Statistical Machine Learning II
Spring 2017, Learning Theory, Lecture 2

Jean Honorio jhonorio@purdue.edu

1 Hoeffding's inequality

We prove Hoeffding's lemma and leave Hoeffding's inequality as an exercise.

Definition 2.1. *Let $\mathcal{X}$ be an arbitrary domain. A function $f : \mathcal{X} \to \mathbb{R}$ is called convex if:*

$$(\forall a, b \in \mathcal{X}, s \in [0,1])\ f((1-s)a + sb) \leq (1-s)f(a) + sf(b)$$

Lemma 2.1 (Hoeffding's lemma)**.** *Assume that the random variable $x \in [0,1]$ has mean $\mathbb{E}[x] = \mu$. We have that:*

$$\mathbb{E}[e^{t(x-\mu)}] \leq e^{\frac{1}{8}t^2}$$

for all $t \in \mathbb{R}$.

Proof. Invoke Definition 2.1 with $f(x) = e^{t(x-\mu)}$, $a = 0$, $b = 1$:

$$\begin{aligned}
&(\forall s \in [0,1])\ f(s) \leq (1-s)f(0) + sf(1)\\
\Rightarrow&(\forall x \in [0,1])\ f(x) \leq (1-x)f(0) + xf(1)\\
\Rightarrow&(\forall x \in [0,1])\ e^{t(x-\mu)} \leq (1-x)e^{-t\mu} + xe^{t(1-\mu)}
\end{aligned}$$

By computing expectations on both sides, we get:

$$\begin{aligned}
\mathbb{E}[e^{t(x-\mu)}] &\leq (1-\mathbb{E}[x])e^{-t\mu} + \mathbb{E}[x]e^{t(1-\mu)}\\
&= (1-\mu)e^{-t\mu} + \mu e^{t(1-\mu)}\\
&= e^{-t\mu}(1-\mu+\mu e^t)\\
&= e^{g(t)}
\end{aligned}$$

where:

$$g(t) = -t\mu + \log\left(1-\mu+\mu e^t\right)$$

It is easy to note that $g(0) = 0$ and that:

$$\frac{\partial g}{\partial t}(t) = -\mu + \frac{\mu e^t}{1-\mu+\mu e^t} \ \Rightarrow\ \frac{\partial g}{\partial t}(0) = 0$$

Let $w = \dfrac{\mu e^t}{1-\mu+\mu e^t}$, then:

$$\begin{aligned}\frac{\partial^2 g}{\partial t^2}(t) &= \frac{\mu e^t\ (1-\mu+\mu e^t) - \mu e^t\ \mu e^t}{(1-\mu+\mu e^t)^2} \\ &= w(1-w) \\ &\leq 1/4\end{aligned}$$

By Taylor's theorem, for every real t there exists a $v \in [0,t]$ such that:

$$\begin{aligned}g(t) &= g(0) + t\frac{\partial g}{\partial t}(0) + \frac{1}{2}t^2\frac{\partial^2 g}{\partial t^2}(v) \\ &\leq \frac{1}{2}t^2\frac{1}{4} \\ &= t^2/8\end{aligned}$$

which proves our claim. □

2 Exercises

a) Prove the following (look at the proofs of Corollaries 1.2 and 1.3, and use Hoeffding's lemma 2.1):

Corollary 2.1 (Hoeffding's inequality)**.** *Assume that $x_1 \dots x_n$ are n independent random variables with support on $[0,1]$ and mean μ. Fix $\varepsilon > 0$. We have that:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2}$$

b) Prove the following (look at the proofs of Corollaries 1.2 and 1.3):

Corollary 2.2 (Hoeffding's inequality)**.** *Assume that $x_1 \dots x_n$ are n independent random variables, where each $x_i \in [a_i, b_i]$. Fix $\varepsilon > 0$. We have that:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]\right| \geq \varepsilon\right] \leq 2e^{\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

3 Application: Empirical Risk Minimization with a Finite Hypothesis Class

One of the main goals of machine learning is to minimize a risk with respect to a data distribution. Unfortunately, we never observe the data distribution directly, but a finite set of samples drawn from it. Assume an algorithm "learns" by minimizing an empirical risk, i.e., a risk that depends on a training set. Here we prove a generalization result of this learning procedure.

Theorem 2.1. *Assume that $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are arbitrary domains. Assume that the pair (x, y) follows an arbitrary distribution $\mathcal{D}$. Assume that $(x_1, y_1) \dots (x_n, y_n)$ are n i.i.d. samples drawn from the distribution $\mathcal{D}$. Assume that $\mathcal{F}$ is a finite set of functions, i.e., $\mathcal{F} = \{f_1 \dots f_k\}$ where $k < +\infty$ and $(\forall j)\ f_j : \mathcal{X} \to \mathcal{Y}$. The expected risk and its minimizer are defined as:*

$$\overline{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[1[f(x) \neq y]]$$
$$\overline{f} = \arg\min_{f\in\mathcal{F}} \overline{R}(f)$$

The empirical risk and its minimizer are defined as:

$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} 1[f(x_i) \neq y_i]$$
$$\widehat{f} = \arg\min_{f\in\mathcal{F}} \widehat{R}(f)$$

Fix $\delta \in (0, 1)$. We have that:

$$\mathbb{P}\left[\overline{R}(\widehat{f}) - \overline{R}(\overline{f}) < \sqrt{\frac{2(\log k + \log(2/\delta))}{n}}\right] \geq 1 - \delta$$

or equivalently, if $n \geq \dfrac{2(\log k + \log(2/\delta))}{\varepsilon^2}$ then:

$$\mathbb{P}\left[\overline{R}(\widehat{f}) - \overline{R}(\overline{f}) < \varepsilon\right] \geq 1 - \delta$$

Proof. Fix a function $f \in \mathcal{F}$. Define the random variable $z = 1[f(x) \neq y] \in [0, 1]$. Note that the expected and empirical risks are:

$$\overline{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[z]$$
$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} z_i$$

and moreover $\mathbb{E}[z_i] = \overline{R}(f)$, thus:

$$\begin{aligned}\mathbb{E}[\widehat{R}(f)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} z_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[z_i] \\ &= \overline{R}(f)\end{aligned}$$

By the Hoeffding's inequality (Corollary 2.1) for a single hypothesis $f \in \mathcal{F}$, we have:

$$\mathbb{P}\left[\left|\widehat{R}(f) - \overline{R}(f)\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2}$$

By applying the union bound for all k functions in $\mathcal{F}$ and by Hoeffding's inequality (Corollary 2.1), we have:

$$\begin{aligned}\mathbb{P}\left[(\exists f \in \mathcal{F}) \ \left|\widehat{R}(f) - \overline{R}(f)\right| \geq \varepsilon\right] &= \mathbb{P}\left[\left|\widehat{R}(f_1) - \overline{R}(f_1)\right| \geq \varepsilon \text{ or } \ldots \text{ or } \left|\widehat{R}(f_k) - \overline{R}(f_k)\right| \geq \varepsilon\right] \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left[\left|\widehat{R}(f) - \overline{R}(f)\right| \geq \varepsilon\right] \\ &\leq 2ke^{-2n\varepsilon^2}\end{aligned}$$

Equivalently:

$$\begin{aligned}\mathbb{P}\left[(\forall f \in \mathcal{F}) \ \left|\widehat{R}(f) - \overline{R}(f)\right| < \varepsilon\right] &= \mathbb{P}\left[\left|\widehat{R}(f_1) - \overline{R}(f_1)\right| < \varepsilon \text{ and } \ldots \text{ and } \left|\widehat{R}(f_k) - \overline{R}(f_k)\right| < \varepsilon\right] \\ &= 1 - \mathbb{P}\left[(\exists f \in \mathcal{F}) \ \left|\widehat{R}(f) - \overline{R}(f)\right| \geq \varepsilon\right] \\ &\geq 1 - 2ke^{-2n\varepsilon^2} \qquad (1)\end{aligned}$$

Let $\delta = 2ke^{-2n\varepsilon^2}$, then $\varepsilon = \sqrt{\dfrac{\log k + \log(2/\delta)}{2n}}$. Finally since $\widehat{f}$ minimizes $\widehat{R}$ we know that $\widehat{R}(\widehat{f}) \leq \widehat{R}(\overline{f})$. From eq.(1) and the above, we have:

$$\begin{aligned}\overline{R}(\widehat{f}) - \overline{R}(\overline{f}) &< \widehat{R}(\widehat{f}) + \varepsilon - \widehat{R}(\overline{f}) + \varepsilon \\ &\leq 2\varepsilon\end{aligned}$$

which proves our claim. □

Expressions of the form of eq.(1) are called *uniform convergence*.

4 Exercises

a) Assume that $\mathcal{X} = \mathbb{R}^p$ for some number of features p. As in binary classification, assume that $\mathcal{Y} = \{-1, +1\}$. First, assume that $\mathcal{F}$ is the set of *linear classifier* functions of the form:

$$f(x) = \begin{cases} +1 & \text{if } \langle w, x\rangle \geq 0 \\ -1 & \text{if } \langle w, x\rangle < 0 \end{cases}$$

for some $w \in \{-1, 0, +1\}^p$. How many vectors w are in the set $\{-1, 0, +1\}^p$? In other words, what is k in Theorem 2.1? Now, assume that $\mathcal{F}$ is the set of *linear classifier* functions where $w \in \{-1, 0, +1\}^p$ and where w has at most s non-zero elements, for some fixed value s. What is k in Theorem 2.1?