Introduction to Statistical Learning Theory

Lecture 11

We have shown the following PAC-Bayes generalization bound:

Theorem 1.1 (Generalization Bound)

Let Q, P be distributions on $\mathcal{H}$ and $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Assume $\ell(h, z) \in [0, 1]$. Let $S \sim \mathcal{D}^m$ be a sample, then with probability greater or equal to $1 - \delta$ over S we have

$$L_{\mathcal{D}}(Q) \leq L_S(Q) + \sqrt{\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{2m}}$$

We will show a few applications.

We will look at a natural posterior - soft-ERM: $Q(h) = \frac{1}{Z_Q} e^{-\beta L_S(h)}$.
Z_Q is the normalization constant (assuming it can be normalized).

For $\beta \to 0$, Q is uniform. For $\beta \to \infty$, Q is concentrated on the ERM.

Its natural counterpart is the prior $P(h) = \frac{1}{Z_P} e^{-\beta L_{\mathcal{D}}(h)}$.

We do not know P, but we only use it for theoretical analysis.

Lemma 1.2

$$KL(Q||P) \leq \beta\left(L_{\mathcal{D}}(Q) - L_S(Q)\right) - \beta\left(L_{\mathcal{D}}(P) - L_S(P)\right)$$

$$KL(Q||P) = \mathbb{E}_Q\left[\ln\left(\frac{Q(h)}{P(h)}\right)\right] = \mathbb{E}_Q\left[\ln\left(\frac{e^{-\beta L_S(h)}}{e^{-\beta L_{\mathcal{D}}(h)}}\right)\right] - \ln\left(\frac{Z_Q}{Z_P}\right)$$
$$= \beta\left(L_{\mathcal{D}}(Q) - L_S(Q)\right) - \ln\left(\frac{Z_Q}{Z_P}\right)$$

We now need to bound $\ln\left(\frac{Z_Q}{Z_P}\right)$:

$$\ln\left(\frac{Z_Q}{Z_P}\right) = \ln\left(\int_{\mathcal{H}} \frac{e^{-\beta L_S(h)}}{Z_P} dh\right) = \ln\left(\int_{\mathcal{H}} p(h) e^{\beta L_{\mathcal{D}}(h)} e^{-\beta L_S(h)} dh\right)$$
$$= \ln\left(\mathbb{E}_P\left[e^{\beta(L_{\mathcal{D}}(h) - L_S(h))}\right]\right) \geq \mathbb{E}_P\left[\beta(L_{\mathcal{D}}(h) - L_S(h))\right]$$

Theorem 1.3 (soft-ERM bound)

Let Q be the soft-ERM posterior, with probability greater of equal to $1-\delta$,

$$KL(L_S(Q)||L_{\mathcal{D}}(Q)) \leq \frac{\sqrt{2}\beta}{m^{3/2}}\sqrt{\ln\left(\frac{2m+2}{\delta}\right)} + \frac{\beta^2}{2m^2} + \frac{\ln\left(\frac{2m+2}{\delta}\right)}{m} \quad (1)$$

It seems like soft-ERM is a universal learner! What doesn't it contradict the fundamental theorem?

We might need β to be large for $L_S(Q)$ to be close to the $L_S(h_{ERM})$.

Proof sketch -

Using the lemma we know that
$KL(Q||P) \leq \beta\left(L_{\mathcal{D}}(Q) - L_S(Q)\right) - \beta\left(L_{\mathcal{D}}(P) - L_S(P)\right)$.

From the PAC-Bayes generalization theorem we have with probability greater or equal to $1 - \delta/2$

$$L_{\mathcal{D}}(Q) - L_S(Q) \leq \sqrt{\frac{KL(Q||P) + \ln\left(\frac{2m+2}{\delta}\right)}{2m}}$$

$$|L_{\mathcal{D}}(P) - L_S(P)| \leq \sqrt{\frac{\ln\left(\frac{2m+2}{\delta}\right)}{2m}}$$

The union bound and some arithmetic finishes the proof. □

We will now show another application - large margin classifiers.

Consider a classifier that returns a real number, whose classification is $sign(h(x))$.

Let $\ell(h(x), y) = \ell^0(h(x), y) = \mathbb{1}\{y \cdot h(x) \leq 0\}$ denote the $0-1$ loss.
Define $\ell^\gamma(h(x), y) = \mathbb{1}\{y \cdot h(x) \leq \gamma\}$ the γ-margin loss.

Theorem 1.4 (linear classifier margin)

Let $\mathcal{X} = [-1, 1]^d$, $\mathcal{H} = \{sign(\langle w, x\rangle) : \ w \in [-1, 1]^d\}$ the hypothesis space of linear classifiers, and let $A : \mathcal{X}^m \to \mathcal{H}$ be any learning algorithm on this space. For any distribution $\mathcal{D}$, and with probability greater of equal to $1 - \delta$ on $S \sim \mathcal{D}^m$

$$L^0_{\mathcal{D}}(A(S)) \leq L^\gamma_S(A(S)) + \sqrt{\frac{d \ln\left(\frac{2d}{\gamma}\right) + \ln\left(\frac{m+1}{\delta}\right)}{2m}}$$

Notice A is a deterministic algorithm, not PAC-Bayesian.

Proof - Define $\bar{w} = A(S)$, $P = U([-1,1]^d)$ and $Q = U\left((\bar{w} + [-\frac{\gamma}{2d}, \frac{\gamma}{2d}]^d]) \cap P\right)$. The following lemma connects A to Q:

Lemma 1.5

$L^0_{\mathcal{D}}(\bar{w}) \leq L^{\frac{\gamma}{2}}_{\mathcal{D}}(Q) \leq L^{\gamma}_{\mathcal{D}}(\bar{w})$ *and* $L^0_S(\bar{w}) \leq L^{\frac{\gamma}{2}}_S(Q) \leq L^{\gamma}_S(\bar{w})$

Proof of lemma: For $w \in support(Q)$ and $x \in \mathcal{X}$ we have

$$|\langle w, x\rangle - \langle \bar{w}, x\rangle| = \left|\sum_{i=1}^{d} x_i(w_i - \bar{w}_i)\right| \leq \sum_{i=1}^{d} |x_i(w_i - \bar{w}_i)| \leq \sum_{i=1}^{d} |(w_i - \bar{w}_i)|$$
$$\leq \sum_{i=1}^{d} \frac{\gamma}{2d} = \frac{\gamma}{2}$$

This proves $L^0_{\mathcal{D}}(\bar{w}) \leq L^{\frac{\gamma}{2}}_{\mathcal{D}}(w) \leq L^{\gamma}_{\mathcal{D}}(\bar{w})$ (same with S) and we finish by taking expectation.

We now need to bound $KL(Q||P)$:

Lemma 1.6

$KL(Q||P) \leq d \ln\left(\frac{2d}{\gamma}\right)$

Proof of lemma:

$$KL(Q||P) = \int_{\mathcal{H}} q(h) \ln\left(\frac{q(h)}{p(h)}\right) dh = \ln\left(\frac{vol(P)}{vol(Q)}\right) \leq \ln\left(\frac{2^d}{(\gamma/d)^d}\right) \quad \square$$

We can now finish the proof of the theorem:

$$L^0_{\mathcal{D}}(\bar{w}) \leq L^{\frac{\gamma}{2}}_{\mathcal{D}}(Q) \leq L^{\frac{\gamma}{2}}_{S}(Q) + \sqrt{\frac{KL(Q||P) + \ln\left(\frac{m+1}{\delta}\right)}{2m}}$$

$$\leq L^{\gamma}_{S}(\bar{w}) + \sqrt{\frac{d \ln\left(\frac{2d}{\gamma}\right) + \ln\left(\frac{m+1}{\delta}\right)}{2m}} \quad \square$$

We will now show a new way to prove generalization - compression bounds.

The idea - If you can define your hypothesis using only a fraction of the data, you will not overfit.

Note - This does not mean the algorithm looks only at a fraction of the data!

Example: Threshold function.

Example: Support vector machines. Only need support vectors to define the classifier.

Definition 2.1 (Compression Scheme)

A size k compression scheme is a pair of two functions:

$$\begin{aligned} Compression\, function: &\quad c:(\mathcal{X}\times\mathcal{Y})^m \to (\mathcal{X}\times\mathcal{Y})^{\leq k} \\ Reconstruction, function: &\quad r:(\mathcal{X}\times\mathcal{Y})^{\leq k} \to \mathcal{H} \end{aligned}$$

Definition 2.2 (Compression algorithm)

A learning algorithm A is a size k compression algorithm if exists a compression scheme c, r such that $A(S) = r(c(S))$.

Notation: The function c picks at most k samples out of S. Denote by I and J the indexes of the chosen samples and its compliment. Denote by S_I and S_J the chosen samples and its compliment.

Theorem 2.3

Let A be a size k compression algorithm with $k < m/2$, and assume that $\ell(h,z) \in [0,L]$. The following holds with probability greater or equal to $1-\delta$:

$$L_{\mathcal{D}}(A(S)) \leq L_{S_J}(A(S)) + L\sqrt{\frac{\ln\left(\frac{1}{\delta}\right) + k\ln\left(\frac{em}{k}\right)}{m}}$$

Proof: For all $I \subset \{1,...,m\}$ denote $h_I = r(S_I)$. As h_I is independent of S_J, (before choosing by c) by Hoeffding

$$L_{\mathcal{D}}(h_I) \leq L_{S_J}(h_I) + L\sqrt{\frac{\ln\left(\frac{1}{\delta'}\right)}{2(m-k)}} \leq L_{S_J}(h_I) + L\sqrt{\frac{\ln\left(\frac{1}{\delta'}\right)}{m}}$$

with probability greater or equal to $1-\delta'$.

The number of candidate index sets I is $\sum_{i=0}^{k} \binom{m}{i} \leq \left(\frac{em}{k}\right)^k$ using the Sauer-Shelah lemma.

If we chose $\delta' = \delta \left(\frac{em}{k}\right)^{-k}$ and use the union bound we get that with probability greater or equal to $1 - \delta' \left(\frac{em}{k}\right)^k = 1 - \delta$ for all possible index set I we have

$$L_{\mathcal{D}}(h_I) \leq L_{S_J}(h_I) + L\sqrt{\frac{\ln\left(\frac{1}{\delta'}\right)}{m}} = L_{S_J}(h_I) + L\sqrt{\frac{\ln\left(\frac{1}{\delta}\right) + k\ln\left(\frac{em}{k}\right)}{m}}$$

This proves the theorem as $A(S)$ is $h_{c(S)}$. $\square$

Note we can replace $L_{S_J}(A(S))$ with $\frac{m}{m-k} L_S(A(S))$.

A note about SVM - The number of support vectors is not known in advance.

We cannot use Theorem 2.3 as is, but it can be fixed using a SRM idea.

For binary classification, does this imply PAC learnability and therefore finite VC dimension?

Almost. We can always vacuously inflate $\mathcal{H}$.

Solution - Assume that for all $h \in \mathcal{H}$ there exists S such that $r(c(S)) = h$. Under this assumption we can conclude $VC(\mathcal{H}) \leq k$.

Open question - If $VC(\mathcal{H}) = d < \infty$, does $\mathcal{H}$ has a compression scheme?