner and Kristin P. Bennett.
Multicategory classification by support vector machines.
Computational Optimizations and Applications, 12, 1999.

Michael Collins.
Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

References II

Koby Crammer and Yoram Singer.
On the algorithmic implementation of multiclass kernel-based vector machines.
Journal of Machine Learning Research, 2:265–292, 2001.

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

T. Hastie and R. Tibshirani.
Classification by pairwise coupling.
The Annals of Statistics, 26(1):451–471, 1998.

Tzu-Kuo Huang, Ruby C. Weng, and Chih-Jen Lin.
Generalized Bradly-Terry models and multi-class probability estimates.
Journal of Machine Learning Research, 7:85–115, 2006.

References III

S. S. Keerthi, K. B. Duan, S. K. Shevade, and A. N. Poo.
A fast dual algorithm for kernel logistic regression.
Machine Learning, 61(1–3):151–165, 2005.

Y. Lee, Y. Lin, and G. Wahba.
Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data.
Journal of the American Statistical Association, 99.

O. L. Mangasarian and D. R. Musicant.
Robust linear and support vector regression.
IEEE Trans. Pattern Analysis Machine Intelligence, 22, 2000.

References IV

Yuichi Shiraishi.
Game-theoretical and statistical study on combination of binary classifiers for multi-class classification.
Ph.D. thesis, Department of Statistical Science, The Graduate University for Advanced Studies, 2008.

B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett.
New support vector algorithms.
Neural Computation, 12:1207–1245, 2000.

Ben Taskar, Carlos Guestrin, and Daphne Koller.
Max-margin markov networks.
In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

References V

J. Weston and C. Watkins.
Multi-class support vector machines.
Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, 1998.

Ji Zhu and Trevor Hastie.
Kernel logistic regression and the import vector machine.
14:1081–1088, 2002.