نمونه‌سوالات امتحان نهایی

آمار و احتمال مهندسی
دانشکده مهندسی کامپیوتر ـ بهمن‌ماه ۱۴۰۲

مدرس: امیر نجفی

توضیحات:

* بخش اول و دوم نمونه سوالات زیر گزیده‌ای از امتحان میان‌ترم دوم و امتحان نهایی در ترم بهار سال ۱۴۰۰ هستند.
* امتحان میان‌ترم دوم آن سال شامل ۶ سوال به صورت take home بوده و دانشجویان ۹ ساعت برای پاسخ به سوالات وقت داشتند. اما امتحان نهایی شامل ۶ سوال و ۱۶۰ دقیقه وقت داشته است. هر دو امتحان مجازی بوده‌اند.
* سطح سختی امتحان نهایی شما نسبت به سوالات میان‌ترم دوم آن سال (بخش ۱) **کمتر** است.
* سطح سختی امتحان نهایی شما کمابیش **مشابه** با امتحان پایانی آن سال (بخش ۲) خواهد بود. اما نحوه پوشش مطالب ممکن است متفاوت باشد.
* در بخش ۳ تعداد کمی نمونه سوال جدید نیز آورده شده است. بندهای این سوالات به صورت **ساده**، **معمولی**، **سخت** و **بسیار سخت** سطح‌بندی شده‌اند. سطح سختی سوالات امتحان نهایی شما نسبت به این مقیاس‌بندی **معمولی** خواهد بود.
* بخشی از هدف نگارش بخش ۳ آموزشی بوده است، و شامل مطالب و مثال‌هایی است که قصد داشتم در کلاس مطرح کنم اما فرصت نشد.

بخش اول
میان‌ترم دوم بهار ۱۴۰۰

سوال ۱ میان‌ترم دوم (۲ نمره):

فرض کنید n نمونه مستقل از توزیع نمایی با پارامتر یقینی ولی نامعلوم $\lambda > 0$ در دسترس باشند. این نمونه‌ها را $X_1, \ldots, X_n$ می‌نامیم.

الف) تخمین بیشینه درست‌نمایی از پارامتر λ را بدست آورید.

ب) تخمین‌گر بیشینه درست‌نمایی (MLE) برای میانگین این توزیع کدام است؟

سوال ۲ میان‌ترم دوم (۴ نمره):

یک توزیع پارامتریزه شده با PMF برابر با $P_X\left(x; \theta_1, \theta_2, \theta_3\right)$ را به شکل زیر تعریف می‌کنیم:

$$P_X\left(x; \theta_{1:3}\right) = \begin{cases} 1/3 & x = \theta_1 \\ 1/3 & x = \theta_2 \\ 1/3 & x = \theta_3 \\ 0 & \text{o.w.} \end{cases}$$

که بردار پارامترهای $\theta = \left(\theta_1, \theta_2, \theta_3\right)$ یقینی ولی نامعلوم است. همچنین حتماً داریم $i \neq j \Rightarrow \theta_i \neq \theta_j$. فرض کنید n نمونه i.i.d. از این توزیع به صورت $X_1, X_2, \ldots, X_n$ در دسترس باشند. می‌خواهیم میانگین این توزیع را تخمین بزنیم:

الف) فرض کنید از تخمین‌گر میانگین نمونه‌ای $\hat{\mu} = \hat{\mu}\left(X_{1:n}\right) = \bar{X}_n$ (sample mean یا population mean) برای این کار استفاده شود. آیا این تخمین‌گر بدون گرایش (unbiased) است؟

ب) میانگین مربعات خطا، یعنی $\mathbb{E}\left[\left(\bar{X}_n-\mu\right)^2\right]$ را برای تخمین‌گر قسمت الف) محاسبه کنید. در اینجا مقصود از μ میانگین واقعی است.

فرض کنید به جای تخمین‌گر میانگین نمونه‌ای، از یک تخمین‌گر جدید به نام $\hat{\mu}_{\mathrm{New}}=\hat{\mu}_{\mathrm{New}}\left(X_{1:n}\right)$ استفاده شود. نحوه محاسبه مقدار این تخمین‌گر به صورت زیر است: ابتدا تمام **مقادیر یکتای** موجود بین $X_1, X_2, \ldots, X_n$ را پیدا کرده و سپس فقط میان این مقادیر یکتا میانگین گرفته می‌شود.

ج) آیا این تخمین‌گر بدون گرایش است؟ در صورتیکه بدون گرایش است نشان دهید که مقدار bias صفر می‌شود. در غیر این صورت مثال نقض بیاورید.

د) مانند قسمت ب)، میانگین مربعات خطای این تخمین‌گر را برای n دلخواه محاسبه کنید. نرخ کاهش خطای آن بر حسب n را با تخمین‌گر میانگین نمونه‌ای مقایسه کنید.

ه) (**امتیازی**) تخمین بیشینه درست‌نمایی برای میانگین این توزیع (یعنی $\hat{\mu}_{\mathrm{ML}}=\hat{\mu}_{\mathrm{ML}}\left(X_1, \ldots, X_n\right)$) را بدست آورید.

سوال ۴ میان‌ترم دوم (۴ نمره):

متغیرهای تصادفی $X_1, \ldots, X_n$ به صورت .i.i.d از چگالی احتمال $f_X(x)$ (و متناظراً توزیع انباشته احتمال $F_X(x)$) حاصل شده‌اند. فرض کنید که از این متغیرهای تصادفی نمونه‌گیری شده و آنان را بر حسب مقدارشان به صورت صعودی مرتب کرده باشیم. مقدار **یکی مانده به بزرگترین** در این دنباله به دلیلی برای ما مهم است. توزیع آماری آن را را بدست آورید.

سوال ۶ میان‌ترم دوم (۳ نمره):

متغیرهای تصادفی X, Y با توزیع مشترک دانسته شده‌ای مفروض هستند. فرض کنید که قصد داریم متغیر تصادفی Y را به صورت یک ترکیب همگن درجه دو به صورت $\hat{Y}=aX^2+bX$ از روی X تخمین بزنیم، به طوری که میانگین مربعات خطا کمینه شود.
الف) ضرایب مجهول a, b را بر حسب گشتاورهای مجزا و یا مشترک X, Y (یعنی $\mathbb{E}\left(X^nY^m\right),\ \ n, m \geq 0$) بدست آورید.

ب) در صورتیکه بخواهیم تخمین ناهمگن درجه دوم به شکل $\hat{Y} = aX^2 + bX + c$ داشته باشیم، ضرایب مجهول را دوباره محاسبه کنید.

توضیحات و راهنمایی‌ها:

* سوالات امتیازی هرکدام ۱ نمره اضافه و مستقل از بارم‌بندی سوالات دارند.

* میانگین توزیع نمایی با پارامتر λ برابر با $\frac{1}{\lambda}$ است.

* تابع error function به صورت زیر تعریف می‌شود: $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t$. هر جاییکه لازم دیدید می‌توانید از این تابع استفاده کنید.

بخش دوم
سوالات پایان‌ترم بهار ۱۴۰۰

سوال ۲ امتحان نهایی (۴ نمره):

یک فرستنده مخابراتی قصد ارسال n بیت اطلاعات به یک گیرنده را دارد. داده‌ها از طریق یک کانال نویزی فرستاده شده و هر بیتی که توسط فرستنده ارسال می‌گردد با احتمال $p = 1/2$ و مستقل از ارسال‌های قبلی دچار خطا می‌شود. به منظور اطمینان از صحت ارسال‌ها، در سمت فرستنده از یک کدگذار (coder) بهره می‌بریم. بدین شکل که به جای ارسال هر کدام از n بیت فوق، تعداد L بیت که به طریقی خاص انتخاب شده‌اند، به نمایندگی از آن ارسال می‌گردند. لذا، درنهایت به جای n بیت اولیه، تعداد کل nL بیت ارسال خواهند شد.
این کار قرار است با چسباندن اطلاعات زائد (Redundancy) تعداد ارسال‌ها را افزایش داده، اما در عوض احتمال بروز خطا در آنان را کاهش دهد. خاصیت هر بلوک L بیتی که نماینده یکی از n بیت اولیه می‌باشد این است که تنها در صورتی بیت اصلی در سمت گیرنده دچار خطا خواهد شد که تمامی L بیتی که آن را نمایندگی می‌کنند در کانال خراب شوند. و حتی اگر صرفاً یکی از آنان سالم به مقصد برسد، بیت اصلی به درستی decode می‌گردد. در صورتیکه علاقه داشتید بدانید مشابه این کار چگونه امکان‌پذیر است، درس «تئوری اطلاعات» را در ترم‌های آینده بگیرید.
در این سوال قصد داریم تا احتمال بروز حداقل یک خطا در ارسال n بیت اصلی را بدست بیاوریم. در آخر نشان خواهیم داد که در صورت $O\left(n\log_2 n\right)$ بار استفاده از کانال به جای n بار، احتمال انتقال کاملاً صحیح کل داده‌ها برای $n \gg 1$ به سمت ۱ میل خواهد نمود.

الف) فرض کنید که اصلاً از کدگذاری استفاده نمی‌شد و در ارسال هر یک از n بیت اصلی، صرفاً همان بیت و به همان شکل اصلی خود ارسال می‌گشت (به عبارتی، داشتیم $L = 1$). احتمال اینکه از میان n بیت اصلی، حداقل یکی دچار خطا شود چقدر است؟
(به اختصار استدلال کنید که با افزایش n این احتمال به سمت یک میل خواهد کرد.)

ب) در صورت کدگذاری با طول بلوک‌های کد L، احتمال اینکه حداقل یکی از n بیت اصلی با خطا در گیرنده decode شود را حساب کنید.

ج) نشان دهید در صورتیکه طول بلوک‌های کد را به صورت $L = (1+\varepsilon)\log_2 n$ انتخاب کنیم (به ازای هر $\varepsilon > 0$)، احتمال بروز خطا با افزایش n به سمت صفر میل خواهد کرد.

سوال ۳ امتحان نهایی (۴ نمره):

یک سکه تصادفی با احتمال شیر یا خط نامعلوم را n بار به صورت مستقل پرتاب کرده و مشاهده می‌کنیم که k بار شیر ظاهر می‌شود.

الف) تخمین MLE از احتمال شیر آمدن سکه را محاسبه کنید.

حال با سازنده سکه ملاقات داشته و اطلاعاتی در مورد سکه کسب می‌کنیم. سازنده سکه می‌گوید که سکه‌های ساخت او به یک سمت بایاس دارند. وی معتقد است که احتمال شیر آمدن سکه‌هایش با احتمال α بین صفر تا ۱/۲ (یعنی در بازه $[0,0.5]$)، و با احتمال $1-\alpha$ بین ۱/۲ تا ۱ (یعنی بازه $[0.5,1]$) است. همچنین، در بازه صفر تا ۱/۲ میان مقادیر برتری نسبت به یکدیگر وجود ندارد. در بازه ۱/۲ تا ۱ نیز به همین شکل، برتری بین مقادیر نیست.
(بدون کاستن از کلیت مسئله، فرض کنید که $\alpha > 0.5$)

ب) با استفاده از این اطلاعات پیشین، تخمین MAP از احتمال شیر آمدن سکه را دوباره محاسبه کنید.

سوال ۴ امتحان نهایی (۴ نمره):

یک سازمان دولتی از شاخص عملکردی خود راضی نیست، و قصد دارد آن را بهبود ببخشد. شاخص عملکردی کل سازمان برابر با میانگین آماری شاخص عملکردی یکایک نیروهایش است، که می‌توان آنان را متغیرهای تصادفی هم‌توزیع ولی مستقل در نظر گرفت. سازمان یک بازرس انتخاب کرده تا نحوه انجام کار نیروهایش را بررسی کرده و شاخص عملکردی تعدادی از آنان را محاسبه کند.
بازرس تعداد ۱۰۰ نیرو را به صورت تصادفی انتخاب کرده و پس از بررسی سوابق کاری یکساله آنان شاخص‌های عملکردیشان را محاسبه نموده است. وی فرض کرده است که توزیع شاخص فوق برای هر یک از نیروها یک متغیر تصادفی گاوسی است. میانگین شاخص‌های عملکردی برای این جمعیت برابر با ۵۴٪ شده. همچنین بازرس انحراف معیار ۵٪ را کران بالایی برای انحراف معیار هر یک از ۱۰۰ اندازه‌گیری خود دانسته است.
سازمان به منظور بهبود عملکرد اخیراً تعدادی از مدیران اجرایی خود را تغییر داده. مدتی پس از این اتفاق، بازرس دوباره مشغول به کار شده و این بار ۱۰ نیرو را به صورت تصادفی انتخاب و میانگین شاخص عملکردی این جمعیت نمونه را محاسبه کرده است. عدد بدست آمده این بار به ۵۷٪ افزایش پیدا کرده، و بازرس به دلیل دقت پایین‌تر در اثر کمبود وقت، این دفعه انحراف معیار ۸٪ را کران بالایی برای انحراف معیار اندازه‌گیری‌هایش گزارش کرده. (دقت کنید که کران‌های بالای انحراف معیارها، یعنی ۵٪ و ۸٪، تخمین‌های شخصی خود بازرس هستند و مستقل از داده‌ها مقداردهی شده‌اند. همچنین وی کماکان فرض کرده که توزیع شاخص‌ها گاوسی است).

حال، سازمان شما را به عنوان یک آماردان استخدام می‌کند و از شما می‌پرسد که آیا افزایش ۳ درصدی در شاخص عملکردی پس از تغییر مدیران اجرایی به لحاظ آماری «معنی‌دار» هست یا خیر؟
الف) فرضیه‌های H_0 و H_1 را در این حالت تعریف کرده و یک آماره مناسب برای انجام کار پیشنهاد دهید. همچنین توضیح دهید که آزمون فرضی که طراحی کرده‌اید دقیق (exact) است یا نادقیق.

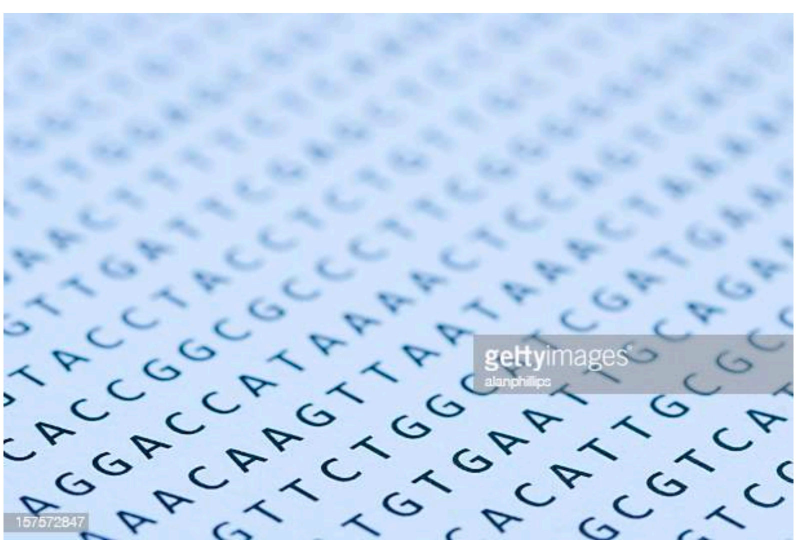
ب) «میزان معنی‌دار بودن» افزایش شاخص عملکردی را بیابید. برای انجام این کار لازم است یک شاخص شناخته شده مانند p–value را محاسبه کنید. (برای محاسبه دم یا tail نمودارها می‌توانید از اینترنت یا زبان‌های برنامه‌نویسی آماری مانند R استفاده کنید (**امتحانات در آن سال مجازی بوده**). همچنین برخی نامساوی‌ها در بخش راهنمایی‌ها نیز آمده که قابل استفاده هستند.)

فرض کنید که سازمان پس از مطالعه گزارش شما، اعلام می‌کند که p–value گزارش شده به اندازه کافی پایین نیست و لذا افزایش شاخص عملکردی سازمان هنوز معنی‌دار نشده است.

ج) اگر می‌توانستید اندازه فقط یکی از جمعیت‌های نمونه ۱۰۰ نفره (قبل از تعویض مدیران) و یا۱۰ نفره (بعد از آن) را به اندازه ۲۰ نفر افزایش داد تا مقدار p–value به بیشترین مقدار کمتر شود، شما پیشنهاد افزایش کدامیک را می‌دادید؟

سوال ۵ امتحان نهایی (۴ نمره):

ملکول DNA در بدن موجودات زنده یک دنباله طولانی به طول N و متشکل از چهار الفبای اصلی A,C,G و T است. در اینجا، الفبای A,C,G,T بیانگر حروف ابتدایی از نام ۴ نوع منحصر به فرد از زیرملکولهایی به نام نوکلئوتید هستند.
فرض کنید که بتوانیم رشته DNA را به صورت تصادفی مدل‌سازی کنیم: فرض کنید هر یک از N حرف رشته مستقل از سایرین و با احتمال‌های یکسان $1/4$ بتواند هر یک از اعضای الفبای A,C,G,T باشد (دقت کنید که در واقعیت اینطور نیست!). در این صورت، رشته DNA مورد بحث در این سوال، یک نمونه (یا realization) از این توزیع خواهد بود.



نمایه‌ای از توالی A,C,G,T رشته DNA انسان. طول رشته DNA در انسان تقریباً $N = 3 \times 10^9$ است.

الف) قصد داریم بدانیم که یک زیررشته (substring) فرضی و ساختگی خاص به طول L (برای مثال، یک زیر رشته ۱۰۰تایی مانند $ACCGTATT...GCC$) با چه احتمالی در حداقل یک جا از رشته DNA دیده خواهد شد. یک کران بالا برای احتمال خواسته شده بیابید.
(راهنمایی: می‌توانید از کران اجتماع یا Union Bound استفاده کنید)

ب) یک کران پایین نیز برای احتمال قسمت الف) پیدا کنید.
(راهنمایی: رشته DNA را به زیررشته‌های بدون همپوشانی با طول L تقسیم کنید)

ج) نشان دهید که در صورت انتخاب $L \geq (1+\varepsilon)\log_4 N$ احتمال دیده شدن یک زیررشته پیش فرض و ساختگی در رشته تصادفی DNA با افزایش N به سمت صفر میل می‌کند (به ازای هر $\epsilon > 0$). همچنین، در صورت انتخاب $L \leq (1-\varepsilon)\log_4 N$ احتمال مشاهده هر زیررشته ساختگی دلخواهی در رشته اصلی به سمت ۱ خواهد رفت.

د) (**امتیازی**) نشان‌دهید احتمال مشاهده زیررشته‌های خودتکرارشونده با طول $L \geq (2+\varepsilon)\log_4 N$ با افزایش N به سمت صفر میل خواهد کرد. رشته‌های خودتکرارشونده زیررشته‌هایی از DNA هستند که حداقل در یک جای دیگر از رشته نیز دوباره دیده می‌شوند.
(برای این بخش لازم است توضیحات مفصل ارائه دهید و صرف نوشتن روابط کافی نیست).

توضیحات و راهنمایی‌ها:

* سوالات امتیازی هر کدام ۱ نمره اضافه و مستقل از بارم‌بندی سوالات دارند.
* کران اجتماع: فرض کنید رویدادهای $A_1, \ldots, A_k$ زیرمجموعه‌هایی از فضای نمونه Ω (یا معادلاً اعضایی از مجموعه وقایع $\mathscr{F}$) باشند. در این صورت، همواره داریم:

$$\mathbb{P}\left(A_1 \cup \ldots \cup A_k\right) \leq \mathbb{P}\left(A_1\right) + \ldots + \mathbb{P}\left(A_k\right)$$

* در صورتیکه $X \sim \mathcal{N}(0,1)$، آنگاه کران زیر برای دم توزیع برقرار است: $\mathbb{P}(X > t) \leq \exp\left(-t^2/2\right)$.

بخش سوم
نمونه‌سوالات جدید

نمونه سوال ۱

فرض کنید که موضوع چالش‌برانگیزی را در حساب کاربری خودتان در یک شبکه اجتماعی به اشتراک می‌گذارید. به دلیل حساسیت بالای موضوع شمار زیادی از افراد که تعدادشان را با n نشان می‌دهیم به پست شما بازخورد نشان داده و علاوه بر دادن یک امتیاز (مثلاً از ۱ تا ۱۰) به نظر شما، یک نظر شخصی یا کامنت هم در پایین پست شما می‌نویسند که می‌تواند شامل تعریف و تمجید، و یا *** باشد. برای سادگی فرض کنید که افراد به صورت غیرهمزمان و پشت‌سرهم پست را رویت کرده و امتیازدهی و کامنت‌دهی می‌کنند. لذا در نهایت شما یک دنباله از امتیازات را به صورت متغیرهای تصادفی $X_1, \ldots, X_n$ را مشاهده می‌کنید.

متغیرهای تصادفی فوق را می‌توان هم‌توزیع فرض کرد، که این توزیع در واقع بیان‌گر فراوانی آماری احساسات و عقاید در کف شبکه اجتماعی است. میانگین نامعلوم این توزیع را μ و واریانس آن را σ^2 فرض کنید. اما مشکل اینجاست که $X_1, \ldots, X_n$ از یکدیگر مستقل نیستند! در واقع، وقتی شخص iام در حال نظر دادن به پست شماست، کامنت تمامی افراد قبل از خود ($j < i$) را نیز می‌بیند، و با توجه به دیدگاه آنان قدری بایاس خواهد شد. لذا میان هر دو متغیر X_i, X_j یک «همبستگی آماری» مثبت وجود دارد. ضریب این همبستگی را مستقل از مقدار i, j $(i \neq j)$ برابر با ρ فرض کنید $(0 < \rho \leq 1)$.

در انتهای روز، بعد از تحمل ضربات روحی فراوان و یک جنگ یک‌تنه سایبری، شما خسته و بی‌رمق قصد دارید که میانگین امتیازهای داده شده به خودتان را تخمین بزنید!

الف) (سطح سختی: ساده) نشان دهید که علیرغم مستقل نبودن مشاهدات، تخمین‌گر میانگین نمونه‌ای کماکان یک تخمین بدون بایاس برای μ است:

$$\hat{\mu} = \frac{X_1 + \ldots + X_n}{n}$$

ب) (سطح سختی: معمولی) مقدار واریانس تخمین‌گر فوق را محاسبه کنید. در محاسبات شما مقادیر نامعلوم ρ و σ نیز ظاهر خواهند شد که مشکلی ندارد.

ج) (سطح سختی: ساده) آیا تخمین‌گر فوق سازگار است؟

د) (سطح سختی: **معمولی**) حال فرض کنید که سیاست گرداننده شبکه اجتماعی فوق که اخیراً توسط یک میلیاردر روان‌پریش خریداری شده تغییر کرده و به هر کسی که پست شما را بازدید می‌کند تنها k تا از جدیدترین کامنت‌ها را نشان می‌دهد که k عددی مشخص و ثابت است. در این صورت جواب شما به سوال ج) چه تغییر می‌کند؟ جوابتان را اثبات کنید.

ه) (سطح سختی: **سخت**) در یک مدلسازی واقعی‌تر، اشخاصی که پست شما را می‌بینند می‌توانند به صورت بالقوه تمامی کامنت‌های قبلی را بخوانند. اما به جدیدترین کامنت‌ها بیشتر توجه می‌کنند. لذا می‌توان مقدار ضریب همبستگی (correlation coefficient) میان X_i, X_j که آن را با $\rho_{i,j}$ نمایش می‌دهیم به صورت زیر زیر در نظر گرفت:

$$\rho_{i,j} = \rho e^{-\alpha|i-j|}$$

که در اینجا α عددی مثبت است. در این صورت جواب شما به قسمت ج) چه تغییر خواهد کرد؟

نمونه سوال ۲

فرض کنید که $X_1, \dots, X_n$ نمونه‌های i.i.d. از یک متغیر تصادفی مانند X با تابع چگالی احتمال $f_X(x)$ باشند. n را عددی فرد فرض کنید. نمونه‌ها را بعد از realization به صورت صعودی مرتب کرده و مقداری که دقیقا در وسط قرار می‌گیرد (یعنی $(n+1)/2$امین نمونه) را انتخاب می‌کنیم و آن را $\hat{X}_{\text{median}}$ می‌نامیم.

الف) (سطح سختی: **معمولی**) توزیع آماری (مثلاً چگالی احتمال) $\hat{X}_{\text{median}}$ را برحسب $f_X(x)$، $F_X(x)$ و n بیابید.
راهنمایی: ابتدا تعدادجایگشت‌هایی که نمونه‌ها را به سه دسته «نیمه کمتر»، «وسطی» و «نیمه بیشتر» تقسیم می‌کند را بیابید. سپس مشابه با نحوه محاسبه توزیع بیشینه و کمینه عمل کنید.

ب) (سطح سختی: **سخت**) قصد داریم $\hat{X}_{\text{median}}$ را به عنوان یک تخمین‌گر برای «میانه» توزیع X معرفی کنیم. به یاد بیاورید که میانه یک توزیع جایی است که نیمی از جرم احتمال در یک سمت و نیم دیگر در سمت دیگرش قرار می‌گیرند. برای این منظور، ابتدا نشان دهید که در حد $n \to \infty$ بیشینه چگالی احتمال $\hat{X}_{\text{median}}$ که در قسمت الف) محاسبه شد به میانه $f_X(x)$ میل خواهد کرد.
برای این منظور، فرض کنید که $f_X(x)$ در محل میانه غیرصفر است و حول آن نقطه تغییرات شدید ندارد.

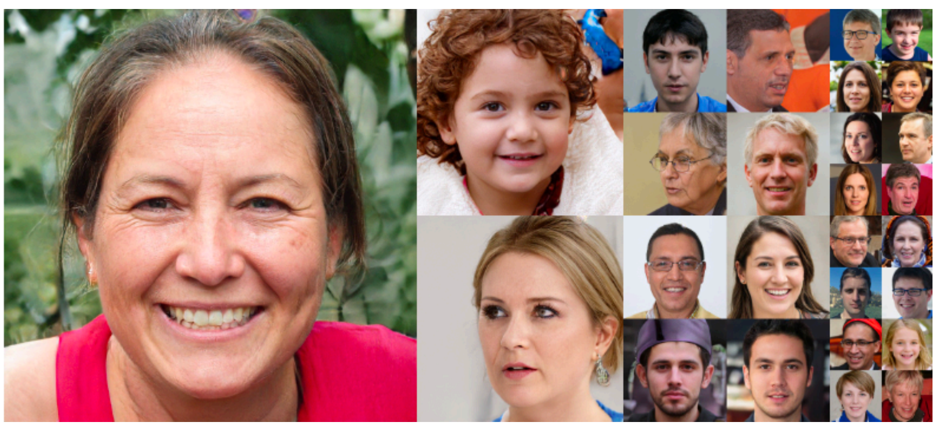
ج) (سطح سختی: **بسیار سخت**) نشان دهید در صورتیکه با سرعتی بیش از $O(n^{-1/2})$ از محل میانه دور شویم، تابع چگالی احتمال $\hat{X}_{\text{median}}$ به صورت نمایی کاهش خواهد یافت. لذا میانگین $\hat{X}_{\text{median}}$ نیز مانند بیشینه چگالی احتمال آن (مود) به سمت میانه $f_X(x)$ میل کرده و حداکثر واریانسی که خواهد داشت $O(n^{-1/2})$ خواهد بود.

راهنمایی: دقت کنید که به واسطه قضیه بسط تیلور، برای x به اندازه کافی کوچک و a دلخواه داریم:

$$F_X(a+x) \simeq F_X(a) + f_X(a)x$$

نمونه سوال ۳

شبکه‌های عصبی GAN در طی کمتر از یک دهه تحول بزرگی در عرصه هوش مصنوعی ایجاد کرده‌اند. عمده تصاویر و صداگذاری‌های فیکی که در طول روز در شبکه‌های اجتماعی دست‌مایه خنده و شادی کاربران هستند زیر سر این شبکه‌هاست! البته سایر استفاده‌ها از آنان مثبت‌تر و در راه تسهیل زندگی ما بوده است، و کاربردهای فراوانی در پزشکی، هنر، فیلمسازی و ... داشته‌اند. این شبکه‌ها با آموزش بر روی طیف وسیعی از دادگان، یاد می‌گیرند که نمونه‌های مشابه با آنان را تولید کنند. مثلاً در صورتیکه تعداد زیادی تصویر چهره به آنان نشان دهید، یاد می‌گیرند چهره‌های طبیعی‌ای تولید کنند که در عالم واقعیت وجود ندارند. یعنی متعلق به یک شخصی که در دنیای فیزیکی ما زندگی می‌کند نیستند، ولی همزمان از یک چهره واقعی نیز قابل تشخیص نمی‌باشند.



در صورتیکه از یک GAN که بر روی دادگان زیادی از تصاویر چهره افراد آموزش دیده است نمونه بگیرید، به نظر همه چیز درست می‌رسد. اما مشکل جایی ظاهر می‌شود که تعداد بسیار بسیار زیادی نمونه بگیرید! مثلاً از مرتبه صدها هزار بار. در آن صورت متوجه خواهید شد که چهره‌ها از جایی به بعد تکراری هستند... البته حضور خروجی‌های تکراری خیلی زودتر نمایان خواهد شد. دقت کنید که دو تصویر تکراری هنوز در جزئیات ریزی تفاوت دارند: مثل وجود یک خال روی صورت، موهای جوگندمی بیشتر/کمتر و ... اما ماهیت تصاویر یکسان است. به این پدیده Mode Collapse اطلاق می‌گردد، و نشان می‌دهد که حتی GANها نیز تخیل محدودی دارند.
دانشمندان برای مقایسه GANهای تولید شده توسط شرکت‌ها و موسسات مختلف، ظرفیت تولید خروجی‌های متمایز آنان را اندازه می‌گیرند و بر همین اساس آنان را رتبه‌بندی می‌کنند. مثلاً در حال حاضر GANهای تولید شده توسط Google و Nvidia بهترین‌های حوزه محسوب می‌شوند و می‌توانند صدها هزار الی میلیون‌ها تصویر فیک تولید کنند (لازم به ذکر

است در طی ۲-۳ سال گذشته مدل‌های جدید روی کار آمده‌اند که احتمالاً در حال کنار زدن GANها هستند). در این سوال قصد داریم ببینیم چگونه می‌توان ظرفیت یک GAN را اندازه‌گیری کرد.

بیایید شبکه‌مان را اینگونه مدلسازی کنیم: فرض کنید که GAN ما توان تولید n خروجی متمایز را داشته باشد. هر بار که شبکه را prompt می‌کنیم (مثلاً از طریق دادن نویز به ورودی)، یکی از این n خروجی به صورت مستقل از خروجی‌های قبلی و به صورت یکنواخت انتخاب شده و نمایش داده خواهد شد. همچنین، هر بار که یک خروجی تولید شود که قبلاً نیز تولید شده بوده، ما متوجه خواهیم شد.
در گام اول، قصد داریم ببینیم به صورت متوسط باید چند بار از GAN نمونه بگیریم تا تمامی n خروجی ممکن حداقل یکبار ظاهر شده باشند. واضح است که بعد از گرفتن اولین خروجی، بلافاصله یکی از n حالت شناسایی می‌گردد (چون قبل از آن اصلاً مشاهده‌ای نکرده بودیم).

الف) (سطح سختی: **ساده**) فرض کنید در لحظه گرفتن دومین خروجی از شبکه هستید. نشان دهید تعداد خروجی‌هایی که لازم است از الآن بگیرید تا یک تصویر غیرتکراری رویت کنید (تصویری مخالف با تصویر اول) از یک توزیع هندسی با احتمال موفقیت $\frac{n-1}{n}$ تبعیت می‌کند.

ب) (سطح سختی: **معمولی**) فرض کنید در میانه کار هستید. آخرین خروجی که گرفته‌اید، یک تصویر جدید و غیرتکراری بوده است و تعداد تصاویر یکتایی که تا الآن مشاهده شده را به عدد i رسانده است. نشان دهید تعداد خروجی‌هایی که لازم است از الآن بگیرید تا $i+1$امین خروجی غیرتکراری ظاهر شود کماکان توزیع هندسی دارد. پارامتر این توزیع را بر حسب n و i بدست بیاورید.

ج) (سطح سختی: **سخت**) تعداد خروجی‌های لازم برای اینکه تمامی n تصویر حداقل یکبار دیده شوند را با M نشان می‌دهیم، که یک متغیر تصادفی صحیح و مثبت است. بدست آوردن توزیع دقیق M کاری بسیار سخت است. اما میانگین M چقدر است؟ فرمول دقیق میانگین را به صورت یک جمع nتایی بیان کنید.

د) (سطح سختی: **سخت**) نشان دهید که مجموع بدست آمده برای میانگین M در قسمت ج) برای nهای بزرگ با $n \log n$ تقریب خواهد خورد (قضیه Coupon collector).

ه) (سطح سختی: **بسیار سخت**) واریانس M حول میانگین‌اش چقدر است؟ نشان دهید که انحراف معیار (جذر واریانس) با نرخی کندتر از $n \log n$ رشد خواهد کرد. لذا مقدار M به ازای nهای بزرگ با احتمال ۱ به $n \log n \left(1+o(1)\right)$ میل خواهد کرد.
($o(1)$ یعنی مقداری که با رشد پارامتر مسئله یعنی n به سمت صفر میل می‌کند. اما نرخ کاهش را مشخص نکرده‌ایم).

به نظر می‌رسید که تنها راه فهمیدن ظرفیت یک GAN این است که آنقدر خروجی گرفته شود تا از جایی به بعد دیگر خروجی جدید نبینیم. برای اینکار لازم است تعداد ۲-۳ برابر $n \log n$ مشاهده و مقایسه انجام بدهیم (در واقع تعداد مقایسه‌ها از مرتبه n^2 خواهد شد!). دقت کنید که خود n از مرتبه میلیون است.
سال ۲۰۱۸، آقای Sanjeev Arora (یک دانشمند شناخته شده در حوزه علوم کامپیوتر) و تیم همکارانشان در دانشگاه پرینستون و شرکت Google یک راه بسیار ساده و زیرکانه برای تخمین نادقیق ظرفیت یک GAN پیدا کردند، که تعداد خروجی‌های لازم از شبکه را به $O(n^{1/2})$ تقلیل داد. ایده استفاده شده از یک مسئله بسیار ساده و قدیمی در آمار و احتمال بیرون آمده بود... در واقع ما نیز در ابتدای همین ترم در درس آمار و احتمال کلیات آن را فرا گرفتیم: Birthday Party Problem!

و) (سطح سختی: **سخت**) ایده آقای Arora و یارانشان این بود: شروع به خروجی گرفتن از شبکه کنید تا اولین خروجی تکراری ظاهر شود. فرض کنید در kامین خروجی این اتفاق رخ دهد. در آن صورت n از قدرمرتبه k^2 خواهد بود (با یک ضریب). با تکرار این آزمایش، میانگین‌گیری، و قدری محاسبات می‌توان ضریب را نیز بدست آورد. به فرمول «مسئله جشن تولد» که ابتدای ترم مطرح شد مراجعه کنید، و ادعای آقای Arora را توجیه نمایید. نیازی به محاسبات طولانی برای محاسبه ضریب نیست.

موفق باشید!