CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Statistical Learning Theory—Finite Case I

Lecturer: Michael I. Jordan *Scribes: Luca Schenato*

1 Chernoff Bounds

Before entering into the details of statistical learning theory, we introduce some useful bounds that will be used extensively in the theory.

The generalized Markov inequality

$$P(h(x) > \epsilon) \leq \frac{E[h(x)]}{\epsilon} \tag{1}$$

where $h(\cdot)$ is a nonnegative function.

Let X a random variable and $g(\cdot)$ a one-to-one function, then by the Markov inequality:

$$\begin{aligned} P(g(X) > g(\epsilon)) &\leq \frac{E[g(X)]}{g(\epsilon)} \\ P(X > \epsilon) &\leq \frac{E[g(X)]}{g(\epsilon)} \end{aligned} \tag{2}$$

Commonly used functions $g(\cdot)$ are the moments of a random variable.

Now, let us consider a random variable X that has all moments, i.e. $E[|X|^n] < \infty$ for all $n \in N$. Define $g(X) \triangleq e^{\lambda X}$. By inequality (**??**) we derive the following inequality:

$$P(X > \epsilon) \leq \frac{E[e^{\lambda X}]}{e^{\lambda \epsilon}} = \frac{M(\lambda)}{e^{\lambda \epsilon}} \tag{3}$$

where $M(\lambda)$ is the *moment generating function* of the r.v. X.

Let $S_n \triangleq \sum_{i=1}^n X_i$ where X_i are i.i.d, and $a \in R$ then

$$\begin{aligned} P(\frac{1}{n}S_n > a) &\leq \frac{E[e^{\lambda S_n})]}{\lambda an} \\ &= \frac{E[e^{\lambda \sum_{i=1}^n X_i})]}{e^{\lambda an}} \\ &= \frac{E[\prod_i e^{\lambda X_i})]}{e^{\lambda an}} \\ &\overset{i.i.d}{=} \frac{\prod_i M(\lambda)}{e^{\lambda an}} = \left(\frac{M(\lambda)}{e^{\lambda a}}\right)^n \\ &= \exp\{-n(\lambda a - \ln M(\lambda))\} \end{aligned} \tag{4}$$

The last inequality holds for any λ, therefore it holds for the λ that minimizes the bound. Let us define the rate function $I(a) \triangleq \sup_\lambda(\lambda a - \ln M(\lambda))$, which corresponds to the Legendre transform of the log moment generating function $\ln M(\lambda)$. Therefore

$$P(\frac{1}{n}S_n > a) \leq \exp\{-nI(a)\} \tag{5}$$

1.1 Bernoulli Distribution

For a Bernoulli random variable $X \sim B(p)$, where $p = P(X = 1)$, the moment generating function is $M(\lambda) = (1-p) + pe^{\lambda}$ and the rate function can be computed explicitly:

$$\begin{aligned}
&\frac{d}{d\lambda}\left[\lambda a - \ln(1-p+pe^{\lambda})\right] = 0 \\
&\Rightarrow \lambda = \ln\left(\frac{a(1-p)}{p(1-a)}\right) \\
&\Rightarrow I(a) = a\ln\frac{a}{p} + (1-a)\ln\frac{1-a}{1-p} = KL(a \;|||\; p)
\end{aligned}$$

where $KL()$ is the KL divergence.

Since we are interested in $P(\frac{1}{n}S_n - p > \epsilon)$, let us define $a \stackrel{\Delta}{=} p + \epsilon$ and then expand $I(a)$ around p as follows:

$$I(p+\epsilon) = (p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-p-\epsilon)\ln\frac{1-p-\epsilon}{1-p} = \frac{\epsilon^2}{2p(1-p)} + O(\epsilon^3) \geq \frac{\epsilon^2}{2p(1-p)} \geq 2\epsilon^2 \tag{6}$$

where we made use of the fact that $O(\epsilon^3)$ is a negative quantity, and that the product $p(1-p)$ is maximized for $p = \frac{1}{2}$. The final form of the inequality above is:

$$P(\frac{1}{n}S_n - p > \epsilon) \leq e^{-2\epsilon^2 n} \tag{7}$$

which is known as the Hoeffding bound, also known as the additive Chernoff bound.

1.2 Multiplicative Chernoff Bounds

It is also possible to derive expressions for multiplicative Chernoff bounds. The derivations are analogous to the additive bounds and we report here only the final expressions:

$$P\left(\frac{\frac{1}{n}S_n - p}{\sqrt{p}} > \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{3}} \tag{8}$$

$$P\left(\frac{p - \frac{1}{n}S_n}{\sqrt{p}} > \epsilon\right) \leq e^{-\frac{\epsilon^2 n}{2}} \tag{9}$$

Analogous bounds exist also for martingales, known as Azuma's inequalities.

2 Goal of Statistical Learning Theory

Our goal is to answer to two questions:

1. How close in probability is the performance of our model based on a specific sample of data to the performance on unseen data? Formally we want to prove that the event:

$$R(\alpha_l) < R_{emp}(\alpha_l) + \sqrt{4\frac{h(\ln\frac{2l}{h}+1) - \ln\frac{\eta}{4}}{l}} + \frac{1}{l} \tag{10}$$

   holds with probability greater or equal than $(1-\eta)$, where l is the sample size, α_l is the minimizer of $R_{emp}(\alpha)$, and h is the VC dimension of the model class.

2. How close in probability is the performance of the best model based on a finite sample of data from the best possible model? Formally we want to prove that:

$$R(\alpha_l) < R(\alpha_0) + \sqrt{4\frac{h(\ln\frac{2l}{h}+1) - \ln\frac{\eta}{4}}{l}} + \frac{1}{l} + \sqrt{\frac{-\ln\eta}{2l}} \tag{11}$$

   holds with probability greater or equal than $(1-\eta)$, where α_0 is the minimizer of $R(\alpha)$.

3 Simplest Model: the pessimistic case

Let us now turn to the *simplest model*, where we consider only a finite number N of models, which are parameterized by $\alpha \in \Lambda$. Recall that the risk functional is defined as $R(\alpha) \triangleq \int Q(z,\alpha)dF(z)$, where $z = (x,\omega)$, ω is the class label, and $Q(z,\alpha) \in \{0,1\}$.

Given a finite set of data $\{z_1, z_2, \ldots, z_l\}$, the empirical risk and its minimizer are defined as:

$$\begin{aligned} R_{emp} &\triangleq \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \\ \alpha_l &\triangleq \arg\min_{\alpha} R(\alpha) \end{aligned}$$

We can now give upper bounds on the following event:

$$\begin{aligned} & P\left(\sup_{1\le k\le N}\left(\int Q(z,\alpha_k)dF(z) - \frac{1}{l}\sum_i Q(z_i,\alpha_k)\right) > \epsilon\right) \\ = \quad & P\left(\bigcup_{1\le k\le N}\left[\int Q(z,\alpha_k)dF(z) - \frac{1}{l}\sum_i Q(z_i,\alpha_k) > \epsilon\right]\right) \\ \overset{union\,bound}{\le} \quad & \sum_{k=1}^{N} P\left(\int Q(z,\alpha_k)dF(z) - \frac{1}{l}\sum_i Q(z_i,\alpha_k) > \epsilon\right) \\ \overset{Chernoff}{\le} \quad & \sum_{k=1}^{N} e^{-2\epsilon^2 l} = Ne^{-2\epsilon^2 l} \end{aligned}$$

where the inequality (**??**) comes from the union bound, i.e. $P(A\cup B) \le P(A) + P(B)$, and the inequality (**??**) comes from the additive Chernoff bounds (Note that $E[R_{emp}(\alpha)] = R(\alpha)$).

The inequality above can be rewritten in terms of risk bounds as follows:

$$\begin{aligned} & \eta \triangleq Ne^{-2\epsilon^2 l} \\ & \Rightarrow \epsilon = \sqrt{\frac{\ln N - \ln\eta}{2l}} \end{aligned}$$

Thus with probability $1-\eta$ we have:

$$\begin{aligned}
&\sup_{1\le k\le N}(R(\alpha_k)-R_{emp}(\alpha_k))\le\epsilon\\
&\Rightarrow R(\alpha_k)-R_{emp}(\alpha_k)\le\epsilon,\ \forall\alpha_k\\
&\Rightarrow R(\alpha_l)\le R_{emp}(\alpha_l)+\sqrt{\frac{\ln N-\ln\eta}{2l}} \qquad (12)
\end{aligned}$$

Now we want to answer the second question. We have:

$$P\left(\frac{1}{l}\sum_i Q(z_i,\alpha_0)-\int Q(z,\alpha_0)dF(z)\right)\le e^{2\epsilon^2 l}$$

Therefore, with probability $1-\eta$:

$$R(\alpha_0)\le R_{emp}(\alpha_0)+\sqrt{\frac{-\ln\eta}{2l}} \qquad (13)$$

Since α_l minimizes $R_{emp}(\alpha)$, we have $R_{emp}(\alpha_0)\ge R_{emp}(\alpha_l)$. As a consequence, with probability $1-2\eta$ the following inequality holds:

$$R(\alpha_0)\le R_{emp}(\alpha_0)+\sqrt{\frac{\ln N-\ln\eta}{2l}}+\sqrt{\frac{-\ln\eta}{2l}} \qquad (14)$$

Proof:

Define:

$$\begin{aligned}
A &= \text{Equation}\,(\textbf{??})\\
B &= \text{Equation}\,(\textbf{??})\\
C &= \text{Equation}\,(\textbf{??})
\end{aligned}$$

Noticing that $R_{emp}(\alpha_l)\le R_{emp}(\alpha_0)\le R(\alpha_0)$, we have:

$$\begin{aligned}
P(C) &= 1-P(\bar{C})\\
&\ge 1-P(\overline{\bar{A}\cap\bar{B}})\\
&= 1-P(\bar{A}\cup\bar{B})\\
&\overset{union\,bound}{\ge} 1-P(\bar{A})-P(\bar{B})\\
&= 1-\eta-\eta=1-2\eta
\end{aligned}$$

The inequality (**??**) is tight in the pessimistic case, i.e. when $p=\frac{1}{2}$. This is the result of the approximation we used in Equation (**??**).