Generalization Bounds

Here we consider the problem of learning from binary labels. We assume training data $D = \langle x_1, y_1\rangle, \ldots \langle x_N, y_N\rangle$ with y_t being one of the two values 1 or -1. We will assume that these training pairs have been drawn independently from a distribution (or density) ρ. Our objective is to construct a predictor for y given x which will work well for a new pair drawn from ρ.

Here we consider a theoretical approach to understanding learning alorithms based on provable guarantees for generalization performance. The primary objective is to gain a better understanding of choice of loss function in the linear classification learning scheme. Ideally, theoretical analysis could be used to design a loss function leading to better generalization behavior. We will see a theoretical analysis that directly supports probit loss.

1 The Occam Bound

The Occam bound is perhaps the simplest generalization guarantee and is the starting point of our analysis. For this theorem we consider a countable class $\mathcal{H}$ of binary predictors $h : \mathcal{X} \to \{-1, 1\}$. We will call $h \in \mathcal{H}$ a hypothesis. We will assume a fixed language, or code, in which each hypothesis can be named. Let $|h|$ be the number of bits needed to name the hypothesis h. Let $L_{01}(h)$ and $\widehat{L}_{01}(h)$ be defined as follows where $I[\Phi]$ is 1 if Φ is true and 0 otherwise.

$$\begin{aligned} L_{01}(h) &\equiv \mathrm{P}_{\langle x,\, y\rangle \sim \rho}\left[h(x) \neq y\right] \\ \widehat{L}_{01}(h) &\equiv \frac{1}{N}\sum_{t=1}^{N} I\left[h(x_t) \neq y_t\right] \end{aligned}$$

The Occam bound states that for IID draws of N training pairs, and for $\delta > 0$, with probability at least $1 - \delta$ over the draw of the training data D, we have the following.

$$\forall h \in H \;\; L_{01}(h) \leq \widehat{L}_{01}(h) + \sqrt{\frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{2N}} \tag{1}$$

This bound is *uniform* in the sense that, with high probability, the bound holds for all hypotheses simultaneously.

2 Bounds as Algorithms

We can convert any uniform bound on generalization loss to a learning algorithm by selecting the hypothesis minimizing the bound.

$$h^* = \underset{h}{\operatorname{argmin}} \;\; \widehat{L}_{01}(h) + \sqrt{\frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{2N}} \tag{2}$$

Note that, because the generalization is uniform, we get the following with probability at least $1 - \delta$ over the draw of the training data.

$$L_{01}(h^*) \leq \widehat{L}_{01}(h^*) + \sqrt{\frac{(\ln 2)|h^*| + \ln \frac{1}{\delta}}{2N}} \tag{3}$$

If H contains a simple hypothesis that fits the training data well then we are guaranteed to be able to make accurate predications. But of course, no such simple rule may exist and different languages yield different classes of simple rules.

3 Consistency

A learning algorithm will be called consistent if, in the limit of infinite data, we have that $L_{01}(w^*)$ approaches lowest possible loss, i.e., the infimum over $h \in H$ of $L_{01}(h)$. We can show that (2) is consistent by first proving that for any $\delta > 0$ we have the the following with probability at least $1 - \delta$ over the draw of the training sample.

$$\forall h \in H \quad \left|\widehat{L}_{01}(h) - L_{01}(h)\right| \leq \sqrt{\frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{2N}} \tag{4}$$

The proof of this is similar to the proof of (1). When (1) and (4) both hold we get the following.

$$\begin{aligned}
L_{01}(h^*) &\leq \widehat{L}_{01}(h^*) + \sqrt{\frac{(\ln 2)|h^*| + \ln \frac{1}{\delta}}{2N}} \\
&\leq \widehat{L}_{01}(h) + \sqrt{\frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{2N}} \\
&\leq L_{01}(h) + 2\sqrt{\frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{2N}} \\
L_{01}(h^*) &\leq \min_{h \in H} L_{01}(h) + 2\sqrt{\frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{2N}} \qquad (5)
\end{aligned}$$

If we take $\delta = 1/N$ we get that as $N \to \infty$ the expected value of $L_{01}(h^*)$ must approaches the infimum of $L_{01}(h)$ for $h \in H$. So (2) is consistent.

4 Proof of the Occam Bound

The Proof is a simple application of the following three inequalities.

- **Chernoff Bound:** $P(L_{01} > \widehat{L}_{01} + \epsilon) \leq e^{-2N\epsilon^2}$

- **Union Bound:** $P(\exists x \Phi[x]) \leq \Sigma_x \;\; P(\Phi[x])$
- **Kraft Inequality:** $\Sigma_h \;\; 2^{-|h|} \leq 1$

We will not prove the Chernoff bound here but is worth noting that it is very similar to the central limit theorem in one dimension. For a 0-1 (Bernoulli) variable we have $\sigma^2 \leq 1/4$ so $N\epsilon/(2\sigma^2)$ must be at least $2N\epsilon^2$. The union bound is a simple generalization of the observation that $P(\Phi \vee \Psi)$ can be no larger than $P(\Phi) + P(\Psi)$. The Kraft inequality holds for prefix codes — a set of code words where no code word is a proper prefix of any other code word. Null terminated character strings (or byte strings) are prefix codes. To prove the Kraft inequality consider randomly generaing one bit at a time and stoping when you have a code for a rule. Then $2^{-|h|} = P(h)$.

To prove the Occam bound we define a hypothesis h to be "bad" (relative to the training data) if it violates the theorem. More specificaly we have the following.

$$\mathrm{bad}(h) \equiv \left[\;\; L_{01}(h) > \widehat{L}_{01}(h) + \sqrt{\frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{2N}} \;\;\right]$$

$$\begin{aligned} P(\mathrm{bad}(h)) &\leq e^{-2m\epsilon^2} \\ &= \delta 2^{-|h|} \\ P(\exists h \; \mathrm{bad}(h)) &\leq \sum h \delta 2^{-|h|} \\ &= \delta \sum h 2^{-|h|} \leq \delta \end{aligned}$$

5 A Bayesian Interpretation

Let P range over probability distributions on rules. Define $|h|_P$ as follows.

$$|h|_P = \log_2 \frac{1}{P(h)}$$

The proof of the Occam bound can be easily modified to prove the following — we simply replace the Kraft inequality with $\sum_h P(h) = 1$. We get that with probability at least $1 - \delta$ over the draw o the training data we have that the following holds.

$$\forall h \in H \quad L_{01}(h) \leq \widehat{L}_{01}(h) + \sqrt{\frac{(\ln 2)|h|_P + \ln \frac{1}{\delta}}{2N}} \qquad (6)$$

This is now a "Bayesian" theorem in the sense that it is based on an arbitrary "prior". Note that the theorem holds for any prior independent of whether the prior is in any sense "correct".

6 The PAC-Bayesian Theorem

We now consider a bound that can be used for continous (uncountable) hypothesis spaces. Let H be a (possibly continuous) hypothesis space. As in the Occam bound, we fix a "prior" distribution (or density) P on H. We do not require that the prior is "correct". Rather than consider individual hypotheses, we now consider selecting a posterior distribution (or density) Q on H. We define $L_{01}(Q)$ and $\widehat{L}_{01}(Q)$ as follows.

$$\begin{aligned} L(Q) &= \mathrm{E}_{[h\sim Q]}\left[L_{01}(h)\right] \;=\; \mathrm{P}_{h\sim Q,\;\langle x,y\rangle\sim\rho}\left[h(x)\neq y\right] \\ \widehat{L}_{01}(Q) &= \frac{1}{N}\sum_{t=1}^{N}\mathrm{P}_{h\sim Q}\left[h(x_t)\neq y_t\right] \end{aligned}$$

These error rates correspond to a prediction process in which we first select a rule h according to the distribution Q and then use h to make the prediction.

The PAC-Bayesian bound states that with probability at least $1-\delta$ over the draw of the training data we have the following.

$$\forall Q \;\; L_{01}(Q) \leq \widehat{L}_{01}(Q) + \sqrt{\frac{KL(Q||P) + \ln\frac{4N}{\delta}}{2N-1}} \tag{7}$$

The "prior" P expresses the "learning bias". It is analogous to the choice of the coding language defining $|h|$ in earling theorems — recall that any prior P on a discrete hypothesis space defines a coding length $|h|_P$. Here $(\ln 2)|h|_P$ is replaced by $KL(Q||P)$ where the "posterior" Q replaces the rule h.

It is interesting to note that (7) can be viewed as a generalization of (1). Suppose that H is discrete (countable) with $P(h) > 0$ for each $h \in H$. Let Q_h be the posterior that has all mass put on rule h.

$$Q_h(g) = \begin{cases} 1 & \text{if } g = h \\ 0 & \text{otherwise} \end{cases}$$

The posterior Q_h satisfies the following.

$$\begin{aligned} KL(Q_h||P) &= \sum_g Q_h(g)\ln\frac{Q_h(g)}{P(g)} \\ &= Q_h(h)\ln\frac{Q_h(h)}{P(h)} = \ln\frac{1}{P(h)} = (\ln 2)|h|_P \end{aligned}$$

We then have that (7) is a version of (1) where we have traded a little tightness for greater generality.

7 Proof of the PAC-Bayesian Theorem

This section is under construction. It is suggested that you skip to the next section.

To prove the PAC-Bayesian bound we first define $\Delta(h)$ to be the following random variable.

$$\Delta(h) = |L_{01}(h) - \widehat{L}_{01}h|$$

A Chernoff bound:

$$P(\Delta(h) > \epsilon) \leq 2e^{-2m\epsilon^2}$$

Lemma (stated without proof):

$$\mathrm{E}\left[S \sim D^n\right] e^{(2m-1)\Delta(h)^2} \leq 4m$$

Therefore:

$$\mathrm{E}\left[S \sim D^n\right] \mathrm{E}\left[h \sim P\right] e^{(2m-1)\Delta(h)^2} \leq 4m$$

Therefore (the lemma):

$$\forall^\delta S \quad \mathrm{E}\left[h \sim P\right] e^{(2m-1)\Delta(h)^2} \quad \leq \frac{4m}{\delta}$$

$$\mathrm{E}_{[h\sim P]}\left[e^{(2m-1)\Delta(h)^2}\right] \quad \leq \quad \frac{4m}{\delta} \text{ (the lemma)}$$

$$\mathrm{E}_{[h\sim Q]}\left[\frac{P(h)}{Q(h)} e^{(2m-1)\Delta(h)^2}\right] \quad \leq \quad \frac{4m}{\delta}$$

$$\ln \mathrm{E}_{[h\sim Q]}\left[\frac{P(h)}{Q(h)} e^{(2m-1)\Delta(h)^2}\right] \quad \leq \quad \ln \frac{4m}{\delta}$$

$$\mathrm{E}_{[h\sim Q]}\left[(2m-1)\Delta(h)^2 + \ln \frac{P(h)}{Q(h)}\right] \quad \leq \quad \ln \frac{4m}{\delta} \text{ (Jensen's inequality)}$$

$$\mathrm{E}_{[h\sim Q]}\left[\Delta(h)^2\right] \quad \leq \quad \frac{KL(Q||P) + \ln \frac{4m}{\delta}}{2m-1}$$

$$|L_{01}(Q) - \widehat{L}_{01}(Q)| \quad \leq \quad \sqrt{\frac{KL(Q||P) + \ln \frac{4m}{\delta}}{2m-1}}$$

8 A Margin Bound

We can use the PAC-Bayesian theorem to prove a generalization bound for a variant of L_{probit}-L_2 regression, also known as probit regression.

We take the prior to be the multivariant Gaussian $\mathcal{N}(0, I)$ and we consider a family of posteriors Q_w where each posterior is defined by a weight vector w.

$$\begin{aligned} P &= \mathcal{N}(0, I) && (8) \\ Q_w &= \mathcal{N}(w, I) && (9) \end{aligned}$$

The Gaussian prior P correspnds to L_2 regularization. By the PAC-Bayesian theorem we have that with probabilty at least $1-\delta$ over the draw of the training data the following holds simultaneously for all $w \in R^D$.

$$L_{01}(Q_w) \quad \leq \quad \widehat{L}_{01}(Q_w) + \sqrt{\frac{KL(Q_w, P) + \ln \frac{4N}{\delta}}{2N-1}} \tag{10}$$

We first consider $KL(Q_w, p)$.

$$\begin{aligned} KL(Q_w, P) &= KL(\mathcal{N}(w, I), \mathcal{N}(0, I)) \\ &= \mathrm{E}_{[x \sim \mathcal{N}(w,I)]}\left[\ln \frac{\mathcal{N}(w, I)(x)}{\mathcal{N}(0, 1)(x)}\right] \\ &= \mathrm{E}_{[x \sim \mathcal{N}(w,I)]}\left[\frac{1}{2}||x||^2 - \frac{1}{2}||x-w||^2\right] \\ &= \frac{1}{2}\mathrm{E}_{[x \sim \mathcal{N}(w,I)]}\left[x \cdot x - (x-w) \cdot (x-w)\right] \\ &= \frac{1}{2}\mathrm{E}_{[x \sim \mathcal{N}(w,I)]}\left[x \cdot x - (x \cdot x - 2x \cdot w + w \cdot w)\right] \\ &= \frac{1}{2}\mathrm{E}_{[x \sim \mathcal{N}(w,I)]}\left[2x \cdot w - w \cdot w\right] \\ &= \frac{||w||^2}{2} && (11) \end{aligned}$$

We next consider $\widehat{L}_{01}(Q_w)$. In the following we assume that feature vectors have been normalized so that $||\Phi(x)|| = 1$.

$$\begin{aligned}
\widehat{L}_{01}(Q_w) &= \frac{1}{N}\sum_{t=1}^{N} \mathrm{P}_{w' \sim \mathcal{N}(w,I)}\left[y_t(w' \cdot \Phi(x_t)) \leq 0\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} \mathrm{P}_{\epsilon \sim \mathcal{N}(0,I)}\left[y_t(w - \epsilon) \cdot \Phi(x_t) \leq 0\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} \mathrm{P}_{\epsilon \sim \mathcal{N}(0,I)}\left[y_t(\epsilon \cdot \Phi(x_t)) \geq y_t(w \cdot \Phi(x_t))\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} \mathrm{P}_{\epsilon \sim \mathcal{N}(0,1)}\left[y_t\epsilon \geq m_t(w)\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} \mathrm{P}_{\epsilon \sim \mathcal{N}(0,1)}\left[\epsilon \geq m_t(w)\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} L_{probit}(m_t(w)) \qquad (12)
\end{aligned}$$

Putting together (10), (11) and (12) we get that with probability at least $1 - \delta$ over the draw of the training data the following holds simultaneously for all weight vectors w.

$$L_{01}(Q_w) \leq \frac{1}{N}\sum_{t=1}^{N} L_{probit}(m_t(w)) + \sqrt{\frac{\frac{1}{2}||w||^2 + \ln\frac{4N}{\delta}}{2N-1}} \qquad (13)$$

We can interpret (13) as the following learning algorithm.

$$w^* = \operatorname*{argmin}_w \sum_{t=1}^{N} L_{probit}(m_t(w)) + N\sqrt{\frac{\frac{1}{2}||w||^2 + \ln\frac{4N}{\delta}}{2N-1}} \qquad (14)$$

This bound can be used to justify L_{probit}-L_2 regression. In particular, as we show below, there exists a value of λ such that w^* as defined by (14) satisfies the following.

$$w^* = \operatorname*{argmin}_w \sum_{t=1}^{N} L_{probit}(m_t(w)) + \frac{1}{2}\lambda||w||^2 \qquad (15)$$

The fact that there exists a λ such that (14) and (15) agree shows that (15) is a kind of generalization of (14). Since we typically use (15) by setting λ with holdout data, for modestly large sample sizes (15) should perform better than (14). The important point is that both optimizations use L_{probit}.

To see that there exists a λ where (14) and (15) agree we first rewrite (14) as follows.

$$w^* \quad = \quad \underset{w}{\operatorname{argmin}} \sum_{t=1}^{N} L_{\mathrm{probit}}(m_t(w)) + R(||w||^2) \qquad (16)$$

The value w^* in both optimization problems is the vector at which the gradiant of the quantity being minimized is zero. Now consider the gradiant of $R(||w||^2)$ at the optimum w^*.

$$\nabla R(||w||^2)|_{w^*} \quad = \quad 2R'(||w^*||^2)w^* \qquad (17)$$

The gradiant of the regularizer $(1/2)\lambda||w||^2$ at the point w^* is λw^*. If we set λ equal to $2R'(||w^*||^2)$ then we get that (15) has the same solution as (14).

9 Problems

1. The following is the "two sided" form of the Chernoff bound.

$$P(|\widehat{L}_{01}(h) - L_{01}(h)| \geq \epsilon) \leq 2e^{-2m\epsilon^2}$$

Use this inequality (and the Union bound and Kraft inequality) to prove that with probability at least $1-\delta$ over the draw of the training data we have the following.

$$\forall h \;\; |\widehat{L}_{01}(h) - L_{01}(h)| \leq \sqrt{\frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{2m}}$$

2. It is sometimes convenient to use a feature map Φ satisfying the following for $j \geq 1$

$$\Phi_{2j}(x) = -\Phi_{2j-1}(x)$$

The advantage of this feature map is that we can assume without loss of generality that $w_i \geq 0$.

Consider a "prior" $P(w)$ in which each parameter w_i is selected independently according to a prior density which is nonzero only for $w_i \geq 0$ in which case we have $P(w_i) = e^{-w_i}$ (note that this integrates to 1). For a given parameter vector w define Q_w to be a density on w' as follows.

$$Q_w(w') \quad = \prod_{i=1}^{D} \begin{cases} 0 & \text{if } w'_i < w_i \\ e^{-(w'_i - w_i)} & \text{otherwise} \end{cases}$$

a. Compute $KL(Q_w, P)$.

b. Let g(m) be a function such that for any x and $y \in \{-1, 1\}$ we have

$$\mathrm{P}_{w' \sim Q_w}\left[y(w' \cdot \Phi(x)) \leq 0\right] \leq g(y(w \cdot \Phi(x)))$$

Use the PAC-Bayesian theorem to derive an upper bound on $L_{01}(Q_w)$ in terms of the "loss function" $g(m)$ and your answer to part a.

c. Assuming $||\Phi(x)||_1 \leq 1$ for all $x \in \mathcal{X}$, find a function $g(m)$ satisfying the requirement in part b.