PAC-Bayes & Margins

John Langford, IBM (jcl+@cs.cmu.edu)
John Shawe-Taylor, Royal Halloway (jst@cs.rhul.ac.uk)

Simple Derandomization

First we note a trivial bound arising from a folk theorem and the relationship to our result.

Lemma 0.1 *(Simple Averaging bound) For any stochastic classifier with distribution Q and true error rate e_Q, the averaging classifier,*

$$c_Q(\mathbf{x}) = sign\left(\int_H h(\mathbf{x})dQ(h)\right)$$

has true error rate:

$$e(c_Q) \leq 2e_Q$$

Proof. For every example $(\mathbf{x}, y)$, every time the averaging classifier errs, the probability of the stochastic classifier erring must be at least $1/2$. ■

This result is interesting and of practical use when the empirical error rate of the original averaging classifier is low. Furthermore, we can prove that $c_Q(\mathbf{x})$ *is* the original averaging classifier.

Lemma 0.2 *For $Q = Q(\mathbf{w}, \gamma, \epsilon)$ derived according to the Stochastic Margin Theorem:*

$$c_Q(\mathbf{x}) = sign\,(v_{\mathbf{w}}(\mathbf{x}))$$

Proof. For every $\mathbf{x}$ this equation holds because of two simple facts:

1. For any $\hat{\mathbf{w}}$ that classifies an input $\mathbf{x}$ differently from the averaging classifier, there is a unique equiprobable paired weight vector that agrees with the averaging classifier.
2. If $v_{\mathbf{w}}(\mathbf{x}) \neq 0$, then there exists a nonzero measure of classifier pairs which always agrees with the averaging classifier.

Condition (1) is met by reversing the sign of $\hat{\mathbf{w}}_\top$ and noting that either the original random vector or the reversed random vector must agree with the averaging classifier.

Condition (2) is met by the randomly drawn classifier $\hat{\mathbf{w}} = \lambda\mathbf{w}$ and nearby classifiers for any $\lambda > 0$. Since the example is not on the hyperplane, there exists some small sphere of paired classifiers (in the sense of condition (1)). This sphere has a positive measure. ■

Complex derandomization

The simple averaging bound is elegant, but it breaks down when the empirical error is large because:

$$e(c) \leq 2e_Q = 2(\hat{e}_Q + \Delta_m) \simeq 2\hat{e}_\gamma(c) + 2\Delta_m$$

where $\hat{e}_Q$ is the empirical error rate of a stochastic classifier and Δ_m goes to zero as $m \to \infty$. Next, there exists a bound of the form $e(c_Q) \leq \hat{e}_\gamma(c) + \Delta'_m$ where $\Delta'_m > \Delta_m$ but $\hat{e}_\gamma(c) \leq 2\hat{e}_\gamma(c)$. For details, see the paper.

Bibliography

REFERENCES

P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, 1998.

Thomas Cover and Joy Thomas, "Elements of Information Theory" Wiley, New York 1991.

Ralf Herbrich and Thore Graepel, A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In Advances in Neural Information Processing Systems 13, pages 224-230. 2001.

T. Jaakkola, M. Meila, T. Jebara, "Maximum Entropy D iscrimination\char" NIPS 1999.

John Langford and Rich Caruana, (Not) Bounding the True Error NIPS2001.

John Langford, Matthias Seeger, and Nimrod Megiddo, "An Improved Predictive Accuracy Bound for Averaging Classifiers" ICML2001.

John Langford and Matthias Seeger, "Bounds for Averaging Classifiers." CMU tech report, CMU-CS-01-102, 2001.

David McAllester, "PAC-Bayesian Model Averaging" COLT 1999.

Yoav Freund and Robert E. Schapire, "A Decision Theoretic Generalization of On-line Learning and an Application to Boosting" Eurocolt 1995.

Matthias Seeger, "PAC-Bayesian Generalization Error Bounds for Gaussian Processes", Tech Report, Division of Informatics report EDI-INF-RR-0094. http://www.dai.ed.ac.uk/homes/seeger/papers/gpmcall-tr.ps.gz

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee, "Boosting the Margin: A new explanation for the effectiveness of voting methods" The Annals of Statistics, 26(5):1651-1686, 1998.

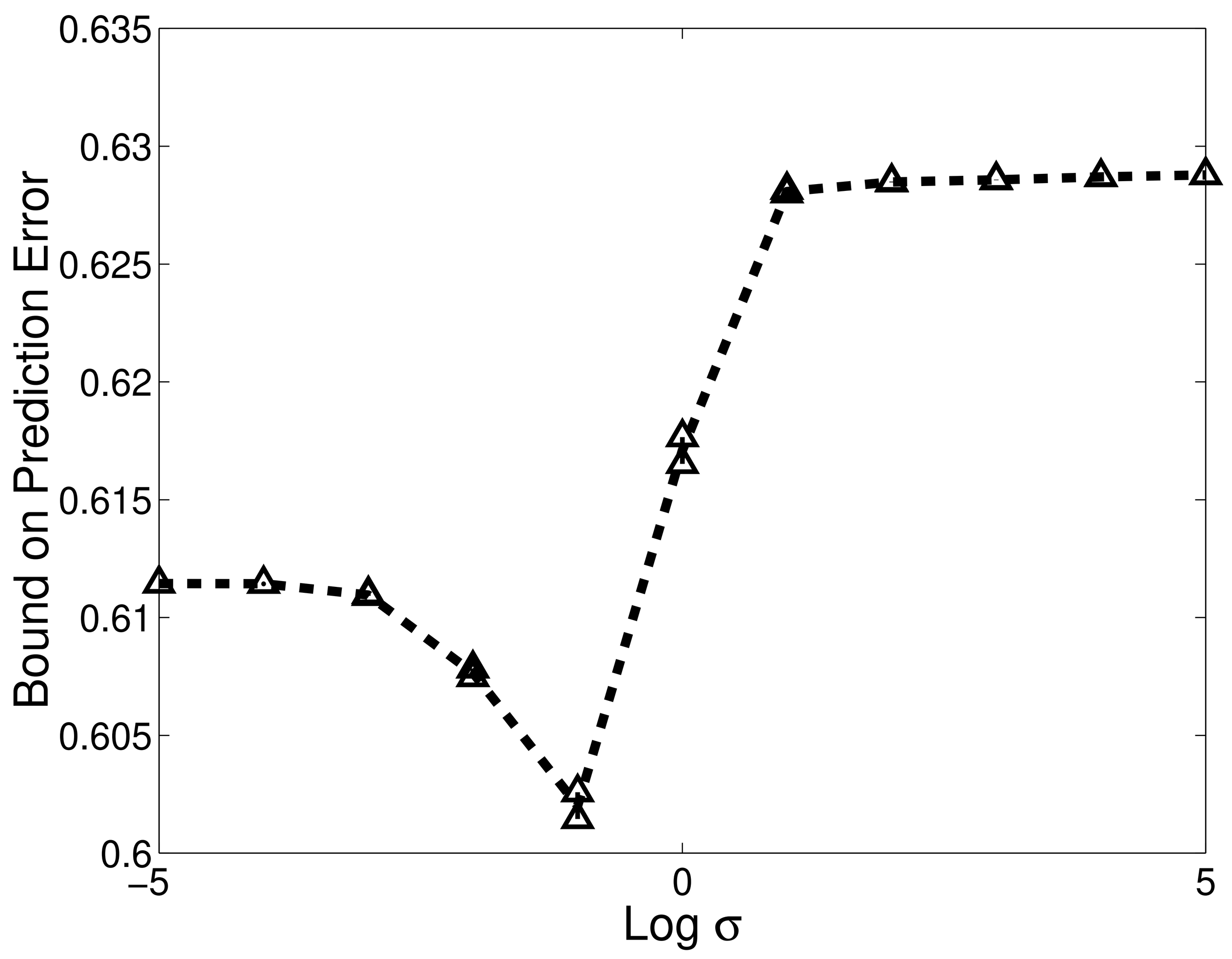
J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

Empirical Results

Thore Graepel has calculated the bound for a Support Vector Machine (with a Radial Basis Function kernel) applied to two data sets (and kernelized the bound). Results are parameterized by σ = the variance in the RBF kernel.

Dataset	Sonar
Train set size	124
Test set size	84
Input dimensions	60

The bound for different kernel variances.



The true error as estimated by a test set and averaging over multiple train/test data splits.