دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

آمار و احتمال

پروژه نهایی

استاد:

دکتر آزادی‌نمین

تدریسیار:

علیرضا سلحشور، آیدین روزبه

پاییز ۱۴۰۳

مهلت ارسال: ۲۴ دی

بخش اول : تفسیر فراوانی

سوال : فرض کنید دو متغیر تصادفی x و y را داریم که هر یک دارای توزیع یکنواخت در (۱ و ۰) می باشند. احتمال اینکه نزدیک ترین عدد صحیح به x/y زوج باشد، چقدر است؟

الف) ابتدا به صورت تئوری این مسئله را مورد بحث قرار دهید و مقدار احتمال را بیابید.

ب) حال با استفاده از زبان مورد نظر به ترتیب با تعداد نمونه ۱۰۰، ۲۵۰، ۵۰۰ و ۱۰۰۰ این مقدار احتمال را بدست آورید.(نمونه های تصادفی را تولید کرده و برای هر یک نزدیکترین عدد صحیح را محاسبه نمایید و نسبت تعداد زوج به کل را به عنوان احتمال گزارش نمایید.)

ج)(امتیازی) روند تغییرات این احتمال را برحسب تعداد نمونه در یک نمودار نمایش دهید.

بخش دوم: قانون بیز (شبیه سازی)

در این بخش شما باید با استفاده از قوانین و قواعدی که درباره قانون بیز یاد گرفتید، سوالات زیر را حل کرده **و شبیه‌سازی مربوطه** را انجام دهید. در **همه** بخش ها نتایج خواسته شده (مثلا توابع توزیع احتمال) **باید** نمودار شوند.

۱) فرض کنید یک سکه داریم که احتمال شیر آمدن آن برای ما نامعلوم است. به همین دلیل فرض می‌کنیم که این احتمال، خود یک متغیر تصادفی مثل x است که در ابتدا (و زمانی که هیچ مشاهده‌ای انجام نشده) یک توزیع یکنواخت بین ۰ و ۱ دارد. حال این سکه را ۲۰ بار پرتاب می‌کنیم و مشاهده می‌کنیم که ۱۴ بار شیر رخ می‌دهد. بعد از این مشاهده، تابع توزیع احتمال x چگونه تغییر می‌کند؟
نکته: برای شبیه‌سازی در زبان مورد نظر، اعداد بین صفر تا یک را با دقت یک صدم شبیه‌سازی کنید. دقت بیشتر از این برای این تمرین نیاز نیست.

۲) حال فرض کنید بعد از انجام آزمایش بالا، دوباره همان سکه را ۲۰ بار دیگر پرتاب می‌کنیم و این بار ۹ بار شیر می‌آید. توزیع x بعد از این مشاهده چگونه تغییر می‌کند؟

۳) تصور کنید ترتیب نتایج در دو آزمایش بالا فرق می‌کرد، یعنی بار اول ۲۰ دفعه پرتاب می‌کردیم و ۹ دفعه شیر مشاهده می‌شد. بار دوم ۲۰ دفعه پرتاب می‌کردیم و ۱۴ دفعه شیر مشاهده می‌شد. آیا نتیجه نهایی در هر دو سناریو یکسان است؟

۴) حال فرض کنید از ابتدا ۴۰ بار پرتاب می‌کنیم و ۲۳ بار شیر رخ می‌دهد. توزیع x بعد از این مشاهده را با حالات قبلی مقایسه کنید و نتیجه گیری خود را بیان کنید.

۵) در مشاهداتی مثل همین مثال، توزیع x در واقع از توزیع بتا پیروی می‌کند. در این توزیع پارامتر های a و b نشان دهنده چه چیزی هستند؟ نتایج بخش های قبل را با استفاده از این موضوع، توجیه کنید.

۶) فرض کنید به جای سکه، یک تاس ۳ وجهی داریم (یعنی سه خروجی ممکن A و B و C دارد). احتمال رخ دادن آن ها را به ترتیب x و y و z فرض می‌کنیم. در ابتدا هیچ تصوری نسبت به این احتمال ها نداریم و مانند حالت قبل، توزیع آن ها را یکنواخت فرض می‌کنیم . حال این تاس را ۲۰ بار پرتاب می‌کنیم. مشاهده می‌شود که ۴ بار A، ۱۰ بار B و ۶ بار C رخ می‌دهد. توزیع x و y و z بعد از این مشاهده به چه صورت تغییر می‌کند؟

۷) فرض کنید یک بار دیگر دست به آزمایش می‌زنیم و ۳۰ بار این تاس را پرتاب می‌کنیم. این بار مشاهده می شود که ۱۰ بار A، ۱۵ بار B و ۵ بار C رخ می‌دهد. توزیع جدید را رسم کنید.

۸) (امتیازی) آیا برای این حالت می‌تواند توزیعی (مانند توزیع بتا برای مثال سکه) پیدا کنید که x و y و z از آن تبعیت کنند؟

بخش سوم: برازش تابع توزیع احتمال (PDF fitting)

در ابتدا لازم است با معیار فاصله KLD (مخفف Kullback-Lieber Divergence) آشنا شویم. این معیار برای سنجش میزان شباهت دو تابع توزیع احتمال به کار گرفته می‌شود. (اگر مقدار آن صفر یا نزدیک صفر باشد یعنی دو تابع خیلی به هم شبیه هستند)

در حالت پیوسته:

$$D_{KL}[P(x)||Q(x)] = \int_{-\infty}^{\infty} P(x)\ln\left(\frac{P(x)}{Q(x)}\right)$$

در حالت گسسته:

$$D_{KL}[P||Q] = \sum_i P_i \ln\left(\frac{P_i}{Q_i}\right)$$

در این بخش قصد داریم تا اندکی با این معیار و خواص آن آشنا شویم.

مسئله ۱) فرض کنید یک رخداد تصادفی به دفعات بسیار زیاد تکرار شده و N خروجی متفاوت در آن ظاهر شده اند. حال ما قصد داریم تا هیستوگرام این رخداد تصادفی را با یک توزیع نرمال تقریب بزنیم. طبیعتا، پارامتر های توزیع نرمال را به گونه ای تنظیم خواهیم کرد که تا حد امکان شبیه به هیستوگرام مشاهده شده باشد. برای این کار نیازمند یک معیار فاصله بین دو توزیع احتمال هستیم که ما بنابر کار خودمان تصمیم گرفته ایم از معیار KLD در حالت گسسته استفاده کنیم. پارامتر های این معیار خطا را به صورت زیر فرض کرده ایم:

مشاهدات تجربی را با تابع P نشان می‌دهیم:

$$P_i = P(x_i) = Emprical$$

و تابع نرمال را با Q:

$$Q_i = Q(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

که در آن میانگین و واریانس مجهول هستند و ما قصد پیدا کردن آن ها را داریم.

به صورت تحلیلی، مقدار این دو مجهول را به گونه‌ای بیابید که خطا را حداقل کنند.

مسئله ۲) همان مسئله ۱ را این بار برای توزیع پواسون انجام دهید. دقت داشته باشید که در این حالت فقط یک مجهول λ وجود دارد.

$$Q(x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

نکته: توزیع پواسون یک توزیع گسسته است ولی نرمال پیوسته. با این حال در حالت اول ما عملا مقدار مطلق توزیع نرمال در یک نقطه خاص مثل x_i را معیار قرار دادیم. در حالت دوم هم فرض را بر این بگذارید که مقادیر x_i اعداد حسابی هستند (...,2 , 1 , 0)

مسئله ۳) (امتیازی) آیا می‌توانید همین فرایند را برای توزیع لاپلاس تکرار کنید و نتیجه را گزارش دهید؟

$$Laplace\ Distribution: f_X(x) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

پارامتر های μ و b مجهول هستند.

مسئله ۴) در ضمیمه این تمرین، یک فایل اکسل مشاهده می‌کنید که شامل ۲ ستون از اعداد است که هر ستون، ۱۰۰۰ سمپل از یک توزیع تصادفی نامعلوم هستند. برای هر یک هیستوگرام را رسم کنید. به نظر شما به توزیع های ستون اول و دوم کدام pdf بهتر نسبت داده می‌شود؟ همان pdf را برای این توزیع از داده های برازش کنید و فاصله KLD را حساب کنید.

برای محاسبه KLD:

در توزیع پواسون، ماکسیمم اعداد خروجی را پیدا کنید و برای $x_i = 0,1,\dots,max\ (x_i)$ مقدار سیگمای مربوطه را حساب کنید. برای توزیع نرمال، بازه هایی به عرض 0.2 تعریف کنید و احتمال نقطه وسط آن را برابر با نسبت تعداد رخداد های داخل آن بازه به کل بازه ها فرض کنید. مثلا اگر در بین مقادیر تصادفی ما، ۵۰ عدد وجود دارد که بین صفر تا *0.2* هستند، مقدار احتمال *0.1* را برابر با *50 / 10000* قرار دهید و به همین ترتیب یک توزیع گسسته تعریف کنید و آن را با توزیع نرمالی که خودتان پیدا کردید مقایسه کنید. همچنین، می‌توانید به جای *0.2* دقت بیشتری را معیار قرار دهید.