CS281B/Stat241B. Statistical Learning Theory. Lecture 4.

Peter Bartlett

1. Concentration inequalities
   (a) Markov, Chebyshev
   (b) Chernoff technique
   (c) Sub-Gaussian
   (d) Sub-Exponential

Risk bounds and uniform convergence

For empirical risk minimization strategies, which choose $f_n \in F$ to minimize

$$\hat{R}(f) = \hat{\mathrm{E}}\ell(f(X), Y) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i),$$

how does the risk $R(f_n) = \mathrm{E}\ell(f_n(X), Y)$ behave?

Does $R(f_n) \to \inf_{f \in F} R(f)$?

How rapidly?

Risk bounds and uniform convergence

If we consider a single prediction rule f, we can appeal to the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i) \to \mathrm{E}\ell(f(X),Y).$$

And, for instance, ℓ bounded implies $\Pr(|\hat{R}(f) - R(f)| > \epsilon)$ decreases exponentially in n.

For this, we'll study *concentration inequalities*, which bound the probability of deviations of random variables from their expectations. But because we use data to choose f_n, we need something stronger than a law of large numbers.

Risk bounds and uniform convergence

Example:

For pattern classification ($\mathcal{Y} = \{0, 1\}$), consider $F = F_+ \cup F_-$ with

$$F_+ = \{1[S] : |S| < \infty\},$$
$$F_- = \{1[S] : |\mathcal{X} - S| < \infty\}$$

Then for a continuous distribution on $\mathcal{X}$ with $P(Y = 1|X) = 0.9$,

$$R(f) = \begin{cases} 0.1 & \text{for } f \in F_-, \\ 0.9 & \text{for } f \in F_+. \end{cases}$$

But for any sample, there is an empirical risk minimizer $f_n \in F_+$ with $\hat{R}(f) = 0$.

Risk bounds and uniform convergence

If the set F is finite, we *can* relate risk to empirical risk:

Theorem: For $\ell(f(x), y) \in \{0, 1\}$,

$$\Pr\left(\exists f \in F \text{ s.t. } \hat{R}(f) = 0 \text{ and } R(f) \geq \epsilon\right) \leq |F| e^{-\epsilon n}.$$

Proof:

$$\begin{aligned}
\Pr\left(\bigcup_{f \in F} \{\hat{R}(f) = 0,\ R(f) \geq \epsilon\}\right) &\leq \sum_{f \in F} \Pr\{\hat{R}(f) = 0,\ R(f) \geq \epsilon\} \\
&\leq |F| \max_{f \in F} \Pr\{\hat{R}(f) = 0,\ R(f) \geq \epsilon\} \\
&\leq |F| (1 - \epsilon)^n \\
&\leq |F| \exp(-n\epsilon).
\end{aligned}$$

Risk bounds and uniform convergence

So any F that is parameterized using a fixed number of bits satisfies this uniform convergence property.

Concentration inequalities

We'll get back to uniform convergence properties later. For now, we'll focus on tail probabilities like $P(T_n \geq t)$ for some statistic T_n. We could consider asymptotic results—like the central limit theorem:

$$\lim_{n\to\infty} P(\bar{X}_n \geq \mu + \sigma\sqrt{n}t) = 1 - \Phi(t).$$

This tells us what happens asymptotically, but we usually have a fixed sample size. What can we say in that case? For example, what is

$$P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right)?$$

These are **concentration inequalities**, i.e., bounds on this kind of probability that $\bar{X}_n$ is concentrated about its mean.

Concentration inequalities

We'll look at several concentration inequalities, that exploit various kinds of information about the random variables.

1. Using moment bounds:
   Markov (first), Chebyshev (second)

2. Using moment generating function bounds, for sums of independent r.v.s:
   Chernoff; Hoeffding; sub-Gaussian, sub-exponential random variables; Bernstein.

3. Martingale methods:
   Hoeffding-Azuma, bounded differences.

Markov's Inequality

Theorem: For $X \geq 0$ a.s., $\mathbf{E}X < \infty$, $t > 0$:

$$P(X \geq t) \leq \frac{\mathbf{E}X}{t}.$$

Proof:

$$\begin{aligned}\mathbf{E}X &= \int X dP \\ &\geq \int_t^\infty x dP(x) \\ &\geq t \int_t^\infty dP(x) \\ &= tP(X \geq t).\end{aligned}$$

Moment Inequalities

Consider $|X - \mathbf{E}X|$ in place of X.

Theorem: For $\mathbf{E}X < \infty$, $f : [0, \infty) \to [0, \infty)$ strictly monotonic, $\mathbf{E}f(|X - \mathbf{E}X|) < \infty, t > 0$:

$$P(|X - \mathbf{E}X| \geq t) = P\left(f(|X - \mathbf{E}X| \geq f(t)\right) \leq \frac{\mathbf{E}f(|X - \mathbf{E}X|)}{f(t)}.$$

Moment Inequalities

e.g., $f(a) = a^2$ gives **Chebyshev's inequality:**

Theorem:

$$P(|X - \mathbf{E}X| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

e.g., $f(a) = a^k$:

Theorem:

$$P(|X - \mathbf{E}X| \geq t) \leq \frac{\mathbf{E}|X - \mathbf{E}X|^k}{t^k}.$$

Chernoff bounds

Use $a \mapsto \exp(\lambda a)$ for $\lambda > 0$:

Theorem: For $\mathbf{E}X < \infty$, $\mathbf{E}\exp(\lambda(X - \mathbf{E}X)) < \infty$, $t > 0$:

$$\begin{aligned} P(X - \mathbf{E}X \geq t) &= P\left(\exp(\lambda(X - \mathbf{E}X)) \geq \exp(\lambda t)\right) \\ &\leq \frac{\mathbf{E}\exp(\lambda(X - \mathbf{E}X))}{\exp(\lambda t)} \\ &= e^{-\lambda t} M_{X-\mu}(\lambda). \end{aligned}$$

$M_{X-\mu}(\lambda) = \mathbf{E}\exp(\lambda(X - \mu))$ (for $\mu = \mathbf{E}X$) is the **moment-generating function** of $X - \mu$.

Example: Gaussian

For $X \sim N(\mu, \sigma^2)$, $M_{X-\mu}(\lambda)$ is

$$\begin{aligned}
\mathbf{E}\exp(\lambda(X-\mu)) &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp(\lambda x - x^2/(2\sigma^2))\,dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left(\lambda^2\sigma^2/2 - (x/\sigma - \lambda\sigma)^2/2\right)\,dx \\
&= \exp(\lambda^2\sigma^2/2)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-(y-\lambda\sigma)^2/2\right)\,dy \\
&= \exp(\lambda^2\sigma^2/2),
\end{aligned}$$

for the change of variable $y = x/\sigma$.

Example: Gaussian

Thus,

$$\begin{aligned}\log P(X-\mu \geq t) &\leq -\sup_{\lambda>0}\left(\lambda t - \log M_{X-\mu}(\lambda)\right) \\ &= -\sup_{\lambda>0}\left(\lambda t - \frac{\lambda^2\sigma^2}{2}\right) \\ &= -\frac{t^2}{2\sigma^2},\end{aligned}$$

using the optimal choice $\lambda = t/\sigma^2 > 0$.

Example: Gaussian

For $X \sim N(\mu, \sigma^2)$, it's easy to check that the Chernoff technique gives a tight bound:

$$\lim_{n \to \infty} \frac{1}{n} \log P(\bar{X}_n - \mu \geq t) = -\frac{t^2}{2\sigma^2}.$$

Example: Bounded Support

Theorem: [Hoeffding's Inequality] For a random variable $X \in [a, b]$ with $\mathbf{E}X = \mu$ and $\lambda \in \mathbb{R}$,

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}.$$

Note the resemblance to a Gaussian: $\lambda^2\sigma^2/2$ vs $\lambda^2(b-a)^2/8$. (And since P has support in $[a, b]$, Var$X \leq (b-a)^2/4$.)

Example: Hoeffding's Inequality Proof

Define

$$A(\lambda) = \log\left(\mathbf{E}e^{\lambda X}\right) = \log\left(\int e^{\lambda x}\, dP(x)\right),$$

where $X \sim P$. Then A is the log normalization of the exponential family random variable X_λ with reference measure P and sufficient statistic x. Since P has bounded support, $A(\lambda) < \infty$ for all λ, and we know that

$$A'(\lambda) = \mathbf{E}(X_\lambda), \qquad\qquad A''(\lambda) = \mathrm{Var}(X_\lambda).$$

Since P has support in $[a, b]$, $\mathrm{Var}(X_\lambda) \le (b-a)^2/4$. Then a Taylor expansion about $\lambda = 0$ (at this value of λ, X_λ has the same distribution as X, hence the same expectation) gives

$$A(\lambda) \le \lambda \mathbf{E}X + \frac{\lambda^2}{2}\frac{(b-a)^2}{4}.$$

Sub-Gaussian Random Variables

Definition: X is **sub-Gaussian** with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}.$$

Note:

- Gaussian is sub-Gaussian.
- X sub-Gaussian iff $-X$ sub-Gaussian.

Sub-Gaussian Random Variables

Note:

- X sub-Gaussian implies

$$\begin{aligned} P(X-\mu \geq t) &\leq \exp(-t^2/(2\sigma^2)), \\ P(X-\mu \leq -t) &\leq \exp(-t^2/(2\sigma^2)), \\ P(|X-\mu| \geq t) &\leq 2\exp(-t^2/(2\sigma^2)). \end{aligned}$$

Sub-Gaussian Random Variables

Note:

- X_1, X_2 independent, sub-Gaussian with parameters σ_1^2, σ_2^2, implies $X_1 + X_2$ sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$.

Indeed, for independent X_1, X_2,

$$\begin{aligned} M_{X_1+X_2} &= \mathbf{E} \exp\left(\lambda(X_1 + X_2)\right) \\ &= \mathbf{E} \exp\left(\lambda X_1\right) \mathbf{E} \exp\left(\lambda X_2\right) \\ &= M_{X_1} M_{X_2}. \end{aligned}$$

So $\log M_{X_1+X_2-\mu} = \log M_{X_1-\mu_1} + \log M_{X_2-\mu_2} \leq \lambda^2(\sigma_1^2 + \sigma_2^2)/2$.

Hoeffding Bound

Theorem: For $X_1, \ldots, X_n$ independent, $\mathbf{E}X_i = \mu_i$, X_i sub-Gaussian with parameter σ_i^2, then for all $t > 0$,

$$P\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

e.g., for $\mathbf{E}X_i = 0$, $X_i \in [a, b]$, we have $\sigma_i^2 = (b-a)^2/4$ so

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Sub-Exponential Random Variables

Definition: X is **sub-exponential** with parameters (σ^2, b) if, for all $|\lambda| < 1/b$,

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}.$$

Examples:

- Sub-Gaussian X with parameter σ^2 is sub-exponential with parameters (σ^2, b) for all $b > 0$.

Sub-Exponential Random Variables

Theorem: For X sub-exponential with parameters (σ^2, b),

$$P(X \geq \mu + t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\sigma^2}\right) & \text{if } 0 \leq t \leq \sigma^2/b, \\ \exp\left(-\frac{t}{2b}\right) & \text{if } t > \sigma^2/b. \end{cases}$$

Sub-Exponential Random Variables

Proof: Assume $\mu = 0$. As before,

$$P(X \geq t) \leq \exp(-\lambda t)\mathbf{E}\exp(\lambda X)$$
$$\leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right)$$

provided $0 \leq \lambda < 1/b$. As before, we optimize the choice of λ. But now, it is constrained to $[0, 1/b)$. Without this constraint, the minimum occurs at $\lambda^* = t/\sigma^2$. So if

$$t/\sigma^2 < 1/b \Longleftrightarrow t < \sigma^2/b,$$

we have

$$P(X \geq t) \leq \exp(-\lambda^* t + {\lambda^*}^2\sigma^2/2) = \exp(-t^2/(2\sigma^2)).$$

Sub-Exponential Random Variables

If t is larger, the minimum occurs at $\lambda = 1/b$ (since the function $t \mapsto -\lambda t + \frac{\lambda^2\sigma^2}{2}$ is monotonically decreasing in $[0, \lambda^*]$, which contains $[0, 1/b]$). Substituting this λ gives

$$P(X \geq t) \leq \exp(-t/b + \sigma^2/(2b^2)) \leq \exp(-t/(2b)),$$

where the second inequality follows from $t \geq \sigma^2/b$.

Sub-Exponential Random Variables

Example: X variance σ^2, bounded: $|X-\mu| \leq b$.

$$\begin{aligned}\mathbf{E}\exp(\lambda(X-\mu)) &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbf{E}(X-\mu)^k}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2}.\end{aligned}$$

And for $|\lambda| < 1/b$, this is no more than

$$\mathbf{E}\exp(\lambda(X-\mu)) \leq 1 + \frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right).$$

Sub-Exponential Random Variables

So if $|\lambda| < 1/(2b)$, $1 - b|\lambda| > 1/2$ and

$$\mathbf{E}\exp(\lambda(X-\mu)) \leq \exp\left(\lambda^2\sigma^2\right).$$

Thus, X is sub-exponential with parameters $(2\sigma^2, 2b)$.

Overview

1. Concentration inequalities
   (a) Markov, Chebyshev
   (b) Chernoff technique
   (c) Sub-Gaussian
   (d) Sub-Exponential