机器学习理论导引
作业三

欧先飞，DZ1833019，ouxianfei@smail.nju.edu.cn

2019 年 5 月 7 日

1 [20pts] Concentration Inequalities

(1) **[5pts]** 请利用高斯函数的性质推导 $\Pr[\bar{X} - \mu \geq \epsilon]$ 上界.

(2) **[5pts]** 请给出 Bernstein 不等式以 $1-\delta$ 概率表达形式.

(3) **[10pts]** 如果随机变量 $X \in [\alpha, \beta]$, 其中 $\beta > \alpha$ 是两个常数。试证明，随机变量 X 是一个亚高斯随机变量 (sub-Gaussian Random Variable)，且满足其亚高斯系数 $b = (\beta - \alpha)^2/4$，即对于任意 $\lambda > 0$，有下式成立

$$\ln \mathbb{E}[\exp\{\lambda(X - \mu)\}] \leq \lambda^2(\beta - \alpha)^2/8, \tag{1.1}$$

其中 $\mu = \mathbb{E}[X]$.

Proof.
(1) $X \sim N(\mu, \sigma^2) \implies \sum_{i=1}^m X_i \sim N(m\mu, m\sigma^2) \implies \bar{X} = \frac{1}{m}\sum_{i=1}^m X_i \sim N(\mu, \frac{\sigma^2}{m})$，所以 $\Pr(\bar{X} - \mu \geq \epsilon) = \Pr(\bar{X} \geq \mu + \epsilon) = \int_{\mu+\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{m}}} \exp(-\frac{(x-\mu)^2}{2\times\frac{\sigma^2}{m}})dx$，令 $\sigma_0^2 = \frac{\sigma^2}{m}$，于是有：

$$\begin{aligned}
\Pr(\bar{X} \geq \mu + \epsilon) &= \int_{\mu+\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{(x-\mu)^2}{2\sigma_0^2})dx \\
&= \int_{\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{x^2}{2\sigma_0^2})dx \\
&= \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{(x+\epsilon)^2}{2\sigma_0^2})dx \\
&= \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{x^2}{2\sigma_0^2}) \exp(-\frac{\epsilon^2}{2\sigma_0^2}) \exp(-\frac{2x\epsilon}{2\sigma_0^2})dx \\
&= \exp(-\frac{\epsilon^2}{2\sigma_0^2}) \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{x^2}{2\sigma_0^2}) \exp(-\frac{2x\epsilon}{2\sigma_0^2})dx \\
&\leq \exp(-\frac{\epsilon^2}{2\sigma_0^2}) \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp(-\frac{x^2}{2\sigma_0^2})dx \\
&= \exp(-\frac{\epsilon^2}{2\sigma_0^2}) = \exp(-\frac{m\epsilon^2}{2\sigma^2})
\end{aligned}$$

(2) $\Pr(\bar{X_n} \geq \mu + \epsilon) \leq \exp(-\frac{n\epsilon^2}{2V+2b\epsilon}) = \delta \implies \epsilon = \frac{1}{n}(\sqrt{b^2\ln^2\delta - 2nV\ln\delta} - b\ln\delta)$，所以 $\Pr(\bar{X_n} \geq \mu + \frac{\sqrt{b^2\ln^2\delta - 2nV\ln\delta} - b\ln\delta}{n}) \leq \delta$。

(3) 令 $Y = X - \mu \in [\alpha - \mu, \beta - \mu]$，其中 $\mu = E(X)$，记 $a = \alpha - \mu$，$b = \beta - \mu$，则 $Y \in [a, b]$。令 $Z = \frac{Y-a}{b-a}$，则 $Y = (b-a)Z + a = Zb + (1-Z)a$。因为 $Z = \frac{Y-a}{b-a} \in [0, 1]$，由 *Jensen* 不等式有 $e^{\lambda Y} = e^{Z\lambda b + (1-Z)\lambda a} \leq Ze^{\lambda b} + (1-Z)e^{\lambda a}$。对不等式两边求取一下期望：

$$\begin{aligned}
\mathbb{E}(e^{\lambda Y}) &\leq e^{\lambda b}\mathbb{E}(Z) + e^{\lambda a}(1 - \mathbb{E}(Z)) \\
&= e^{\lambda b}\frac{-a}{b-a} + e^{\lambda a}\frac{b}{b-a} \\
&= e^{\lambda a}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right) \\
&= e^{\frac{a}{b-a}\lambda(b-a)}\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right)
\end{aligned}$$

对两边求一下对数，则有：

$$\ln\mathbb{E}(e^{\lambda Y}) \leq \frac{a}{b-a}\lambda(b-a) + \ln\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right)$$

令 $x = \lambda(b-a), t = \frac{a}{b-a}$，则 $\frac{a}{b-a}\lambda(b-a) + \ln\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{\lambda(b-a)}\right) = tx + \ln(1 + t - te^x)$，令 $h(x) = tx + \ln(1+t-te^x)$，则有 $h'(x) = t + \frac{-te^x}{1+t-te^x}$，进一步有 $h''(x) = \frac{-te^x(1+t-te^x) - (-te^x)^2}{(1+t-te^x)^2} = -\left(\frac{te^x}{1+t-te^x}\right) - \left(\frac{te^x}{1+t-te^x}\right)^2 = -\left(\frac{te^x}{1+t-te^x} + \frac{1}{2}\right)^2 + \frac{1}{4} \leq \frac{1}{4}$，由泰勒展开定理知存在 $\xi \in [0, x]$ 使得：

$$\begin{aligned}
h(x) &= h(0) + xh'(0) + \frac{x^2}{2}h''(\xi) \\
&= 0 + x \cdot 0 + \frac{x^2}{2}h''(\xi) \\
&\leq \frac{x^2}{2} \cdot \frac{1}{4} = \frac{\lambda^2(b-a)^2}{8}
\end{aligned}$$

代回则有：

$$\ln\mathbb{E}(e^{\lambda Y}) = \ln\mathbb{E}(e^{\lambda(X-\mu)}) \leq h(x) \leq \frac{\lambda^2(b-a)^2}{8} = \frac{\lambda^2(\beta-\alpha)^2}{8}$$

$\square$

2 [30pts] McDiarmid Inequality

请利用 Chernoff 引理证明如下的 McDiarmid 不等式:

定理 1 (McDiarmid Inequality). $X_1, X_2, \ldots, X_m$ 为定义在空间 $\mathcal{X}$ 上的 m 个独立的随机变量. 假设 $f\colon \mathcal{X}^m \mapsto \mathbb{R}$ 是一个关于 $X_1, \ldots, X_m$ 的实值函数, 并满足对任意的 $x_1, x_2, \ldots, x_m, x_i' \in \mathcal{X}$, 都有

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i$$

成立, 其中 $c_i > 0$. 那么对任意 $\epsilon > 0$, 有

$$\Pr[f(X_1, \ldots, X_m) - \mathbb{E}[f(X_1, \ldots, X_m)] \geq \epsilon] \leq e^{-2\epsilon^2 / \sum_{i=1}^m c_i^2}.$$

请写出详细的证明过程。

Proof. *slides* 中的证明跳了一些步骤我没能补齐，这里的证明方法参考自一篇 *stack exchange* 上的回答 *[1]*。

记 $f(X_{1:m}) = f(X_1, \ldots, X_m)$。令 $V_i = \mathbb{E}[f(X_{1:m})|X_1, \ldots, X_i] - \mathbb{E}[f(X_{1:m})|X_1, \ldots, X_{i-1}], 1 < i \leq m$，对于边界情况有 $V_1 = \mathbb{E}[f(X_{1:m})|X_1] - \mathbb{E}[f(X_{1:m})]$，易见 $f(X_{1:m}) - \mathbb{E}[f(X_{1:m})] = \sum_{i=1}^m V_i$。由马尔科夫不等式有:

$$\begin{aligned}
\Pr[f(X_{1:m}) - \mathbb{E}[f(X_{1:m})] \geq \epsilon] &= \Pr[\lambda\left(f(X_{1:m}) - \mathbb{E}[f(X_{1:m})]\right) \geq \lambda\epsilon] \\
&= \Pr[\exp\left\{\lambda\left(f(X_{1:m}) - \mathbb{E}[f(X_{1:m})]\right)\right\} \geq \exp(\lambda\epsilon)] \\
&= \Pr\left[\exp\left\{\lambda \sum_{i=1}^m V_i\right\} \geq \exp(\lambda\epsilon)\right] \\
&\leq e^{-\lambda\epsilon}\mathbb{E}\left[\exp\left\{\lambda \sum_{i=1}^m V_i\right\}\right]
\end{aligned}$$

先给 $\mathbb{E}\left[e^{\lambda V_i}\right]$ 一个上界，令

$$\begin{aligned}
H_i &= \sup V_i &= \sup_u(\mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1}, u)] - \mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1})]) \\
L_i &= \inf V_i &= \inf_v(\mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1}, v)] - \mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1})])
\end{aligned}$$

对两者逐差有:

$$\begin{aligned}
H_i - L_i &= \sup_u \mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1}, u)] - \inf_v \mathbb{E}[f(X_{1:m}|X_1, \ldots, X_{i-1}, v)]) \\
&= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \sup_u f(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_m) f_{X_{i+1}}(x_{i+1}) \ldots f_{X_m}(x_m) dx_{i+1} \ldots dx_m \\
&\quad - \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \inf_v f(x_1, \ldots, x_{i-1}, v, x_{i+1}, \ldots, x_m) f_{X_{i+1}}(x_{i+1}) \ldots f_{X_m}(x_m) dx_{i+1} \ldots dx_m \\
&= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \sup_{u,v}(f(x_1, \ldots, x_{i-1}, u, x_{i+1}, \ldots, x_m)
\end{aligned}$$

$$
\begin{aligned}
& -f(x_1,\ldots,x_{i-1},v,x_{i+1},\ldots,x_m))f_{X_{i+1}}(x_{i+1})\ldots f_{X_m}(x_m)dx_{i+1}\ldots dx_m \\
\leq\ & \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} c_i f_{X_{i+1}}(x_{i+1})\ldots f_{X_m}(x_m)dx_{i+1}\ldots dx_m \\
=\ & c_i
\end{aligned}
$$

又由 *chernoff* 引理，有 $\ln\mathbb{E}[e^{\lambda V_i}] \leq \frac{\lambda^2(H_i-L_i)^2}{8} \leq \frac{\lambda^2 c_i^2}{8}$，于是：

$$
\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{m}V_i\right\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{m}V_i\right\}|X_1,\ldots,X_{m-1}\right]\right] \\
&= \mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{m-1}V_i\right\}\mathbb{E}\left[e^{\lambda V_m}|X_1,\ldots,X_{m-1}\right]\right] \\
&\leq \mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{m-1}V_i\right\}\exp[\frac{\lambda^2 c_m^2}{8}]\right] \\
&\leq \cdots \leq \exp(\frac{\lambda^2}{8}\sum_{i=1}^{m}c_i^2)
\end{aligned}
$$

令 $\lambda = \frac{4\epsilon}{\sum_{i=1}^{m}c_i^2}$，则有：

$$
\begin{aligned}
\Pr[f(X_{1:m})-\mathbb{E}[f(X_{1:m})]\geq\epsilon] &\leq e^{-\lambda\epsilon}\mathbb{E}\left[\exp\left\{\lambda\sum_{i=1}^{m}V_i\right\}\right] \\
&\leq e^{-\lambda\epsilon}\exp(\frac{\lambda^2}{8}\sum_{i=1}^{m}c_i^2) \\
&= e^{-2\epsilon^2/\sum_{i=1}^{m}c_i^2}
\end{aligned}
$$

□

3 [20pts] Basic Inequalities

假设 $\xi_1, \xi_2, \ldots, \xi_n$ 是 n 个伯努利随机变量, 且满足 $\xi_i \sim B(1/i)$, 设 $T_n = \sum_{i=1}^n \xi_i$, 试证明

$$\mathbb{E}[T_n] = \sum_{i=1}^n \frac{1}{i} \geq \ln n.$$
$$\mathbb{V}[T_n] = \sum_{i=2}^n \frac{1}{i}\left(1 - \frac{1}{i}\right) \leq \ln n + 1.$$

Proof.
(1) $\mathbb{E}(T_n) = \mathbb{E}(\sum_{i=1} n\xi_i) = \sum_{i=2} n\mathbb{E}(\xi_i) = \sum_{i=1} n\frac{1}{i}$，而 $\ln n = \int_1^n \frac{1}{x}dx \leq \sum_{i=1}^n \frac{1}{i}$，所以有 $\mathbb{E}(T_n) = \sum_{i=1} n\frac{1}{i} \geq \ln n$。

(2) 令 $\mu = \mathbb{E}(T_n)$,

$$\begin{aligned}
\mathbb{V}(T_n) &= \mathbb{E}[(T_n - \mathbb{E}(T_n))^2] \\
&= \mathbb{E}[T_n^2 - 2\mu T_n + \mu^2] \\
&= \mathbb{E}(T_n^2) - 2\mu\mathbb{E}(T_n) + \mathbb{E}(\mu^2) \\
&= \mathbb{E}[\sum_{i=1}^n \sum_{j=1}^n \xi_i \xi_j] - \mu^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n [\mathbb{I}_{i=j}\frac{1}{i} + \mathbb{I}_{i\neq j}\frac{1}{ij}] - \mu^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n [\frac{1}{ij}] + \sum_{i=1}^n [\frac{1}{i} - \frac{1}{i^2}] - \mu^2 \\
&= \sum_{i=1}^n \frac{1}{i} \sum_{j=1}^n [\frac{1}{j}] + \sum_{i=1}^n [\frac{1}{i} - \frac{1}{i^2}] - \mu^2 \\
&= \sum_{i=1}^n \frac{1}{i}\mu + \sum_{i=1}^n [\frac{1}{i} - \frac{1}{i^2}] - \mu^2 \\
&= \mu^2 + \sum_{i=1}^n [\frac{1}{i} - \frac{1}{i^2}] - \mu^2 \\
&= \sum_{i=1}^n [\frac{1}{i} - \frac{1}{i^2}] = \sum_{i=1}^n \frac{1}{i}[1 - \frac{1}{i}] \\
&\leq \sum_{i=2}^n \frac{1}{i} < \int_1^n \frac{1}{x}dx = \ln n < \ln n + 1
\end{aligned}$$

□

4 [30pts] Rademacher Complexity based Margin Bounds

记 $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ 是一个大小为 m 的训练集, 其中每个样本 (x_i,y_i) 是独立同分布地从分布 $\mathcal{D}$ 中采样得到. 试证明, 对任意 $\delta>0$ 和 $\theta>0$, 任意分类器 $H\in\mathcal{C}(\mathcal{H})$ 至少以 $1-\delta$ 的概率满足:

$$\Pr_{(x,y)\sim\mathcal{D}}[yH(x)<0]\le\sum_{i=1}^{m}\mathbb{I}[y_iH(x_i)\le\theta]+\frac{2}{\theta}\widehat{\mathfrak{R}}_S(\mathcal{H})+3\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{4.1}$$

其中, 函数 $\mathbb{I}[\,\cdot\,]$, 仅当 $\cdot$ 为真时取值为 1, 否则为 0; $\mathcal{C}(\mathcal{H})$ 为函数空间 $\mathcal{H}$ 的凸包, 具体形式为

$$\mathcal{C}(\mathcal{H})=\left\{H=\sum_i\alpha_ih_i:\ h_i\in\mathcal{H},\alpha_i\ge0\text{ and }\sum_i\alpha_i=1\right\}.$$

Proof.

令 $\mathcal{H}_\mathcal{C}=\mathcal{C}(\mathcal{H})$, $\widetilde{\mathcal{H}}_\mathcal{C}=\{(x,y)\mapsto yh(x),h\in\mathcal{H}_\mathcal{C}\}$, 记 $\Phi_\rho(x)=\mathbb{I}_{x\le0}+(1-\frac{x}{\rho})\mathbb{I}_{0<x\le\rho}$, 同时记 $\widetilde{\mathcal{H}}_{\mathcal{C},\rho}=\Phi_\rho\circ\widetilde{\mathcal{H}}_\mathcal{C}$, 易见 $\widetilde{\mathcal{H}}_{\mathcal{C},\rho}$ 中函数均是 $\frac{1}{\rho}-lipshitz$ 的, 且 $\Phi_\rho(x)\in[0,1]$, 由 *slides* 泛化性一章结论可知:

$$\mathbb{E}[\Phi_\rho(yh(x))]\le\frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(yh(x))+2\widehat{\mathfrak{R}}(\widetilde{\mathcal{H}}_{\mathcal{C},\rho})+3\sqrt{\frac{\ln 1/\delta}{2m}}$$

对于要证的不等式左侧, $\Pr_{(x,y)\sim\mathcal{D}}[yH(x)<0]=\mathbb{E}[\mathbb{I}_{yH(x)<0}]\le\mathbb{E}[\Phi_\rho(yH(x))]$。

对于不等式右侧的第一项, $\frac{1}{m}\sum_{i=1}^m\Phi_\rho(yh(x))\le\frac{1}{m}\sum_{i=1}^m\mathbb{I}_{yh(x)\le\rho}\le\sum_{i=1}^m\mathbb{I}_{yh(x)\le\rho}$

对于不等式右侧的第二项,

$$\begin{aligned}
\widehat{\mathfrak{R}}(\widetilde{\mathcal{H}}_{\mathcal{C},\rho}) &= \widehat{\mathfrak{R}}(\Phi_\rho\circ\widetilde{\mathcal{H}}_\mathcal{C})\le\frac{1}{\rho}\widehat{\mathfrak{R}}(\widetilde{\mathcal{H}}_\mathcal{C})\\
&= \frac{1}{\rho}\mathbb{E}_\sigma[\sup_{h\in\widetilde{\mathcal{H}}_\mathcal{C}}\frac{1}{m}\sum_{i=1}^m\sigma_ih(x_i)y_i]\\
&= \frac{1}{\rho}\mathbb{E}_\sigma[\sup_{h\in\mathcal{H}_\mathcal{C}}\frac{1}{m}\sum_{i=1}^m\sigma_ih(x_i)]\\
&= \frac{1}{\rho}\mathbb{E}_\sigma[\sup_{h_1\in\mathcal{H},\ldots,h_d\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\sigma_i\sum_{j=1}^d\alpha_jh_j(x_i)]\\
&= \frac{1}{\rho}\mathbb{E}_\sigma[\sum_{j=1}^d\alpha_j\sup_{h_j\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\sigma_ih_j(x_i)]\\
&= \frac{1}{\rho}\sum_{j=1}^d\alpha_j\mathbb{E}_\sigma[\sup_{h_j\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^m\sigma_ih_j(x_i)]\\
&= \frac{1}{\rho}\sum_{j=1}^d\alpha_j\widehat{\mathfrak{R}}(\mathcal{H})\\
&= \frac{1}{\rho}\widehat{\mathfrak{R}}(\mathcal{H})
\end{aligned}$$

将上述三个结合，即有

$$
\begin{aligned}
\Pr_{(x,y)\sim\mathcal{D}}[yH(x)<0] &\leq \mathbb{E}[\Phi_\rho(yH(x))] \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\Phi_\rho(yh(x)) + 2\widehat{\mathfrak{R}}(\widetilde{\mathcal{H}}_{\mathcal{C},\rho}) + 3\sqrt{\frac{\ln 1/\delta}{2m}} \\
&\leq \sum_{i=1}^{m}\mathbb{I}_{yh(x)\leq\rho} + 2\frac{1}{\rho}\widehat{\mathfrak{R}}(\mathcal{H}) + 3\sqrt{\frac{\ln 1/\delta}{2m}}
\end{aligned}
$$

□

参考文献

[1] Anand (https://stats.stackexchange.com/users/2513/anand). Understanding proof of mcdiarmid39;s inequality. Cross Validated. URL:https://stats.stackexchange.com/q/21513 (version: 2017-04-13).