Machine Learning Theory

FENG JIANTING

This is a collection of summary of resources in theoretical machine learning. [1]

Contents

Part I. High Dimensional Statistics

This part is mainly focus on preliminary knowledge for statistics in high dimensional space.

1 Concentration Inequalities

A simple way to control a tail probability $\mathbb{P}X \geq t$ is by its moments. Higer-order moments always leads to a sharper bounds. A basic result is Markov's ineuqality

$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}X}{t}, \quad \text{for all } t > 0.$$

which requires finite mean $\mathbb{E}X$. The proof is quiet intuitive,

Proof. For all $t > 0$

$$\mathbb{P}[X \geq t] = \int_{x \geq t} dP \tag{1}$$

$$\leq \int_{x \geq t} \frac{x}{t} dP \tag{2}$$

$$\leq \int_{\mathbb{R}} \frac{x}{t} dP \tag{3}$$

$$= \frac{\mathbb{E}X}{t} \tag{4}$$

□

Another similar result is Chebyshev's inequality

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathrm{Var}X}{t^2}, \quad \text{for all } t > 0.$$

where $\mu = \mathbb{E}X$.

Similarly, for any $k \in \mathbb{N}_+$, we always have

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}\left[|X - \mu|^k\right]}{t^k}$$

once $\mathbb{E}\left[|X - \mu|^k\right]$ exists.

Of course, other strictly increase functions can be applied to $|X - \mu|^k$ other than just polynomials. Consider the moment generating function $\varphi(\lambda) = \mathbb{E}\left[e^{\lambda(X-\mu)}\right]$ exists for $\lambda \leq |b|$ where $b > 0$ is some constant. We can apply Markov's inequality to $Y = \exp(\lambda(X - \mu))$, thereby, we obtain

$$\mathbb{P}[(X - \mu) \geq t]\, \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

take logarithm, and optimizing the RHS, which yields the *Chernoff* bound

$$\log \mathbb{P}[(X - \mu) \geq t] \leq \inf_{\lambda \in [0,b]} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\}$$

1.1 Sub-Gaussian variables and Hoeffding bounds

Base on the discussion above, it's easy to come up with the idea of classify the r.v. based on their MGF. here we give an example, which consider all the r.v. with tail "lower" than gaussian r.v.

Example 1.1. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian r.v. with mean μ and variance σ^2. By straightforward calculation, we get

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] = e^{\frac{\lambda^2\sigma^2}{2}}$$

for all $\lambda \in \mathbb{R}$. Then take inf,

$$\inf_{\lambda\geq 0}\left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right\} = \inf_{\lambda\geq 0}\left\{\frac{\sigma^2\lambda^2}{2} - \lambda t\right\} = -\frac{t^2}{2\sigma^2}$$

The tail probability is bounded by

$$\mathbb{P}\left[X - \mu \geq t\right] \leq e^{-\frac{t^2}{2\sigma^2}}$$

Motivated by this, we can define

Definition 1.2. A r.v. X with $\mu = \mathbb{E}X$ is sub-Gaussian if there exists a positive $\sigma > 0$ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}$$

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

That is, the MGF of the given r.v. is bounded by the MGF of Gaussian r.v..The constant σ is referred to as the sub-Gaussian parameter. Based on the definition, we can easily get the concentration inequality for sub-Gaussian r.v.

$$\mathbb{P}\left[|X - \mu| \geq t\right] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

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

Example 1.3. A Rademacher r.v. ε takes the value $\{-1, 1\}$ with the same probability. Indeed, it's a sub-Gaussian r.v. with parameter $\sigma = 1$.

Claim 1. *Any r.v. supported finitely are sub-Gaussian.*

Example 1.4. Let X be zero-mean, and supported on some intervals $[a, b]$. Letting X' be an independent copy of X, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_X\left[e^{\lambda X}\right] = \mathbb{E}_X\left[e^{\lambda(X-\mathbb{E}_{X'}[X'])}\right] \leq \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]$$

the inequality is given by the convexity of $e^{-\lambda x}$. Let ε be Rademacher r.v., note that $\varepsilon(X - X')$ has the same distribution as $X - X'$. So we have

$$\mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_\varepsilon\left[e^{\lambda\varepsilon(X-X')}\right]\right] \leq \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right]$$

the inequality is given by the fact that Rademacher r.v. is sub-Gaussian. Since $|X - X'| \leq b - a$, we have

$$\mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right] \leq e^{\frac{\lambda^2(b-a)^2}{2}}$$

Which shows that bounded r.v. is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$.

Here we propose the general Hoeffding bound,

Proposition 1.5. *Suppose that the variables $X_i, i = 1, \cdots, n$ are independent, and X_i has mean μ_i with sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have*

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

Proof.

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] &\leq \exp\left\{\frac{\lambda^2}{2}\sum_{i=1}^{n}\sigma_i^2 - \lambda t\right\} \\ &\leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}\end{aligned}$$

□

Following theorem gives us the equivalent characterization of sub-Gaussian r.v.

Theorem 1.6. *Given any zero-mean r.v., the following properties are equivalent:*

1. *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}$$

2. *There is a constant $c \geq 0$ and Gaussian r.v. $Z \sim \mathcal{N}(0, \tau)$ such that*

$$\mathbb{P}\left[|X| \geq s\right] \leq c\mathbb{P}\left[|Z| \geq s\right], \quad \text{for all } s \geq 0$$

3. *There is a constant $\theta \geq 0$ such that*

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \quad \text{for all } k = 1, 2, \cdots$$

4. *There is a constant $\sigma \geq 0$ such that*

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \quad \text{for all } \lambda \in [0, 1)$$

1.2 Sub-exponential variables and Bernstein bounds

The notion of sub-Gaussianity is fairly restrictive, so it's natrual to relax it. Accordingly, we now turn to sub-exponential r.v. with milder condition.

Definition 1.7. A random variable X with $\mu = \mathbb{E}X$ is sub-exponential if there are non-negative parameters (ν, α) such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}.$$

It immediately follows that all sub-Gaussian r.v. are sub-Exponential, take $(\nu, \alpha) = (\sigma, 0)$. However, the converse doesn't hold.

Example 1.8. Let $Z \sim \mathcal{N}(0,1)$, and consider the r.v. $X = Z^{@}$. For $\lambda < \frac{1}{2}$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-1)}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\lambda(z^2-1)e^{-z^2/2}}dz \\ &= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\end{aligned}$$

for $\lambda > \frac{1}{2}$, the MGF doesn't exists. For all $|\lambda| < \frac{1}{4}$,

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}$$

therefore, Z is a $(\nu, \alpha) = (4, 2)$ sub-exponential r.v.

Proposition 1.9. *Suppose that X is sub-exponential with parameters (ν, α). Then*

$$\mathbb{P}\left[X - \mu \geq t\right] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ e^{-\frac{t}{2\alpha}} & \text{if } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Proof. Follow the Chernoff bound, we have

$$\mathbb{P}\left[X - \mu \geq t\right] \leq \exp\left(\log \mathbb{E}[e^{\lambda X}] - \lambda t\right) \leq \exp\left(\frac{\nu^2\lambda^2}{2} - \lambda t\right)$$

for $\lambda < \frac{1}{\alpha}$. Take $g(\lambda, t) = \frac{\nu^2\lambda^2}{2} - \lambda t$, consider the unconstrained optimization, the function takes minimum at $\lambda^* = \frac{t}{\nu^2}$.

If $\frac{t}{\nu^2} < \frac{1}{\alpha}$, then $\inf_{\lambda \in [0, \alpha^{-1})} g(\lambda, t) = -\frac{t^2}{2\nu^2}$.

Otherwise, if $t \geq \frac{\nu^2}{\alpha}$, the minimum achieves at $\lambda^* = \alpha^{-1}$, and

$$\inf_{\lambda \in [0, \alpha^{-1})} g(\lambda, t) = -\frac{t}{\alpha^2} + \frac{\nu^2}{2\alpha^2} \leq -\frac{t}{2\alpha}$$

the inequality is derived from $\frac{\nu^2}{\alpha} \leq t$. □

Based on the example of X^2, tsub-exponential r.v. can be verified explicitly with boundingthe MGF, however, in many settings, it's impracticable. An alternative approach is based on the control of polynomial moments, called *Bernstein's condition*. Given a r.v. with $\mu = \mathbb{E}X$ and $\sigma^2 = \mathrm{Var}X$, we say Bernstein's condition with parameter b holds if

$$|\mathbb{E}\left[(X-\mu)^k\right]| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \text{for } k \geq 2.$$

For any bounded r.v., the Bernstein condition is obvious.

When this condition holds, we can expand the MGF

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\lambda^k\frac{\mathbb{E}\left[(X-\mu)^k\right]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2},\end{aligned}$$

Sum the geometric series so as to obtain

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}$$

Consequently, we conclude that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{2b}.$$

which is a sub-exponential r.v. with $(\nu, \alpha) = (\sqrt{2}\sigma, 2b)$.

Proposition 1.10. *For any r.v. satisfying Bernstein condition, we have*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}$$

for all $|\lambda| < \frac{1}{b}$. *Moveover,*

$$\mathbb{P}\left[|X-\mu| \geq t\right] \leq 2e^{\frac{t^2}{2(\sigma^2+bt)}} \quad \textit{for all } t \geq 0$$

Take $\lambda = \frac{t}{bt+\sigma^2} \in [0, b^{-1})$ in Chernoff bound, the tail bound can be derived.

Similar to sub-Gaussian r.v., the sum of sub-exponential r.v. is also a sub-exponential r.v. with the following bound, take $\{X_k\}_{k=1}^n$ with

$$\alpha_* = \max_{k\in[n]}\alpha_k \quad \text{and} \quad \nu_* = \sqrt{\sum_{k=1}^n v_k^2}.$$

then

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n(X_k - \mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{nt^2}{2(\nu_*^2/n)}}, & \text{for } 0 \leq t \leq \frac{\nu_*^2}{n\alpha_*}, \\ e^{-\frac{nt}{2\alpha_*}}, & \text{for } t > \frac{\nu^2}{n\alpha_*} \end{cases}$$

Example 1.11. A chi-squared (χ^2) r.v. with n degrees of freedom, denoted by $Y \sim \chi^2_n$, can be represented as the sum

$$Y = \sum_{k=1}^{n} Z_k^2, \text{ where } Z_k \sim \mathcal{N}(0,1)$$

as preceding discussion, Z_k is $(2,4)$ sub-exponential r.v., therefore, Y is $(2\sqrt{n},4)$ sub-exponential, we can easily deduced that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| \geq 1\right] \leq 2e^{-nt^2/8}, \quad \text{for all } t \in (0,1)$$

The concentration of χ^2 r.v. takes an important role in the following Johnson-Lindenstrauss theorem for random projections.

Example 1.12. Suppose given $N \geq 2$ distinct vectors $\{u^1, \cdots, u^N\} \subset \mathbb{R}^d$, if the data dimension d is large, then it might be expensive to store and manipulate the data set. The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ with $m \ll d$, which guarantee that

$$1 - \delta \leq \frac{\|F(u^i) - F(u^j)\|_2^2}{\|u^i - u^j\|_2^2} \leq 1 + \delta, \quad \text{for all pairs } u^i \neq u^j.$$

The construction is probabilistic: we implement random projection with a random matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ filled with independent $\mathcal{N}(0,1)$ elements. And define

$$\begin{aligned} F : \mathbb{R}^d &\to \mathbb{R}^m \\ u &\mapsto \boldsymbol{X}u/\sqrt{m} \end{aligned}$$

Now, we verify that F satisfies the given condition with high probability. Let $x_i \in \mathbb{R}^d$ denote the i-th row of $\boldsymbol{X}$, fix $u \in \mathbb{R}^d$, we know that $\langle x_i, \frac{u}{\|u\|_2}\rangle \sim \mathcal{N}(0,1)$. Therefore,

$$Y = \frac{\|\boldsymbol{X}u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle x_i, u/\|u\|\rangle^2,$$

follows a χ^2_m r.v., where the d.f. is m. Therefore, applying the previous bound for χ^2 distribution,

$$\mathbb{P}\left[\left|\frac{\|\boldsymbol{X}u\|_2^2}{m\|u\|_2^2} - 1\right| \geq \delta\right] \leq 2e^{-m\delta^2/8} \quad \text{for all } \delta \in (0,1).$$

Rearranging and recalling the definition of F yields that

$$\mathbb{P}\left[\frac{\|F(u)\|_2^2}{|u\|_2^2} \notin [1-\delta, 1+\delta]\right] \leq 2e^{-m\delta^2/8} \quad \text{for any fixed } 0 \neq u \in \mathbb{R}^d.$$

Note that there exists $\binom{N}{2}$ pairs of distinct u, we apply the union bound and get

$$\mathbb{P}\left[\frac{\|F(u^i-u^j)\|_2^2}{\|u^i-u^j\|_2^2} \notin [1-\delta,1+\delta] \text{ for some } u^i \neq u^j\right] \leq 2\binom{N}{2}e^{-m\delta^2/8}$$

for any $\epsilon \in (0,1)$, this probability can be bound below ϵ by choosing $m \geq \frac{C}{\delta^2}\log N$.

Theorem 1.13. *For a zero-mean r.v. X, the following statements are equivalent:*

1. *There are non-negative numbers ν, α such that*
$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}$$
2. *There is a positive number c_0 such that $\mathbb{E}[e^{\lambda X}] < \infty$ for all $\lambda < c_0$*
3. *There are constants $c_1, c_2 > 0$ such that*
$$\mathbb{P}^{|X|\geq t} \leq c_1 e^{-c_2 t} \quad \text{for all } t > 0.$$
4. *The quantity $\gamma = \sup_{k\geq 2}\left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ is finite.*

One condition of two-sided Bernstein bound holds is $|X| \leq b$ almost surely, but if we only have one-sided inequality, we can still get some one-sided bounds.

1.3 Martingale-based methods

Let $\{X_k\}_{k=1}^n$ be a sequence of independent r.v., consider the random variable $f(X) = f(X_1, \cdots, X_n)$ for some funciton $f : \mathbb{R}^n \to \mathbb{R}$. Define

$$Y_k = \mathbb{E}[f(X)|X_1, \cdots, X_k] \quad \text{for } k = 1, \cdots, n-1$$

and $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$. Based on telescoping decomposition

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{Y_k - Y_{k-1}}_{=D_k}$$

References

[1] Ali Rahimi and Benjamin Recht. "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Ed. by J. Platt et al. Vol. 20. Curran Associates, Inc., 2007.