Classification, Penalized Likelihood, and the Support Vector Machine

Grace Wahba

Based on Lee, Lin and Wahba, JASA 2004;
Lee, Wahba and Ackerman, J. Atmos. Ocean Technology, 2004;
Wahba, PNAS 2002;
Xiwu Lin, Thesis 1998 (TR 1003).
Yoonkyung Lee won the ”Best Student Poster” award at the American Meteorological Society 2003 Satellite and Oceanography Session.
References will be up via the TALKS link on my website: `http://www.stat.wisc.edu/˜wahba`

Chicago Chapter of ASA Dinner
East Bank Club, Chicago
December 1, 2004

Abstract

This talk will be a short tutorial on the recently popular support vector machine, (SVM), a relatively recent technique for (nonparametric) classification. We will explain how it relates to penalized likelihood estimates, and how its popularity is well deserved.

OUTLINE

1. Optimal classification and the Neyman-Pearson Lemma.

2. The Penalized Likelihood Estimate, two classes.

3. The Support Vector Machine (SVM), two classes.

4. Tuning the estimates.

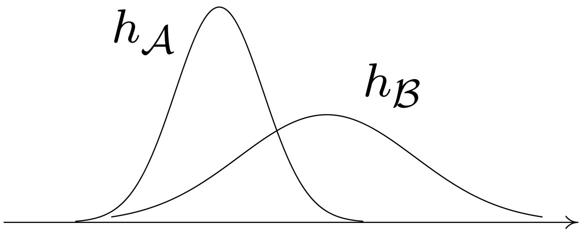
5. The Multicategory Support Vector Machine (MSVM).

6. Application to cloud classification from MODIS data.

7. The Multicategory Penalized Likelihood Estimate.

Closing remarks, more closing remarks.

♣♣ 1. Optimal Classification and the Neyman-Pearson Lemma:



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

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

♣♣ 1.Optimal Classification and the Neyman-Pearson Lemma (cont.).

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

$\phi(t)$: $\quad t \rightarrow \{{}^{\mathcal{A}}_{\mathcal{B}}\}$

Optimum (Bayes) classifier: (Neyman-Pearson Lemma) Minimizes the expected cost:

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

♣♣ 2. Penalized Likelihood Estimation, Two Classes.

Let $f(t) = log \frac{p(t)}{1-p(t)}, \quad p(t) = e^{f(t)}/(1+e^{f(t)})$. For comparison to the SVM coming up next, use the unusual coding:

$$y = \begin{array}{l} +1 = \mathcal{A} \\ -1 = \mathcal{B} \end{array}$$

Then the log likelihood can be written:

$$\mathcal{L}(y, f) = log(1 + e^{-yf}).$$

Given $\{y_i, t_i, i = 1, ..., n\}$, the penalized log likelihood estimate of f is the solution to the problem: Find $f(t) = d + h(t)$ with $h \in \mathcal{H}_K$ to min

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

Then (Kimeldorf & Wahba 1971)

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

$$\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(t_i, t_j). \qquad (**)$$

Substitute (*,**) into (***), choose λ, given λ, find c and d numerically. The estimate $p_\lambda(t)$ of $p(t)$ is recovered from $f_\lambda(t)$.

Since $p = 1/2$ corresponds to $f = 0$, it is possible, if desired, to use f_λ as a classifier, via

$$f_\lambda(t) > 0 \rightarrow \mathcal{A}$$

$$f_\lambda(t) < 0 \rightarrow \mathcal{B}$$

Plot of a penalized likelihood estimate of 19 year risk of a heart attack as a function of cholesterol and diastolic blood pressure, based on data from the Western Electric Health Study (O'Sullivan, Yandell and Raynor, JASA 1986) goes here.

♣♣ 3. The Support Vector Machine, two classes.

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

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

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

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

Then

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

$$\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} c_i c_j K(t_i, t_j). \qquad (**)$$

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

$$f_\lambda(t) > 0 \rightarrow \mathcal{A}$$

$$f_\lambda(t) < 0 \rightarrow \mathcal{B}$$

Numerically, must solve a mathematical programming problem.

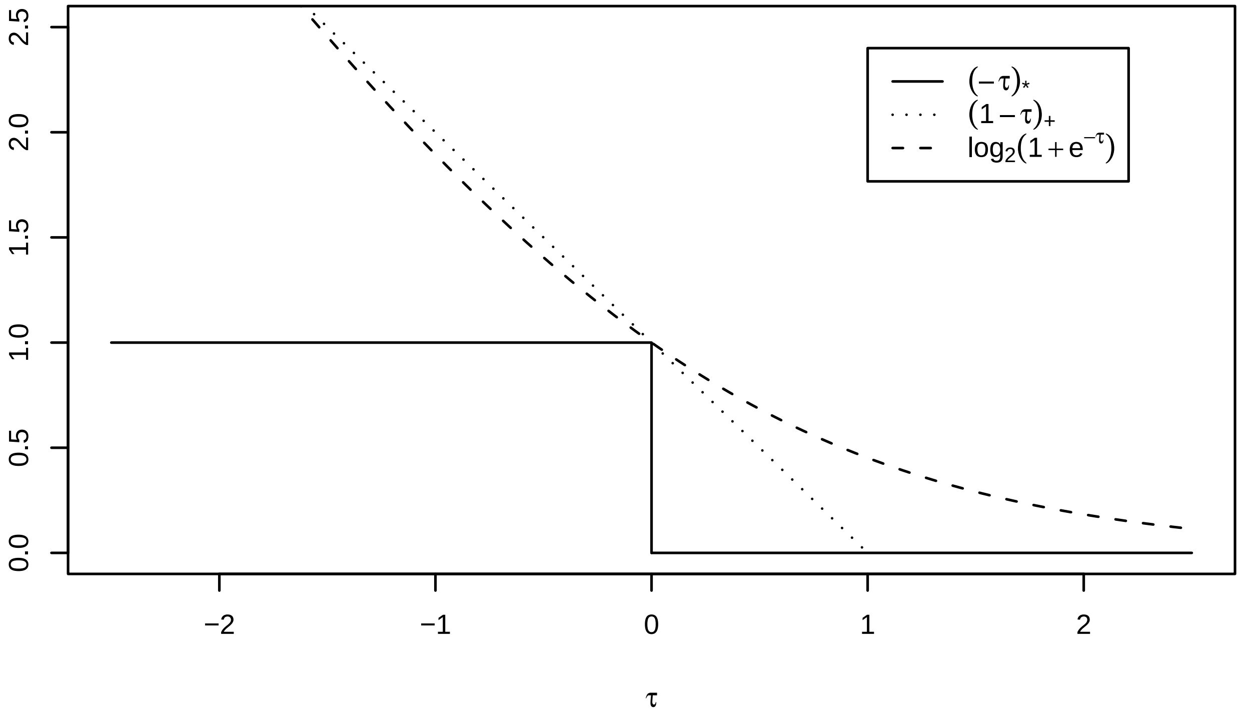


Figure 1. Let $\mathcal{C}(y_i, f(t_i)) = c(y_i f(t_i)) = c(\tau)$. Comparison of $c(\tau) = (-\tau)_*$, $(1-\tau)_+$ and $log_2(1+e^{-\tau})$, the log likelihood function. Any strictly convex function that goes through 1 at $\tau = 0$ will be an upper bound on the misclassification counter $(-\tau_*)$ and will be a looser bound than some SVM (hinge) function $(1-\theta\tau)_+$. $\tau = yf$ is known as the margin - there are many other ”large margin” classifiers....

Figures of several large margin classifiers (from Yi Lin) go here.

♣♣ 3.The SVM (cont.) What is the SVM estimating?.

What is the SVM estimating?

Lemma (Yi Lin 2002) (two category version)

The minimizer of $E(1 - y_{new} f(t))_+$ is $sign\ f(t)$ $(= sign\ (p(t) - \frac{1}{2}) = sign\ (2p(t) - 1))$

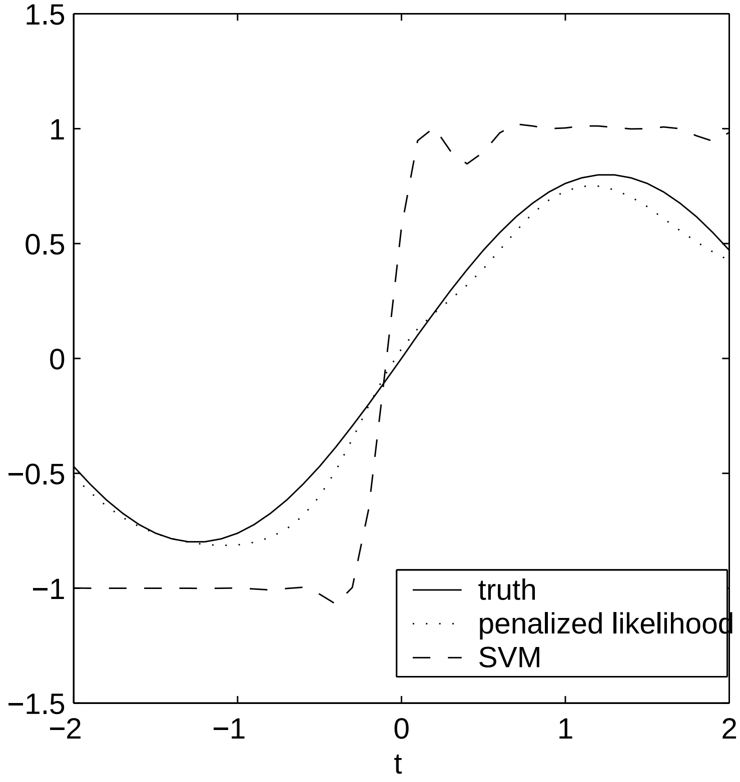
where $f(t) = \log p(t)/(1 - p(t))$.

So the SVM, the solution of the problem: Find $f_\lambda = d + h$ which minimizes

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

where λ is chosen to minimize (a proxy for) $R(\lambda)$, is estimating sign $f(t)$ - not $f(t)$ itself, but just what you need to minimize the misclassification rate.

♣♣ 3. The SVM (cont.).The SVM is not estimating a probability.



300 Bernoulli random variables were generated, equally spaced t from $p(t) = 0.4 sin(0.4\pi t) + 0.5$. Solid line: $(2p(t) - 1)$. Dotted line:$(2p_\lambda - 1)$, p_λ is (optimally tuned) penalized likelihood estimate of p. Dashed line: $f_{svm\ \lambda}$, is (optimally tuned) SVM. Observe $f_{svm\ \lambda} \sim \pm 1$, thus p_λ is estimating $p(t)$, whereas $f_{svm\ \lambda}$ is estimating $sign(2p-1)$ = $sign(p-1/2)$= $sign\ f$. (based on Gaussian K) (plot: Yoonkyung Lee)

♣♣ 4.Tuning the estimates.

The smoothing parameter λ must be chosen. If the Gaussian kernel, $K(s,t) = \exp -\frac{\|s-t\|^2}{\sigma^2}$ is used then σ^2 must also be chosen. λ and σ^2 can be jointly chosen by GACV (Generalized Approximate Cross Validation, 5-fold crossvalidation, or, if copious data is available, by test-tune-train data sets.)

♣♣ 5. The Multicategory Support Vector Machine (MSVM).

From [LeeLinWahba04],[LeeWahbaAckerman04], earlier reports. $k > 2$ categories.
Coding:

$$y_i = (y_{i1}, \cdots, y_{ik}), \sum_{j=1}^{k} y_{ij} = 0,$$

in particular $y_{ij} = 1$ if the ith subject is in category j and $y_{ij} = -\frac{1}{k-1}$ otherwise. $y_i = (1, -\frac{1}{k-1}, \cdots, -\frac{1}{k-1})$ indicates y_i is from category 1. The MSVM produces $f(t) = (f^1(t), \cdots f^k(t))$, with each $f^j = d^j + h^j$ with $h^j \in \mathcal{H}_K$, *required to satisfy a sum-to-zero constraint*

$$\sum_{j=1}^{k} f^j(t) = 0,$$

for all t in $\mathcal{T}$. The largest component of f indicates the classification.

♣♣ 5. The Multicategory Support Vector Machine (MSVM)(cont.).

The MSVM is defined as the vector of functions $f_\lambda = (f_\lambda^1, \cdots, f_\lambda^k)$, with each h^k in $\mathcal{H}_K$ satisfying the sum-to-zero constraint, which minimizes

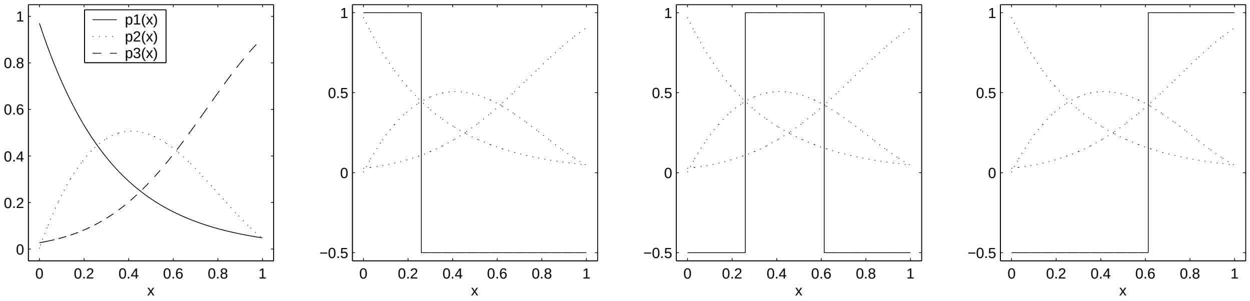
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)}(f^r(t_i)+\frac{1}{k-1})_+ + \lambda\sum_{j=1}^{k}\|h^j\|^2_{\mathcal{H}_K}$$

where $cat(i)$ is the category of y_i. (So, there is no cost term for $r = cat(i)$ but a cost in the other terms unless $f^r(t_i) \leq -\frac{1}{k-1}$)

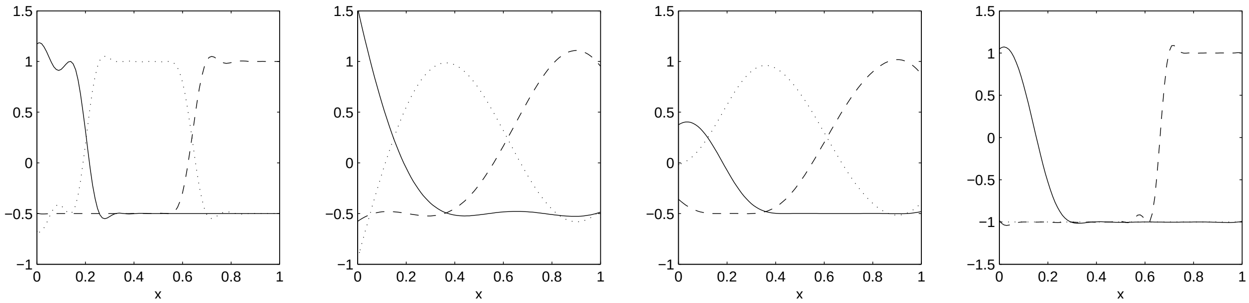
The $k = 2$ case reduces to the usual 2-category SVM.

The target for the MSVM is $f(t) = (f^1(t), \cdots, f^k(t))$ with $f^j(t) = 1$ if $p_j(t)$ is bigger than the other $p_l(t)$ and $f^j(t) = -\frac{1}{k-1}$ otherwise.

♣♣ 5. The Multicategory Support Vector Machine (MSVM)(cont.).



Above: Probabilities and target f^j's for three category SVM demonstration.(Gaussian Kernel)



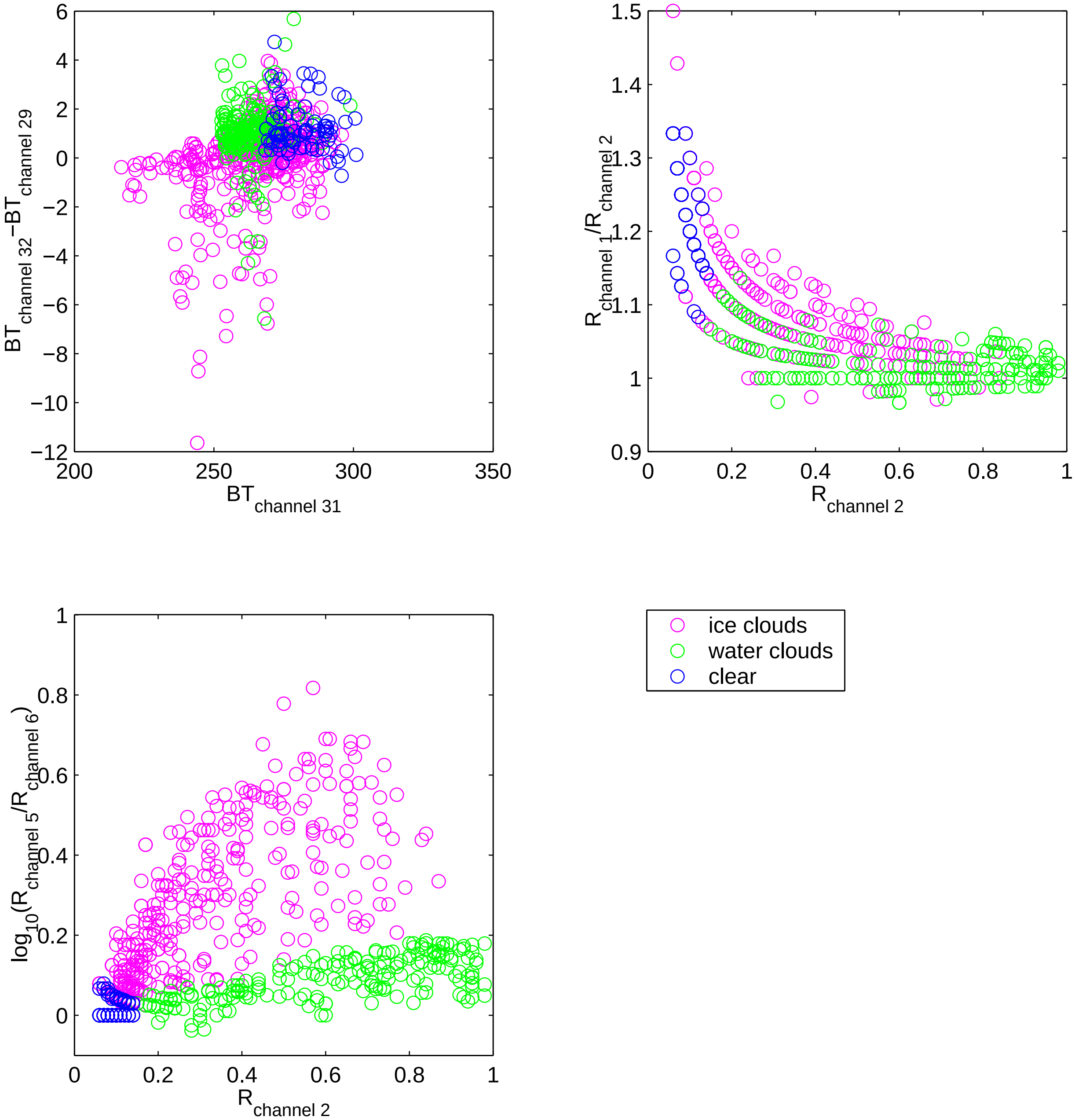
The left panel above gives the estimated f^1, f^2 and f^3. λ and σ were optimally tuned. (i. e. with the knowledge of the 'right' answer). In the second from left panel both λ and σ were chosen by 5-fold cross validation in the MSVM and in the third panel they were chosen by GACV. In the rightmost panel the classification is carried out by a one-vs-rest method.

♣♣ 6. Application to the classification of upwelling MODIS radiance data to clear sky, water clouds or ice clouds.

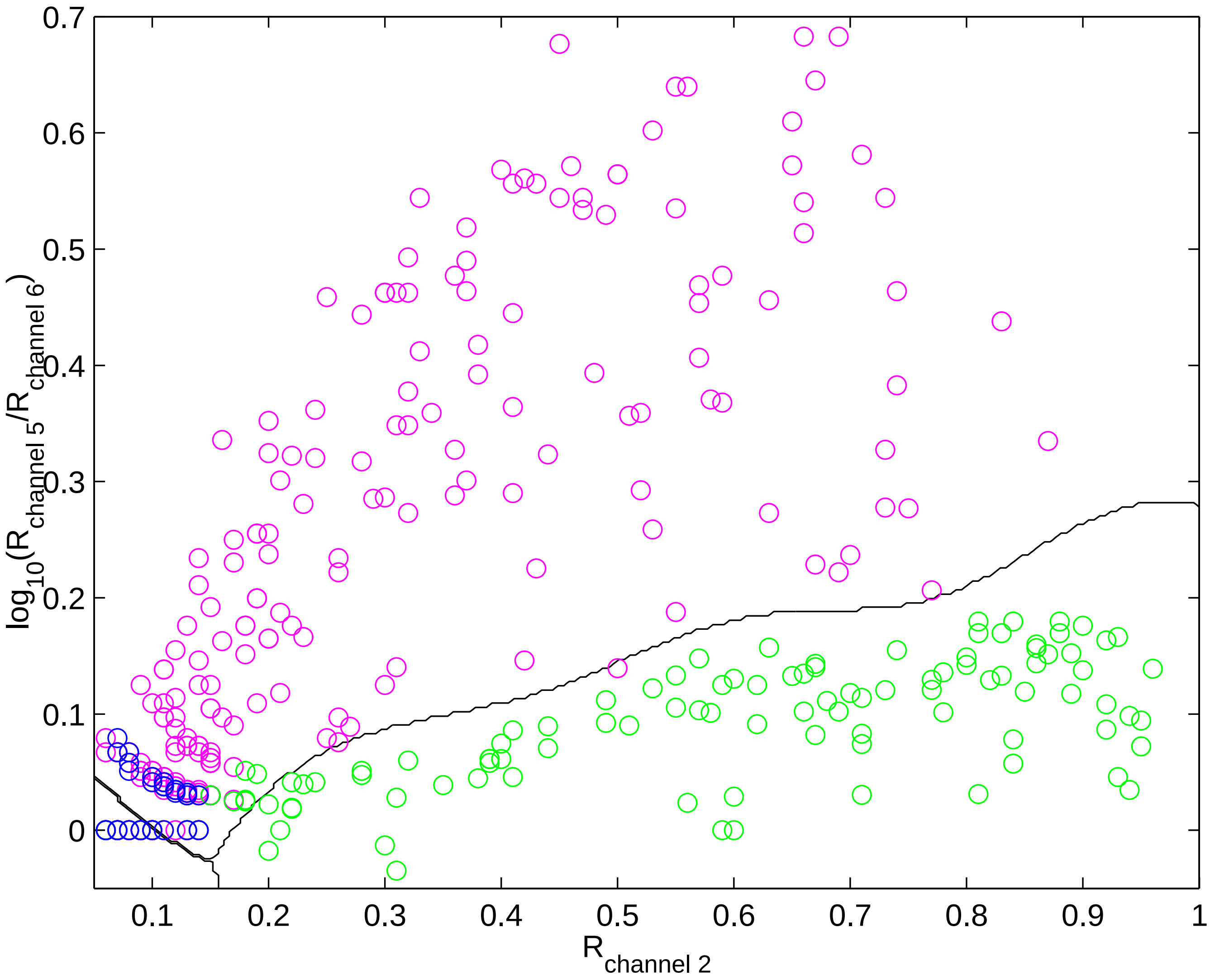
From [LWA04]. Classification of 12 channels of upwelling radiance data from the satellite- borne MODIS instrument. MODIS is a key part of the Earth Observing System (EOS).

Classify each vertical profile as coming from clear sky, water clouds, or ice clouds.

744 simulated radiance profiles (81 clear-blue, 202 water clouds-green, 461 ice clouds-purple).



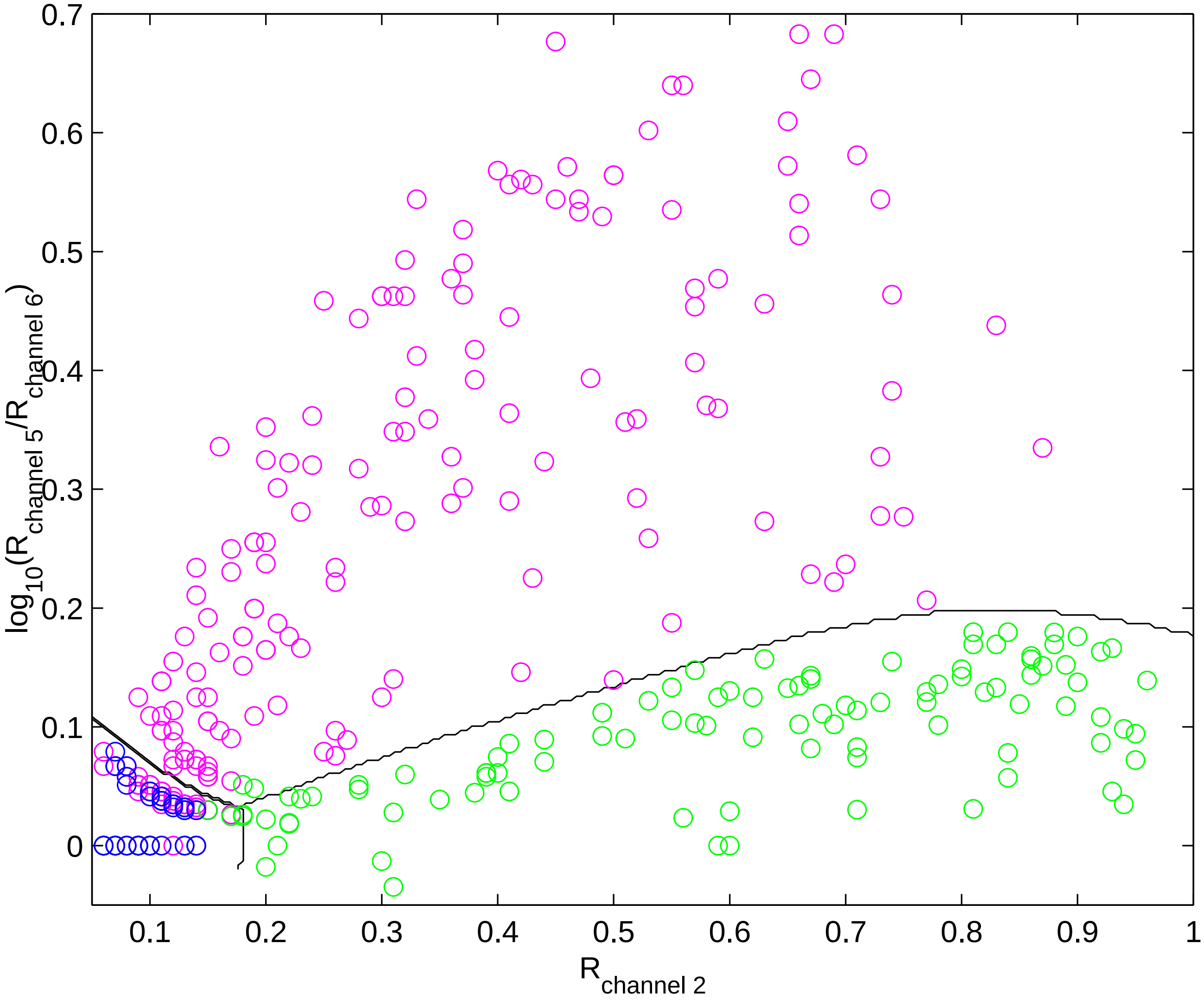
Pairwise plots of three different variables (including composite variables.(purple = ice clouds, green = water clouds, blue = clear)



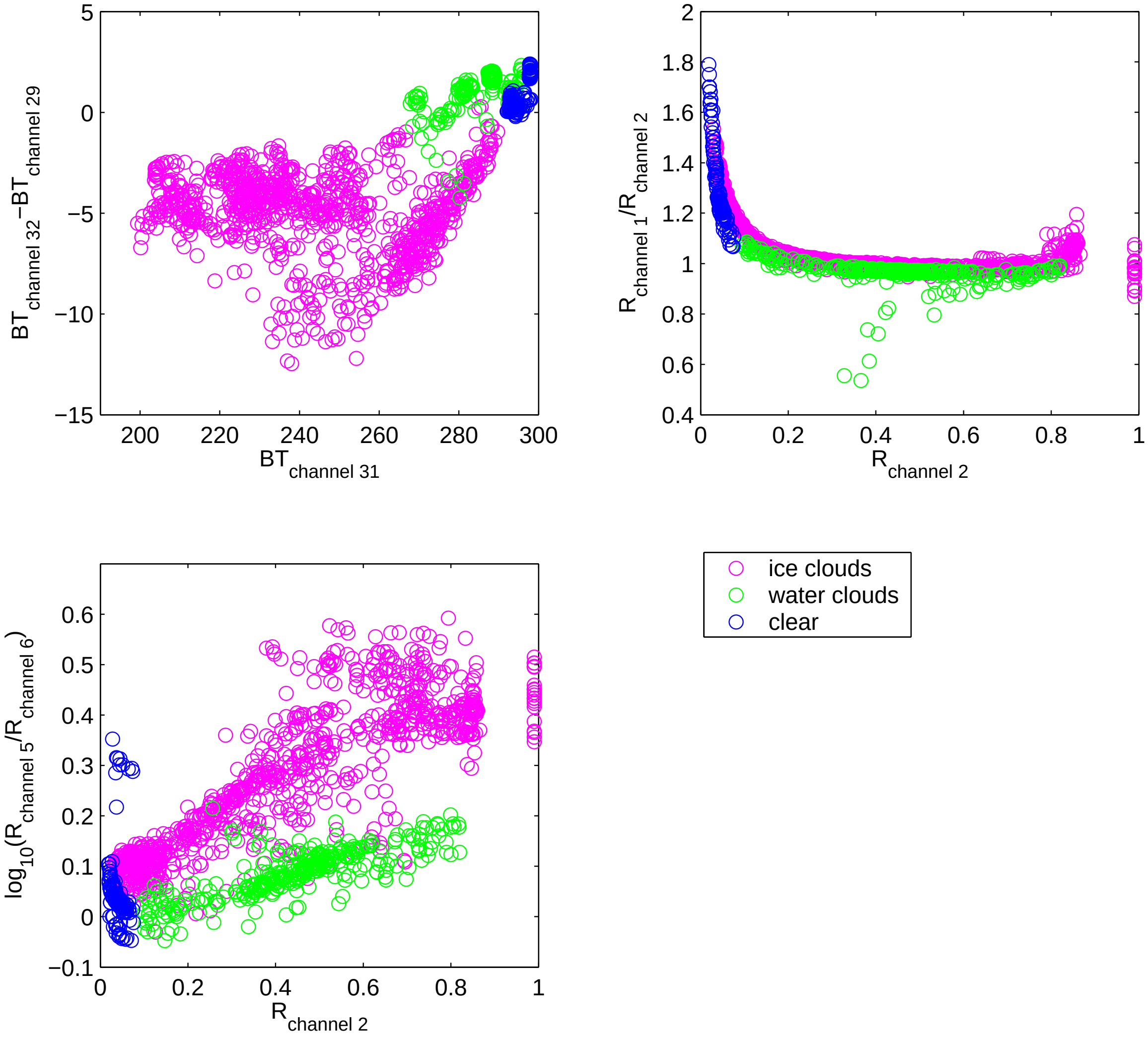
Classification boundaries on the 374 test set determined by the MSVM using 370 training examples, two variables, one is composite.

MSVM test error rates for the combinations of variables and classifiers.

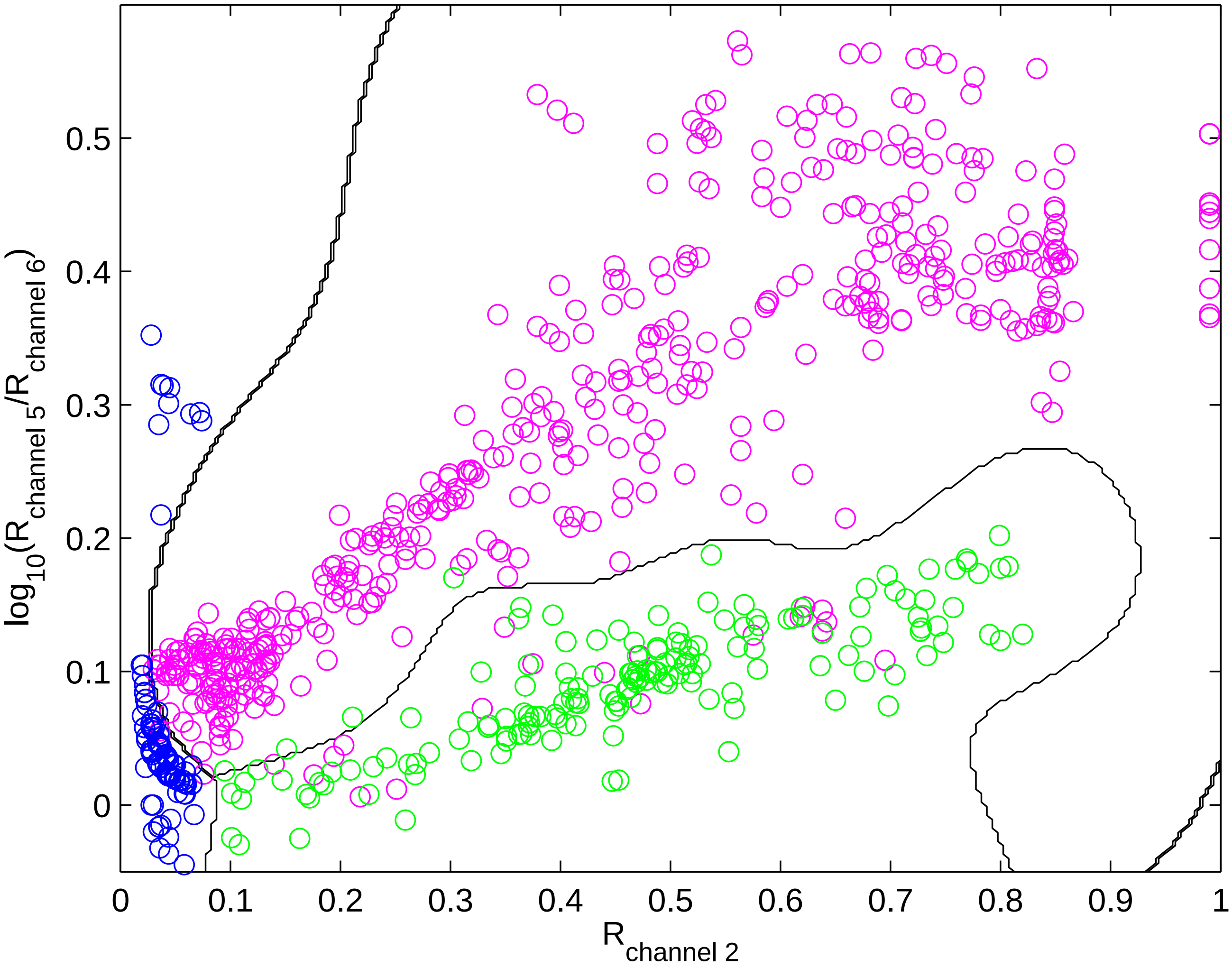
Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	11.50
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	10.16
12	(ii) original 12 variables	12.03
12	log transformed (ii)	9.89



Classification boundaries determined by the nonstandard MSVM when the cost of misclassifying clear clouds is 4 times higher than other types of misclassifications.



Real Data: Pairwise plots of three different variables (including composite variables). (purple = ice clouds, green = water clouds, blue = clear) 1536 profiles ”Labeled by an expert.” Note remarkable similarity to simulated data!



Real Data: Classification boundaries on the test set determined by the MSVM using training examples, two variables, one is composite.

MSVM test error rates for the combinations of variables and classifiers.

Simulated Data:

Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	11.50
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	10.16
12	(ii) original 12 variables	12.03
12	log transformed (ii)	9.89

Real data:

Number of variables	Variable descriptions	Err rates (%)
2	(i) R_2, $\log_{10}(R_5/R_6)$	4.69
5	(i)+R_1/R_2, BT_{31}, $BT_{32} - BT_{29}$	0.26
12	(ii) original 12 variables	0.78
12	log transformed (ii)	0.65

Test error rate of the MODIS cloud masking algorithm on the real data: 18% (!)

♣♣ 7. Multicategory penalized likelihood[XLin98].

$k+1$ categories, $k>1$. Let $p_j(t)$ be the probability that a subject with attribute vector t is in category j, $\sum_{j=0}^{k} p_j(t) = 1$. From [XLin98]: Let

$$f^j(t) = \log p_j(t)/p_0(t), j = 1, \cdots, k.$$

Then:

$$\begin{aligned} p_j(t) &= \frac{e^{f^j(t)}}{1+\sum_{j=1}^{k} e^{f^j(t)}},\ j = 1, \cdots, k \\ p_0(t) &= \frac{1}{1+\sum_{j=1}^{k} e^{f^j(t)}} \end{aligned}$$

Coding:

$$y_i = (y_{i1}, \cdots, y_{ik}),$$

$y_{ij} = 1$ if the ith subject is in category j and 0 otherwise.

♣♣ 8. Remarks

It has been recognized by other authors that when the data is coded as ± 1, that the likelihood function as well as quadratic loss (ridge regression) are large margin classifiers, and have given them new names - e. g. xxx-vector machines. Other large margin classifiers have appeared under various names. In some sense, the hinge function associated with the SVM is the nearest convex upper bound to the misclassification counter.

SVM's are very desirable and popular in higher dimensions, and when the classes are (nearly) separable.

The SVM's tend to be sparse, as many coefficients corresponding to correctly classified data points away from the boundary will be 0.

Penalized likelihood estimates are more appropriate when there is large overlap between the classes and/or you want a probability.

♣♣ 8. More Remarks

Experimental software for the MSVM is available on a limited basis from Yoonkyung Lee yklee@stat.ohio-state.edu. Public code under development.

Simulated MODIS Data for the conditions studied here is reasonably realistic, and may provide a useful rough cut when real labeled training data is not available.

The tuned MSVM is amazingly good at ’learning’ how an expert labels MODIS radiance profiles.

The MSVM may be adjusted to reflect different costs for different kinds of misclassifications.

Interesting questions arise with regard to choosing important variables or combinations of variables.

The MSVM as well as the SVM is highly appropriate for many classification problems.