CSE522, Winter 2011, Learning Theory Lecture 12 - 02/10/2011

PAC-Bayes

Lecturer: Ofer Dekel *Scribe: Aniruddh Nath*

1 Setting

Assume **binary classification**: $\mathcal{Y} = \{+1, -1\}$, loss is 0-1.

Algorithm:

1. Define *prior distribution* P on $\mathcal{H}$
2. Get sample $S \sim \mathcal{D}^m$
3. Define *posterior distribution* Q on $\mathcal{H}$

Note that distributions play two different semantic roles:

- $\mathcal{D}$ is a model of the world;
- P, Q express our beliefs about the correct answer.

Definition 1. *The* ***expected risk*** *of Q is:* $\ell(Q; \mathcal{D}) = \mathsf{E}_{h\sim\mathcal{D}}(\ell(h; \mathcal{D}))$

Definition 2. *The* ***Gibbs classifier*** $h_{Gibbs(Q)}(x)\rightarrow y$ *is defined by the following procedure:*

1. Sample $h \sim Q$

2. Get x

3. Output $h(x)$

$\mathsf{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h_{Gibbs(Q)}; (x, y))\right] = \ell(Q; \mathcal{D})$

Example 1

- $\mathcal{H} = \{h_1, \ldots, h_k\}$,
- $P =$ uniform over $\mathcal{H}$,
- $Q = 1$ if $h = h_{ERM}$, and 0 otherwise.

Example 2 Bayesian algorithms:

$$\text{Applying Bayes rule,} \qquad \mathsf{P}(h|S) = \frac{\mathsf{P}(S|h)\mathsf{P}(h)}{\mathsf{P}(S)}$$

where

- $\mathsf{P}(h|S)$ is the posterior,

- $\mathsf{P}(S|h)$ is the *likelihood* of the data (S) given the hypothesis (h),
- $\mathsf{P}(h)$ is the prior probability of hypothesis h, and
- $\mathsf{P}(S)$ can be thought of as a normalization constant, whose purpose is to make the probabilities add up to 1.

Example 3

- $\mathcal{H}$ is the set of linear classifiers in the n-dimensional unit ball,
- P is the uniform distribution over $\mathcal{H}$,
- Q is the uniform distribution on all $w \in \mathcal{H}$ such that $\langle w, \tilde{w} \rangle \geq 0$, where $\tilde{w} \in \mathcal{H}$ is the output of your favorite learning algorithm (e.g. SVM, ERM).

2 Kullback-Leibler divergence

This is our new complexity measure, roughly analogous to Rademacher complexity and VC dimension.

Definition 3. *If Q and P are two distributions on the same space, the* ***Kullback-Leibler divergence*** *between them is:*

$$KL(Q \parallel P) = \mathsf{E}_{z\sim Q} \ln \frac{Q(z)}{P(z)}$$

Origins of KL (Information Theory)

Alice is sending a binary message to Bob over a digital channel. They each have a copy of a *codebook*, which maps symbols in the alphabet $\{a, \ldots, z\}$ to binary strings.

A *variable length prefix-free code* uses shorter strings to encode more frequent letters. Since no string in the code is a prefix of any other string, there is never ambiguity about where one string ends and the next begins. The codebook is chosen to minimize $\mathsf{E}_{x\sim P}$ [#bits for x], where P is a distribution over $\{a, \ldots, z\}$.

Theorem 4. ***Shannon's coding theorem:*** *the best thing to do is to use $\log_2 \frac{1}{P(x)}$ bits to encode x. The expected number of bits per letter is then:*

$$\mathsf{E}_{x\sim P}\left[\#bits\right] = \mathsf{E}_{x\sim P}\left[\log_2 \frac{1}{P(x)}\right] = \sum_{x=a}^{z} P(x) \log_2 \frac{1}{P(x)} \triangleq H(P)$$

where $H(P)$ is the ***entropy*** *of P.*

What happens if the codebook was created assuming that the symbols were distributed according to P, but the real distribution turns out to be Q instead?

$$\mathsf{E}_{x\sim Q}\left[\#\text{bits}\right] = \mathsf{E}_{x\sim Q}\left[\log_2 \frac{1}{P(x)}\right] = \mathsf{E}_{x\sim Q}\left[\log_2 \frac{Q(x)}{P(x)} + \log_2 \frac{1}{Q(x)}\right] = KL(Q \parallel P) + H(Q)$$

$KL(Q \parallel P)$ is the extra number of bits expected per letter from using P instead of Q to create the codebook.

(Note that in information theory, KL divergence is defined in base 2 rather than base e. The units of information in base e are *nats*.)

KL Divergence in PAC-Bayes

(Back to base e.)

Example 4 P is uniform over $\{h_1, \ldots, h_k\}$; $Q_{\tilde{h}}(h) = 1$ if $h = \tilde{h}$, and 0 otherwise.

$$KL(Q \parallel P) = \mathsf{E}_Q \ln \frac{Q(h)}{P(h)} = 1. \ln \frac{1}{1/k} = \ln k$$

Example 5 P is uniform over $\{h_1, \ldots, h_{k-1}\}$; Q is a distribution such that $Q(h_k) > 0$.

$$KL(Q \parallel P) = \mathsf{E}_Q \ln \frac{Q(h)}{P(h)} = \ldots + Q(h_k) \ln \frac{Q(h_k)}{0} = \infty$$

(By convention, $\frac{1}{0} = \infty$ and $0. \ln \frac{0}{0} = 0$.)

Example 6 P is the uniform distribution over linear classifiers in the n-dimensional unit ball, and Q is the uniform distribution over linear classifiers in some n-dimensional unit hemisphere.

$$KL(Q \parallel P) = \mathsf{E}_Q \ln \frac{Q(h)}{P(h)} = \mathsf{E}_Q \ln \frac{2P(h)}{P(h)} = \ln 2$$

Special case: $\alpha, \beta \in [0, 1]$

$$KL(\alpha \parallel \beta) \leftrightarrow KL(Bernoulli(\alpha) \parallel Bernoulli(\beta)) = \alpha \ln \frac{\alpha}{\beta} + (1-\alpha) \ln \frac{1-\alpha}{1-\beta} = 0$$

Theorem 5. $KL(Q||P) \geq 0$

$$\begin{aligned} KL(1 \parallel 0) &= 1 \ln \frac{1}{0} + 0 \ln \frac{0}{1} = \infty \\ KL(0 \parallel 1) &= 0 \ln \frac{0}{1} + 1 \ln \frac{1}{0} = \infty \\ KL(1 \parallel \frac{1}{2}) &= 1 \ln \frac{1}{\frac{1}{2}} + 0 \ln \frac{0}{\frac{1}{2}} = \ln 2 \end{aligned}$$

Theorem 6. *(McAllester 2003/1999)*
$\forall \mathcal{D}, \forall \mathcal{H}, \forall P$ (prior on $\mathcal{H}$), with probability $\geq 1 - \delta$ over the sampling of $S \sim \mathcal{D}^m$,

$$\forall Q \text{(distribution on } \mathcal{H}) \quad KL(\ell(Q; S) \parallel \ell(Q; \mathcal{D})) \leq \frac{KL(Q \parallel P) + \ln \frac{m+1}{\delta}}{m}$$

Corollary 7.

$$\ell(Q; \mathcal{D}) \leq \ell(Q; S) + \sqrt{\frac{2\ell(Q; S). \left(KL(Q \parallel P) + \ln \frac{m+1}{\delta}\right)}{m}} + \frac{2\left(KL(Q \parallel P) + \ln \frac{m+1}{\delta}\right)}{m}$$

Proof. (of theorem 6)

$$\textbf{Define: } Z = \mathsf{E}_{h\sim P} e^{m.KL(\ell(h;S)\|\ell(h;\mathcal{D}))}$$

$$\text{Part I)} \qquad KL(\ell(Q;S) \parallel \ell(Q;\mathcal{D})) \leq \frac{KL(Q \parallel P) + \ln\left(\frac{1}{\delta}\mathsf{E}_S[Z]\right)}{m}$$

$$\text{Part II)} \qquad \mathsf{E}_S[Z] \leq m+1$$

I)

$$\begin{aligned}
\text{By Markov's inequality, } \forall a, \quad & \mathsf{P}[Z > a] \leq \frac{\mathsf{E}_S[Z]}{a} \\
\text{Plugging in } a = \frac{\mathsf{E}_S[Z]}{\delta}, \quad & \mathsf{P}[Z > a] \leq \delta \\
\text{With probability } \geq 1-\delta, \quad & Z \leq a = \frac{\mathsf{E}_S[Z]}{\delta} \\
\Rightarrow \quad & \ln(Z) \leq \ln\left(\frac{\mathsf{E}_S[Z]}{\delta}\right)
\end{aligned}$$

$$\begin{aligned}
\ln(Z) &= \ln\left(\mathsf{E}_{h\sim P} e^{m.KL(\ell(h;S)\|\ell(h;\mathcal{D}))}\right) \\
&= \ln\left(\mathsf{E}_{h\sim Q} \frac{P(h)}{Q(h)} e^{m.KL(\ell(h;S)\|\ell(h;\mathcal{D}))}\right)
\end{aligned}$$

Upper bound using Jensen's inequality + concavity of ln:

$$\begin{aligned}
\ln(Z) &\geq \mathsf{E}_{h\sim Q}\left(\ln\frac{P(h)}{Q(h)} + \ln e^{m.KL(\ell(h;S)\|\ell(h;\mathcal{D}))}\right) \\
&= -KL(Q \parallel P) + \mathsf{E}_{h\sim Q}\left[m.KL(\ell(h;S) \parallel \ell(h;\mathcal{D}))\right] \\
&\geq = -KL(Q \parallel P) + m.KL(\ell(Q;S) \parallel \ell(Q;\mathcal{D}))
\end{aligned}$$

II) *(See lecture 13.)*

□