Statistics 860. Lecture 14

OUTLINE

1. Review of Optimal Classification.

2. Comparison of penalized likelihood and SVM classifiers.

3. The standard case SVM – equal cost of misclassification and representative training set. GACV and $\xi\alpha$ tuning for the standard case.

4. Yi Lin's theorem: The (tuned) SVM is estimating the sign of the log-odds ratio and minimizing the expected misclassification rate.

5. Extension to the non-standard case:
Non-representative training set, unequal costs.

google - 600,000 hits in 2011, 14,000,000 in 2016.

Y. Lin, Y. Lee, and G. Wahba, Support vector machines for classification in nonstandard situations, Technical Report 1016, `tr1016.pdf` Has appeared, *Machine Learning, 46, 191-202, 2002*.

Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines, Technical Report 1022 `tr1022.pdf` Has appeared *Machine Learning, 48, 115-136,2002*.

G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–311. MIT Press, 2000. (`svm.pdf`)

Y. Lin. A note on margin based classifiers. *tr1044r.pdf, 2002*. Has appeared in Statistics and Probability Letters.

Short selection of books on Support Vector Machines. See also `kernel-machines.org, amazon.com`

- . Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

- B. Scholkopf and A. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

- B. Scholkopf, C. Burges, and A. Smola. *Advances in Kernel Methods-Support Vector Learning*. MIT Press, 1999.

- B. Scholkopf, K. Tsuda, and J-P.Vert. *Kernel Methods in Computational Biology*. MIT Press, 2004.

- Statistica Sinica. Challenges in statistical machine learning. v. 16, 2006. Special Issue.

- A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans. *Advances in Large Marin Classifiers*. MIT Press, 2000.

The Multicategory SVM-next lecture:

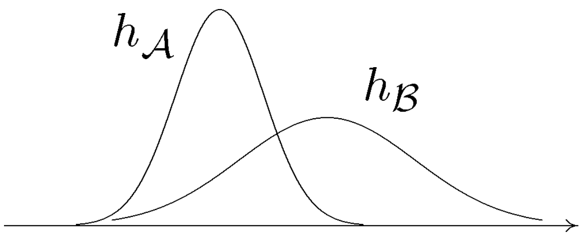
In 860/pdf1:
lee.lee.pdf
lee.lin.wahba.04.pdf
lee.wahba.ackerman.04.pdf
lee.wahba.ackerman.corr.04.pdf
lee.kim.lee.koo.2006.pdf

Where to go for software

- SVM-Light, Thorsten Joachims
  `svmlight.joachims.org`
  In C.

- MSVM-Multicategory SVM, Yoonkyung Lee
  `www.stat.osu.edu/~yklee/software.html`
  Addon to R.

♣♣ Optimal Classification and the Neyman-Pearson Lemma:



$h_{\mathcal{A}}(\cdot), h_{\mathcal{B}}(\cdot)$ densities of x

for class $\mathcal{A}$ and class $\mathcal{B}$.

NOTATION:

$\pi_{\mathcal{A}} =$ prob. next observation (Y) is an $\mathcal{A}$

$\pi_{\mathcal{B}} = 1 - \pi_{\mathcal{A}} =$ prob. next observation is a $\mathcal{B}$

$$\begin{aligned} p(x) &= prob\{Y = \mathcal{A}|x\} \\ &= \frac{\pi_{\mathcal{A}} h_{\mathcal{A}}(x)}{\pi_{\mathcal{A}} h_{\mathcal{A}}(x) + \pi_{\mathcal{B}} h_{\mathcal{B}}(x)} \end{aligned}$$

1

Let $c_{\mathcal{A}} =$ cost to falsely call a $\mathcal{B}$ an $\mathcal{A}$

$c_{\mathcal{B}} =$ cost to falsely call an $\mathcal{A}$ a $\mathcal{B}$

Bayes classification rule: Let

$\phi(x)$: $\quad x \to \{{\mathcal{A} \atop \mathcal{B}}\}$

Expected cost:

$E\{c_{\mathcal{A}}[1-p(x)]\ I(\phi(x)=\mathcal{A})\}$

get a $\mathcal{B}$ and call it an $\mathcal{A}$

$+E\{c_{\mathcal{B}}[p(x)]\ I(\phi(x)=\mathcal{B})\}$

get an $\mathcal{A}$ and call it $\mathcal{B}$

Optimum (Bayes) classifier:

$$\phi_{\text{OPT}}(x) = \begin{cases} \mathcal{A} & \text{if } \frac{p(x)}{1-p(x)} > \frac{c_{\mathcal{A}}}{c_{\mathcal{B}}}, \\ \mathcal{B} & \text{otherwise.} \end{cases}$$

To estimate $p(x)$, alternatively let $f(x) = \log p(x)/(1-p(x))$, the log odds ratio a.k.a. the logit. "Standard" case: Training set

$$\{y_i, x_i\} \qquad \begin{array}{l} y_i \in \{\mathcal{A}, \mathcal{B}\} \\ x_i \in \mathcal{T}, \text{ some index set} \end{array}.$$

Relative frequency of $\mathcal{A}$'s in the training set is about the same as in the general population.

Penalized log likelihood estimation:

Estimate f by penalized likelihood. If $c_{\mathcal{A}}/c_{\mathcal{B}} = 1$, then the optimal classifier is

$$\begin{array}{l} f(x) > 0 \text{ (equivalently, } p(x) - \frac{1}{2} > 0) \rightarrow \mathcal{A} \\ f(x) < 0 \text{ (equivalently, } p(x) - \frac{1}{2} < 0) \rightarrow \mathcal{B} \end{array}$$

♣♣ Penalized log likelihood estimation of the logit $f = \log[p/(1-p)]$.

$$y = \begin{matrix} 1 & = \mathcal{A} \\ 0 & = \mathcal{B} \end{matrix} \text{ (important)}$$

The probability distribution function (likelihood) for $y \mid p$ is: $\mathcal{L} = p^y(1-p)^{1-y} = \begin{cases} p & \text{if } y = 1 \\ (1-p) & \text{if } y = 0 \end{cases}$
and the negative log likelihood is

$$-\log \mathcal{L} = -\log[p^y(1-p)^{1-y}]$$
$$= -y \log p - (1-y)\log(1-p).$$

Using $p = e^f/(1+e^f)$ gives
$-\log \mathcal{L} = -yf + \log(1+e^f)$

♣♣ Penalized log likelihood estimation of f (continued) (special case).

$$\{y_i, x_i\}, \;\; y_i = \begin{matrix} 1 \\ 0 \end{matrix} \;\; , x_i \in \mathcal{T}$$

Find $f(x) = d + h(x)$ with $h \in \mathcal{H}_K$ to min

$$\frac{1}{n}\sum_{i=1}^{n}\left[-y_i f(x_i) + \log(1 + e^{f(x_i)})\right] + \lambda\|h\|^2_{\mathcal{H}_K}$$

where $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) with reproducing kernel

$$K(s,t), \;\; s,t, \in \mathcal{T}.$$

Theorem: (Special case of second variational problem)

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i).$$

♣♣ Penalized log likelihood estimation of f (continued)

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i)$$

Find $d, c = (c_1, \ldots, c_n) = c_\lambda$ to minimize

$\frac{1}{n}\sum_{i=1}^{n}\left[-y_i f(x_i) + \log(1 + e^{f(x_i)})\right] + \lambda\|h\|^2_{\mathcal{H}_K}.$

Here

$$\|h\|^2_{\mathcal{H}_K} \equiv \sum_{i,j=1}^{n} c_i c_j \ K(x_i, x_j).$$

Given λ, this is a nice strictly convex optimization problem. Choose λ by GACV. Target for GACV is to minimize the Comparative Kullback-Liebler (CKL) distance of the estimate from the true distribution:

$$R(\lambda) = E_{f_{true}} \sum_{i=1}^{n} -y_{new.i} f_\lambda(x_i) + \log(1 + e^{f_\lambda(x_i)}).$$

♣♣ Support Vector Machines

$$y = \begin{matrix} +1 = \mathcal{A} \\ -1 = \mathcal{B} \end{matrix} \text{ (note different coding)}$$

Find $f(x) = d + h(x)$ with $h \in \mathcal{H}_K$ to min

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

where $(\tau)_+ = \tau, \tau > 0, = 0$ otherwise.

Then

$$f_\lambda(x) = d + \sum_{i=1}^{n} c_i K(x, x_i). \qquad (*)$$

Substitute (*) into (**), choose λ, given λ, find c and d. The classifier is

$$f_\lambda(x) > 0 \to \mathcal{A}$$

$$f_\lambda(x) < 0 \to \mathcal{B}$$

♣♣ Comparison of the penalized log likelihood estimate f_λ of the log odds ratio $\log p/(1-p)$ and f_λ, the SVM classifier:

Suspicion: They are related...

Let us relabel y in the likelihood –

$$\tilde{y} = \begin{cases} +1 & \text{if } \mathcal{A}, \\ -1 & \text{if } \mathcal{B}. \end{cases}$$

Then

$$-yf + \log(1 + e^f) \to \log(1 + e^{-\tilde{y}f})$$

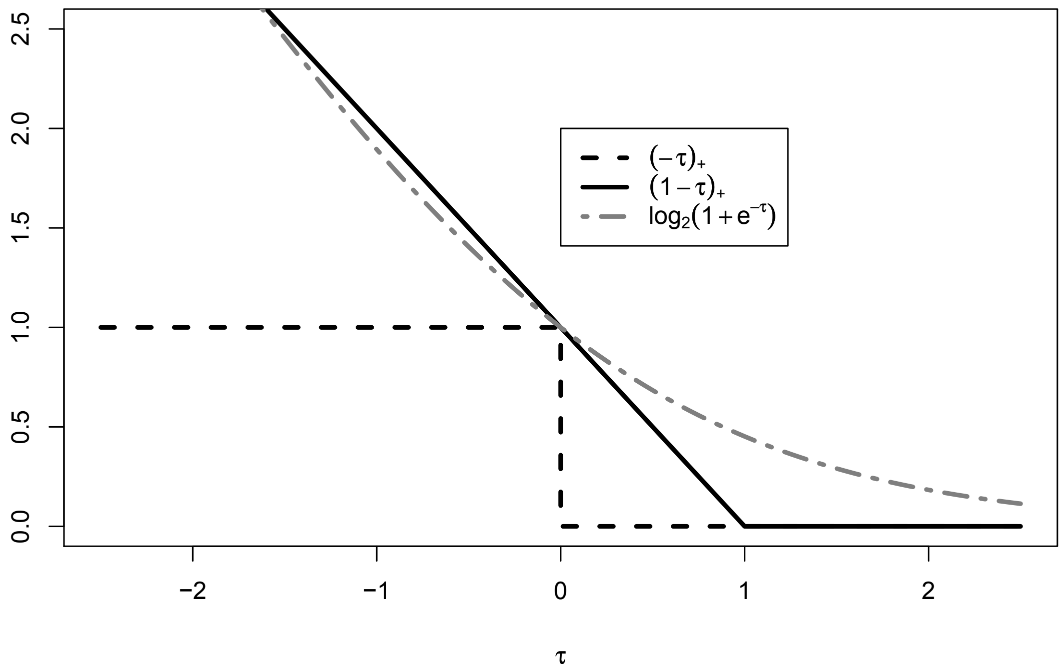
Figure 1 compares

$$\log(1 + e^{-yf}), \quad (1 - yf)_+ \text{ and } (-yf)_*$$

where

$$(\tau)_* = \begin{cases} 1 & \text{if } \tau > 0, \\ 0 & \text{otherwise.} \end{cases}$$

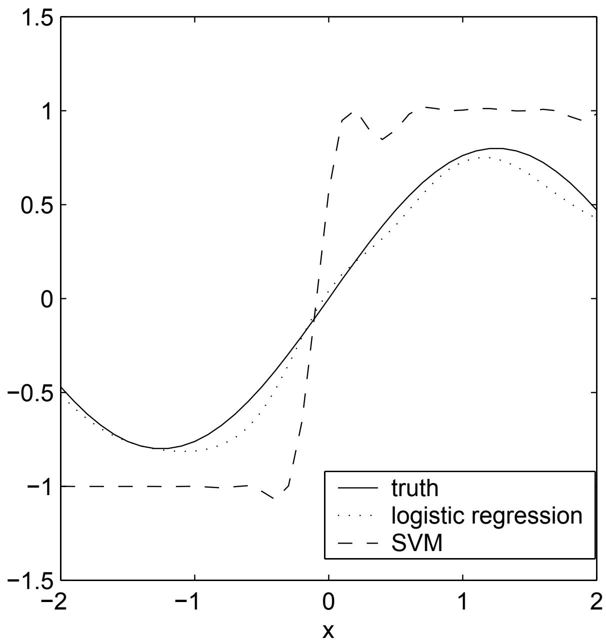
($(-yf)_*$ is the misclassification counter).



[Let $\tau = yf$]. Comparison of $(-\tau)_*$, $(1-\tau)_+$ and $log_e(1+e^{-\tau})$. Bin Yu observed at the talk that $log_2(1+e^{-\tau})$ goes through 1 at $\tau = 0$. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on the missclassification function and will be a looser bound than some SVM function.

The SVM is estimating the sign of the log odds ratio, just what you need for classification

SVM and Penalized Likelihood Estimates Compared:
true: $p(x) = .4(sin 0.4\pi x) + .5$ on [-2,2]
The plots are: $2p - 1$ for 'true' and $2p_{\hat{\lambda}} - 1$ for the penalized likelihood estimate, for comparison to the SVM estimate.



$n = 300, x_i$ equally spaced on $[0, 1]$, y_i simulated according to $p(x)$, coded to ± 1 for the SVM. They give nearly identical classification rules, as determined by the $sign$ of the estimate.