Statistics 860. Lecture 14, part 2 Tuning the SVM:

Recall that the penalized log likelihood estimate was tuned by a criteria which chose λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} -y_{new\cdot i} f_\lambda(x_i) + \log(1 + e^{f(x_i)}).$$

$R(\lambda)$ is the expected 'distance' or negative log likelihood for a new observation with the same x_i. For the SVM classifier it is possible to follow an alagous route if we have a criteria which chooses λ to minimize a proxy for

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n} (1 - y_{new\cdot i} f_\lambda(x_i))_+.$$

That is, it is choosing λ (and possibly other parameters in K) to minimize a proxy for an upper bound on the misclassification rate- although the real goal is to minimize the misclassification rate.

There is a GACV version for the SVM which tunes to optimize $R(\lambda)$. Details can be found in `tr1016.pdf`, `tr1022.pdf`, `svm.pdf`. Thus, it will be targeted at an upper bound for the misclassification rate.

Yi Lin's Lemma:

The minimizer of $E(1 - y_{new} f(x))_+$ is sign $f(x)$

$= \text{sign}\ (p(x) - \frac{1}{2})$

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

AS A CONSEQUENCE: Find $f_\lambda = d + h$ which minimizes

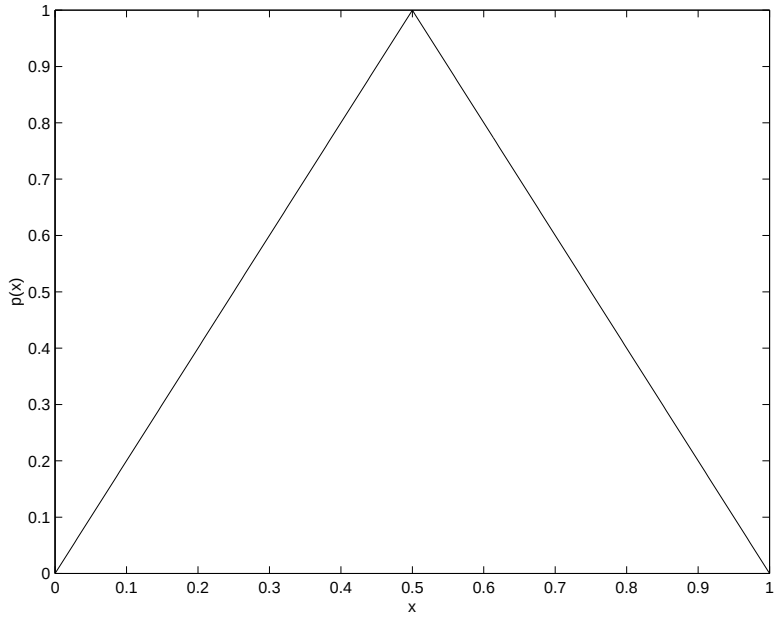
$$\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|h\|^2_{\mathcal{H}_K}$$

where λ is chosen to minimize (a proxy for) $R(\lambda)$,

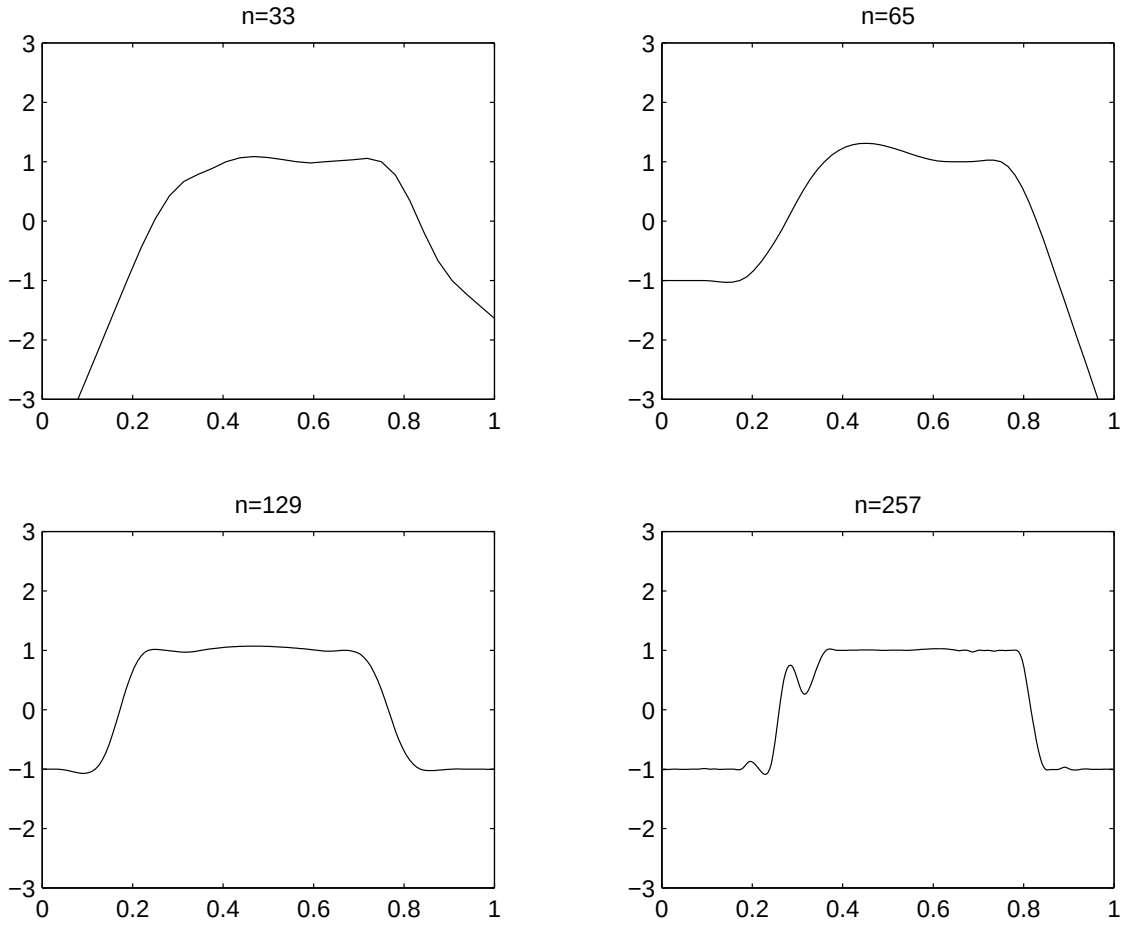
is estimating sign $f(x)$ – EXACTLY WHAT YOU NEED

to minimize the misclassification rate!

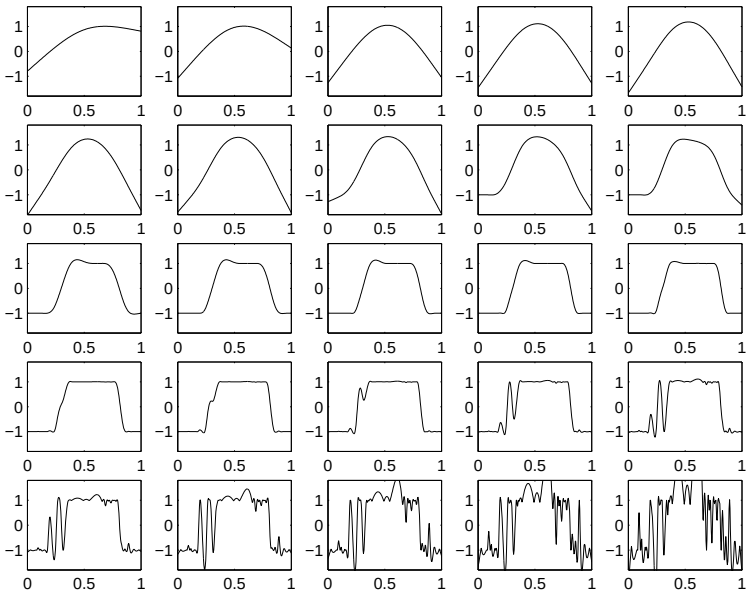
Tuning experiments to follow are courtesy Yi Lin, reprinted from [`tr1014.pdf`].



From Yi Lin. The underlying conditional probability function $p(x) = Prob\{y = 1|x\}$ in our simulation. The function sign $[p(x) - 1/2]$ is 1, for $0.25 < x < 0.75$; -1 otherwise.



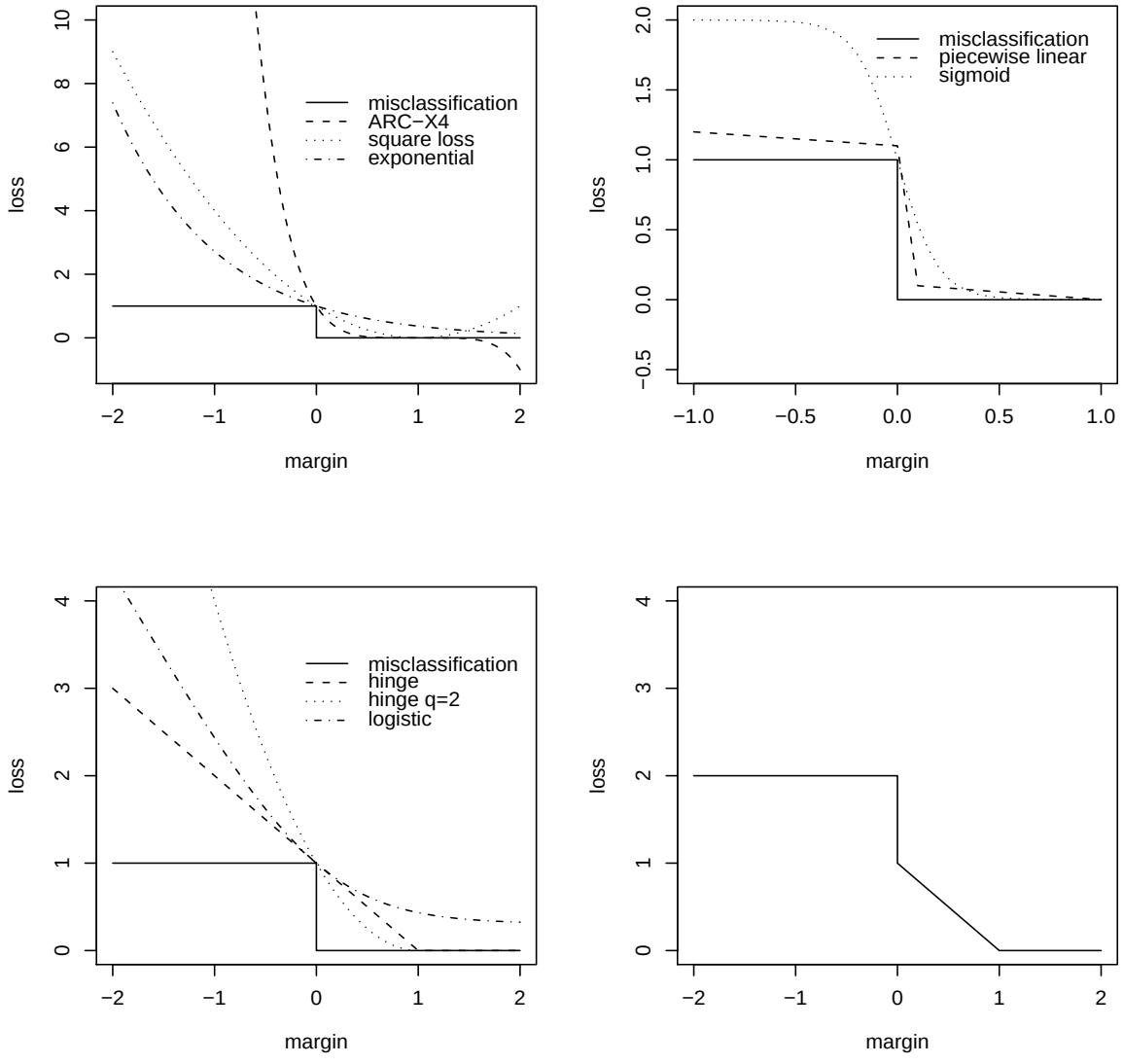
From Yi Lin. SVM estimates, Sobolev Hilbert space kernel (spline kernel), for samples of size 33, 65, 129, 257. The training set is generated using p from the preceeding slide and the x_i equally spaced on $[0, 1]$. The tuning parameter λ is chosen to minimize the $GCKL$ in each case. Note that as the sample size becomes larger, the curve becomes more like the step function sign $(p(x) - 1/2)$.



From Yi Lin. For the same $n = 257$ sample as in the preceeding figure- the solutions to the SVM regularization $n\lambda = 2^{-1}, 2^{-2}, \ldots, 2^{-25}$, left to right starting with the top row. . We see that solution is close to *sign*$[p(x) - 1/2]$ when $n\lambda$ is in the neighborhood of 2^{-18}. 2^{-18} was the minimizer of the $GCKL$, suggesting that it is necessary to tune the SVM to estimate sign $(p - 1/2)$ well.

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.

Next slide is from Yi Lin, A note on margin-based classifiers, `tr1044r.`



Examples of margin-based loss functions.

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.

Thorsten Joachim's $\xi - \alpha$ method for tuning the SVM: Roughly speaking, it replaces $R(\lambda)$ (the hinge function) from p 18 with the misclassification counter. The term which corresponds to the degrees of freedom for signal ($traceA(\lambda)$ in penalized least squares is similar in GACV and the $\xi - \alpha$ method.
See `svmlight.joachims.org`. The $\xi - \alpha$ method is built into the svmlight code for computing support vector machines, and is very popular. GACV behaves similarly to the $\xi - \alpha$ method, see `tr1039.pdf`

When large data sets are available 10-fold cross validation is a popular approach.

♣♣ The Nonstandard Situation

$\pi_{\mathcal{A}} =$ prob. an observation in the population is an $\mathcal{A}$

$\pi_{\mathcal{B}} = 1 - \pi_{\mathcal{A}} =$ prob. an observation in the population is a $\mathcal{B}$ (as before)

$\pi^s_{\mathcal{A}} =$ fraction of training set that are $\mathcal{A}$'s

$\pi^s_{\mathcal{B}} = 1 - \pi^s_{\mathcal{A}} =$ fraction of training set that are $\mathcal{B}$'s

Let

$$p_s(x) = \frac{\pi^s_{\mathcal{A}} h_{\mathcal{A}}(x)}{\pi^s_{\mathcal{A}} h_{\mathcal{A}}(x) + \pi^s_{\mathcal{B}} h_{\mathcal{B}}(x)}$$

$$= \text{Prob.}\{y^s = \mathcal{A} | x\}$$

y^s=element of training set

♣♣ The Nonstandard Situation (continued)

Since p_s is more directly accessible we re-express the Bayes classification rule to minimize the expected cost for a random sample from the population: to get

$$\phi_{\text{OPT}}(x) = \left\{ \begin{array}{ll} \mathcal{A} & \text{if} \frac{p_s(x)}{1-p_s(x)} > \frac{c_{\mathcal{A}}}{c_{\mathcal{B}}} \frac{\pi^s_{\mathcal{A}}}{\pi^s_{\mathcal{B}}} \frac{\pi_{\mathcal{B}}}{\pi_{\mathcal{A}}} \\ \mathcal{B} & \text{otherwise} \end{array} \right\}$$

Letting $\begin{array}{rcl} L(\mathcal{B}) & = & c_{\mathcal{A}} \pi^s_{\mathcal{A}} \pi_{\mathcal{B}} \\ L(\mathcal{A}) & = & c_{\mathcal{B}} \pi^s_{\mathcal{B}} \pi_{\mathcal{A}} \end{array}$

gives

$$\begin{array}{rcl} \phi_{\text{OPT}}(x) & = & \mathcal{A} \quad \text{if } p_s(x) - \frac{L(-1)}{L(-1)+L(1)} > 0 \\ & = & \mathcal{B} \quad \text{if } p_s(x) - \frac{L(-1)}{L(-1)+L(1)} < 0 \end{array}$$

♣♣ The Nonstandard Situation (continued).

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

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

(only the ratio $L(\mathcal{A})/L(\mathcal{B})$ counts if a constant is absorbed in λ).

Lemma [tr1045.pdf]

The minimizer of

$$E\ L(y^s_{new})(1 - y^s_{new} f(x))_+ \text{ is}$$

$$\text{sign}\quad \left(p_s(x) - \frac{L(-1)}{L(-1) + L(1)}\right)$$

"s" – training set

↗ replaces $\frac{1}{2}$