دانشکده‌ی مهندسی کامپیوتر

تحلیل آماری در ابعاد بالا	زمستان ۱۳۹۸
تمرین سری دوم	
مدرس: ابوالفضل مطهری، محمدحسین رهبان	موعد تحویل: پایان روز ۲۲ فروردین

توجه: هفت سوال از سوالات زیر را به دلخواه انتخاب کرده و حل نمایید. حل سوال‌های بیشتر جنبه‌ی امتیازی خواهد داشت.

۱ دو کاربرد از نامساوی‌های مبتنی بر مارتینگل

۱.۱ عدد رنگی گراف‌های تصادفی

یک گراف تصادفی $G_{(n,p)}$ را یک گراف تصادفی با n رأس تعریف می‌کنیم که بین هر دو رأس آن، با احتمال p یک یال وجود دارد.
برای یک گراف داده‌شده‌ی G، عدد رنگی گراف برابر است با کمینه‌ی تعداد رنگ‌های لازم برای رنگ‌آمیزی رأس‌های G، به این ترتیب که هیچ دو رأس مجاور گراف، هم‌رنگ نباشند. این عدد را با $\chi(G)$ نمایش می‌دهیم. می‌دانیم که پیدا کردن عدد رنگی گراف، یک مسئله‌ی NP-Hard است. در این تمرین قصد داریم نشان دهیم که $\chi(G_{(n,p)})$ حول میانگین خود، تجمع دارد. برای این کار، مارتینگل Doob زیر را در نظر بگیرید:

$$Z_i \triangleq \mathbb{E}[\chi(G_{(n,p)})|G_1, \dots, G_i]$$

که در آن G_i زیرگراف با رأس‌های $\{1, \dots, i\}$ از گراف اصلی است. توجه کنید که $Z_0 = \mathbb{E}[\chi(G)]$.

۱. نشان دهید که $|Z_{i+1} - Z_i| < 1$ است.

۲. به کمک نامساوی Azuma-Hoeffding، نشان دهید که عدد رنگی این گراف با احتمال بالا در فاصله‌ی $O(\sqrt{n\log(n)})$ از امید ریاضی‌اش قرار دارد.

۲.۱ Balls and Beans!

فرض کنید n توپ را در n سبد پرتاب می‌کنیم و قصد داریم تعداد سبدهای خالی را مطالعه کنیم. متغیر X_i را شماره‌ی سبدی که توپ i در آن افتاده است بگیرید. همچنین Y را برابر تعداد سبدهای خالی تعریف کنید.

۱. با تعریف مارتینگل Doob مناسب، کرانی برای $\mathbb{P}(Y - \mathbb{E}[Y] \geq \epsilon n)$ ارائه دهید.

۲. امید ریاضی Y را بیابید.

۲ خواص ابتدایی پیچیدگی راداماخر

برای مجموعه‌ی توابع $\mathcal{F}$ و $\mathcal{G}$ خواص زیر را برای پیچیدگی راداماخر اثبات کنید:

۱. اگر $\mathcal{F} \subseteq \mathcal{G}$، آن‌گاه $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{G})$.

۲. برای یک $\alpha \in \mathbb{R}$ دلخواه، $\mathcal{R}_n(\alpha\mathcal{F}) = |\alpha|\mathcal{R}_n(\mathcal{F})$.

۳. $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n\Big(\mathrm{conv}(\mathcal{F})\Big)$، که در آن $\mathrm{conv}(\mathcal{F})$، پوش محدب (Convex Hull) است.

۴. نشان دهید $\mathcal{R}_n(\mathcal{F}+\mathcal{G}) \le \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$. همچنین نشان دهید که این کران در حالت کلی Tight است.

۵. به ازای یک تابع g کران‌دارِ داده‌شده، نشان دهید:

$$\mathcal{R}_n(\mathcal{F}+g) \le \mathcal{R}_n(\mathcal{F}) + \frac{||g||_\infty}{\sqrt{n}}$$

۳ لِم مازارت

فرض کنید که $A \subseteq \mathbb{R}^m$ یک مجموعه با کاردینال محدود باشد. همچنین تعریف کنید $r \triangleq \max_{\mathbf{x}\in A} ||\mathbf{x}||_2$. در این شرایط، نشان دهید:

$$\mathbb{E}_\sigma\Big[\frac{1}{m}\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\Big] \le \frac{r\sqrt{2\log(|A|)}}{m}$$

که در آن $\sigma = [\sigma_i]_{i=1}^m$ متغیرهای تصادفی σ_i، متغیرهای مستقل راداماخر هستند و $x_1, x_2, \dots, x_m$ مولفه‌های بردار $\mathbf{x}$ هستند.

راهنمایی: سعی کنید کرانی برای

$$\exp\Big(t\mathbb{E}_\sigma\Big[\frac{1}{m}\sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i\Big]\Big)$$

ارائه دهید و برای این کار از نامساوی Jensen و لم هوفدینگ استفاده کنید.

۴ فاصله‌ی همینگ

متغیرهای تصادفی مستقل از هم $X_1, X_2, ..., X_n$ را در نظر بگیرید که مقادیر خود را از مجموعه‌ی کراندار S اخذ می‌کنند. بردار $X_1^n = [X_1, X_2, ..., X_n]$ مقادیر خود را از S^n می‌گیرد. احتمال یک مجموعه $A \subset S^n$ به این صورت تعریف می‌شود:

$$P(A) \triangleq P\left[X_1^n \in A\right]$$

فاصله‌ی همینگ بین دو بردار عضو S^n برابر با تعداد درآیه‌های متفاوت این دو بردار تعریف می‌شود. فاصله همینگ بین بردار x_1^n و مجموعه A نیز بدین صورت تعریف می‌کنیم:

$$d(x_1^n, A) \triangleq \min_{y_1^n\in A}\ d(x_1^n, y_1^n)$$

الف) برای هر $t > 0$ نشان دهید:

$$P\left[d(x_1^n, A) \ge t + \sqrt{\frac{n}{2}\log\left(\frac{1}{P(A)}\right)}\right] \le e^{-2t^2/n}$$

ب) برای یک مجموعه A با احتمال 10^{-6}، احتمال آنکه فاصله‌ی همینگ یک نقطه تا آن، بیش از $10\sqrt{n}$ باشد را حساب کنید.

راهنمایی: از نابرابری Mc-Diarmid و این حقیقت که $d(x^n, B) \le 0$ اگر و تنها اگر $x^n \in B$ استفاده کنید.

۵ ماتریس تصادفی گوسی

فرض کنید درایه‌های ماتریس تصادفی $A \in \mathbb{R}^{m\times N}$ گاوسی استاندارد و مستقل از هم و $\vec{u}, \vec{v} \in \mathbb{R}^N$ دو بردار دلخواه داده‌شده باشند:
الف) با استفاده از صورت‌های معادل متغیرهای زیرگوسی ثابت کنید:

$$\begin{cases} \mathbb{P}\big[\vec{u}^\top A^\top A\vec{v} > a\big] \le & \dfrac{e^{-sa}}{(1-2s\vec{u}^\top\vec{v})^{\frac{m}{2}}} \qquad s>0 \\ \mathbb{P}\big[\vec{u}^\top A^\top A\vec{v} \le a\big] \le & \dfrac{e^{sa}}{(1+2s\vec{u}^\top\vec{v})^{\frac{m}{2}}} \qquad s>0 \end{cases}$$

راهنمایی: می‌توانید از رابطه‌ی $\det\{I - \vec{u}\vec{v}^\top\} = 1 - \vec{u}^\top\vec{v}$ استفاده نمایید.
ب) ثابت کنید:

$$\mathbb{P}\big[|\frac{1}{m}\vec{u}^\top A^\top A\vec{v} - \vec{u}^\top\vec{v}| > \epsilon\vec{u}^\top\vec{v}\big] \leq 2e^{-m(\frac{\epsilon^2}{4} - \frac{\epsilon^3}{6})}$$

۶ بی‌نام

متغیرهای تصادفی $X_1, \cdots, X_n$ را درنظر بگیرید که از هم مستقل هستند و روی $\mathbb{R}^d$ مقدار می‌گیرند و همچنین فرض‌کنید که برای همه‌ی متغیرهای فوق، توزیع $-X$ و X یکسان است، یا به عبارتی متقارن هستند. بگیرید:

$$\begin{cases} S_n \triangleq \sum_{i=1}^n X_i \\ \Sigma^2 \triangleq \mathbb{E}\max_{j=1,\cdots,d}\sum_{i=1}^n X_{i,j}^2 \end{cases}$$

تعریف کنید:

$$C(n,d) \triangleq \sup\frac{\mathbb{E}\|S_n\|_\infty^2}{\Sigma^2}$$

که سوپریمم روی تمام توزیع‌های مستقل و متقارن متغیر تصادفی‌های $X_1, \cdots, X_n$ با Σ^2 متناهی گرفته‌شده‌است. ثابت کنید $C(n,d)$ یک تابع غیر نزولی برحسب n, d است و

$$C(n,d) \leq 2(1 + \log(2d)).$$

همچنین تعریف کنید:

$$C(\infty,d) = \lim_{n\to\infty} C(n,d).$$

ثابت کنید برای $d \geq 2$،

$$C(\infty,d) \geq \left(\Phi^{-1}\left(1 - \frac{1}{2(d+1)}\right)\right)^2$$

و

$$\lim_{d\to\infty}\frac{C(\infty,d)}{2\log(d)} = 1$$

توجه: $\Phi(x) := \mathbb{P}(Z \leq x)$ که $Z \sim \mathcal{N}(0,1)$

۷ کلاس توابع خطی

کلاس توابع زیر را در نظر بگیرید:

$$\mathcal{F} \triangleq \{x \to \operatorname{sign}(\langle\theta,\ x\rangle)\ \big|\theta \in \mathbb{R}^d, ||\theta||_2 = 1\}$$

فرض کنید که $\{x_1.x_2, \ldots, x_n\}$ نمونه‌های ما در فضا هستند که در آن بردارهای $x_i \in \mathbb{R}^d$ مستقل خطی می‌باشند. همچنین $d \geq n$. نشان دهید که پیچیدگی راداماخر تجربی برابر خواهد بود با:

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = \mathbb{E}_\epsilon\big[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n\epsilon_i f(x_i)|\big] = 1$$

۸ اشتراک دو کلاس

فرض کنید که $\mathcal{H}_1,\ \mathcal{H}_2$ دو کلاس از توابع از $\mathcal{X}$ به $\{0,\ 1\}$ باشند. کلاس توابعِ $\mathcal{H} = \{h_1h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$ را در نظر بگیرید. نشان دهید که پیچیدگی راداماخر تجربی هر مجموعه از داده‌ها در کران‌های زیر صدق می‌کند:

$$\mathcal{R}(\mathcal{H}(x_1^n)/n) \leq \mathcal{R}(\mathcal{H}_1(x_1^n)/n) + \mathcal{R}(\mathcal{H}_2(x_1^n)/n)$$

راهنمایی: نامساوی Talagrand برای هر کلاس توابع $\mathcal{H}$ و هر تابع Φ که L-Lipschitz است بیان می‌کند که:

$$\mathbb{E}_\epsilon\Big[\sup_{h\in\mathcal{H}}|\frac{1}{n}\sum_{i=1}^n\epsilon_i\ \Phi\big(h(x_i)\big)|\Big] \leq \mathrm{L}\mathbb{E}_\epsilon\Big[\sup_{h\in\mathcal{H}}|\frac{1}{n}\sum_{i=1}^n\epsilon_i h(x_i)|\Big]$$

۹ نامساوی برای مقایسه‌ی دُم‌ها

لمِ Panchenko را ثابت کنید: فرض کنید که دو متغیر تصادفی حقیقی $X,\ Y$ داشته باشیم که به ازای هر a:

$$\mathbb{E}[(X-a)_+] \le \mathbb{E}[(Y-a)_+]$$

که در آن $x_+ \triangleq \max\{0, x\}$. از طرفی فرض کنید که برای Y به ازای یک $\kappa \ge 1$ و $b > 0$، برای هر $t \ge 0$ داشته‌باشیم که:

$$P\{Y \ge t\} \le \kappa e^{-bt}$$

در این صورت نشان دهید که:

$$P\{X \ge t\} \le \kappa e^{1-bt}$$

۱۰ گوی‌های واحد

کلاس توابع زیر را در نظر بگیرید:

$$\mathcal{F} \triangleq \{x \to \mathrm{sign}(\langle\theta,\ x\rangle) | \theta \in \Theta\}$$

فرض کنید که $\{x_1.x_2, \dots, x_n\}$ نمونه‌های ما در فضا باشند که در آن $x_i \in \mathbb{R}^d$، همچنین $\|x_i\|_\infty \le b$.

۱. فرض کنید $\Theta = \left\{\theta \in \mathbb{R}^d \middle| \|\theta\|_2 \le r\right\}$. در این صورت ثابت کنید:

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \le \frac{rb\sqrt{d}}{\sqrt{n}}$$

۲. فرض کنید $\Theta = \left\{\theta \in \mathbb{R}^d \middle| \|\theta\|_1 \le r\right\}$. در این صورت ثابت کنید:

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) \le \frac{rb\sqrt{2\log(2d)}}{\sqrt{n}}$$

۱۱ بزرگترین زیردنباله مشترک

فرض کنید $\mathbf{a} = a_1 \dots a_n$ و $\mathbf{b} = b_1 \dots b_n$ دو دنباله دودویی به طول n باشند که ارقام $\mathbf{a}, \mathbf{b}$ به صورت مستقل و یکنواخت از $\{0, 1\}$ انتخاب شده‌اند. متغیر تصادفی X_n را طول بزرگترین زیردنباله‌ی مشترک آنها بگیرید. ثابت کنید مقدار X_n حول میانگین آن متمرکز است، در واقع ثابت کنید:

$$P\Big(\ |X_n - \mathbb{E}\left[X_n\right]| \ge \delta\Big) \le 2\exp\left(-\delta^2/8n\right)$$

یادداشت: محاسبه‌ی میانگین X_n یک مسئله‌ی باز است، اما می‌دانیم:

$$0.788071 \le \lim_{n\to\infty} \mathbb{E}\frac{X_n}{n} = \gamma \le 0.826280$$

۱۲ مستقل در مدل اردوش

یک گراف G را با مجموعه رأس‌های V و مجموعه یال‌های E در نظر بگیرید. یک مجوعه $I \subset V$ از رأس‌ها مستقل نامیده می‌شود اگر هیچ یالی بین راس‌های آن وجود نداشته باشد. فرض کنید Z تعداد مجموعه‌های مستقل گراف باشد (مجموعه‌ی تهی را مستقل فرض می‌کنیم). فرض کنید: $G = G(n, nd)$ ($1 \le d \le \frac{n-1}{2}$) که منظور از $G(n, M)$ گرافی n رأسی است که یال‌های ان از بین زیرمجموعه‌های M عضوی از جفت‌راس‌ها به صورت یکنواخت انتخاب شده‌اند (این صورت دوم مدل Erdős–Rényi است که صورت اول آن‌را در سوال اول دیدیم). فرض کنید Z_n تعداد مجموعه‌های مستقل آن باشند. نامساوی زیر را ثابت کنید:

$$\mathbb{P}\left(\log Z_n \ge \mathbb{E}\left[\log Z_n\right] + t\right) \le 2\exp\left(-\frac{t^2}{Cn}\right)$$

که در آن C یک ثابت مستقل از n است.