PAC-Bayes & Margins

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

abstract

We show two related things:

(1) Given a classifier which consists of a weighted sum of features with a large margin, we can construct a stochastic classifier with negligibly larger training error rate. The stochastic classifier has a future error rate bound that depends on the margin distribution and is independent of the size of the base hypothesis class.

(2) A new true error bound for classifiers with a margin which is simpler, functionally tighter, and more data-dependent than all previous bounds.

Summary

We are concerned with voting classifiers of the form:

$$\operatorname{sign}(\sum_i x_i w_i)$$

Voting classifiers are quite common. These include "Bayes optimal" classifiers, Maximum Entropy classifiers (4), Support Vector Machines, Adaboost (9), and others which has motivated analysis of true error bounds specialized to voting classifiers.

Our goal is to construct a true error bound on the future error rate of a learned averaging classifier.

In particular, we show that the "build a stochastic classifier" approach (5) to constructing a non-vacuous PAC-Bayes bound is always possible given that some margin γ holds for the training set. This also shows that PAC-Bayes bounds are capable of capturing the low effective "sample complexity" which exists when a large margin classifier is used. This construction can be derandomized to yield a margin bound.

Prior Work

There are two mathematical flavors of margin bound dependent upon the weights w_i of the vote and the features x_i that the vote is taken over.

1. Those ((12), (1)) with a bound on $\sum_i w_i^2$ and $\sum_i x_i^2$ ("l_2/l_2" bounds).
2. Those ((11), (6)) with a bound on $\sum_i w_i$ and $\max_i x_i$ ("l_1/l_∞" bounds).

The results here are of the "l_2/l_2" form. We improve on Shawe-Taylor et al. (12) and Bartlett (1) by a $\log(m)^2$ sample complexity factor and much tighter constants (1000 or unstated versus 9 or 18). In addition, the bound here covers margin errors without weakening the error-free case.

Herbrich and Graepel (3) moved significantly towards the approach adopted here, but the methodology adopted meant that their results do not scale well to high dimensional feature spaces as the bound here (and earlier results) do.

Vocabulary

Vocab	Meaning
$\mathbf{x} = (x_1, ..., x_N)$	The prediction input
$y \in \{-1, 1\}$	The prediction output
$(\mathbf{x}, y)$	A labeled example
D	An unknown distribution over $(\mathbf{x}, y)$
$v_{\mathbf{w}}(\mathbf{x}) = \sum_i \mathbf{w}_i \mathbf{x}_i$	The vote of a voting classifier
$c(\mathbf{x}) = \operatorname{sign}(v_{\mathbf{w}}(\mathbf{x}))$	A voting classifier
S	The set of training examples
$m = \|S\|$	The count
$e_D(c)$	$\Pr_D(c(\mathbf{x}) \neq y)$ The true error rate of c.
$e_{Q(\mathbf{w},\mu,\epsilon)}$	$E_{c\sim Q} e_D(c)$ = the true error rate of a stochastic classifier with distribution Q
$\bar{F}(x)$	$= 1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ = the tail probability of a Gaussian with mean 0 and variance 1

Fundamental Assumption: All labeled examples are drawn independently from D.

The core theorem

Here we state a bound which can take into account the distribution of the training set. This theorem demonstrates the flexibility of the technique since it incorporates significantly more data-dependent information into the bound calculation. When applying the bound one would choose μ to satisfy equality (1).

Given a weight vector $\mathbf{w}$, any particular feature vector $\mathbf{x}$ decomposes into a portion $\mathbf{x}_\|$ which is parallel to $\mathbf{w}$ and a portion $\mathbf{x}_\top$ which is perpendicular to $\mathbf{w}$. Hence, we can write $\mathbf{x} = x_\| \mathbf{e}_\| + x_\top \mathbf{e}_\top$, where $\mathbf{e}_\|$ is a unit vector in the direction of $\mathbf{w}$ and $\mathbf{e}_\top$ is a unit vector in the direction of $\mathbf{x}_\top$. Note that we may have $y x_\| < 0$, if $\mathbf{x}$ is misclassified by $\mathbf{w}$.

Theorem 0.1 *(Stochastic Margin) For all averaging classifiers c with normalized weights $\mathbf{w}$ and for all $\epsilon > 0$ stochastic error rates, If we choose $\mu > 0$ such that*

$$E_{\mathbf{x},y\sim S} \bar{F}\left(\frac{y x_\|}{x_\top}\mu\right) = \epsilon \tag{1}$$

then there exists a posterior distribution $Q(\mathbf{w}, \mu, \epsilon)$ such that

$$\Pr_{S\sim D^m}\left(\forall \epsilon, \mathbf{w}, \mu :\ \mathrm{KL}(\epsilon \| e_{Q(\mathbf{w},\mu,\epsilon)}) \leq \frac{\ln \frac{1}{\bar{F}(\mu)} + \ln \frac{m+1}{\delta}}{m}\right) \geq 1 - \delta$$

where $\mathrm{KL}(q\|p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p}$ = the Kullback-Leibler divergence between two coins of bias $q < p$.

The Proof

The proof uses the PAC-Bayes bound, which states that for all prior distributions P,

$$\Pr_{S\sim D^m}\left(\forall Q :\ \mathrm{KL}(\hat{e}_Q \| e_Q) \leq \frac{\mathrm{KL}(Q\|P) + \ln \frac{m+1}{\delta}}{m}\right) \geq 1 - \delta$$

We choose $P = N(0, I)$, an isotropic Gaussian[a].

A choice of the "posterior" Q completes the proof. The Q we choose depends upon the direction $\mathbf{w}$, the margin γ, and the stochastic error ϵ. In particular, Q equals P in every direction perpendicular to $\mathbf{w}$, and a rectified Gaussian tail in the $\mathbf{w}$ direction[b]. The distribution of a rectified Gaussian tail is given by $R(\mu) = 0$ for $x < \mu$ and $R(\mu) = \frac{1}{\bar{F}(\mu)\sqrt{2\pi}} e^{-x^2/2}$ for $x \geq \mu$.

The chain rule for relative entropy (Theorem 2.5.3 of (2)) and the independence of draws in each dimension implies that:

$$\begin{aligned}
\mathrm{KL}(Q\|P) &= \mathrm{KL}(Q_\| \| P_\|) + \mathrm{KL}(Q_\top \| P_\top) \\
&= \mathrm{KL}(R(\mu) \| N(0,1)) + \mathrm{KL}(P_\top \| P_\top) \\
&= \mathrm{KL}(R(\mu) \| N(0,1)) + 0 \\
&= \int_\mu^\infty \ln \frac{1}{\bar{F}(\mu)} R(x) dx \\
&= \ln \frac{1}{\bar{F}(\mu)}
\end{aligned}$$

Thus, our choice of posterior implies the theorem if the empirical error rate is $\hat{e}_{Q(\mathbf{w},\mathbf{x},\epsilon)} \leq E_{\mathbf{x},y\sim S} \bar{F}\left(\frac{\mathbf{x}_\|}{\mathbf{x}_\top}\mu\right) \leq \epsilon$ which we show next.

Given a point $\mathbf{x}$, our choice of posterior implies that we can decompose the stochastic weight vector, $\hat{\mathbf{w}} = \hat{w}_\| \mathbf{e}_\| + \hat{w}_\top \mathbf{e}_\top + \tilde{\mathbf{w}}$, where $\mathbf{e}_\|$ is parallel to $\mathbf{w}$, $\mathbf{e}_\top$ is parallel to $\mathbf{x}_\top$ and $\tilde{\mathbf{w}}$ is a residual vector perpendicular to both. By our definition of the stochastic generation $\hat{\mathbf{w}}_\| \sim R(\mu)$ and $\hat{\mathbf{w}}_\top \sim N(0,1)$. To avoid an error, we must have:

$$\begin{aligned}
y &= \operatorname{sign}(v_{\hat{\mathbf{w}}}(\mathbf{x})) \\
&= \operatorname{sign}(\hat{w}_\| x_\| + \hat{w}_\top x_\top).
\end{aligned}$$

Then, since $\hat{w}_\| \geq \mu$, no error occurs if:

$$y(\mu x_\| + \hat{w}_\top x_\top) > 0$$

Since $\hat{w}_\top$ is drawn from $N(0,1)$ the probability of this event is:

$$\Pr\left(y(\mu x_\| + \hat{w}_\top x_\top) > 0\right) \geq 1 - \bar{F}\left(\frac{y x_\|}{x_\top}\mu\right)$$

And so, the empirical error rate of the stochastic classifier is bounded by:

$$\hat{e}_Q \leq E_{\mathbf{x},y\sim S} \bar{F}\left(\frac{y x_\|}{x_\top}\mu\right) = \epsilon$$

as required.

[a] Later, the fact that an isotropic Gaussian has the same representation in all rotations of the coordinate sytem will be useful.

[b] Note that we use the invariance under rotation of $N(0, I)$ here to line up one dimension with $\mathbf{w}$.