**باسمه تعالی**

**دانشگاه صنعتی شریف**

**دانشکده مهندسی برق**

**نظریه اطلاعات، آمار و یادگیری**

**استاد: دکتر یاسائی**

**نیم‌سال اول ۱۴۰۱ـ۱۴۰۲**

# تمرین سری دوم

---

## ۱ برخی خواص انحراف $\chi^2$

فرض کنید یک خانواده ی پارامتری از توزیع ها به صورت $\{\mathcal{P}_\theta : \theta \in \Theta\}$ داریم و $\pi$ یک توزیع روی فضای $\Theta$ است. توزیع مخلوط $\mathcal{P}_\pi$ را به صورت زیر تعریف می کنیم:

$$\mathcal{P}_\pi = \int \mathcal{P}_\theta \pi(d\theta)$$

برای توزیع دلخواه $\mathcal{Q}$ تعریف کنید: $G(\theta, \tilde{\theta}) = \mathbb{E}_{\mathcal{Q}}[\frac{\mathcal{P}_\theta \mathcal{P}_{\tilde{\theta}}}{\mathcal{Q}^2}]$

**(۱)** ثابت کنید: $\chi^2(\mathcal{P}_\pi || \mathcal{Q}) = \mathbb{E}_{\theta, \tilde{\theta} \overset{i.i.d.}{\sim} \pi}[G(\theta, \tilde{\theta})] - 1$

**(۲)** ثابت کنید اگر $\chi^2(\mathcal{P}_n || \mathcal{Q}_n) = O(1)$ داریم: $\mathbb{P} \lhd \mathbb{Q}$ ( که در آن $\mathbb{P} = \{\mathcal{P}_n\}_{n=1}^\infty$ و $\mathbb{Q} = \{\mathcal{Q}_n\}_{n=1}^\infty$ )

## ۲ مسئله ی تشخیص در SBM

یک Planted Partition Model یک مدل برای تولید گراف تصادفی است. فرض کنید $\sigma \in \{-1, 1\}^n$ باشد. در این صورت گراف تصادفی به صورت زیر تولید می شود:

$$A_{ij} \sim \begin{cases} \mathcal{P} & \sigma_i = \sigma_j \\ \mathcal{Q} & \sigma_i \neq \sigma_j \end{cases}$$

که در آن $A = [A_{ij}]$ ماتریس مجاورت وزن دار گراف است. این توزیع را با $G(\sigma, \mathcal{P}, \mathcal{Q})$ نمایش می دهیم. در حالتی که $\mathcal{P} \sim Ber(p)$ و $\mathcal{Q} \sim Ber(q)$ به این مدل Stochastic Block Model گفته می شود و آنرا با $SBM(\sigma, p, q)$ نمایش می دهیم. حال مسئله آزمون فرض دوتایی زیر را درنظر بگیرید:

$$H_0 : \mathcal{G} \overset{i.i.d.}{\sim} \mathcal{P}_0 = G(n, \frac{\mathcal{P} + \mathcal{Q}}{2})$$
$$H_1 : \mathcal{G} \overset{i.i.d.}{\sim} \mathcal{P}_1 = G(\mathcal{P}, \mathcal{Q})$$

که در آن منظور از توزیع $G(\mathcal{P}, \mathcal{Q})$ این است که ابتدا بردار $\sigma$ با توزیع $\sigma_i \sim radmacher(i.i.d)$ (با احتمال برابر $\pm 1$ ) تولید می شود و سپس $\mathcal{G}$ از توزیع $G(\sigma, \mathcal{P}, \mathcal{Q})$ تولید می شود. منظور از $G(n, \frac{\mathcal{P}+\mathcal{Q}}{2})$ نیز این است که وزن همه ی یال ها از توزیع $\frac{\mathcal{P}+\mathcal{Q}}{2}$ بیاید. حال میخواهیم در چند گام قضیه زیر را ثابت کنیم:

**قضیه:** در حالت $SBM$ اگر $p = \frac{a}{n}, q = \frac{b}{n}$ باشد، و $\frac{(a-b)^2}{2(a+b)} < 1$ در اینصورت تشخیص بین دو فرض بالا غیر ممکن می شود، یعنی احتمال خطای تشخیص نمیتواند به صفر همگرا شود وقتی $n \to \infty$.

**(۱)** اگر $P_\sigma = G(\sigma, \mathcal{P}, \mathcal{Q})$ باشد، ثابت کنید: $G(\sigma, \hat{\sigma}) = \int \frac{P_\sigma P_{\hat{\sigma}}}{\mathcal{P}_0} \leq exp(\frac{\rho}{2}\langle \sigma, \hat{\sigma} \rangle^2)$ . ( که در آن $\rho = \int \frac{(\mathcal{P}-\mathcal{Q})^2}{2(\mathcal{P}+\mathcal{Q})}$).

**(۲)** ثابت کنید در حالت $SBM$ که $p = \frac{a}{n}, q = \frac{b}{n}$ داریم: $\rho = \frac{\tau + o(1)}{n}$ (که در آن $\tau = \frac{(a-b)^2}{2(a+b)}$ )

**(۳)** با استفاده از قضیه ی حد مرکزی حکم قضیه را ثابت کنید.(فرض کنید در اینجا همگرایی در توزیع همگرایی MGF را نتیجه می دهد، نیازی به اثبات این مورد نیست.)

## ۳ یافتن تطابق کامل پنهان شده!

مسئله ی یافتن تطابق کامل به این صورت تعریف می شود که ابتدا در یک گراف دو بخشی(که هر بخش شامل $n$ راس است) ، از بین تمام تطابق های کامل به طور یونیفرم یک تطابق انتخاب می شود( $M^*$ ). سپس گرافی که به ما نمایش داده می شود گراف $K_{n,n}$ است، که در آن وزن یال هایی که درون $M^*$ هستند از توزیع $\mathcal{P}$ و وزن دیگر یال ها از توزیع $\mathcal{Q}$ می آید. هدف ما تخمین زدن تطابق $M^*$ است. تابع هزینه را به شکل $\ell(\hat{M}, M^*) = \frac{1}{n}|\hat{M} \triangle M^*|$ تعریف می کنیم و هدف آن است که تخمینگری پیدا شود که برای آن داشته باشیم: $\mathbb{E}[\ell(\hat{M}, M^*)] \longrightarrow 0$ در این سوال می خواهیم قضیه زیر را ثابت کنیم:

**قضیه:** برای مدل با پارامتر های $(n, \mathcal{P}, \mathcal{Q})$ اگر داشته باشیم:

$$\sqrt{n}B(\mathcal{P}, \mathcal{Q}) \le 1 + \varepsilon$$

نتیجه می شود : $\mathbb{E}[\ell(\hat{M}_{ML}, M^*)] \longrightarrow 0$ ( که در آن $\hat{M}_{ML}$ تخمینگر Maximum Likelihood است ).

**(۱)** ثابت کنید: $\hat{M}_{ML} \in arg\max_{M \in \mathcal{M}} \sum_{e \in M} log \frac{\mathcal{P}}{\mathcal{Q}}(W_e)$

**(۲)** با فرض های قضیه بالا و با توجه به اینکه حداکثر $\binom{n}{k} k!$ تطابق کامل وجود دارد که با $M^*$ در $2k$ یال تفاوت دارد، ثابت کنید:

$$\forall \beta \ge 8\log(1+\varepsilon) \quad \mathbb{P}\{|M^* \triangle \hat{M}_{ML}| \ge \beta n\} \le e^{\frac{1}{2}} \frac{e^{\frac{-\beta^2 n}{4}}}{1 - e^{-\frac{\beta}{4}}}$$

راهنمایی: از سوال (۲) تمرین قبل استفاده کنید، همچنین برای باند کردن $\binom{n}{k} k!$ از نامساوی $(1-x) \le e^{-x}$ استفاده کنید.

**(۳)** با انتخاب $\beta = max\{8\log(1+\varepsilon), 2\sqrt{\frac{\log n}{n}}\}$ ثابت کنید:

$$\mathbb{E}[\ell(M^*, \hat{M}_{ML})] \le \beta + e^{\frac{1}{2}} \frac{\beta^2/4}{\beta/8} \le 5\beta$$

سپس حکم قضیه را نتیجه بگیرید.

**(۴)** (امتیازی) فرض کنید مدل به این صورت تغییر کند که بعد از انتخاب $M^*$ همه ی یال های آن با احتمال ۱ در گراف دوبخشی حاضر باشند، اما بقیه ی یال ها هر کدام با احتمال $\frac{d}{n}$ حضور داشته باشند و سپس مانند قبل وزن یال های درون $M^*$ از توزیع $\mathcal{P}$ و وزن یال های خارج از آن از توزیع $\mathcal{Q}$ انتخاب شود. ثابت کنید مشابه قضیه ی بالا با شرط زیر برقرار است.

$$\sqrt{d}B(\mathcal{P}, \mathcal{Q}) \le 1 + \varepsilon$$

## ۴ Chernoff-Rubin-Stein

فرض کنید $\mathcal{P}_\theta \overset{i.i.d.}{\sim} X_1, ..., X_n$ و $\theta \in [-a, a]$.

**(۱)** نابرابری زیر را ثابت کنید:

$$\inf_{\hat{\theta}} \sup_{\theta \in [-a,a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \ge \min_{0<\varepsilon<1} \max\{\varepsilon^2 a^2, \frac{(1-\varepsilon)^2}{n\bar{I}}\}$$

که در آن $\bar{I} = \frac{1}{2a}\int_{-a}^{a} I(\theta) d\theta$ میانگین اطلاعات فیشر است.

راهنمایی: از اثبات نامساوی کرامر_رائو بیزی که در لکچرنت درس آمده استفاده کنید.

**(۲)** باند بالا را ساده کنید و نشان دهید:

$$\sup_{\theta \in [-a,a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \ge \left(\frac{1}{a^{-1} + \sqrt{n\bar{I}}}\right)^2$$

## ۵ ساختار های مولد خصمانه و f- انحراف ها

یکی از مسائل مهم در حوزه علوم داده مدلسازی نحوه شکل گرفتن داده و ایجاد داده جدید به نحوی که داده های تولید شده تا حد خوبی شبیه داده های واقعی باشد. به اینگونه مدل ها مدل های مولد می گویند. یک روش معروف برای طراحی چنین مدل های مولدی، استفاده از روش های خصمانه است. روش های خصمانه به اینصورت است که ابتدا یک بردار $Z \sim N(0, I)$ وارد یک شبکه عصبی می شود و در خروجی شبکه داده های $X$ ظاهر می شود. که اصطلاحا به این شبکه مولد می گویند. حال داده تولید شده از شبکه مولد در کنار داده واقعی به دست آمده از توزیع $P$ قرار داده می شود. حال یک شبکه عصبی دیگر داده ها با دریافت یک داده $X$ تلاش می کند که تصمیم بگیرد که آیا داده ورودی از توزیع واقعی به دست آمده یا خروجی شبکه مولد بوده است. به این شبکه اصطلاحا شبکه متمایز کننده می گویند. حال دو شبکه به صورت خصمانه آموزش داده می شوند که همدیگر را شکست دهند. شبکه مولد تلاش می کند که شبکه متمایز کننده را فریب دهد و شبکه متمایز کننده تلاش می کند که داده واقعی را از داده های ساختگی به خوبی تشخیص دهد. نتیجه این بازی برای شبکه مولد آن است که تا جای ممکن داده های خروجی خود را شبیه به داده های واقعی کند یا اصطلاحا توزیع داده خروجی خود را شبیه به توزیع $P$ کند. حال می خواهیم نشان دهیم که چگونه در این ساختار های به صورت ذاتی توابع f- انحراف ها ظاهر می شوند.

**(۱)** ابتدا فرض کنید که یک داده $X$ را مشاهده می کنیم که این داده یا یک داده واقعی از توزیع $P$ است یا از توزیع پارامتری $Q_\theta$ حاصل شده است. یک روند کلی برای تصمیم گیری در این مورد پیدا کردن توزیع شرطی به صورت زیر است

$$P(Z|X), Z \in \{0, 1\}$$

حال نشان دهید که تلاش برای پیدا کردن بهترین $P(Z|X)$ به منظور رسیدن کمترین خطا (یا بیشترین دقت) در تعیین نوع داده، معادل به دست آوردن مقدار یک انجرافf- بین توزیع $P$ و $Q_\theta$ است. تابع f مذکور را به دست آورید.(توزیع prior روی فرض اول و دوم را یکنواخت در نظر بگیرید.)

**(۲)** حال متغییر تصادفی باینری $Z$ را در نظر بگیرید که $Z = 0$ به معنای $X \sim P$ یا $P(X|Z = 0) = P$ و $Z = 1$ به معنای $X \sim Q_\theta$ یا $P(X|Z = 1) = Q_\theta$ است. در این صورت یک معیار دیگر برای تعیین سیاست تصمیم گیری $P_\theta(Z|X)$ به صورت زیر است

$$argmin_\theta E_{x,z}[2log(P_\theta(Z|X))]$$

نشان دهید که مقدار بهینه تابع هدف فوق یک باند پایینی از یک انحراف-f است. تابع f را به دست آورید.

از این رو، اگر در ساختار های خصمانه، با پرتاب یک سکه سالم داده ها از توزیع واقعی یا خروجی شبکه مولد به شبکه متمایز کننده بدهیم و پارامتر های شبکه متمایز کننده را با تابع هدف فوق بهینه سازی کنیم. مقدار تابع هدف به مقدار یک انحراف-f نزدیک شود.

**(۳)** حال شبکه مولد اگر بخواهد که شبکه متمایز کننده را فریب دهد چگونه باید روی تابع هدف شبکه متمایز کننده که دارای فرم قسمت (۲) است، اثر بگذارد. سعی کنید رقابت بین این دو شبکه به صورت یک مسئله minmax بنویسید.

## ۶ اطلاعات آماری

در مسئله طبقه بندی در حالت کلی داده های ورودی به صورت زوج مرتب $(X, Y)$،$X \in \mathbb{R}^n$،$Y \in \{-1, 1\}$ هستند و تابع هزینه به صورت $L : \mathbb{R} \times y \implies \mathbb{R}$ است که در این صورت مسئله طبقه بندی به صورت مسئله بهینه سازی زیر حاصل می شود

$$argmin_f R(f), R(f) = E_{X,Y}(L(f(X), Y))$$

یک مشکل آن است که تابع هزینه $L$ لزوما محدب نیست. از این رو، مسئله به فرم زیر به دست می آورند

$$argmin_f R_\Phi(f), R_\Phi(f) = E_{X,Y}(\Phi(f(X), Y))$$

که تابع $\Phi$ تابع محدب و باند بالایی برای تابع $L$ است

**(۱)** اگر تابع $L(f(X), Y) = 1\{f(X)Y < 0\}$ باشد یک فرم کلی انتخاب برای $\Phi$ به صورت $\Phi(f(X), Y) = \phi(f(X)Y)$ تعریف می شود. نشان دهید

$$R_\Phi(f) = E_X(l_\phi(f(X), \eta(X)))$$

که در آن $\eta(x) = p(Y = 1|X = x)$ است. فرم تابع $l_\phi$ را به دست آورید.

**(۲)** حال اگر بخواهیم قبل از مشاهده داده، تابع ثابت $f(x) = \alpha$ به عنوان خروجی مسئله انتخاب کنیم. مقدار بهینه تابع هزینه به فرم زیر حاصل می شود

$$R^*_{prior,\Phi} = inf_\alpha(P(Y = 1)\phi(\alpha) + (P(Y = -1)\phi(-\alpha)))$$

نشان دهید که

$$R^*_{prior,\Phi} - R^*_\Phi = D_f(P_1||P_{-1})$$

که f در آن به صورت زیر تعریف می شود

$$f(t) = \sup_\alpha \left[ l^*_\phi(\pi) - \frac{\pi\phi(\alpha)t + (1-\pi)\phi(-\alpha)}{\pi t + (1-\pi)} \right](t\pi + 1 - \pi), \pi = P(Y = 1), l^*_\phi(\pi) = \inf_\alpha l_\phi(\alpha, \pi)$$

**(۴)** به عبارت $R^*_{prior,\Phi} - R^*_\Phi$ اطلاعات آماری می گویند. به نوعی عبارت فوق میزان اطلاعات را که $X$ در مورد $Y$ حمل می کند، نشان می دهد. برای اینکه این مفهوم مشخص تر شود نشان دهید که

$$R^*_{prior,\Phi} - R^*_\Phi = I(X;Y), \phi(\alpha) = log(1 + e^{-\alpha})$$

**(۵)**(امتیازی) حال اگر در ساختار خصمانه مطرح شده در سوال ۵ ، شبکه متمایز کننده را به صورت یک شبکه عصبی در نظر بگیریم که در لایه آخر تنها یک سلول عصبی دارد و تابع فعالی سازی آن به صورت $\phi(\alpha) = log(1 + e^{-\alpha})$ باشد. مسئله minmax به دست آورده در قسمت ۳ سوال ۵ را بر حسب توابع $R_\Phi$،$R_{prior,\Phi}$ باز نویسی کنید. در این حالت به صورت شهودی توضیح دهید که از لحاظ تئوری اطلاعاتی دو شبکه متمایز کننده و شبکه مولد چگونه علیه یکدیگر باز متخاصمانه را اجرا می کنند.

## ۷ دنباله گوسی و minimax

فرض کنید دنباله از داده های گوسی به فرم زیر داریم

$$X_i = \theta_i + Z_i, i \in [p], Z_i \sim N(0, 1)$$

حال می خواهیم مقدار $\theta_{max} = \max_{i\in[p]} \theta_i$ را تخمین بزنیم و مقدار تابع هزینه نیز به صورت $L(T, \theta) = (T - \theta)^2$ که $T : X^p \in \mathbb{R}^p \implies \mathbb{R}$ تابع تخمینگر است.

**(۱)** نشان دهید که عدد ثابت c وجود دارد که

$$\inf_T \sup_{\theta\in\mathbb{R}^p} E_\theta[(\theta_{max} - T)^2] \leq c \log(p)$$

**(۲)** تخمینگری را طراحی کنید که

$$\inf_T \sup_{\theta\in\mathbb{R}^p} E_\theta[(\theta_{max} - T)^2] \geq C \log(p)$$

دقت کنید که $C$ عدد ثابت محدود و مستقل از مقدار $p$ است.

## ۸ تخمین گر های نرم و سخت

**(۱)** نشان دهید که پارامتر $\hat{\theta}^{HT}$ یکی از جواب مسئله $l_0$ **squared least** است. ارتباط $\lambda$ و $\tau$ را به دست بیاورید.

$$\hat{\theta} = argmin_\theta||y - \theta||^2_2 + \lambda||\theta||_0$$

$$\hat{\theta}_i^{HT} = \begin{cases} y_i & |y_i| > \tau \\ 0 & |y_i| \leq \tau \end{cases}$$

**(۲)** نشان دهید که پارامتر $\hat{\theta}^{ST}$ یکی از جواب مسئله $l_1$ **squared least** است. ارتباط $\lambda$ و $\tau$ را به دست بیاورید.

$$\hat{\theta} = argmin_\theta ||y - \theta||_2^2 + \lambda ||\theta||_1$$

$$\hat{\theta}_i^{ST} = \begin{cases} y_i - \tau & y_i > \tau \\ 0 & |y_i| \leq \tau \\ y_i + \tau & y_i < -\tau \end{cases}$$

**(۳)** نشان دهید که پارامتر $\hat{\theta}^{HT}$ یکی از جواب مسئله زیر نیز هست.

$$\hat{\theta} = argmin_{\theta:||y-\theta||_\infty < \tau} ||\theta||_0$$

**(۴)** نشان دهید که پارامتر $\hat{\theta}^{ST}$ یکی از جواب مسئله زیر نیز هست.

$$\hat{\theta} = argmin_{\theta:||y-\theta||_\infty < \tau} ||\theta||_1$$

## ۹ ابعاد بالا

بردار تصادفی $p$ بعدی گاوسی را به شکل $X \sim N(\theta, I_p)$ در نظر بگیرید. که در آن پارامتر $\theta$ متعلق به مجموعه زیر است.

$$\Theta = \{\theta \in \mathbb{R}^p : |\theta_1| \leq p^{1/4}, ||\theta_{/1}||_2 \leq 2(1 - p^{-1/4}|\theta_1|)\}$$

که در آن $\theta_{/1} = (\theta_2, ..., \theta_p)$

برای $p$ به اندازه کافی بزرگ نشان دهید:

**(۱)** خطای minmax کران زیر را دارد.

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} E_\theta ||\hat{\theta} - \theta||_2^2 \lesssim 1$$

**(۲)** بدترین خطا برای MLE باندی از بالا ندارد

$$\theta_{MLE} = argmin_{\hat{\theta}} ||X - \theta||_2$$

$$\sup_{\theta \in \Theta} E_\theta [||\hat{\theta}_{MLE} - \theta||_2^2] \gtrsim \sqrt{p}$$