A PAC-Bayesian Bound for Dropouts

David McAllester
TTI-Chicago

An SVM-Like Generalization Bound

Draw m pairs (x, y) IID from a data distribution with $x \in R^d$, $||x|| = 1$ and $y \in \{-1, 1\}$. We consider $w \in R^d$.

$$L_{01}(w, x, y) = \begin{cases} 0 & \text{if } yw^\intercal x > 0 \\ 1 & \text{otherwise} \end{cases}$$

Let Q_w be a an isotropic Gaussian centered at w.

$$L_{\mathrm{probit}}(w) \;=\; \mathrm{E}_{(x,y)\sim D,\ \epsilon\sim\mathcal{N}(0,I)}\left[L(w+\epsilon, x, y)\right] \;=\; \mathrm{E}_{(x,y)\sim D}\left[L_{\mathrm{probit}}(yw^\intercal x)\right]$$

$$\hat{L}_{\mathrm{probit}}(w) \;=\; \frac{1}{m}\sum_{i=1}^m \mathrm{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[L(w+\epsilon, x, y)\right] \;=\; \frac{1}{m}\sum_{i=1}^m L_{\mathrm{probit}}(yw^\intercal x)$$

$$L_{\mathrm{probit}}(w) \leq \left(\frac{1}{1-\frac{1}{2\lambda}}\right)\left(\hat{L}_{\mathrm{probit}}(w) + \frac{\lambda}{m}\left(\frac{1}{2}||w||^2 + \ln\frac{1}{\delta}\right)\right)$$

McAllester 99, Langford and Shawe-Taylor 02, McAllester 03, Catoni 07.

Awkward Observations

- Finding the exact solution for a fixed sample is not interesting.

- For L_2 reguluraization we have that $||w^*||^2$ grows linearly with m.

- The strong convexity is actually weak — λ/m rather than $\lambda/\sqrt{m}$.

$$A + 2B \;\; \leq \;\; \inf_{\lambda > 1/2} \left(\frac{1}{1 - \frac{1}{2\lambda}}\right)(A + \lambda B) \;\; \leq \;\; A + \sqrt{2AB} + 2B$$

- Robustness trumps convexity.

The PAC-Bayesian Theorem (Catoni's Version)

Let P be a fixed prior distribution or density on models.

Let $L(h, x, y) \in [0, L_{\max}]$ be the loss of model h on training pair (x, y).

Theorem: For $\lambda > 1/2$ selected before seeing the training data we have that with probabiltiy $1 - \delta$ over the draw of the training data the following holds simultanously for all "posterior" distributions Q.

$$L(Q) \leq \left(\frac{1}{1 - \frac{1}{2\lambda}}\right)\left(\hat{L}(Q) + \frac{\lambda L_{\max}}{m}\left(\mathcal{KL}(Q, P) + \ln\frac{1}{\delta}\right)\right)$$

$$\mathcal{KL}(Q_w, P) = \frac{1}{2}||w||^2$$

A Simpler Theorem

- Let $\mathcal{H}$ be a discrete but possibly infinite set of "rules".
- Let $|h|$ be the number of bits it takes to write rule h.
- Let $L(h, x, y) \in [0, L_{\max}]$ be a loss.

Theorem: With probability at least $1 - \delta$ over the draw of the sample we have that the following holds simultaneously for all h.

$$L(h) \leq \inf_{\lambda > \frac{1}{2}} \frac{1}{1 - \frac{1}{2\lambda}} \left(\widehat{L}(h) + \frac{\lambda L_{\max}}{m} \left((\ln 2)|h| + \ln \frac{1}{\delta} \right) \right)$$

Proof

We consider $L_{\max} = 1$. From the Chernoff bound

$$P_{S \sim D^N}\left(\hat{L}(h) \leq L(h) - \epsilon(h)\right) \leq e^{-m\frac{\epsilon(h)^2}{2L(h)}}$$

and a union bound over h we get

$$L(h) \leq \widehat{L}(h) + \sqrt{L(h)\left(\frac{2\left((\ln 2)|h| + \ln\frac{1}{\delta}\right)}{m}\right)}.$$

We then use

$$\sqrt{ab} = \inf_{\lambda > 0} \frac{a}{2\lambda} + \frac{\lambda b}{2}$$

and solve for $L(h)$.

Dropout Training

We assume a labeled training set (x_1, y_1), ..., (x_n, y_n).

We assume some measure of error or loss $L(N_\omega, x_i, y_i) \in [0, L_{\max}]$

We will train the weight vector ω by a form of stochastic gradient descent.

$$\omega_{t+1} = \omega_t - \eta \; \nabla_{\omega_t} \; L(N_{\omega_t \circ s_t}, x_t, y_t)$$

$(x_t, y_t) = (x_i, y_i)$ for some random $i \in \{1, \ldots, n\}$.

$s_t \in \{0, 1\}^d$ is a random mask.

A Message from Hinton et al.

A way to view the dropout procedure is as a very efficient way of performing model averaging. — Hinton, Srivastava, Krizhevsky, Sutskever and Salakhutdinov, arXiv:1207.0580v1

Dropouts with Gaussian Noise

Now let Q_ω be the ensemble $N_{(\omega+\epsilon)\circ s}$ with Gaussian noise ϵ and random dropout mask s under preservation rate α. Let the prior P be Q_0.

$$\begin{aligned}
\mathcal{KL}(Q_\omega,\ P) &= \mathrm{E}_{s,\epsilon}\left[\ln\left(\frac{Q(s)Q(\epsilon \circ s|s)}{P(s)P((w+\epsilon)\circ s|s)}\right)\right] \\
&= \mathrm{E}_s\left[\mathrm{E}_\epsilon\left[\ln\frac{Q(\epsilon \circ s|s)}{P((w+\epsilon)\circ s|s)}\right]\right] \\
&= \mathrm{E}_s\left[KL(Q, P|s)\right] \\
&= \mathrm{E}_s\left[\frac{1}{2}||w \circ s||^2\right] \\
&= \frac{\alpha}{2}||w||^2
\end{aligned}$$

$$L(Q_\omega) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left(\hat{L}(Q_\omega) + \frac{\lambda L_{\max}}{N} \left(\frac{\alpha}{2} ||\omega||^2 + \ln \frac{1}{\delta} \right) \right)$$

The PAC-Baysian Posterior

$$L(Q) \leq \left(\frac{1}{1 - \frac{1}{2\lambda}}\right)\left(\hat{L}(Q) + \frac{\lambda L_{\max}}{N}\left(\mathcal{KL}(Q, P) + \ln\frac{1}{\delta}\right)\right)$$

$$Q^*_\lambda(N) = \frac{1}{Z_\lambda}\, P(N)\, e^{-\frac{N}{\lambda L_{\max}}\hat{L}(N)}$$

The PAC-Bayesian Variance Bound

Fix a learning algorithm $\mathcal{A}$ such that for any sample S we have that $\mathcal{A}(S)$ is a model ensemble.

Using "Langford's Prior" $P = \mathrm{E}_S\left[\mathcal{A}(S)\right]$ we get

$$\mathrm{E}_S\left[L(\mathcal{A}(S))\right] \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\mathrm{E}_S\left[\hat{L}(\mathcal{A}(S))\right] + \frac{\lambda L_{\max}}{N}\,\mathrm{E}_S\left[\mathcal{KL}(\mathcal{A}(S), \mathrm{E}_S\left[\mathcal{A}(S)\right])\right]\right)$$

Summary

- Dropouts optmize the loss of an ensemble of models.

- PAC-Bayesian theory governs the performance of ensembles of models.

- A preservation rate of α reduces the regularization penalty in PAC-Bayesian generalization bounds by a factor of α for a variety of regularizers.

- The optimal PAC-Bayesian posterior has different dropout rates for different units.

- The variance bound appears to be much tighter but is inscrutable.