Machine Learning Lecture 2

Group 2

1 第二讲 Concentration Inequality(续)

1.1 Chernoff Bound

在上一讲中，我们已经学习了 Chernoff Bound 的在两种条件下是成立的，这里两种条件分别是独立同分布伯努利随机变量与在 [0,1] 区间独立同分布的随机变量这两种情况，对于这两种情况具体的表述如下：

1.1.1 情况 1：独立同分布伯努利随机变量

定理 1 (Chernoff Bound - 情况 1)**.** 设 $X_1, X_2, ..., X_n$ 为独立同分布的伯努利随机变量，$\mathbb{E}[X_i] = p$，则对于任意 $\delta > 0$，有：

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geqslant \delta\right) \leq e^{-O(n)}$$

1.1.2 情况 2：独立同分布有界随机变量

定理 2 (Chernoff Bound - 情况 2)**.** 设 $X_1, X_2, ..., X_n$ 为独立同分布的随机变量，$X_i \in [0, 1]$，$\mathbb{E}[X_i] = p$，则：

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geqslant \delta\right) \leq e^{-O(n)}$$

对于这两种情况更详细的讨论以及证明可以参见上一讲的笔记。现在我们进一步放宽条件，考虑 $X_1, X_2, ..., X_n$ 为独立的随机变量，$X_i \in [0, 1]$，$\mathbb{E}[X_i] = p_i$ 的情况，这时对于情况二的证明中的

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = \left(\mathbb{E}[e^{tX}]\right)^n \leq \left(pe^t + (1-p)\right)^n$$

的第二个等式是不成立的，于是我们需要改进一下证明，值得欣喜的是通过一步放缩我们就可以解决这个问题，从而对这种情况得到同样的结论，表述如下：

1.1.3 情况 3：独立非同有界随机变量

定理 3 (Chernoff Bound - 情况 3)**.** 设 $X_1, X_2, ..., X_n$ 为独立的随机变量，$X_i \in [0,1]$，$\mathbb{E}[X_i] = p_i$，令 $p = \frac{1}{n}\sum_{i=1}^{n} p_i$，则：

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \geqslant \delta\right) \leq e^{-O(n)}$$

Proof. 正如前面所说的，虽然 $\prod_{i=1}^{n}\mathbb{E}[e^{tX_i}] = \left(\mathbb{E}[e^{tX}]\right)^n$ 不成立了，但如果我们还能保证

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

，那么我们还是可以解决这种情况，后面的证明与情况二的证明一致。

注意情况二中的证明中我们用 Jensen 不等式证明了 $\mathbb{E}[e^{tX}] \leq pe^t + (1-p)$

于是实际上

$$\prod_{i=1}^{n}\mathbb{E}[e^{tX_i}] \leq \prod_{i=1}^{n}(p_i e^t + (1-p_i))$$

又注意到，利用 Jensen 不等式和对数函数的凹性，可得：

$$\frac{1}{n}\sum_{i=1}^{n}\ln(p_i e^t + (1-p_i)) \leq \ln\left(\sum_{i=1}^{n}\frac{1}{n}(p_i e^t + (1-p_i))\right) = \ln(pe^t + (1-p))$$

即

$$\prod_{i=1}^{n}(p_i e^t + (1-p_i)) \leq (pe^t + (1-p))^n$$

结合这个放缩与前面的不等式，我们就证明

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

，从而对于这个情况我们就可以得到定理 3 描述的结果。 □

1.2 Chernoff Bound 的两种形式

注意到：

$$e^{-nD_B(p+\delta||p)} \leq e^{-2n\delta^2} \iff D_B(p+\delta||p) = (p+\delta)\ln\frac{p+\delta}{p} + (1-p-\delta)\ln\frac{1-p-\delta}{1-p} \geq 2\delta^2$$

而对于右边这个不等式是 Pinsker 不等式的一个直接结论，不过通过初等的求导与不等式也是可以证明在 $p, p+\delta$ 都属于 (0,1) 时是成立的，由于步骤较为繁琐，证明放在最后仅供参考。

这样我们就得到下面关于 Chernoff Bound 的两个不同的表述:

- **additive Chernoff Bound**

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \delta\right) \leq e^{-2n\delta^2}$$

- **Relative Entropy Chernoff Bound**

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \delta\right) \leq e^{-nD_B(p+\delta\|p)}$$

1.3 Hoeffding 不等式

定理 4 (Hoeffding 不等式)**.** 设随机变量 $X_1, X_2, ..., X_n$ 相互独立, $X_i \in [a_i, b_i]$, $-\infty < a_i < b_i < \infty$, $\mathbb{E}[X_i] = p_i$, $p = \frac{1}{n}\sum_{i=1}^{n} p_i$, 则:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \delta\right) \leq e^{\frac{-2n^2\delta^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

对于这个不等式的证明，框架基本上与定理 3 是一致的，但要稍作修改且需要一个更强的放缩，故略去。同时格外要强调的是对于这几种情况独立条件都是尤为重要，如果没有独立条件的话一般情况下没有这样的不等式。

1.4 无放回抽样的集中不等式

考虑从有限总体 $a_1, a_2, ..., a_N \in \{0, 1\}$ 中抽样，有两种策略:

- draw with replacement(有放回抽样)
- draw without replacement(无放回抽样)

对于第一种有放回的情况，显然它等价于独立同分布伯努利随机变量，也就是满足定理 1 的条件，自然是有 Chernoff Bound 的，关键是第二种情况，这里前面强调的独立条件丧失了，导致我们无法直接利用前面的结果。

记 $X_1,...,X_n$ 来自有放回抽样，即为独立同分布伯努利随机变量，$Y_1,...,Y_n$ 来自无放回抽样。由定理 1 的证明过程知利用独立性的关键步骤是证明

$$\mathbb{E}[e^{t(X_1+...+X_n)}] = (pe^t + (1-p))^n$$

，虽然因为独立性的丧失，我们不太可能对 $Y_1,...,Y_n$ 也得到这个等式，不过我们可以考虑证明 $\mathbb{E}[e^{t(Y_1+...+Y_n)}] \leq \mathbb{E}[e^{t(X_1+...+X_n)}]$，如果这个不等式成立则我们也可以得到一样的结论。

幸运地，这个不等式也是对的，而且证明也很直接，就是通过比较它们泰勒展开的各项系数:

$$\mathbb{E}[e^{t(X_1+...+X_n)}] = 1 + t\mathbb{E}\left[\sum_{i=1}^n X_i\right] + \frac{t^2}{2!}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] + \cdots$$

$$\mathbb{E}[e^{t(Y_1+...+Y_n)}] = 1 + t\sum_{i=1}^n \mathbb{E}[Y_i] + \frac{t^2}{2!}\left(\sum_{i=1}^n \mathbb{E}[Y_i^2] + 2\sum_{1\leq i<j\leq n} \mathbb{E}[Y_iY_j]\right) + \cdots$$

易知 $\mathbb{E}[X_i] = \mathbb{E}[Y_i]$，而且对于 $i = j$，有 $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] = \mathbb{E}[Y_iY_j]$。

当 $i \neq j$ 时, 不妨让 $i < j$，因为

$$\mathbb{E}(X_iX_j) = \mathbb{P}(X_i = 1,\ X_j = 1) = \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1)$$

$$\mathbb{E}(Y_iY_j) = \mathbb{P}(Y_i = 1,\ Y_j = 1) = \mathbb{P}(Y_i = 1)\mathbb{P}(Y_j = 1|Y_i = 1)$$

显然 $\mathbb{P}(X_i = 1) = \mathbb{P}(Y_i = 1)$，而因为 Y_i, Y_j 负相关，在 Y_i 抽出球下，Y_j 抽出球的几率是变小的，$\mathbb{P}(Y_j = 1|Y_i = 1) < \mathbb{P}(Y_j = 1)$，所以我们就可以得到 $\mathbb{E}(Y_iY_j) \leq \mathbb{E}(X_iX_j)$

对于更高阶项的系数我们也有类似的结论，若对于 $n = k$ 成立，我们可以证明 $n = k+1$ 成立:

$$\mathbb{E}(X_{i_1},...,X_{i_{k+1}}) = \mathbb{P}(X_{i_{k+1}} = 1|X_{i_1} = 1,...,X_{i_k} = 1) = \mathbb{P}(X_{i_{k+1}} = 1)\mathbb{P}(X_{i_1} = 1,...,X_{i_k} = 1)$$

$$\mathbb{E}(Y_{i_1},...,Y_{i_{k+1}}) = \mathbb{P}(Y_{i_{k+1}} = 1|Y_{i_1} = 1,...,Y_{i_k} = 1) < \mathbb{P}(Y_{i_{k+1}} = 1)\mathbb{P}(Y_{i_1} = 1,...,Y_{i_k} = 1)$$

$$\mathbb{P}(Y_{i_{k+1}} = 1)\mathbb{P}(Y_{i_1} = 1,...,Y_{i_k} = 1) < \mathbb{P}(X_{i_{k+1}} = 1)\mathbb{P}(X_{i_1} = 1,...,X_{i_k} = 1)$$

从而我们便可以证明 $Ee^{t(X_1+...+X_n)} \geq Ee^{t(Y_1+...+Y_n)}$, 也就可以得到下面的定理。

定理 5 (无放回抽样的集中不等式)**.** 设 $Y_1,...,Y_n$ 是从 $a_1,...,a_N \in \{0,1\}$ 中无放回抽取的样本，令 $p = \frac{1}{N}\sum_{j=1}^N a_j$，则:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Y_i - p \geq \delta\right) \leq e^{-2n\delta^2}$$

2 第三讲：VC Theory for Generalization

2.1 Occam 剃刀原理

简单的模型通常具有更好的泛化能力，而复杂的模型在训练数据上可能拟合得更好，但泛化能力较差。

2.2 监督学习框架

1. 训练数据：$(x_1, y_1), ..., (x_n, y_n)$，其中 $x_i \in \mathcal{X}$，$y_i \in \mathcal{Y}$，独立同分布于 D_{XY}
2. 训练过程：
   - 从 hypothesis space$\mathcal{F}$ 中选择 $\hat{f}$，使得在训练数据上的损失较小
   - Training error $= \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$
3. 推断过程：
   - Population error / Generalization error $= \mathbb{P}(Y \neq \hat{f}(X))$

2.3 Generalization Gap

Generalization Gap 定义为：

$$\mathbb{P}(Y \neq \hat{f}(X)) - \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i))$$

上式后半段表示训练误差；

令 $Z_i = I(y_i \neq \hat{f}(x_i))$，则泛化差距可写为：

$$\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i$$

注意，这里的 $Z_1, ..., Z_n$ 不是独立的，因此不能直接应用前面的 Chernoff 界。

2.4 有限假设空间的情况（简化设置）

假设假设空间 $\mathcal{F}$ 是有限的，即 $|\mathcal{F}| < \infty$。考虑最坏情况：

$$\mathbb{P}\left(\mathbb{P}(Y \neq \hat{f}(X)) - \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)) \geq \epsilon\right)$$
$$\leq \mathbb{P}\left(\exists f \in \mathcal{F}, \mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq f(x_i)) \geq \epsilon\right)$$

应用联合界（union bound）：

$$\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(\mathbb{P}(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq f(x_i)) \geq \epsilon\right)$$

对于每个固定的 $f \in \mathcal{F}$，$I(y_i \neq f(x_i))$ 是独立同分布的伯努利随机变量，因此可以应用 Chernoff Bound：

$$\leq |\mathcal{F}|e^{-2n\epsilon^2}$$

这一结果表明：

- 简单的模型（假设空间小）会有更好的泛化能力，但在训练数据上的拟合可能较差
- 复杂的模型（假设空间大）在训练数据上可能拟合得更好，但泛化能力较差

VC 理论则是在假设空间无限大时，对这种权衡关系进行定量分析的理论框架。

2.5 附注

我们旨在证明以下不等式成立：

$$D_B(p+\delta||p) = (p+\delta)\ln\frac{p+\delta}{p} + (1-p-\delta)\ln\frac{1-p-\delta}{1-p} \geq 2\delta^2$$

其中 $0 < p < 1$，且 δ 的取值需保证 $0 < p+\delta < 1$。

证明. 为了证明此不等式，我们构造一个辅助函数 $g(\delta)$：

$$g(\delta) = \left((p+\delta)\ln\frac{p+\delta}{p} + (1-p-\delta)\ln\frac{1-p-\delta}{1-p}\right) - 2\delta^2$$

我们的目标是证明对于所有有效的 δ，都有 $g(\delta) \geq 0$。

第一步：分析函数在 $\delta = 0$ 的取值

首先，计算当 $\delta = 0$ 时函数的值：

$$\begin{aligned} g(0) &= \left(p \ln \frac{p}{p} + (1-p) \ln \frac{1-p}{1-p} \right) - 2(0)^2 \\ &= (p \ln(1) + (1-p) \ln(1)) - 0 \\ &= 0 \end{aligned}$$

这表明在 $\delta = 0$ 时，不等式取等号。

第二步：计算一阶导数 $g'(\delta)$

接下来，我们对 $g(\delta)$ 求一阶导数。令 $f(\delta) = (p+\delta) \ln \frac{p+\delta}{p} + (1-p-\delta) \ln \frac{1-p-\delta}{1-p}$。通过求导法则，我们得到 $f'(\delta)$：

$$f'(\delta) = \ln \left(\frac{p+\delta}{p} \right) - \ln \left(\frac{1-p-\delta}{1-p} \right)$$

因此，$g(\delta)$ 的一阶导数为：

$$g'(\delta) = f'(\delta) - 4\delta = \ln \left(\frac{p+\delta}{p} \right) - \ln \left(\frac{1-p-\delta}{1-p} \right) - 4\delta$$

在 $\delta = 0$ 处，一阶导数为：

$$g'(0) = \ln(1) - \ln(1) - 0 = 0$$

这表明 $\delta = 0$ 是函数 $g(\delta)$ 的一个驻点（critical point）。

第三步：计算二阶导数 $g''(\delta)$

为了判断驻点的性质，我们继续求二阶导数 $g''(\delta)$：

$$\begin{aligned} g''(\delta) &= \frac{d}{d\delta} g'(\delta) \\ &= \frac{1}{p+\delta} - \left(\frac{-1}{1-p-\delta} \right) - 4 \\ &= \frac{1}{p+\delta} + \frac{1}{1-p-\delta} - 4 \end{aligned}$$

第四步：证明 $g''(\delta) \geq 0$

证明的关键在于证明 $g''(\delta) \geq 0$。这等价于证明：

$$\frac{1}{p+\delta} + \frac{1}{1-p-\delta} \geq 4$$

我们可以利用一个基本不等式：对于任意两个正数 a 和 b，有 $\frac{1}{a} + \frac{1}{b} \geq \frac{4}{a+b}$。

令 $a = p + \delta$ 和 $b = 1 - p - \delta$。由于 $0 < p + \delta < 1$，所以 $a > 0$ 且 $b > 0$。它们的和是：

$$a + b = (p + \delta) + (1 - p - \delta) = 1$$

因此，

$$\frac{1}{p+\delta} + \frac{1}{1-p-\delta} \geq \frac{4}{(p+\delta)+(1-p-\delta)} = \frac{4}{1} = 4$$

这就证明了 $f''(\delta) \geq 4$（其中 $f''(\delta)$ 是 $g''(\delta)$ 的第一部分），所以

$$g''(\delta) = \left(\frac{1}{p+\delta} + \frac{1}{1-p-\delta}\right) - 4 \geq 4 - 4 = 0$$

第五步：结论

我们已经证明了 $g''(\delta) \geq 0$ 恒成立，这意味着 $g(\delta)$ 是一个凸函数（convex function）。对于凸函数，一阶导数为零的点是其全局最小值点。由于 $g'(0) = 0$，所以 $\delta = 0$ 是 $g(\delta)$ 的全局最小值点。该全局最小值为 $g(0) = 0$。

因此，对于所有有效的 δ，我们有 $g(\delta) \geq g(0) = 0$。即：

$$(p+\delta)\ln\frac{p+\delta}{p} + (1-p-\delta)\ln\frac{1-p-\delta}{1-p} - 2\delta^2 \geq 0$$

移项后即得原不等式：

$$(p+\delta)\ln\frac{p+\delta}{p} + (1-p-\delta)\ln\frac{1-p-\delta}{1-p} \geq 2\delta^2$$

□