Support Vector Machine II

Statistical Data Analysis with Positive Definite Kernels

Kenji Fukumizu

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

October 6-10, 2008, Kyushu University

Outline

Generalization ability of SVM

- Framework of risk bound
- Risk bound of SVM

Extension of SVM

- Multiclass classification with SVM
- Combination of binary classifiers
- Structured output and others

Risk and empirical risk : Terminology

Supervised learning:

- $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$: data. i.i.d. sample.
- $X_i \in \mathcal{X}$: input, $Y_i \in \mathcal{Y}$: output.
- $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y}\}$: function class.

Risk and empirical risk

- Loss function $\ell(y, f)$: measure discrepancy of Y_i and $f(X_i)$.
- Risk: the purpose of learning is to minimize the risk;

$$L(f) = E[\ell(Y, f(X))] \qquad (f \in \mathcal{F}).$$

- Empirical risk:

$$L_n(f) = \widehat{E}_n[\ell(Y, f(X))] = \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, f(X_i)) \qquad (f \in \mathcal{F}).$$

- Learning must be done with data:

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} L_n(f).$$

Loss function

- Mean square error.
  - $\ell(y, f) = (y - f)^2$.
  - Empirical risk: $\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$ (least mean square).
  - Risk = $E[(Y - f(X))^2]$.
- 0-1 loss. $y, f(x) \in \{\pm 1\}$.
  - $\ell(y, f) = \frac{1 - yf(x)}{2}$.
  - Empirical risk = ratio of errors:
    $$\widehat{E}_n[\ell(Y, f(X))] = \frac{1}{n} |\{i \mid Y_i \neq f(X_i)\}|.$$
  - Risk = mean error rate: $E[\ell(Y, f(X))] = \Pr(Y \neq f(X))$.
- Log likelihood
  - $\ell(y, f) = -\log p(y|f)$.
  - Empirical risk = - Empirical log likelihood.
  - Risk = - Expected log likelihood.

Bounding risk I

- Goal: What can we say about $L(\widehat{f})$?

$$L(\widehat{f}) - \underbrace{\widehat{L}_n(\widehat{f})}_{\text{known}} = \underbrace{E[\ell(Y,\widehat{f}(X))|\mathcal{D}] - \widehat{E}_n[\ell(Y,\widehat{f}(X))]}_{?}.$$

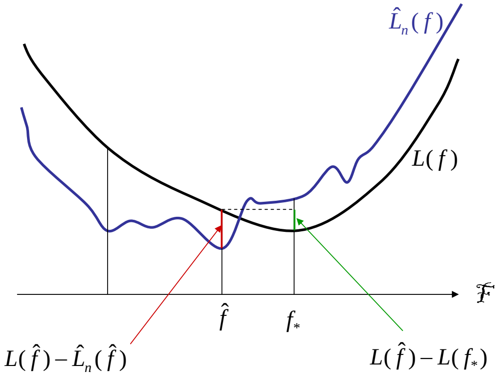
- Approaches to analysis.
  - Asymptotic expansion of the expectation:

    e.g. $$E_{\mathcal{D}}\big[E[\ell(Y,\widehat{f}(X))] - \widehat{E}_n[\ell(Y,\widehat{f}(X))]\big] = \frac{A}{n} + \dots$$

    $\Longrightarrow$ AIC, GIC.
  - Bounding risk:

    e.g. $$\begin{aligned}&\Pr\big(E[\ell(Y,\widehat{f}(X))|\mathcal{D}] \le \widehat{E}_n[\ell(Y,\widehat{f}(X))] + \varepsilon\big)\\ &\le \Pr\Big(\sup_{f\in\mathcal{F}}\big(E[\ell(Y,f(X))] - \widehat{E}_n[\ell(Y,f(X))]\big) \le \varepsilon\Big) \quad \le \alpha e^{-\beta\varepsilon^2 n}.\end{aligned}$$

Bounding risk II



Techniques

- How can we obtained a bound? (not explained in this course)
  - Symmetrization argument
  - Concentration inequality (Hoeffding, Azuma's inequality)
  - Complexity bound (e.g. VC-dimension)

- For basic approach, see e.g. [Vap98].
- More recent approach by Rademacher average [BBM02, BM02].

Surrogate loss I

- Risk is often evaluated by 0-1 loss (error rate)

$$\ell_{01}(y, f) = (1 - y\,\mathrm{sgn}(f))/2.$$

$$L(f) = E[\ell_{01}(y, f(X))] = E[Y \neq \mathrm{sgn}(f(X))].$$

- SVM uses hinge loss for learning:

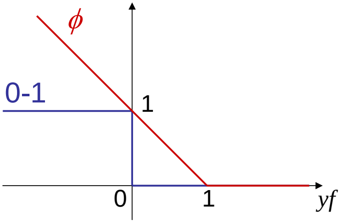
$$\ell_{hinge}(y, f) = \phi(fy), \quad \phi(t) = (1 - t)_+$$

$$\min \widehat{E}_n[\phi(Y_i f(X_i))] + \frac{\lambda}{2}\|f\|^2.$$

- Hinge loss is a surrogate loss function.

$$\ell_{01}(y, f(x)) \leq \phi(yf(x)).$$

Surrogate loss II



Uniform risk bound for SVM I

- Recall margin = $1/\|w\|$ (w: weight of linear classifier).
- Let $R > 0$. Consider

$$\widehat{E}_n[\phi(Yf(X))] \qquad \text{subj. to } \|f\|_{\mathcal{H}_k} \le R.$$

Note: Slightly different from the original SVM.

Theorem 1

Let $\mathcal{F}_R = \{f \in \mathcal{H}_k \mid \|f\|_{\mathcal{H}_k} \le R\}$. For any $\delta > 0$,

$$\Pr\left(\sup_{f \in \mathcal{F}_R} \left| L(f) - \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ \right| \le 2R\sqrt{\frac{E[k(X,X)]}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \right) \ge 1 - \delta$$

Uniform risk bound for SVM II

Theorem 2

Let $\mathcal{F}_R = \{f \in \mathcal{H}_k \mid \|f\|_{\mathcal{H}_k} \leq R\}$. *With probability* $\geq 1 - \delta$,

$$L(f) \leq \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+ + 2R\sqrt{\frac{E[k(X,X)]}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

for any $f \in \mathcal{F}_r$.

- The risk is smaller for a class of larger margin (smaller R), given that the empirical error is the same.
- The complexity term of the function class does not depend on the dimensionality ($\approx$ number of parameters), but only on the norm.

More on the bound for SVM.

- The previous theorem does not reflect the learning of SVM rigorously:

  The margin (norm) is determined as a result of learning, not *a priori*.

- More rigorous approaches to the risk bound of SVM:
  - Bound by fat shattering dimension [BST99].
  - Luckiness framework [Her01].

Multiclass classification - overview - I

- Multiclass classification:
  $(X_1, Y_1), \ldots, (X_N, Y_N)$: data
  - X_i: explanatory variable
  - $Y_i \in \{C_1, \ldots, C_L\}$: 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

Various methods (incomplete list).

- 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)
    - one-vs-one
    - 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.

- Primal problem of multiclass SVM:

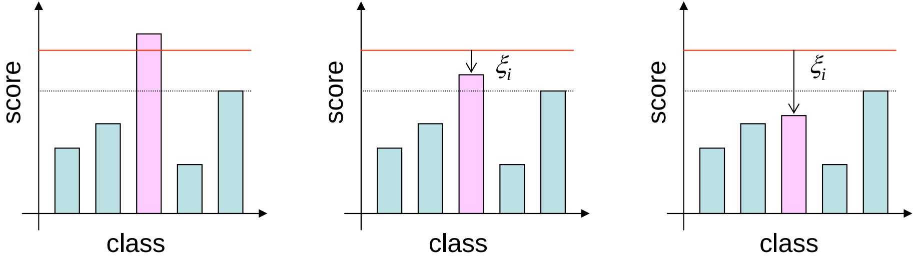
$$\min_{W,\xi} \frac{\beta}{2}\|W\|^2 + \sum_{i=1}^{N} \xi_i \qquad \text{subj. to} \quad w_{Y_i}^T X_i + \delta_{\ell Y_i} - w_\ell^T X_i \geq 1 - \xi_i \quad (\forall \ell, i).$$

  Note: ξ_i represents the break of separability.

- # dual variable = NL. Computational cost must be reduced by some methods.

Multiclass SVM III

Meaning of margin



Generalization ability of SVM

- Framework of risk bound
- Risk bound of SVM

Extension of SVM

- Multiclass classification with SVM
- Combination of binary classifiers
- Structured output and 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

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

- 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]):

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

Generalization ability of SVM

Framework of risk bound

Risk bound of SVM

Extension of SVM

Multiclass classification with SVM

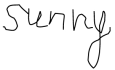
Combination of binary classifiers

Structured output and others

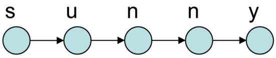
Structured output

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

X: image



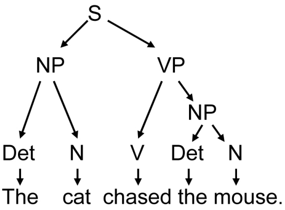
Y: label sequence



X: sentence

The cat chased the mouse.

Y: parsing tree



Large margin approach to structured output I

References

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

Approach

- $(X_1, Y_1), \ldots, (X_N, Y_N)$: 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

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

- The computational cost must be reduced by some methods (e.g. [TGK04, ATH03]).

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

[ATH03] Y. Altun, I. Tsochantaridis, and T. Hofmann.
Hidden markov support vector machines.
In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[BB99] Erin J. Bredensteiner and Kristin P. Bennett.
Multicategory classification by support vector machines.
Computational Optimizations and Applications, 12, 1999.

[BBM02] P. Bartlett, O. Bousquet, and S. Mendelson.
Localized rademacher complexities.
In *Proceedings of the 15th annual conference on Computational Learning Theory*, pages 44–58, 2002.

[BM02] Peter L. Bartlett and Shahar Mendelson.
Rademacher and gaussian complexities: Risk bounds and structural results.
Jounral of Machine Learning Research, 3:463–482, 2002.

References II

[BST99] Peter Bartlett and John Shawe-Taylor.
Generalization performance of support vector machines and other pattern classifiers.
pages 43–54, 1999.

[Col02] 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.

[CS01] Koby Crammer and Yoram Singer.
On the algorithmic implementation of multiclass kernel-based vector machines.
Journal of Machine Learning Research, 2:265–292, 2001.

[DB95] Thomas G. Dietterich and Ghulum Bakiri.
Solving multiclass learning problems via error-correcting output codes.
Journal of Artificial Intelligence Research, 2:263–286, 1995.

References III

[Her01] Ralf Herbrich.
Learning Kernel Classifiers: Theory and Algorithms.
Cambridge, MA, USA, 2001.

[HT98] T. Hastie and R. Tibshirani.
Classification by pairwise coupling.
The Annals of Statistics, 26(1):451–471, 1998.

[HWL06] 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.

[KDSP05] 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.

[LLW] 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.

References IV

[MM00] O. L. Mangasarian and D. R. Musicant.
Robust linear and support vector regression.
IEEE Trans. Pattern Analysis Machine Intelligence, 22, 2000.

[Shi08] 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.

[SSWB00] B. Schölkopf, A. Smola, R. C. Williamson, and P. L. Bartlett.
New support vector algorithms.
Neural Computation, 12:1207–1245, 2000.

[TGK04] 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

[Vap98] Vladimir N. Vapnik.
Statistical Learning Theory.
Wiley-Interscience, 1998.

[WW98] 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.

[ZH02] Ji Zhu and Trevor Hastie.
Kernel logistic regression and the import vector machine.
14:1081–1088, 2002.