

ECE Statistical Learning Theory - Lecture 3

Prof. Shai Ben-David
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853
shai@ece.cornell.edu

October 7, 2003

Let us recall the basic generalization bounds that we have seen so far.

- If we fix any h before we see the sample, then

$$E^P(h) \leq E^S(h) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

where

$$E^P(h) = P[(x, y) : h(x) \neq y]$$
$$E^S(h) = \frac{|\{(x, y) \in S : h(x) \neq y\}|}{|S|}.$$

This bound assumes that the choice of h was done independently of the sample.

- One can apply this bound to select between several candidate h 's by preassigning a subset of the training examples to serve as a 'test set', and then use this subsample to evaluate the error of each of these hypotheses. However, as more hypotheses are being evaluated over the same test set, the probability that one of them will incur a large gap between its evaluated error and true error increases. To overcome this, the test set should be increased with the number of hypotheses h that are going to be evaluated over it. This is captured by our second basic generalization bound. Namely, given a finite set H of hypotheses, For any distribution P over $X \times \{0, 1\}$, we have :
 $\forall \delta > 0$ and $\forall m$, the following relation holds with probability $1 - \delta$ over samples of size m drawn i.i.d. by P :

$$\forall h \in H \quad E^P(h) \leq E^S(h) + \sqrt{\frac{\ln 2|L(h)| + \ln \frac{1}{\delta}}{2m}},$$

- the next bound we developed applies to a wider setting, allowing for classes of hypotheses H that may be infinite. It involves the appealing idea of Occam's Razor -

judging the quality of hypothesis by their description size - preferring those with concise descriptions.

Fix a description language $L : H \rightarrow \{0, 1\}^*$ (that is $\forall h : X \rightarrow \{0, 1\}$, $h \in H$, $L(h)$ is a binary string). Assume that L is prefix-free and let $|L(h)|$ denote the length of the string $L(h)$.

For any distribution P over $X \times \{0, 1\}$, we have :

$\forall \delta > 0$ and $\forall m$, the following relation holds with probability $1 - \delta$ over samples of size m drawn i.i.d. by P :

$$\forall h \in H \quad E^P(h) \leq E^S(h) + \sqrt{\frac{\ln 2 |L(h)| + \ln \frac{1}{\delta}}{2m}},$$

In this lecture we shall develop yet another bound that, assuming the existence of a special type of descriptions, results in stronger compression guarantees.

Compression Schemes

A compression scheme is a function from finite subsets of X to 2^X . Let us fix a parameter d and consider only subsets of size d .

$$T : \{A \subseteq X : |A| = d\} \rightarrow \{f : X \rightarrow \{0, 1\}\}$$

Furthermore, we require that the hypotheses that can be generated on the basis of a sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ are all of the form $T(A)$ for some $A \subseteq \{x_1, \dots, x_m\}$. That is, any candidate hypothesis arising from an input sample has a short description using examples from this sample.

Examples

- (1) $X = \mathbb{R}$; $T_{\text{intervals}}(\{a, b\}) = \mathbf{1}_{[a, b]}$ (here, $d = 2$). So $T_{\text{intervals}}$ maps pairs of real numbers to the intervals they enclose.
- (2) $X = \mathbb{R}^n$; $d = n$;

$$T_{\text{halfspaces}}(\{a_1, \dots, a_n, a_{n+1}\})(x) = \begin{cases} 1 & \text{if } \langle \mathbf{a}, \mathbf{x} \rangle \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

This compression scheme maps a set of points to the halfspace they naturally define.

- (3) $X = \mathbb{R}^n$; $d = 2n$

$$T_{\text{rectangles}}(\{a_1, \dots, a_k\})(x) = \mathbf{1}_{R(a_1, \dots, a_k)},$$

where $R(a_1, \dots, a_k) =$ minimal axis aligned rectangle containing a_1, \dots, a_k . Note, that for any set of points $B \subset \mathbb{R}^n$ there exists a subset of size at most $k \leq 2n$ such that the minimal axis aligned rectangle defined by this subset equals the one defined by the original set B .

Theorem 1 *Let T be a d -size compression scheme over some domain X . Given a sample*

$$S = \{(x_1, y_1), \dots, (x_m, y_m)\}$$

we consider all h 's of the type

$$h = T(\{x_{i_1}, \dots, x_{i_d}\}).$$

For every distribution P over $X \times \{0, 1\}$ and $\forall \delta > 0$, $\forall m$, and $\forall h$ as above, we have

$$\left| E^P(h) - \hat{E}^S(h) \right| \leq \sqrt{\frac{\ln \binom{m}{d} + \ln \frac{1}{\delta}}{2(m-d)}}.$$

Proof:

$\forall i_1, \dots, i_d$ (fixed before we see the sample),

$T(x_{i_1}, \dots, x_{i_d})$ is independent from $S \setminus \{x_{i_1}, \dots, x_{i_d}\}$ (since S is assumed to be sampled i.i.d.).

So the test set bound applies

$$\left| E^P(T(x_{i_1}, \dots, x_{i_d})) - E^{S \setminus \{x_{i_1}, \dots, x_{i_d}\}} \right| < 2 \sqrt{\frac{\ln \frac{1}{\delta}}{2(m-d)}}.$$

We need a uniform bound over $\binom{m}{d}$ such hypotheses. So we replace the δ in the bound by $\frac{\delta}{\binom{m}{d}}$. This allows the description language to depend on the sample.

This is the bound behind the Support Vector Machines (SVM). □

Example: CDF estimation using unlabeled data

Given some $h : X \rightarrow \{0, 1\}$, define

$$X_i = \begin{cases} 1 & h(x) \neq y \\ 0 & \text{otherwise} \end{cases}.$$

Instead we could consider r.v.'s over X , *i.e.*,

$$X_i = \begin{cases} 1 & h(x) = 1 \\ 0 & \text{otherwise} \end{cases}.$$

We have

$$\begin{aligned} E^S(h) &= \frac{|S \cap h|}{|S|} \\ E^P(h) &= P(x \in \{x : h(x) = 1\}). \end{aligned}$$

Corollary 1 Glivenko-Cantelli Theorem

For any distribution P over $(\mathbb{R}, \mathcal{B})$,

$\forall \delta > 0, \forall m$ sample size, and $\forall I$ interval we have

$$\left| \frac{|S \cap I|}{|S|} - P(I) \right| \leq 2 \sqrt{\frac{2 \ln m + \ln \frac{1}{\delta}}{2(m-2)}}$$

□

Corollary 2 Consider a sample X_1, \dots, X_m , *i.i.d.* from any unknown P over \mathbb{R} .

What is the probability that if one more sample r is obtained, it is greater than all the previous points:

$$r > \max \{X_1, \dots, X_m\}?$$

This probability is at most

$$\sqrt{\frac{2 \ln m + \ln \frac{1}{\delta}}{2m}}.$$