信息学中的概率统计：作业六

截止日期：2024 年 12 月 13 日（周五）下课前。**如无特殊情况，请不要提交电子版！**
注意：本次作业第六题为附加题，正确解决该题目可以得到额外 30% 的分数。

第一题

令 $X \sim \mathrm{Exp}(\lambda)$，$\lambda > 0$。本题中，我们将对 $a > 1$ 给出 $P(X \geq a/\lambda)$ 的上界。

(1) 使用马尔可夫不等式，给出 $P(X \geq a/\lambda)$ 的上界。

(2) 使用切比雪夫不等式，证明 $P(X \geq a/\lambda) \leq \frac{1}{(a-1)^2}$。

(3) 使用 Chernoff Bound，证明 $P(X \geq a/\lambda) \leq a \cdot e^{-a+1}$。

(4) 计算 $P(X \geq a/\lambda)$ 的准确值。

答案：

(1) 注意到 $E(X) = 1/\lambda$，因此 $P(X \geq a/\lambda) \leq 1/a$。

(2) 注意到 $\sigma(X) = 1/\lambda$，因此 $P(X \geq a/\lambda) = P(X - 1/\lambda \geq (a-1)/\lambda) \leq 1/(a-1)^2$。

(3) 当 $0 < t < \lambda$，

$$M_X(t) = E(e^{tX}) = \int_0^\infty \lambda \cdot e^{-\lambda x} \cdot e^{tx} dx = \frac{\lambda}{\lambda - t}。$$

由 Chernoff Bound，

$$P(X \geq a/\lambda) \leq M_X(t) \cdot e^{-at/\lambda}。$$

取 $t = \lambda - \lambda/a$，注意到 $0 < t < \lambda$，有 $P(X \geq a/\lambda) \leq a \cdot e^{-a+1}$。

(4) $P(X \geq a/\lambda) = e^{-a}$。

第二题

在课上，我们介绍了随机变量的收敛性。设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果对于任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

则称 $\{X_n\}$ 依概率收敛于 X，写作 $X_n \xrightarrow{P} X$。在本题中，我们将介绍随机变量的另一种收敛性。

设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果 $P(\lim_{n\to\infty} X_n \to X) = 1$，也即对于任意 $\epsilon > 0$，

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_n - X| \geq \epsilon\right) = 0,$$

则称称 $\{X_n\}$ 几乎必然收敛于 X，写作 $X_n \xrightarrow{a.s.} X$。

(1) 令 $\{X_n\}$ 为一列相互独立的随机变量，且 $X_n \sim B(n/1, 1)$。证明 $\{X_n\}$ 依概率收敛于 0，但 $\{X_n\}$ 不几乎必然收敛于 0。

(2) 令 $\{X_n\}$ 为一列独立同分布的随机变量，$X_n \sim B(1, p)$。令 $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_n$。证明 $Y_n \xrightarrow{a.s.} p$。

答案：

(1) 对于任意 $\epsilon > 0$，$P(|X_n - 0| > \epsilon) = 1/n \to 0$。令 $\epsilon = 1$，

$$P\left(\bigcup_{m=n}^{\infty} |X_m - X| \geq \epsilon\right) = 1 - \prod_{m=n}^{\infty}(1 - 1/m) \geq 1 - (1 - 1/n)^n \to 1 - 1/e。$$

因此，$P\left(\bigcup_{m=n}^{\infty} |X_m - X| \geq \epsilon\right) \to 0$ 不成立。

(2) 由 Chernoff bound，对于任意 $\epsilon > 0$，

$$P(|Y_m - p| \geq \epsilon) \leq 2e^{-2m\epsilon^2}。$$

因此，对于任意 $\epsilon > 0$，

$$P\left(\bigcup_{m=n}^{\infty} |X_m - X| \geq \epsilon\right) \leq \sum_{m=n}^{\infty} 2e^{-2m\epsilon^2} \to 0。$$

第三题

某个不使用随机性的计算机程序 A，为了输出正确结果，该程序需要调用另一计算机程序 B 进行 T 次调用，每次调用使用可能不同的输入，且每次调用使用的输入可能依赖于之前对程序 B 的调用返回的结果。程序 A 使用程序 B 在 T 次调用输出的结果以输出最终结果 θ。具体来说，假设对程序 B 进行 T 次调用的结果为 $\omega_1, \omega_2, \ldots, \omega_T$，在正确得到 $\omega_1, \omega_2, \ldots, \omega_T$ 的前提下，程序 A 总是能输出正确的结果 θ。

现有计算机程序 B'。在同样的输入下，程序 B' 以 2/3 的概率输出与程序 B 相同的结果，以 1/3 的概率输出不同的结果。现在，在没有程序 B，仅有程序 A 和程序 B' 的情况下，设计一个方案，以 $1 - \delta$ 的概率得到正确结果 θ。该方案对程序 A 和程序 B' 的调用次数应与 T 和 $\log(1/\delta)$ 为多项式关系。

答案： 当对程序 B 进行第 t 次调用时时，以同样的输入和独立的随机性对程序 B' 进行 $c \cdot \log(T/\delta)$ 次调用，令 ω'_t 为 $c \cdot \log(T/\delta)$ 次调用中程序 B' 输出频率最高的结果并将 ω'_t 返回给程序 A。令事件 E_t 表示 ω'_t 与程序 B 在同样的输入下的输出相同，根据课上的分析，$P(E_t \geq 1 - \delta/T)$。由 Union bound，

$$P\left(\bigcap_{t=1}^{T} E_t\right) \geq 1 - \delta。$$

而当 $E_1, E_2, \ldots, E_T$ 均发生时，有 $\omega_1 = \omega'_1, \omega_2 = \omega'_2, \ldots, \omega_T = \omega'_T$，也即程序 A 输出正确的结果。

第四题

在课上，我们用 Chernoff bound 证明了下述不等式：若 $X \sim B(n, p)$，则

$$P(X \geq E(X) + n\epsilon) \leq e^{-2n\epsilon^2},$$

$$P(X \leq E(X) - n\epsilon) \leq e^{-2n\epsilon^2}。$$

在本题中，我们将对二项分布证明另一版本的 Chernoff bound。

(1) 证明 $M_X(t) \leq e^{(e^t - 1) \cdot E(X)}$。提示：使用不等式 $1 + x \leq e^x$。

(2) 证明对于任意 $\epsilon > 0$,

$$P(X \geq (1+\epsilon)E(X)) \leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{E(X)};$$

对于任意 $0 < \epsilon < 1$,

$$P(X \leq (1+\epsilon)E(X)) \leq \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{E(X)}。$$

提示：参考作业二第六题。

(3) 利用（2）中的结论，重新证明作业二第二题（3）。也即，有 n 个球，每个球都等可能被放到 $m = n$ 个桶中的任一个。令 X_i 表示第 i 个桶中球的数量，$Y = \max\{X_1, X_2, \ldots, X_n\}$。证明 $P(Y \geq 4\log_2 n) \leq 1/n$。

答案：

(1)

$$M_X(t) = (1 - p + pe^t)^n = (1 + (e^t - 1)p)^n \leq e^{(e^t-1)\cdot np}。$$

(2) 注意到 $e^{(e^t-1)\cdot np}$ 为服从 $\lambda = np$ 的泊松分布的矩生成函数。使用和作业二第六题一样的计算过程（注意到 $\lambda = np = E(X)$），并令 $x = (1+\epsilon)E(X) = (1+\epsilon)np$,

$$P(X \geq (1+\epsilon)E(X)) \leq e^{\epsilon E(X)} \cdot (1+\epsilon)^{-(1+\epsilon)E(X)} = \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{E(X)}。$$

类似的，

$$P(X \leq (1-\epsilon)E(X)) \leq e^{-\epsilon E(X)} \cdot (1-\epsilon)^{-(1-\epsilon)E(X)} = \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{E(X)}。$$

(3) 注意到 $E(X_i) = 1$。使用（2）中的第一个不等式，并取 $\epsilon = 4\log_2 n - 1$，有

$$P(X_i \geq 4\log_2 n) \leq \frac{e^{4\log_2 n - 1}}{(4\log_2 n)^{4\log_2 n}}。$$

当 $n = 2$，由于 $e < 4$,

$$\frac{e^{4\log_2 n-1}}{(4\log_2 n)^{4\log_2 n}} \leq \frac{1}{4\log_2 n} = \frac{1}{4} = \frac{1}{n^2}。$$

当 $n \geq 3$，注意到 $4\log_2 n > 2e$，因此

$$\frac{e^{4\log_2 n-1}}{(4\log_2 n)^{4\log_2 n}} \leq 2^{-4\log_2 n} \leq \frac{1}{n^2}。$$

因此，总有 $P(X_i \geq 4\log_2 n) \leq 1/n^2$。使用 Union bound 即可。

第五题

在课上，我们证明了下述结论：对于任意向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, 令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$，则以至少 1/2 的概率，对于任意 $1 \leq i, j \leq n$,

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left\|\frac{1}{\sqrt{k}}A(x_i - x_j)\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

也即令 $F(x) = \frac{1}{\sqrt{k}}Ax$ 为一随机线性变换，则以至少 1/2 的概率，$F(x)$ 保持了每一对 x_i 和 x_j 之间的距离。

证明该结论的核心工具是下述引理：对于任意 $x \in \mathbb{R}^d$，

$$P\left((1-\epsilon)\|x\|^2 \le \left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \le (1+\epsilon)\|x\|^2\right) \ge 1 - 2e^{-k\epsilon^2/8}。 \tag{1}$$

为了证明原结论，对所有可能的 $x = x_i - x_j$ 使用上述结论，并使用 Union bound。

在本题中，我们将证明随机线性变换 $F(x) = \frac{1}{\sqrt{k}}Ax$ 不仅可以保持每一对 x_i 和 x_j 之间的距离，还可以保持每一对 x_i 和 x_j 之间的点积。在本题中，对于向量 $a, b \in \mathbb{R}^d$，$\langle a, b\rangle = a^\top b$ 为向量 a 与 b 的点积。

(1) 考虑向量 $y_1, y_2, \ldots, y_n \in \mathbb{R}^d$，对于全部 $1 \le i \le n$，满足 $\|y_i\| = 1$。令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$ 。证明以至少 1/2 的概率，下述事件成立：

- 对于任意 $1 \le i \le n$，$(1-\epsilon/4)\|y_i\|^2 \le \left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2 \le (1+\epsilon/4)\|y_i\|^2$；
- 对于任意 $1 \le i, j \le n$ 且 $i \ne j$，$(1-\epsilon/4)\|y_i + y_j\|^2 \le \left\|\frac{1}{\sqrt{k}}A(y_i+y_j)\right\|^2 \le (1+\epsilon/4)\|y_i+y_j\|^2$

(2) 在（1）中结论的基础上，证明以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，

$$\left|\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j \right\rangle - \langle y_i, y_j\rangle\right| \le \epsilon。$$

(3) 考虑向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$。注意 x_i 不一定满足 $\|x_i\| = 1$。证明以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，

$$\left|\left\langle \frac{1}{\sqrt{k}}Ax_i, \frac{1}{\sqrt{k}}Ax_j \right\rangle - \langle x_i, x_j\rangle\right| \le \epsilon\|x_i\|\|x_j\|。$$

答案：

(1) 对于 $y_1, y_2, \ldots, y_n$ 以及 $y_i + y_j$（$1 \le i, j \le n$ 且 $i \ne j$）使用(1)中的结论，并使用 Union bound。注意到共有 $n + n(n-1)/2 = n(n+1)/2$ 个向量，因此事件成立的概率至少为 $1 - 2e^{-k\epsilon^2/32} \cdot n(n+1)/2$。取 $k = O(\log n/\epsilon^2)$ 即可保证该概率至少为 1/2。

(2) 注意到

$$\langle y_i, y_j\rangle = \frac{\|y_i+y_j\|^2 - \|y_i\|^2 - \|y_j\|^2}{2},$$

$$\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j \right\rangle = \frac{\left\|\frac{1}{\sqrt{k}}Ay_i + \frac{1}{\sqrt{k}}Ay_j\right\|^2 - \left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2 - \left\|\frac{1}{\sqrt{k}}Ay_j\right\|^2}{2}。$$

因此

$$\begin{aligned}\left|\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j \right\rangle - \langle y_i, y_j\rangle\right| \le& \frac{\left|\left\|\frac{1}{\sqrt{k}}Ay_i + \frac{1}{\sqrt{k}}Ay_j\right\|^2 - \|y_i+y_j\|^2\right|}{2} + \frac{\left|\left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2 - \|y_i\|^2\right|}{2} + \frac{\left|\left\|\frac{1}{\sqrt{k}}Ay_j\right\|^2 - \|y_j\|^2\right|}{2} \\ \le& \frac{\epsilon/4 \cdot \|y_i+y_j\|^2 + \epsilon/4 \cdot \|y_i\|^2 + \epsilon/4 \cdot \|y_j\|^2}{2} \le \epsilon\end{aligned}$$

(3) 令

$$\hat{x}_i = \begin{cases} x_i/\|x_i\| & x_i \ne 0 \\ 0 & x_i = 0 \end{cases}。$$

对 $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n\}$ 中不为零向量的元素使用 (2) 中的结论，则以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，若 $\hat{x}_i \neq 0$ 且 $\hat{x}_j \neq 0$，

$$\left| \left\langle \frac{1}{\sqrt{k}} A\hat{x}_i, \frac{1}{\sqrt{k}} A\hat{x}_j \right\rangle - \langle \hat{x}_i, \hat{x}_j \rangle \right| \le \epsilon。$$

考虑 $1 \le i, j \le n$，若 $x_i \neq 0$ 且 $x_j \neq 0$，在上述不等式左右两侧同时乘 $\|x_i\|\|x_j\|$ 即可。若 $x_i = 0$ 或 $x_j = 0$，则有

$$\left\langle \frac{1}{\sqrt{k}} Ax_i, \frac{1}{\sqrt{k}} Ax_j \right\rangle = \langle x_i, x_j \rangle = 0。$$

第六题

在课上，我们证明了对于任意 $S_1, S_2, \ldots, S_m \subseteq \{1, 2, \ldots, n\}$，存在 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，使得对于任意 $1 \le i \le m$，

$$\mathrm{disc}_\chi(S_i) = \left| \sum_{j \in S_i} \chi(j) \right| \le O(\sqrt{n \log m})。$$

在本题中，我们将证明存在 $S_1, S_2, \ldots, S_n \subseteq \{1, 2, \ldots, n\}$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在 $1 \le i \le n$ 使得

$$\mathrm{disc}_\chi(S_i) = \left| \sum_{j \in S_i} \chi(j) \right| \ge \Omega(\sqrt{n}),$$

也即课上给出的上界 $O(\sqrt{n \log m})$ 几乎是最优的。

(1) 证明下述反集中不等式：$X \sim B(n, 1/2)$，存在常数 $c_1, c_2 > 0$，使得

$$P(X \ge n/2 + c_1 \cdot \sqrt{n}) \ge c_2。$$

提示：该不等式有多种证明方法。一种可能的思路是首先使用定量化的中心极限定理（课上提到的 Berry–Esseen 定理）建立二项分布与标准正态分布的联系，之后对标准正态分布证明反集中不等式。

(2) 令 S 为 $\{1, 2, \ldots, n\}$ 的子集，对于每个 $j \in \{1, 2, \ldots, n\}$，$P(j \in S) = 1/2$，且不同 j 是否被包含在 S 中相互独立。利用（1）中的结论，证明对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在常数 $c_3, c_4 > 0$，

$$P\left(\left| \sum_{j \in S} \chi(j) \right| \ge c_3 \sqrt{n} \right) \ge c_4。$$

(3) 证明存在 $S_1, S_2, \ldots, S_m \subseteq \{1, 2, \ldots, n\}$ 和常数 $c > 0$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在 $1 \le i \le m$ 使得

$$\left| \sum_{j \in S_i} \chi(j) \right| \ge c\sqrt{n}。$$

这里，$m \le C \cdot n$，C 是某个常数。提示：考虑使用概率证法，将 $S_1, S_2, \ldots, S_m$ 取为 $\{1, 2, \ldots, n\}$ 独立同分布的随机子集，并扩展（2）中的分析。

(4) 证明当 $m = n$ 时，（3）中的结论同样成立。

答案：

(1) 令 $Y \sim N(0,1)$，

$$P\left(Y \geq \frac{1}{4}\right) = \frac{1}{2}P\left(|Y| \geq \frac{1}{4}\right) = \frac{1}{2}\left(1 - P\left(|Y| \leq \frac{1}{4}\right)\right) = \frac{1}{2}\left(1 - \int_{-1/4}^{1/4} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx\right),$$

注意到 $\int_{-1/4}^{1/4} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \leq 1/2$，因此 $P\left(Y \geq \frac{1}{4}\right) \geq 1/4$。由 Berry–Esseen，存在常数 c，使得

$$\left|P((X - n/2)/\sqrt{n/4} \geq 1/4) - P(Y \geq 1/4)\right| = \left|P((X - n/2)/\sqrt{n/4} \leq 1/4) - P(Y \leq 1/4)\right| \leq c/\sqrt{n}。$$

当 $n \leq 100c^2$，有 $P(n \geq n/2 + \sqrt{n}/8)$ 为固定常数，因此结论成立。当 $n \geq 100c^2$，

$$\left|P((X - n/2)/\sqrt{n/4} \geq 1/4) - P(Y \geq 1/4)\right| \leq 1/10,$$

而 $P(Y \geq 1/4) \geq 1/4$，因此 $P((X - n/2)/\sqrt{n/4} \geq 1/4) \geq 1/4 - 1/10 \geq 1/10$，也即

$$P(X \geq n/2 + \sqrt{n}/8) \geq 1/10。$$

(2) 令 T 为 $\chi(j) = -1$ 的 $j \in \{1, 2, \ldots, n\}$ 的数量。注意到随机变量 $\sum_{j \in S} \chi(j) + T$ 服从 $B(n, 1/2)$。当 $T \leq n/2$，

$$P\left(\left|\sum_{j \in S} \chi(j)\right| \geq c_1\sqrt{n}\right) \geq P\left(\sum_{j \in S} \chi(j) + T \geq c_1\sqrt{n} + T\right) \geq c_2$$

当 $T \geq n/2$，令 $\chi'(j) = -j$，有

$$P\left(\left|\sum_{j \in S} \chi(j)\right| \geq c_1\sqrt{n}\right) = P\left(\left|\sum_{j \in S} \chi'(j)\right| \geq c_1\sqrt{n}\right) \geq P\left(\sum_{j \in S} \chi'(j) + (n - T) \geq c_1\sqrt{n} + (n - T)\right) \geq c_2,$$

而 $\sum_{j \in S} \chi'(j) + n - T$ 同样服从 $B(n, 1/2)$ 且 $n - T \leq n/2$，因此总有

$$P\left(\left|\sum_{j \in S} \chi(j)\right| \geq c_1\sqrt{n}\right) \geq c_2。$$

(3) 取 $m = n/c_2$。令 $S_1, S_2, \ldots, S_m$ 为 $\{1, 2, \ldots, n\}$ 的独立同分布的随机子集，且分布均与（2）中 S 的分布相同。令事件 $A_{\chi,i}$ 表示

$$\left|\sum_{j \in S_i} \chi(j)\right| < c_1\sqrt{n},$$

则对于固定的 χ，$A_{\chi,i}$ 相互独立，且 $P(A_{\chi,i}) \leq 1 - c_2$。对于固定的 χ，令事件 $B_\chi = \bigcap_{i=1}^m A_{\chi,i}$，有 $P(B_\chi) = (1 - c_2)^m \leq e^{-n}$。由 Union bound，注意到共有 2^n 个不同的 χ，有

$$P\left(\bigcup_\chi B_\chi\right) \leq 2^n \cdot e^{-n} < 1,$$

也即存在 $S_1, S_2, \ldots, S_m \subseteq \{1, 2, \ldots, n\}$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在 $1 \leq i \leq m$ 使得

$$\left|\sum_{j \in S_i} \chi(j)\right| \geq c_1\sqrt{n}。$$

(4) 令 $n' = n/C$ 并向下取整, 应用 (3) 中的结论, 得到 $m' \le Cn' \le n$ 个 $\{1, 2, \ldots, n'\}$ 的子集 $S_1, S_2, \ldots, S_{m'}$, 满足对于任意 $\chi : \{1, 2, \ldots, n'\} \to \{-1, +1\}$, 存在 $1 \le i \le m'$ 使得

$$\left|\sum_{j \in S_i} \chi(j)\right| \ge c_1\sqrt{n'} \ge \sqrt{n} \cdot c_1/\sqrt{C}。$$

将 $S_1, S_2, \ldots, S_{m'}$ 视为 $\{1, 2, \ldots, n\}$ 的子集即可。