لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید باید آن را ذکر کنید.
- اگر با افرادی همفکری کرده‌اید، نام ایشان را ذکر کنید.
- پاسخ ارسالی باید توسط خود شما نوشته شده باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت RL_HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می توانید تا سقف ۲ روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید و در مجموع ۵ روز تاخیر مجاز برای تمارین در اختیار دارید.

سوال ۱: (نظری) نظریه اطلاعات (۲۰ نمره)

برای یک توزیع احتمال متغیری به اسم آنتروپی به شکل زیر تعریف می شود که برای بررسی عدم قطعیت یک توزیع از آن استفاده می‌شود.

$$H(x) = -\sum_{x} P(x) log P(x)$$

همچنین یکی از ابزارها برای مقایسه میزان اطلاعات بین دو توزیع آنتروپی نسبی است و برای دو توزیع P،Q به صورت زیر تعریف می‌شود:

$$D_{KL}(P,Q) = -\sum_{x} P(x) log(\frac{P(X)}{Q(x)})$$

اطلاعات متقابل دو متغیر تصادفی X و Y نیز به صورت زیر تعریف می‌شود:

$$I(X;Y) = H(X) - H(X|Y)$$

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

حال با توجه به اطلاعات داده‌شده به سوال‌های زیر پاسخ دهید:

الف)برای متغیرهای تصادفی X،Y،Z مثالی بیاورید که هر یک از نامساوی‌های زیر برقرار باشد:

۱) $I(X;Y \mid Z) < I(X;Y)$

۲) $I(X;Y \mid Z) > I(X;Y)$

ب) هر یک از نامساوی زیر را ثابت و شرایطی که نامساوی به مساوی تبدیل می‌شود را توضیح دهید

۱) $H(X,Y,Z) - H(X,Y) \leq H(X,Z) - H(X)$

۲) $I(X;Z|Y) \geq I(Z;Y|X) - I(Z;Y) + I(X;Z)$

ج) فرض کنید شما یک ماجراجو هستید که در جستجوی گنجی افسانه‌ای هستید که در یک جزیره دورافتاده قرار دارد. شما به تنهایی در این ماجراجویی شرکت می‌کنید و باید از بین مسیرهای مختلفی که به گنج منتهی می‌شوند، مسیر بهینه را انتخاب کنید چهار مسیر A،B،C،D برای رسیدن به گنج وجود دارد. در ابتدا هیچ اطلاعاتی در خصوص اکشن بهینه نداریم.پس از انتخاب مسیر a و دریافت پاداش حال احتمال اینکه هر کدام از مسیرها مسیر بهینه باشد به صورت زیر تغییر پیدا می‌کند:

$$P(Z=A|Y=y) = \frac{1}{1+e^{-0.5(y-5)}}$$

$$P(Z=B|Y=y) = \frac{1-P(Z=A|Y=y)}{2}$$

$$P(Z=C|Y=y) = \frac{1-P(Z=A|Y=y)}{6}$$

$$P(Z=D|Y=y) = \frac{1-P(Z=A|Y=y)}{3}$$

۱) فرض کنید منظور از متغیر Z اکشن بهینه باشد و منظور از Y میزان پاداشی است که پس از انجام آن عمل a دریافت می شود. مقدار $I(Z;Y)$ را محاسبه کنید می‌توانید تابع توزیع y دلخواه را در نظر بگیرید
۲) $I(Y;Z)$ در این مساله نشان دهنده چیست؟ در صورتی که مقدار آن بزرگ یا کوچک باشد چه مفهمومی دارد؟
د) فرض کنید X،Y،Z سه متغیر تصادفی باشند حال تابع توزیع توام این سه متغیر را به صورت $p(x,y,z)$ در نظر بگیرید و ثابت کنید رابطه زیر برقرار است. همچنین بررسی کنید در چه شرایطی مقدار $DKL(p(x,y,z)||p(x)p(y)p(z))$ برابر صفر می‌گردد

$$DKL(p(x,y,z)||p(x)p(y)p(z)) = -H(X,Y,Z) + H(X) + H(Y) + H(Z)$$

ه) فرض کنید $X_1 \to X_2 \to X_3 \to X_4$ یک زنجیره مارکوف باشند ثابت کنید:

$$I(X1;X3) \ + I(X2;X4) \ \leq I(X1;X4) + I(X2;X3)$$

پاسخ:
حالتی را در نظر بگیرید که در آن X نیز یک متغیر تصادفی باینری یکنواخت باشد. همچنین فرض کنید $Y = X$ و $Z = Y$ باشد آنگاه:

$$I\left(X;Y\right) = H\left(X\right) - H\left(X|Y\right) = H\left(X\right) = 1$$
$$I\left(X;Y|Z\right) = H\left(X|Z\right) - H\left(X|Y,Z\right) = 0$$

b) فرض کنید X ، Y هردو یک متغیرتصادفی باینری یکنواخت و مستقل از یکدیگر باشند و Z=X+Y آنگاه:

$$I\left(X;Y\right) = 0$$
$$I\left(X;Y|Z\right) = H\left(X|Z\right) = 0.5$$

ب)
۱) می توانیم رابطه زیر را بنویسیم:

$$H(X,Y,Z) - H(X,Y) = H(Z|X,Y) = H(Z|X) - I(Y;Z|X) \leq H(Z|X) = H(X,Z) - H(X)$$

و مشخص است در حالتی نامساوی به مساوی تبدیل می شود که I(Y؛Z|X)=۰ صفر باشد.
۲) در این حالت نیزرابطه زیر برقرار است

$$I(X;Z|Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y|X) + I(X;Z)$$
$$I(X;Z|Y) = I(Z;Y|X) - I(Z;Y) + I(X;Z)$$

ج) برای راحتی کار، Y را شامل سه مجموعه به صورت {۵،۰،۱} که هر کدام با احتمال برابر رخ می‌دهد، قرار دهیم.

$$I(Z;Y) = H(Z) - H(Z|Y)$$

حال ابتدا $H(Z)$ را محاسبه می‌کنیم:

$$H(Z) = -\sum_z p(z)\log p(z)$$

لازم است P(Z) را محاسبه کنیم

$$P(Z = A|Y = 0) = 0.075$$
$$P(Z = A|Y = 1) = 0.1192$$
$$P(Z = A|Y = 5) = 0.5$$
$$P(Z = A) = (0.5 + 0.075 + 0.1192)\left(\frac{1}{3}\right)$$
$$= 0.2317$$

به همین ترتیب سایر احتمال ها نیز محاسبه می‌کنیم:

$$P(Z = B) = 0.38416$$
$$P(Z = C) = 0.1280$$
$$P(Z = D) = 0.2561$$

حال H(Z) برابر مقدار زیر می شود:

$$H(Z) = 0.9020$$

حال لازم است H(Z | Y) را محاسبه کنیم

$$H(Z|Y) = \sum_y P(y)H(Z|Y=y) = 1.7592$$
$$I(Z;Y) = 0.1427$$

$I(Z;Y)$ نیز میزان اطلاعات بدست آمده نسبت به مسیر بهینه براساس پاداش دریافتی را نشان می‌دهد. اگر $I(Z;Y)$ بزرگ باشد، دانستن پاداش دریافتی اطلاعات زیادی در مورد مسیر بهینه به ما می‌دهد و بالعکس.
د) رابطه را به صورت زیر می‌نویسیم

$$\begin{aligned} D(p(x,y,z)||p(x)p(y)p(z)) &= E\left[\log \frac{p(x,y,z)}{p(x)\,p(y)\,p(z)}\right] \\ &= E\left[\log p(x,y,z)\right] - E\left[\log p(x)\right] - E\left[\log p(y)\right] - E\left[\log p(z)\right] \\ &= -H(X,Y,Z) + H(X) + H(Y) + H(Z) \end{aligned}$$

و همچنین تنها در شرایطی برابر صفر می‌شود که p(x،y،z)=p(x)p(y)p(z)
ه) روابط را به صورت زیر بسط می دهیم

$$\begin{aligned} &I(X_1;X_4) + I(X_2;X_3) - I(X_1;X_3) - I(X_2;X_4) \\ &= H(X_1) - H(X_1|X_4) + H(X_2) - H(X_2|X_3) \\ &\quad - (H(X_1) - H(X_1|X_3)) - (H(X_2) - H(X_2|X_4)) \\ &= H(X_1|X_3) - H(X_1|X_4) + H(X_2|X_4) - H(X_2|X_3) \\ &= H(X_1,X_2|X_3) - H(X_2|X_1,X_3) - H(X_1,X_2|X_4) + H(X_2|X_1,X_4) \\ &\quad + H(X_1,X_2|X_4) - H(X_1|X_2,X_4) - H(X_1,X_2|X_3) + H(X_1|X_2,X_3) \\ &= -H(X_2|X_1,X_3) + H(X_2|X_1,X_4) - H(X_2|X_1,X_4) - H(X_2|X_1,X_3,X_4) \\ &= I(X_2;X_3|X_1,X_4) \geq 0 \end{aligned}$$

سوال ۲: (نظری) بهینه سازی (۱۵ نمره)

هدف از بهینه سازی پیدا کردن نقطه یا نقاطی است که مقدار یک تابع را روی فضای مشخصی کمینه یا بیشینه کند. فرض کنیم میخواهیم مقدار تابع f را کمینه کنیم

$$x^* = argmin_{x\in U} f(x) \tag{۱}$$

با فرض اینکه تابع f کران پایین دارد و مشتق پذیر است و U نیز کران دار است، $x^* \in \partial(U) \cup Cr(f,U)$ که $\partial(U)$ نقاط مرزی U می باشد و $Cr(f,U)$ نقاط بحرانی f در U می باشند که مشتق f در این نقاط صفر می شود.
از بین تمام نقاط بحرانی، نقاطی که مقدار ماتریس هسیان [۱] در آن ها مثبت معین می شود کمینه محلی می باشند. اما همچنان کمینه بودن آن ها در کل فضای U مشخص نیست. این موضوع برای توابع محدب ساده تر است. زیرا در صورت محدب بودن f هر نقطه بحرانی یک نقطه کمینه برای f است.
برای مسائل بهینه سازی با قید نیز می توان از روش ضرایب لاگرانژ استفاده کرد. فرض کنید با فرض $g(x) = 0$ می خواهیم تابع $f(x)$ را کمینه کنیم. کافیست معادله $\nabla f(x) = \lambda \nabla g(x)$ با فرض $g(x) = 0$ را حل کنیم. جواب مسئله اصلی داخل جواب های معادله جدید می باشد. در واقع مسئله مقید اولیه به مسئله بدون قید $f(x) + \lambda g(x)$ تبدیل می شود.
روش ضرایب لاگزانژ حالت خاص روشی برای حل مسائل بهینه سازی است که بر پایه مفهومی به نام مسئله دوگان استوار است. فرض کنید علاوه بر قیدهای برابر، قیدهایی به صورت نابرابری نیز وجود دارند

$$\min_x f(x) \;\; s.t. \;\; g_i(x) \leq 0 \;\; , \;\; h_j(x) = 0 \tag{۲}$$

دوگان این مسئله به صورت زیر تعریف می شود،

$$\max_{\mu,\lambda} L(\mu,\lambda) \;\; s.t. \;\; \mu \geq 0 \tag{۳}$$

$$L(\mu,\lambda) = \inf_x \left\{ f(x) + \mu^T g(x) + \lambda^T h(x) \right\} \tag{۴}$$

[۱] Hessian Matrix

عبارت دوگان یک کران پایین برای تابع اصلی بهینه سازی ارائه می دهد و حل مسئله دوگان لزوما منجر به حل مسئله اصلی نمی شود. در صورت برقراری شرایط KKT جواب مسئله دوگان معادل با جواب مسئله اصلی می شود. یکی از ویژگی های فرم دوگان این است که تابع L یک تابع مقعر است و بیشینه سازی آن همانند کمینه سازی یک تابع محدب ساده تر می باشد.
گاهی در بهینه سازی به جای پیدا کردن نقاط بهینه به دنبال توابع بهینه هستیم. حل چنین مسائلی موضوع حوزه حساب تغییرات [۲] می باشد. برابری لاگرانژ اویلر یک شرط لازم برای بهینه بودن تابع معرفی می کند.
فرض کنیم $y = y(x)$ تابعی از x است و هدف کمینه یا بیشینه کردن $\int F(x, y, y')dx$ است که F یک تابع دلخواه مشتق پذیر است. برای $y(x)$ بهینه داریم،

$$\frac{d}{dx}\frac{\partial F}{\partial y'} = \frac{\partial F}{\partial y} \tag{۵}$$

(آ) نشان دهید در صورت محدب بودن f، هر نقطه بحرانی ، یک نقطه کمینه می باشد.

(ب) کمینه و بیشینه تابع $f(x, y, z) = x^2 + y^2 + z^2$ را با شرایط $z^2 = x^2 + y^2$ و $z = x + y + 1$ بیابید.

(ج) ثابت کنید

$$L(\mu, \lambda) \leq \min_x f(x)$$

و مثالی ارائه دهید که برابری برای هیچ μ, λ ای رخ ندهد.

(د) مسئله بهینه سازی زیر را حل کنید،

$$\min_x \left\{ -\sum_{i=1}^{n} \log(\alpha_i + x_i) \right\} \qquad s.t. \qquad x \geq 0, 1^T x = 1 \tag{۶}$$

(ه) توزیع احتمال با بیشینه آنتروپی پیوسته ای را بیابید که میانگین و واریانس مشخصی داشته باشد،

$$argmax_P \left\{ -\int P(x) \log P(x) dx \right\} \qquad \mathbb{E}_P[X] = \mu, \quad \mathbb{V}_P(X) = \sigma^2 \tag{۷}$$

سوال ۳: (نظری) زنجیره مارکوف (۱۵ نمره)

یکی از کاربردهای زنجیره مارکوف استفاده از آن‌ها در مسائلی است که نمونه برداری با یک توزیع احتمال مشخص در آن‌ها دشوار می‌باشد در این تمرین یکی از این مسائل را بررسی می‌کنیم.
وقتی می گوییم یک زنجیره مارکوف غیر قابل کاهش است منظور این است که تمامی وضعیت‌ها به یکدیگر دسترسی داشته باشند به عبارتی دیگر رابطه زیر برقرار باشد:

$$\forall i, j \in S \; \exists n \geq 0 \text{ such that } P(X_n = j \mid X_0 = i) > 0$$

حالت i دارای دوره تناوب k است اگر هر مسیر بازگشت به حالت i به طول مضارب k باشد. به زبان دیگر، دوره تناوب یک حالت برابر است با

$$k = \gcd\{n : \Pr(X_n = i | X_0 = i) > 0\}$$

حال در صورتی که دوره تناوب تمام حالت ها برابر یک باشد می گوییم زنجیره مارکوف بدون تناوب است
الف) زنجیره مارکوفی را در نظر بگیرید که در آن به ازای هر وضعیت x و y اگر $x \neq y$ و $y \in N(x)$ باشد انگاه:
$P_{(x,y)} = 1/M$ و در غیر این صورت $P_{(x,y)} = 0$ همچنین $P_{(x,x)} = 1 - \frac{N(x)}{M}$.

$$\text{N(x): set of neighbors of x}$$

$$M \geq max_{x \in \Omega} |N(x)|.$$

حال اگر این زنجیره مارکوف غیرقابل کاهش و بدون تناوب باشد، ثابت کنید توزیع مانا[۳] این زنجیره مارکوف یک توزیع یکنواخت است.
ب) مسئله شماردن تعداد حالت‌های مختلف در یک کوله پشتی را درنظر بگیرید. در این مسئله تعدادی شی در اختیار داریم که وزن هرکدام به صورت a_n $a_1, a_2, a_3, \ldots\ldots,$ می‌باشد و وزن کل کوله پشتی نیز b می‌باشد. هدف تخمین تعداد بردارهایی به صورت $(x1, x2, x3, \ldots .xn) \in \{0, 1\}^n$ طوری که $\sum_{i=1}^{n} a_i x_i < b$.
حال فرض کنید برای حل این مسئله به این صورت عمل می کنیم که به صورت یکنواخت یک بردار

[۲] Variational Calculus
[۳] stationary distribution

$(x1, x2, x3, \ldots .xn) \in \{0,1\}^n$ را انتخاب کرده و این کار را به تعداد زیادی تکرارN) (می‌کنیم . تعداد حالت‌هایی که در آن شرط کوله پشتی نقض نمی‌شود را می شماریم و نام این مجموعه را p می نامیم. در نهایت مقدار $\frac{|P|\times 2^n}{N}$ به عنوان جواب نهایی برمی‌گردانیم این روش چه اشکالی دارد؟

ج) حال فرض کنید برای مسئله بالا زنجیره مارکوفی را به صورت زیر می‌سازیم که در آن هر وضعیت Xj را به این صورت $(x1, x2, x3, \ldots .xn)$ تعریف می‌کنیم. حال در هر گام زنجیره مارکوف یک متغیر $i \in [1, n]$ را به صورت یکنواخت انتخاب می‌کند : اگر $x_i = 1$ بود آن را صفر می‌کند و به حالت $Xj + 1$ می‌رود و در صورتی که $x_i = 0$ باشد آن را یک می کند در این حالت در صورتی که با یک کردن x_i شرط کوله پشتی نقض نشود به حالت $Xj + 1$ می رود و در غیر این صورت $X_j = X_{(j+1)}$

۱) نشان دهید در صورتی که $\sum_{i=1}^{n} a_i < b$ آنگاه این زنجیره مارکوف دارای یک توزیع مانا به صورت توزیعی یکنواخت است.

۲) توضیح دهید چگونه می‌توان از این زنجیره مارکوف جهت تخمین تعداد حالات برای مسئله کوله پشتی استفاده کرد.

پاسخ:

الف) اولا نشان می دهیم در یک زنجیره مارکوف محدود ، غیرقابل کاهش و بدون تناوب با n وضعیت و ماتریس انتقالی P، اگر $\pi = (\pi_0, \ldots, \pi_n)$ وجود داشته باشد که $\sum_{i=0}^{n} \pi_i = 1$ و برای هر i, j رابطه زیر برقرار باشد:

$$\pi_i P_{(i,j)} \ = \ \pi_j P_{(j,i)}$$

آنگاه π توزیع پایا و یکتا این زنجیره مارکوف می باشد.
اثبات :

$$\sum_{i=0}^{n} \pi_i P_{i,j} \ = \ \sum_{i=0}^{n} \pi_j P_{j,i} = \pi_j$$

از این رو چون π در رابطه $\pi = \pi P$ و $\sum_{i=0}^{n} \pi_i = 1$ صدق می کند آنگاه π باید حتما توزیع پایا و یکتا این زنجیره مارکوف باشد. حال در زنجیره مارکوف بیان شده طبق صورت مسئله برای $x \neq y$ اگر $\pi_y \pi_x =$ آنگاه:

$$\pi_x P_{(x,y)} \ = \ \pi_y P_{(y,x)}$$

حال در صورتی که π یک توزیع یکنواخت باشد طوری که $\sum_{i=0}^{n} \pi_i = 1$ با توجه به اینکه زنجیره مارکوف محدود ، غیرقابل کاهش و بدون تناوب است آنگاه می توان گفت π توزیع مانا و یکتا این زنجیره می‌باشد.

ب) مشکل این روش این است که توزیع حالت‌های مختلف این مسئله مشخص نیست و نمی توان به آسانی از یک توزیع مشخص جواب این مسئله را نمونه برداری کرد و در برخی حالات نیاز به تعداد نمایی نمونه برداری می‌باشد تا تخمین بدست آمده تخمین مناسبی باشد. به طور مثال، در حالتی که یک آیتم دارای سایز $b - 1$ و سایر آیتم‌ها دارای سایز $2b$ باشد، تنها یک جواب ممکن برای مسئله وجود دارد و برای تخمین این عدد حدود 2^n نمونه برداری لازم است

ج) با توجه به تعریف P واضح است که اگر گراف معادل این زنجیره مارکوف را رسم کنیم، تنها $X = (x_1, x_2, \ldots, x_n)$ هایی در این گراف وجود دارند که شرط $\sum_{i=1}^{n} a_i x_i < b$ را ارضا می‌کنند. اگر دو گره X و Y که تنها در بیت i ام اختلاف دارند را در نظر بگیرید که در آن $x_i = 1$ و $y_i = 0$، مشخص است که یالی از Y به X وجود دارد. همچنین با توجه به اینکه یک کردن y_i منجر به جود آمدن X می‌شود و X نیز شرط $\sum_{i=1}^{n} a_i x_i < b$ را ارضا می‌کند، واضح است یالی از X به Y هم وجود دارد یا به عبارت دیگر یال‌های گراف دو طرفه هستند. حالت $z \in 0^n$ را در نظر بگیرید. از همه حالات X با تغییر x_i هایی که دارای مقدار ۱ هستند می‌توان به حالت z رسید و با توجه به اینکه گراف شرط $\sum_{i=1}^{n} a_i x_i < b$ را ارضا می‌کند، در نتیجه از z نیز به همه حالات Y می‌توان رسید. در نتیجه گراف همبند است. همچنین می‌توان نشان داد گراف حتما دارای طوقه نیز می‌باشد زیرا اگر حالت بهینه کوله پشتی مانند w را در نظر بگیریم، با یک کردن هر آیتمی که برابر صفر می‌باشد (دقت شود با توجه به فرض $\sum_{i=1}^{n} a_i > b$) حتما آیتمی وجود دارد که برداشته نشده باشد، حتما طبق تعریف مسئله به خودش برمی‌گردد. حال با توجه به اینکه زنجیره مورد نیاز غیرقابل کاهش و بدون تناوب است و تابع انتقال مانند حالت الف می‌باشد توزیع مانا این زنجیره یکنواخت است. در نتیجه به استفاده از این زنجیره می‌توان به صورت یکنواحت از این مسیله نمونه برداری کرد و مشکل قسمت ب حل می‌شود زیرا می‌توان هر جواب valid را با احتمال برابر نمونه برداری کرد.

سوال ۴: (نظری) آمار بیزی (۱۵ نمره)

در یک دهه پرفراز و نشیب، جزیره ای به نام "آرکادیا" وجود دارد. مردم این جزیره تصمیم می‌گیرند زندگی را از سر بگیرند و یک شرکت به نام "Reminisce" راه اندازی که با استفاده از فناوری‌های جدید، یادگیری را بازیابی کند.

شرکت Rminisce یک فناوری به نام Mirror Memory را توسعه می‌دهد. این سیستم، به شما اجازه می‌دهد تا لحظات خوب و خاطره‌انگیز زندگی‌تان را بازگو کنید. اما این سیستم نیازمند شناسایی افراد است تا بتواند به درستی خاطرات را به هر فرد متصل کند.

سیستم Mirror Memory شامل یک دوربین و نرم‌افزاری است که ورودی اتاق را کنترل می‌کند. زمانی که شما می‌خواهید هویت خود را به سیستم ثبت کنید، ابتدا عکسی از خودتان گرفته می‌شود،. سپس این عکس با گالری عکس‌های ذخیره شده در سیستم مقایسه می‌شود.

حال برای هر عکس موجود در گالری یک امتیاز بین صفر و یک در نظر گرفته می‌شود:

اگر عکس گرفته شده و عکس گالری نشان دهنده همان فرد باشند، امتیاز از توزیع زیر محاسبه می‌شود:

$$p\left(s|same\right) = \alpha_{ss}\ exp(\lambda_s s)$$

اما اگر این عکس گالری نشان دهنده فردی متفاوت باشند، امتیاز از توزیع زیر محاسبه می‌شود

$$p\left(s|different\right) = \alpha_{ds}\ exp(-\lambda_d s)$$

الف) توضیح دهید آیا توابع در نظر گرفته شده برای توزیع امتیازها منطقی می‌باشند یا خیر؟ علت استفاده از ضرایب α_{ss} و α_{ds} چیست؟
ب) فرض کنید N عکس از N فرد مختلف در سیستم موجود باشد. حال با فرض داشتن N امتیاز مشخص به صورت s۱،s۲،...sj،..sN احتمال اینکه تصویر j ، تصویر فرد مورد نظر باشد چقدر است؟ (فرض کنید si به صورت صعودی داده شده اند و احتمال prior همه اشخاص مشابه می‌باشد.)
ج) حال فرض کنید هیچ امتیاز مشخصی در اختیار نداریم. سیستم صرفا عکسی را برمی‌گرداند که بیشترین امتیاز را دارد. احتمال اینکه این عکس فرد درست تشخیص داده شود چقدر است؟ برای این موضوع یک فرمول عمومی ارائه دهید

پاسخ:

الف) بله، توابع در نظر گرفته شده برای امتیازها منطقی می‌باشند. زیرا همانطور که مشخص است، در صورتی که افراد مشابه باشند، شما با احتمال زیاد امتیاز بالایی می‌گیرید و در صورتی که افراد متفاوت باشند، با احتمال زیاد امتیاز خیلی کمی گرفته می‌شود. با توجه به اینکه جمع احتمال‌ها باید برابر یک باشد، جهت نرمال‌سازی مقادیر امتیازها نیز از ضرایب α_s و α_d استفاده شده است.
ب)

$$\begin{aligned}
P\left(same_j, different_{i\neq j}|s_1,\ldots,s_N\right) &= \frac{p(s_1,\ldots,s_N|same_j, different_{i\neq j})}{p\left(s_1,\ldots,s_N\right)} \\
&= \frac{p(s_j|same)\left(\prod_{i\neq j} p\left(s_i|different\right)\left(\frac{1}{N}\right)\right)}{p\left(s_1,\ldots,s_N\right)} \\
&= \frac{\alpha_s exp\left(\lambda_s s_j\right)\prod_{i\neq j}\alpha_d exp\left(-\lambda_d s_i\right)\left(\frac{1}{N}\right)}{p\left(s_1,\ldots,s_N\right)} \\
&= \frac{\alpha_s \alpha_d^{(N-1)} exp\left(\lambda_s s_j\right) exp\left(-\lambda_d(s_j - \sum_s s_i)\right)\left(\frac{1}{N}\right)}{p\left(s_1,\ldots,s_N\right)} \\
&= \frac{\alpha_s \alpha_d^{(N-1)} exp((\lambda_s+\lambda_d)s_j) exp\left(-\lambda_d(\sum_s s_i)\right)\left(\frac{1}{N}\right)}{\sum_{j'}\alpha_s \alpha_d^{(N-1)} exp((\lambda_s+\lambda_d)s_{j'}) exp\left(-\lambda_d(\sum_s s_i)\right)\left(\frac{1}{N}\right)} \\
&= \frac{exp((\lambda_s+\lambda_d)s_j)}{\sum_{j'} exp((\lambda_s+\lambda_d)s_{j'})}
\end{aligned}$$

ج)

$$\begin{aligned}
p\left(correctrecognition\right) &= \int_0^1 p\left(s'|same\right) P(s<s'|different)^{(N-1)} ds' \\
&= \int_0^1 \alpha_s exp\left(\lambda_s s'\right)\left(\frac{exp(-\lambda_d s')-1}{exp\left(-\lambda_d\right)-1}\right)^{(N-1)} ds' \\
&= \int_0^1 \frac{\lambda_s exp\left(\lambda_s s'\right)}{exp\left(\lambda_s - 1\right)}\left(\frac{exp(-\lambda_d s')-1}{exp\left(-\lambda_d\right)-1}\right)^{(N-1)} ds' \\
&= \frac{1}{\left(exp\left(\lambda_s\right)-1\right)\left(exp\left(-\lambda_d\right)-1\right)^{(N-1)}}\int_0^1 exp(\lambda_s s')(exp\left(-\lambda_d s'\right)-1)^{(N-1)} ds'
\end{aligned}$$

سوال ۵: (نظری) تئوری تخمین (۲۰ نمره)

فرض کنید متغیر تصادفی $X : \mathcal{X} \to \mathbb{R}$ از توزیع P_θ پیروی می کند که θ یک پارامتر نامعلوم است. هدف از تخمین پیدا کردن θ از روی نمونه های مشاهده شده از X است،

$$\mathrm{X} = X_1, X_2, ..., X_n \sim P_\theta(X) \tag{۸}$$

به هر تابعی مانند $W = W(X_1, X_2, ..., X_n)$ یک تخمینگر می گوییم. تخمینگرها بر اساس ویژگی های مختلفی که دارند مقایسه و بررسی می شوند. هدف اصلی یک تخمینگر، ارائه یک تخمین مناسب از پارامتر θ می باشد. با توجه به اینکه خروجی تخمینگر خود یک متغیر تصادفی است، برای مقایسه آن با پارامتر حقیقی از معیار میانگین مجذور خطا [۴] یا MSE استفاده می شود،

$$MSE(W(\mathrm{X}),\theta) = \mathbb{E}[(W(\mathrm{X})-\theta)^2] = (\mathbb{E}[W(\mathrm{X})]-\theta)^2 + \mathbb{E}\left[(W(\mathrm{X})-\mathbb{E}[W(\mathrm{X})])^2\right] \tag{۹}$$

[۴] Mean Squared Error

که بخش اول مجذور بایاس و بخش دوم واریانس تخمینگر است. به یک تخمینگر نااریب می گویند اگر بایاس آن صفر باشد. یکی ار تخمینگرهای معروف تخمینگر بیشینه درست نمایی می باشد،

$$W(\mathrm{X}) = \hat{\theta}_{MLE} = argmax_\theta P(X_1, ..., X_n|\theta) = argmax_\theta \mathcal{L}(\theta; \mathrm{X}) \tag{۱۰}$$

(آ) فرض کنید که $\theta = (\mu, \sigma^2)$ و $P_\theta(X) = \mathcal{N}(X; \mu, \sigma^2)$. تحمینگر بیشینه درست نمایی را برای θ پیدا کنید. آیا این تخمینگر نااریب است؟

(ب) برای تابع دلخواه τ نشان دهید اگر $\hat{\theta}$ تخمین بیشینه درست نمایی از θ باشد، $\tau(\hat{\theta})$ تخمین بیشینی درست نمایی برای $\tau(\theta)$ است. سپس تخمین بیشینه درست نمایی را برای $\frac{1}{\mu^2+1}$ و σ با توجه به تعریف های قسمت قبل بدست آورید.

(ج) یکی از کاربردهای تخمین، بدست آوردن امیدریاضی تابعی از یک متغیر تصادفی است. فرض کنید برای متغیر تصادفی $X \sim p(X)$ و تابع $f(x)$ درصدد تخمین امیدریاضی این تابع تحت توزیع p هستیم،

$$\mu = \mathbb{E}_{X \sim P(X)}[f(X)] \tag{۱۱}$$

فرض کنید نمونه های $x_1, ..., x_n$ از توزیع متفاوت $q(X)$ داریم. در روش Importance-Weighted Sampling (IS) برای تخمین μ از تخمینگر زیر استفاده می شود،

$$\hat{\mu}_{IS} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{p(x_i)}{q(x_i)} = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i) \tag{۱۲}$$

در واقع به جای میانگین ساده، میانگین وزن دار $f(x_i)$ ها محاسبه می شود.

۱. نشان دهید تخمینگر IS نااریب است.

۲. عبارتی برای واریانس تخمینگر IS پیدا کنید. توزیع q را برحسب p و f طوری انتخاب کنید که واریانس $\hat{\mu}_{IS}$ کمینه شود و مقدار کمینه را بیابید. همچنین نشان دهید p و q را می توان طوری تعیین کرد که واریانس این تخمینگر هر اندازه ای بزرگ شود.

۳. مشکل بزرگ شدن واریانس تخمینگر IS در نهایت باعث بزرگ شدن MSE می شود و صرف نااریب بودن، باعث مطلوب بودن این تخمینگر نمی شود. برای حل این مشکل، تخمینگر Normalized Importance-Weighted Sampling (N-IS) ارائه شده است. این تخمینگر به جای تعداد نمونه ها روی مجموع وزن ها نرمال می شود.

$$\hat{\mu}_{N-IS} = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i} \tag{۱۳}$$

نشان دهید تخمینگر N-IS نااریب نیست اما

$$\lim_{n\to\infty} \hat{\mu}_{N-IS} = \mu \tag{۱۴}$$

در نتیجه N-IS در بی نهایت نااریب است.

۴. با فرض کراندار بودن $f(x)$ نشان دهید،

$$\mathbb{V}(\hat{\mu}_{N-IS}) \leq \frac{(M-m)^2}{4} \tag{۱۵}$$

که

$$M = \sup f(x) \qquad\qquad m = \inf f(x) \tag{۱۶}$$

در نتیجه واریانس تخمینگر N-IS کران بالایی مستقل از انتخاب p و q دارد.

۵. یکی از کاربردهای IS تخمین مقدار انتگرال است. فرض کنید می خواهیم انتگرال زیر را تخمین بزنیم:

$$H = \int_2^\infty e^{-\frac{x^2}{2}}\, dx \tag{۱۷}$$

فرض کنید نمونه هایی از توزیع نرمال استاندارد به صورت زیر داریم:

$$S = \{0.26, -0.27, -1.56, 0.41, 0.40, -0.02, 0.10, -1.69, -0.28, -2.53\} \tag{۱۸}$$

آ. ابتدا به روش monte-carlo تخمینی برای H بیابید.

ب. حال به روش IS و در نظر گرفتن توزیع $q = \mathcal{N}(3,1)$ تخمین دیگری از H بیابید. مقدار این تخمین و تخمین بخش قبل را با مقدار واقعی H مقایسه کنید. آیا به تخمین بهتری برای H می رسیم؟ علت آن چیست؟

سوال ۶: (نظری) Variational Inference (۱۵ نمره)

یکی از چالش های موجود در استنباط آماری، تخمین توزیع پسین متغیرهای پنهان می باشد. فرض کنیم Z متغیر پنهان مسئله و X متغیر مشاهده شده باشد. توزیع پیشین Z، $P(Z)$ و توزیع درست نمایی $P(X|Z)$ در مسئله داده شده اند. توزیع پسین Z با قاعده بیز به صورت زیر بدست می آید،

$$P(Z|X) = \frac{P(X|Z)P(Z)}{\int_{Z'} P(X|Z')P(Z')} \tag{۱۹}$$

محاسبه انتگرال مخرج در اکثر مواقع امکان پذیر نیست. از Variational Inference برای تخمین توزیع پسین استفاده می شود. برای این کار سعی می شود توزیع پسین با یک توزیع ساده شده $q(Z)$ تخمین زده شود. فاصله KL میان دو توزیع دلخواه p, q به صورت زیر تعریف می شود

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{۲۰}$$

فاصله KL معیاری برای تفاوت دو توزیع است برای تخمین زدن توزیع پسین، می توان $D_{KL}(q(Z)||P(Z|X))$ را کمینه کرد که با بیشینه کردن $L(q)$ معادل است. حال فرض کنید توزیع q به تعدادی مولفه تجریه شود،

$$q(Z) = q(Z_1, Z_2, ..., Z_n) = q_1(Z_1)q_2(Z_2)...q_n(Z_n) \tag{۲۱}$$

(آ) نشان دهید فاصله KL در $p = q$ مقدار کمینه خود را که صفر است، می گیرد.

(ب) نشان دهید

$$D_{KL}(q(Z)||P(Z|X)) = \log P(X) - \mathbb{L}(q) = \log P(X) - (\mathbb{E}_q[\log P(X,Z)] - \mathbb{E}_q[\log q(Z)]) \tag{۲۲}$$

(ج) نشان دهید $argmax_q \mathcal{L}(q)$ در معادلات زیر صدق می کند،

$$\forall 1 \leq i \leq n : \log q_i(Z_i) = \mathbb{E}_{q_j, j \neq i}[\log P(X,Z)] + const. \tag{۲۳}$$

(د) فرض کنید برای $1 \leq n \leq N$ ، $\mathbf{z}_n \in \mathbb{R}^K$ یک بردار K-of-1 دودویی تصادفی است و $\sum_{i=1}^{K} \pi_i = 1$ متغیر های تصادفی با مجموع یک هستند. متغیر X مشاهده شده و متغیرهای $Z, \pi, \mu, \mathbf{\Lambda}$ پنهان هستند. همچنین داریم:

$$\begin{aligned}
P(X|Z, \mu, \mathbf{\Lambda}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n | \mu_k, \mathbf{\Lambda}_k^{-1})^{z_{nk}} \\
P(Z|\pi) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \\
P(\pi) &\propto \sum_{k=1}^{K} \pi_k^{\alpha_0 - 1} \\
P(\mu, \Lambda) &= \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k \mid \mathbf{m}_0, (\beta_0 \mathbf{\Lambda}_k)^{-1}\right) \mathcal{W}\left(\mathbf{\Lambda}_k \mid \mathbf{W}_0, \nu_0\right)
\end{aligned}$$

که $\mathcal{W}$ توزیع ویشارت [۵] است. تخمین Variational را برای توزیع پسین متغیرهای نهان با فرض زیر بدست آورید. الگوریتم EM برای انجام این تخمین را توصیف کنید.

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) = q(\mathbf{Z}) q(\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Lambda}) \tag{۲۴}$$

[۵] Wishart