IST Austria: Statistical Machine Learning 2015/16

Christoph Lampert <chl@ist.ac.at>

Lecture 7 – Notes

Coordinate classifiers

- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$, $\ell(y, y') = [\![y, y']\!]$, $\quad \mathcal{H} = \{h_1, \dots, h_d\}$ with $h_i(x) = \operatorname{sign} x[i]$

Lemma 1. *If p is uniform in $[-1, 1]^d$, ERM works for $m_0(\epsilon, \delta) = \lceil \log_2 \frac{d-1}{\delta} \rceil$*

Proof:

1. let true labeling function be h_j, it has $\mathcal{R}(h_j) = 0$
2. all other labeling function have $\mathcal{R}(h_k) = \frac{1}{2}$
3. what's the probability that ERM returns a hypotheses h_k with $k \neq j$? Since there exists a hypothesis with 0 error on every training set, any hypothesis that ERM returns will have 0 training error.
4. what's the probability that at least one of the hypotheses h_k with $k \neq j$ have 0 training error?
5. Fix h_k with $k \neq j$. Training examples are i.i.d. evaluations:
$$\Pr_{(x_i, y_i)}(\ y_i = \operatorname{sign} x_i[k]\ ) = \frac{1}{2} \qquad \rightarrow \qquad \Pr_{\mathcal{D}_m}(\ \hat{\mathcal{R}}(h_k) = 0\ ) = \frac{1}{2^m}$$
6. Union bound: $\Pr(A_1 \vee A_2 \vee \dots \vee A_d) \leq \sum_k \Pr(A_k)$
$$\Pr_{\mathcal{D}_m}(\ \exists k \neq j : \hat{\mathcal{R}}(h_k) = 0\ ) \leq \sum_{k \neq j} \frac{1}{2^m} = \frac{d-1}{2^m}$$
7. We want r.h.s. to be no bigger than δ. Solve for m: $m \geq \log_2 \frac{d-1}{\delta}$. Next biggest integer: $m_0 = \lceil \log_2 \frac{d-1}{\delta} \rceil$.

Finite hypothesis classes are PAC learnable

Theorem 2. *Let $\mathcal{H} = \{h_1, \dots, h_K\}$ be a finite hypothesis class and $f \in \mathcal{H}$ (i.e. the true labeling function is one of the hypotheses). Then $\mathcal{H}$ is PAC-learnable by the ERM algorithm with $m_0(\epsilon, \delta) = \lceil \frac{1}{\epsilon}(\ \log(|\mathcal{H}| + \log(1/\delta)\ ) \rceil$*

Proof:

We have to show: the probability that ERM on $m \geq m_0$ samples returns a hypothesis with generalization error bigger than ϵ is not bigger than δ.

1. denote by $e_1, \dots, e_K$ the generalization errors of $h_1, \dots, h_K$.
2. denote by $\mathcal{H}_\epsilon = \{h_i : e_i > \epsilon\} \subset \mathcal{H}$ be the subset of hypotheses with error bigger than ϵ (the ones we don't want).
3. what's the probability that ERM returns a hypotheses $h_j \in \mathcal{H}_\epsilon$? Since there exists a hypothesis with 0 error on every training set, any hypothesis that ERM returns will have 0 training error.
4. what's the probability that at least one of the hypotheses in $\mathcal{H}_\epsilon$ have 0 training error?

5. First, for any fixed $h_j \in \mathcal{H}_\epsilon$, training examples are i.i.d. evaluations:

$$\Pr(\ \hat{\mathbb{R}}_m(h_j) = 0\ ) = (1 - e_j)^m \leq (1 - \epsilon)^m$$

6. Apply a union bound

$$\Pr(\ \exists h_j \in \mathcal{H}_\epsilon\ :\ \hat{\mathbb{R}}_m(h_j) = 0\ ) \leq \sum_{h_j \in \mathcal{H}_\epsilon} \Pr(\ \hat{\mathbb{R}}_m(h_j) = 0\ ) \leq (K-1)(1-\epsilon)^m$$

7. how large is the r.h.s. for $m \geq m_0 = \lceil \frac{1}{\epsilon}(\ \log(|\mathcal{H}| + \log(1/\delta)\ ) \rceil$?

$$\begin{aligned}
(K-1)(1-\epsilon)^m &\leq (K-1)(1-\epsilon)^{m_0} \\
&\leq (K-1)(1-\epsilon)^{\frac{1}{\epsilon}(\log(K+\log(1/\delta))} \\
&= (K-1)e^{\frac{\log(1-\epsilon)}{\epsilon}(\log(K+\log(1/\delta)))} \\
&\leq (K-1)e^{-(\log(K+\log(1/\delta))} \qquad \text{because } \log(1-t) \leq -t \text{, so } \frac{\log(1-\epsilon)}{\epsilon} \leq \frac{-\epsilon}{\epsilon} = -1 \\
&= (K-1)e^{-\log K}\ e^{-\log(1/\delta)} \\
&= \frac{K-1}{K}\frac{1}{1/\delta} \\
&< \delta
\end{aligned}$$

□

Finite hypothesis classes are agnostic PAC learnable

Theorem 3. *Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite hypothesis class.*
Then $\mathcal{H}$ is agnostic PAC-learnable by ERM with $m_0(\epsilon, \delta) = \lceil \frac{2}{\epsilon^2}(\ \log(|\mathcal{H}| + \log(2/\delta)\ ) \rceil$

Proof. Let

- $h_{\text{ERM}} \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \hat{\mathcal{R}}_m(\bar{h})$ (result of ERM)
- $h^* \in \mathbf{argmin}_{\bar{h} \in \mathcal{H}} \mathcal{R}_p(\bar{h})$ (if exists, otherwise use argument of arbitrarily close approximation)

From the following lemma (proved later):

Lemma 4. *For any $\epsilon > 0$, $\delta > 0$, the following inequality hold uniformly in $h \in \mathcal{H}$ with probability at least $1 - \delta$ w.r.t. $\mathcal{D}_m$:*

$$|\mathcal{R}_p(h) - \hat{\mathcal{R}}_m(h)| \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

it follows that with prob. at least $1 - \delta$, it holds at the same time:

$$\mathcal{R}_p(h_{\text{ERM}}) - \hat{\mathcal{R}}_m(h_{\text{ERM}}) \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} \qquad \text{and} \qquad \hat{\mathcal{R}}_m(h^*) - \mathcal{R}_p(h^*) \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}$$

Adding the two inequalities we obtain

$$\begin{aligned}
\mathcal{R}_p(h_{\text{ERM}}) - \mathcal{R}_p(h^*) &\leq \overbrace{\hat{\mathcal{R}}_m(h_{\text{ERM}}) - \hat{\mathcal{R}}_m(h^*)}^{\leq 0} + 2\sqrt{\frac{\log|\mathcal{H}|\ +\ \log\frac{2}{\delta}}{2m}} \\
&\leq 2\sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}} \quad \overset{m \geq m_0}{\leq} \quad \epsilon
\end{aligned}$$

Proof of the lemma

Lemma 5 (Hoeffding's Inequality)**.** *Let $Z_1, \dots, Z_m$ be i.i.d. random variables that take values in the interval $[a, b]$. Let $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and denote $\mathbb{E}[\bar{Z}] = \mu$. Then, for any $\epsilon > 0$,*

$$\Pr[\left|\bar{Z} - \mu\right| > \epsilon] \leq 2e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Proof of uniform bound Lemma:

1. for any fix $h \in \mathcal{H}$, let $Z_i := \ell(y_i, h(x_i))$. These are i.i.d. random variables in the interval $[0, 1]$.
2. then $\bar{Z} = \frac{1}{m}\sum_i Z_i = \hat{\mathcal{R}}_m(h)$ and $\mathbb{E}[\bar{Z}] = \mathcal{R}(h)$, such that

$$\Pr[\left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| > \epsilon] \leq 2e^{-2m\epsilon^2}.$$

3. by a union bound, we obtain

$$\Pr[\exists h \in \mathcal{H} : \left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| > \epsilon] \leq 2|\mathcal{H}|e^{-2m\epsilon^2}.$$

4. calling the right hand side δ, we obtain

$$\Pr\left[\exists h \in \mathcal{H} : \left|\hat{\mathcal{R}}_m(h) - \mathcal{R}(h)\right| > \sqrt{\frac{\log(\frac{2|\mathcal{H}|}{\delta})}{2m}}\right] \leq \delta.$$

which is equivalent to the statement of the lemma. $\square$