机器学习 课程笔记

酥雨
zusuyu@stu.pku.edu.cn

April 25, 2022

目录

1 Inequalities

定理 1.1 (Markov Inequality). 如果非负随机变量 X 期望存在, 则对于任意 $k>0$,

$$\mathbb{P}[X \geqslant k] \leqslant \frac{\mathbb{E}[X]}{k}$$

进一步地, 如果 r 阶矩 $\mathbb{E}[X^r]$ 存在, 则对于任意 $k>0$,

$$\mathbb{P}[X \geqslant k] \leqslant \min_{j \leqslant r} \frac{\mathbb{E}\left[X^j\right]}{j}$$

定理 1.2 (Chebyshev Inequality). 如果随机变量 X 方差存在, 则对于任意 $\varepsilon>0$,

$$\mathbb{P}[|X-\mathbb{E}[X]| \geqslant \varepsilon] \leqslant \frac{\operatorname{Var}[X]}{\varepsilon^2}$$

定义 1.1 (矩生成函数, Moment Generating Function, MGF). 如果随机变量 X 的任意 $n \in \mathbb{N}$ 阶矩存在, 则定义其矩生成函数为

$$\psi_X(t)=\mathbb{E}\left[\mathrm{e}^{tX}\right]=\sum_{i \geqslant 0} t^i \frac{\mathbb{E}\left[X^i\right]}{i!}$$

定理 1.3 (Chernoff Inequality).

$$\mathbb{P}[X \geqslant k] \leqslant \inf_{t>0} \mathrm{e}^{-tk} \psi_X(t)$$

定理 1.4. $X_1, X_2, \cdots, X_n \sim$ i.i.d. $\mathcal{B}(1,p)$, 对于任意 $\varepsilon>0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

其中 $D_B(p\|q)$ 是两个 Bernoulli distribution $P=(p,1-p), Q=(q,1-q)$ 之间的相对熵.

证明.

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] &= \mathbb{P}\left[\sum_{i=1}^n X_i \geqslant n(p+\varepsilon)\right] \\
&\leqslant \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} \mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^n X_i}\right] \\
&= \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} \prod_{i=1}^n \mathbb{E}\left[\mathrm{e}^{tX_i}\right] \\
&= \inf_{t>0} \mathrm{e}^{-tn(p+\varepsilon)} (p\mathrm{e}^t + 1 - p)^n \\
&= \inf_{t>0} \left(\frac{p\mathrm{e}^t + 1 - p}{\mathrm{e}^{t(p+\varepsilon)}}\right)^n
\end{aligned}$$

通过 “简单” 求导, 取 $t=\ln \frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}$ 时上式右边取最小值, 从而有

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] &\leqslant \left(\frac{\frac{(1-p)(p+\varepsilon)}{1-p-\varepsilon} + 1 - p}{\left(\frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}\right)^{p+\varepsilon}}\right)^n = \left(\frac{\frac{1-p}{1-p-\varepsilon}}{\left(\frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}\right)^{p+\varepsilon}}\right)^n \\
&= \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^n = \mathrm{e}^{-nD_B(p+\varepsilon\|p)}
\end{aligned}$$

□

定理 1.5. $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个期望相同的独立随机变量, $\mathbb{E}[X_i]=p$, 对于任意 $\varepsilon>0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

证明. 注意到指数函数是下凸的, 根据 Jensen Inequality, 有

$$\mathbb{E}\left[\mathrm{e}^{tX}\right] \leqslant \mathbb{E}\left[X\mathrm{e}^{t} + (1-X)\mathrm{e}^{0}\right] = p\mathrm{e}^{t} + 1 - p$$

从而

$$\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right] \leqslant (p\mathrm{e}^{t} + 1 - p)$$

沿用定理 1.4 的证明即可. □

定理 1.6 (Chernoff Bound). $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个独立随机变量, $\mathbb{E}[X_i] = p_i$, 记 $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-nD_B(p+\varepsilon\|p)}$$

证明. 注意到对数函数是上凸的, 从而函数 $f(x) = \ln(x\mathrm{e}^{t} + 1 - x)$ 也是上凸的, 同样根据 Jensen Inequality, 有

$$\frac{1}{n}\sum_{i=1}^{n} \ln(p_i\mathrm{e}^{t} + 1 - p_i) \leqslant \ln(p\mathrm{e}^{t} + 1 - p)$$

从而

$$\mathbb{E}\left[\mathrm{e}^{t\sum_{i=1}^{n} X_i}\right] \leqslant \prod_{i=1}^{n}(p_i\mathrm{e}^{t} + 1 - p_i) \leqslant (p\mathrm{e}^{t} + 1 - p)^n$$

□

定理 1.7 (Additive Chernoff Bound). $X_1, X_2, \cdots, X_n \in [0,1]$ 是 n 个独立随机变量, $\mathbb{E}[X_i] = p_i$, 记 $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}$$

证明. 只需要证明 $D_B(p+\varepsilon\|p) \geqslant 2\varepsilon^2$ 即可. 听说可以暴力求导. □

定理 1.8 (Hoeffding Bound). $X_1, X_2, \cdots, X_n$ 是 n 个独立随机变量, $X_i \in [a_i, b_i]$, 记 $p = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i+b_i}{2}$, 对于任意 $\varepsilon > 0$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - p \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-\frac{2n\varepsilon^2}{\left(\frac{1}{n}\sum_{i=1}^{n} b_i - a_i\right)^2}} \leqslant \mathrm{e}^{-\frac{2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

定理 1.9 (McDiarmid Inequality). $X_1, X_2, \cdots, X_n \in \mathcal{X}$ 是 n 个独立随机变量, 如果对于 $f: \mathcal{X}^n \to \mathbb{R}$ 存在常数 $c_1, c_2, \cdots, c_n$ 使得

$$|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leqslant c_i$$

对于任意 $i \in [n], x_1, \cdots, x_n, x_i'$ 成立, 则对于任意 $\varepsilon > 0$, 有

$$\mathbb{P}\left[f(x_1, \cdots, x_n) - \mathbb{E}\left[f(x_1, \cdots, x_n)\right] \geqslant \varepsilon\right] \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right)$$

2 VC Theory

对一个分类器 f, 通常有两种评价指标: training error $err_S(f) = \mathbb{P}_{(x,y)\in S}[y \neq f(x)]$ 与 generalization error $err_D(f) = \mathbb{P}_{(x,y)\sim D}[y \neq f(x)]$. 接下来可能会不加声明地用 S 表示从数据集 D 中 sample 出来的训练集.

称 $err_D(f) - err_S(f)$ 为分类器 f 的 generalization gap. 我们提出**一致收敛 (uniformly converge)** 的概念, 它表示随着训练集 S 的增大, hypothesis space $\mathcal{F}$ 中的所有分类器 f 的 generalization gap 都会 "一致" 地被 bound 住.

定理 2.1 (Uniform Convergence when $|\mathcal{F}| < \infty$). S 是从数据集 D 中随机采样的训练集, $|S| = n$, 有

$$\mathbb{P}\left[\forall f \in \mathcal{F}, err_D(f) - err_S(f) \geqslant \varepsilon\right] \leqslant |\mathcal{F}| \mathrm{e}^{-2n\varepsilon^2}$$

证明. 对于某个确定的 $f \in \mathcal{F}$, 注意到 $err_S(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i \neq f(x_i)]$, $\mathbb{E}\left[y_i \neq f(x_i)\right] = err_D(f)$, 故根据 Chernoff Bound 有 $\mathbb{P}\left[err_D(f) - err_S(f) \geqslant \varepsilon\right] \leqslant \mathrm{e}^{-2n\varepsilon^2}$. 再结合 Union Bound 即得结论. □

定理 2.2 (VC Theorem). 对于 VC-dimension (会在接下来定义) 为 d 的 hypothesis space $\mathcal{F}$, 从数据集 D 中随机采样大小为 n 的训练集 S, 则

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} |err_D(f) - err_S(f)| \geqslant \varepsilon\right] \leqslant 2\left(\frac{2\mathrm{e}n}{d}\right)^d \mathrm{e}^{-cn\varepsilon^2}$$

其中 c 是常数. 或者等价地, 有至少 $1-\delta$ 的概率,

$$err_D(f) \leqslant err_S(f) + O\left(\sqrt{\frac{d\ln n - \ln\delta}{n}}\right)$$

对所有 $f \in \mathcal{F}$ 成立.

3 Lagrange Duality

4 Boosting

5 PAC-Bayesian Theory

定理 5.1 (PAC-Bayesian Theorem). 对于给定的 prior distribution of classifiers $\mathcal{P}$, 从数据集 D 中随机抽取大小为 n 的训练集 S, 有至少 $1-\delta$ 的概率, 对于任意 distribution of classifiers $\mathcal{Q}$ 有如下不等式成立

$$\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)] \leqslant \mathbb{E}_{h\sim\mathcal{Q}}[err_S(h)] + \sqrt{\frac{D_{KL}(\mathcal{Q}\|\mathcal{P}) + \log(3/\delta)}{n}}$$

其中 $err_X(f)$ 表示 classifier f 在数据集 X 上的错误率, 即 $\mathbb{P}_{(x,y)\in X}[y \neq f(x)]$, $D_{KL}(\mathcal{Q}\|\mathcal{P}) = \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{P}_h}\right]$ 为概率分布 $\mathcal{Q}$ 与 $\mathcal{P}$ 的 KL 散度.

引理 5.1. 对于任意在 hypothesis space $\mathcal{F}$ 上的概率分布 $\mathcal{P}, \mathcal{Q}$, 以及任意函数 $f : \mathcal{F} \to \mathbb{R}$, 都有

$$\mathbb{E}_{h\sim Q}[f(h)] \leqslant \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + D_{KL}(\mathcal{Q}\|\mathcal{P})$$

证明.

$$\begin{aligned}
\text{RHS} - \text{LHS} &= \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + D_{KL}(\mathcal{Q}\|\mathcal{P}) - \mathbb{E}_{h\sim Q}[f(h)] \\
&= \ln \mathbb{E}_{h'\sim P}[\exp(f(h'))] + \mathbb{E}_{h\sim\mathcal{Q}}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{P}_h}\right] - \mathbb{E}_{h\sim Q}[f(h)] \\
&= \mathbb{E}_{h\sim Q}\left[\ln \frac{\mathcal{Q}_h}{\frac{\mathcal{P}_h \exp(f(h))}{\mathbb{E}_{h'\sim\mathcal{P}}[\exp(f(h'))]}}\right] \\
&= \mathbb{E}_{h\sim Q}\left[\ln \frac{\mathcal{Q}_h}{\mathcal{R}_h}\right] \\
&= D_{KL}(\mathcal{Q}\|\mathcal{R}) \\
&\geqslant 0
\end{aligned}$$

其中 $\mathcal{R}$ 也是一个 $\mathcal{F}$ 上的概率分布, $\mathcal{R}_h = \frac{\mathcal{P}_h \exp(f(h))}{\mathbb{E}_{h'\sim\mathcal{P}}[\exp(f(h'))]}$. □

引理 5.2. 对于任意 $\delta > 0$, 有

$$\mathbb{P}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[e^{n(err_D(h)-err_S(h))^2}] \geqslant 3/\delta\right) \leqslant \delta$$

证明. 先证明对于某个固定的 $h \sim \mathcal{P}$, 有

$$\mathbb{E}_{S\sim D^n}[e^{n(err_D(h)-err_S(h))^2}] \leqslant 3$$

记 $\Delta = |err_D(h) - err_S(h)|$, 根据 Chernoff bound, 有

$$\mathbb{P}_{S\sim D^n}(\Delta \geqslant \varepsilon) \leqslant 2\exp(-2n\varepsilon^2)$$

于是

$$\begin{aligned}
\mathbb{E}_{S\sim D^n}\left[e^{n\Delta^2}\right] &= \int_0^{+\infty} \mathbb{P}_{S\sim D^n}\left(e^{n\Delta^2} \geqslant t\right) \mathrm{d}t \\
&= \int_1^{+\infty} \mathbb{P}_{S\sim D^n}\left(\Delta \geqslant \sqrt{\frac{\ln t}{n}}\right) \mathrm{d}t + 1 \\
&\leqslant \int_1^{+\infty} 2e^{-2\ln t} \mathrm{d}t + 1 \\
&= 3
\end{aligned}$$

随后, 使用 Markov Inequality 得到

$$\mathbb{P}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[\mathrm{e}^{n\Delta^2}]\geqslant 3/\delta\right)\leqslant\frac{\mathbb{E}_{S\sim D^n}\left(\mathbb{E}_{h\sim\mathcal{P}}[\mathrm{e}^{n\Delta^2}]\right)}{3/\delta}=\frac{\mathbb{E}_{h\sim\mathcal{P}}\left(\mathbb{E}_{S\sim D^n}[\mathrm{e}^{n\Delta^2}]\right)}{3/\delta}\leqslant\frac{\mathbb{E}_{h\sim\mathcal{P}}(3)}{3/\delta}=\delta$$

□

我们利用上述两个引理证明定理 5.1. 有至少 $1-\delta$ 的概率,

$$\begin{aligned}\left(\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)-err_S(h)]\right)^2&\leqslant\mathbb{E}_{h\sim\mathcal{Q}}[\Delta^2]\\&=\frac{1}{n}\mathbb{E}_{h\sim\mathcal{Q}}[n\Delta^2]\\&\leqslant\frac{1}{n}\left(\ln\mathbb{E}_{h\sim P}\left[\mathrm{e}^{n\Delta^2}\right]+D_{KL}(\mathcal{Q}\|\mathcal{P})\right)\\&\leqslant\frac{1}{n}\left(\ln(3/\delta)+D_{KL}(\mathcal{Q}\|\mathcal{P})\right)\end{aligned}$$

其中第一行等号使用了 Cauchy Inequality, 第三行使用了引理 5.1 代入 $f(h)=n\Delta^2$, 第四行使用了引理 5.2, with probability at least $1-\delta$.

5.1 PAC-Bayesian Bound for SVM

命题 5.1. 对于任意的 distribution of classifiers $\mathcal{Q}$, 令 $g_{\mathcal{Q}}$ 为一个确定性二分类器, $g_{\mathcal{Q}}(x)=\mathrm{sgn}(\mathbb{E}_{h\sim\mathcal{Q}}h(x))$, 则

$$err_D(g_{\mathcal{Q}})\leqslant 2\mathbb{E}_{h\sim\mathcal{Q}}[err_D(h)]$$

证明. 如果 $g_{\mathcal{Q}}$ 在一个数据点 x 上出错, 则说明 $\mathcal{Q}$ 中至少一半的 classifier 都在 x 上出错. □

考虑两个 distribution of classifiers $\mathcal{P}=\mathcal{N}(\mathbf{0},I_d),\mathcal{Q}=\mathcal{N}(\mu\mathbf{w},I_d)$, 其中 $\|\mathbf{w}\|_2=1$, μ 是缩放系数. 此时 $g_{\mathcal{Q}}$ 就是传统理解下的 linear classifier $\mathbf{w}$ (这里不考虑常数 b).

根据定理 5.1 的结论, 我们有

$$err_D(g_{\mathcal{Q}})\leqslant 2\left[\mathbb{E}_{h\sim\mathcal{Q}}err_S(h)+\sqrt{\frac{D_{KL}(\mathcal{Q}\|\mathcal{P})+\log(3/\delta)}{n}}\right]$$

$$\begin{aligned}D_{KL}(\mathcal{Q}\|\mathcal{P})&=\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|\mathbf{x}-\mu\mathbf{w}\|^2\right]\frac{1}{2}\left(\|\mathbf{x}\|^2-\|\mathbf{x}-\mu\mathbf{w}\|^2\right)\mathrm{d}\mathbf{x}\\&=\int_{\lambda}\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}\|\lambda\mathbf{w}+\mathbf{y}-\mu\mathbf{w}\|^2\right]\frac{1}{2}\left(\|\lambda\mathbf{w}+\mathbf{y}\|^2-\|\lambda\mathbf{w}+\mathbf{y}-\mu\mathbf{w}\|^2\right)\mathrm{d}\lambda\mathrm{d}\mathbf{y}\\&=\int_{\lambda}\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{d/2}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2-\frac{1}{2}\|\mathbf{y}\|^2\right]\frac{1}{2}\left(\lambda^2+\|\mathbf{y}\|^2-(\lambda-\mu)^2-\|\mathbf{y}\|^2\right)\mathrm{d}\lambda\mathrm{d}\mathbf{y}\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right]\frac{1}{2}(2\lambda\mu-\mu^2)\mathrm{d}\lambda\left[\int_{\mathbf{y}\in\mathbb{R}^{d-1},\mathbf{y}\perp\mathbf{w}}\frac{1}{(2\pi)^{(d-1)/2}}\exp\left(-\frac{1}{2}\|\mathbf{y}\|^2\right)\mathrm{d}\mathbf{y}\right]\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right](\lambda\mu-\mu^2)\mathrm{d}\lambda+\frac{\mu^2}{2}\\&=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(\lambda-\mu)^2\right]\mu\mathrm{d}\frac{(\lambda-\mu)^2}{2}+\frac{\mu^2}{2}\\&=\frac{\mu^2}{2}\end{aligned}$$

6 Algorithmic Stability

定义 6.1 (一致稳定, Uniform Stability). $\mathcal{A}$ 是输入训练集 $S=(z_1,\cdots,z_n)$, 输出一个分类器 $\mathcal{A}(S)$ 的学习算法. 记 $S^i=(z_1,\cdots,z_{i-1},z_i',z_{i+1},\cdots,z_n)$ 是与 S 只相差第 i 个数据点的相邻训练集, $\ell(\cdot,\cdot)$ 是损失函数, 即 $\ell(f,z)$ 是在分类器 f 下, 数据点 z 产生的损失.

称学习算法 $\mathcal{A}$ 关于 $\ell(\cdot,\cdot)$ 满足 $\beta(n)$-一致稳定性, 如果对于任意大小为 n 的训练集 S 及其相邻训练集 S^i, 以及任意数据点 z, 都有

$$|\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S^i),z)|\leqslant\beta(n)$$

定义 6.2 (Risk & Empirical Risk). 分别类似于 test error 与 training error, 定义 risk 与 empirical risk 为

$$R(\mathcal{A}(S))=\mathbb{E}_{z\sim D}[\ell(\mathcal{A}(S),z)]$$

$$R_{emp}(\mathcal{A}(S))=\frac{1}{n}\sum_{i=1}^{n}\ell(\mathcal{A}(S),z_i)$$

以下讨论中不会出现超过一个学习算法, 故简记 $\Phi(S)=R(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S))$.

定理 6.1 (一致稳定能说明泛化). 对于一个关于 $\ell(\cdot,\cdot)$ 满足 $\beta(n)$-一致稳定性的学习算法 $\mathcal{A}$, 其中 $|\ell(\cdot,\cdot)|\leqslant M$ 有上界, 有

$$\mathbb{P}\left[\Phi(S)\leqslant\varepsilon+\beta(n)\right]\leqslant\exp\left(-\frac{n\varepsilon^2}{2(n\beta(n)+M)^2}\right)$$

或者等价的, 有至少 $1-\delta$ 的概率下式成立

$$R(\mathcal{A}(s))\leqslant R_{emp}(\mathcal{A}(s))+\beta(n)+(n\beta(n)+M)\sqrt{\frac{2\ln(1/\delta)}{n}}$$

证明. 先证明两个引理.

引理 6.1. 假设 $\mathcal{A}$ 是对称的, 即对于任意 n 元置换 σ, 有 $\mathcal{A}(\{z_1,\cdots,z_n\})=\mathcal{A}(\{z_{\sigma_1},\cdots,z_{\sigma_n}\})$, 则

$$\mathbb{E}_S[\Phi(S)]\leqslant\beta(n)$$

证明.

$$\mathbb{E}_S[\Phi(S)]=\mathbb{E}_{S,z}[\ell(\mathcal{A}(S),z)]-\mathbb{E}_S[\ell(\mathcal{A}(S),z_1)]=\mathbb{E}_{S,S^1}[\ell(\mathcal{A}(S^1),z_1)-\ell(\mathcal{A}(S),z_1)]\leqslant\beta(n)$$

□

引理 6.2. 如果 $|\ell(\cdot,\cdot)|\leqslant M$ 有上界, 则对于任意 S,S^i, 有

$$|\Phi(S)-\Phi(S^i)|\leqslant 2\left(\beta(n)+\frac{M}{n}\right)$$

证明. 除了 $\ell(\mathcal{A}(S),z_i)-\ell(\mathcal{A}(S^i),z_i')$ 一项外, 其余所有项都可以被 $\beta(n)$-稳定性限制住.

$$\begin{aligned}
|\Phi(S)-\Phi(S^i)|&=|R(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S))-R(\mathcal{A}(S^i))+R_{emp}(\mathcal{A}(S^i))|\\
&\leqslant|R_{emp}(\mathcal{A}(S))-R_{emp}(\mathcal{A}(S^i))|+|R(\mathcal{A}(S))-R(\mathcal{A}(S^i))|\\
&=\frac{1}{n}|\ell(\mathcal{A}(S),z_i)-\ell(\mathcal{A}(S^i),z_i')|+\frac{1}{n}\sum_{j\neq i}|\ell(\mathcal{A}(S),z_j)-\ell(\mathcal{A}(S^i),z_j)|+\left|\mathbb{E}_{z\sim D}[\ell(\mathcal{A}(S),z)-\ell(\mathcal{A}(S^i),z)]\right|\\
&\leqslant\frac{2M}{n}+\frac{n-1}{n}\beta(n)+\beta(n)\\
&\leqslant 2\left(\beta(n)+\frac{M}{n}\right)
\end{aligned}$$

□

考虑 McDiarmid's 不等式, 把 Φ 视作一个关于 $z_1, \cdots, z_n$ 的多元函数, 则引理 6.1 与引理 6.2 分别给出了 Φ 的期望以及在相邻输入上的差的上界. 于是

$$\mathbb{P}\left[\Phi(S) \geqslant \beta(n) + \varepsilon\right] \leqslant \mathbb{P}\left[\Phi(S) - \mathbb{E}\left[\Phi(S)\right] \geqslant \varepsilon\right] \leqslant \exp\left(-\frac{2n\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp\left(-\frac{n\varepsilon^2}{2(n\beta(n) + M)^2}\right)$$

□

7 Unsupervised Learning

前面讨论的都是监督学习. 现在我们讨论一下无监督学习.
无监督学习其实主要在做两件事情: Clustering, 以及 Dimensionality Reduction.

7.1 Clustering

对于一组 $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$, 需要把这些数据点划分成 k 个 cluster $S_1, \cdots, S_k$.
可以如下定义一种划分的损失函数: 记 $\mu_i = \frac{1}{|S_i|} \sum_{j \in S_i} \mathbf{x}_j$ 为第 i 个 cluster 的中心, 损失函数为

$$L(\{S_1, \cdots, S_k\}) = \sum_{i=1}^{k} \sum_{j \in S_i} \|\mathbf{x}_j - \mu_i\|^2$$

7.1.1 K-means

7.1.2 K-means++

定理 7.1. K-means++ 算法给出的损失 L 与最优解 L_{opt} 满足

$$\mathbb{E}[L] \leqslant 8(\ln k + 2) L_{opt}$$

7.2 Dimensionality Reduction