دانشگاه تهران، دانشکده مهندسی برق و کامپیوتر
آمار و احتمال مهندسی

تمرین کامپیوتری دوم ـ موضوع تمرین

طراح: متین بذرافشان ـ علی آریایی
سوپروایزر: علی محمدی

تاریخ تحویل: ۲۴ آذر ۱۴۰۲

۱. توزیع شرطی (۴۰) نمره

دیتاست tarbiat.csv شامل تعداد گذرهای اتوبوس BRT و مترو از ایستگاه های تربیت مدرس در بازه‌های زمانی مشخص است.

۱. برای هر دو ستون، با استفاده از $matplotlib$ یک هیستوگرام رسم کنید.

۲. اگر متغیر تصادفی X را تعداد گذرهای مترو و متغیر تصادفی Y را تعداد گذرهای BRT تعریف کنیم، با توجه به آموخته‌هایتان از درس و ماهیت مسئله و همچین شکل نمودارهایی که رسم کرده‌اید، این متغیرها از چه توزیعی پیروی می‌کنند؟ پارامتر(پارامترهای) هرکدام را محاسبه کنید.

۳. برای مشاهده توزیع مقادیر گسسته، می‌توان از density histogram استفاده کرد. این نمودار را برای ستون metro رسم کنید.

۴. حال با استفاده از $Scipy$، نمودار توزیع X را با پارامترهایی که بدست آورده اید را به نمودار بالا اضافه کنید و از درستی توزیع و پارامترهایی که مسئله‌تان را با آن مدل کرده‌اید، مطمئن شوید.

۵. متغیر تصادفی $Z = X + Y$ تعریف می‌کنیم، این متغیر از چه توزیعی برخوردار است؟ پارامتر(پارامترهای) ان را بدست آورید. با استفاده از $Scipy$ نمودار توزیع Z را رسم کنید. به این نمودار، نمودار density histogram مجموع دو ستون مترو و BRT را اضافه کنید و از درستی توزیع و پارامترهای بدست آمده اطمینان حاصل کنید.

۶. متغیر تصادفی $W \sim P(X|X + Y = n)$ را تعریف می‌کنیم. این توزیع و پارامترهای آن را به صورت تئوری محاسبه کنید.

۷. تابع جرمی احتمال W به ازای $n = ۸$ ، با انتخاب توزیع مناسبی در $Scipy$ و دادن پارامترهایی که به‌دست آورده اید به آن، رسم کنید.

۸. حال چگالی تعداد متروها را به شرط آنکه در مجموع ۸ مترو و BRT از ایستگاه های تربیت مدرس گذشته باشد را به نمودار بالا اضافه کنید. چه مشاهده می‌کنید؟

۲. تابع مولد گشتاور **(۴۰) نمره**

مسئله جمع کننده کالابرگ

مسئله‌ی جمع‌کننده‌ی کالابرگ [۱] یکی از مسائل معروف در آمار و احتمال است. فرض کنید n نوع کالابرگ داریم و از هرکالابرگ نیز به اندازه کافی موجود است. حال متغیر تصادفی X را اینگونه تعریف می‌کنیم: تعداد کالابرگ هایی که باید مشاهده کنیم به طوری که از هر نوع کالابرگ حداقل یکی دیده باشیم. برخلاف ظاهر ساده مسئله، توزیع متغیر تصادفی X پیچیده می باشد.

روش مونته کارلو [۲]: روش مونته کارلو، یک روش عددی در حل مسائل دارای محاسبات سخت و طولانی است. این روش با استفاده از نمونه گیری های تصادفی متعدد، پاسخ مسئله را پیدا می کند.

۱ـ می توان برای بدست آمدن میانگین X از روش مونته کارلو استفاده کرد. بدین منظور، تابعی بنویسید که n و k را دریافت کند، سپس k بار مسئله جمع کننده کالابرگ را به ازای n حل کند، سپس میانگین جواب ها را برگرداند.

۲ـ به ازای $n = ۱۰$ و $k = ۱۰, ۱۰۰, ۱۰۰۰$ ، تابع را اجرا کنید. مقادیر به چه عددی همگرا می شوند؟

Intro to SymPy

کتابخانه SymPy یکی از کتابخانه های قوی پایتون در زمینه ریاضیات و معادلات آن است. به‌وسیله این کتابخانه می‌توانید یک متغیر را به عنوان یک نماد [۳] تعریف کنید. یک نماد مقدار مشخصی ندارد در نتیجه می‌توان به کمک آن عملیات های ریاضی مانند مشتق و انتگرال و ... را صورت غیر عددی را انجام داد.

یک نماد را اینگونه تعریف میکنیم:

```
import sympy
x = sympy.symbols('x')
```

حال می‌توان به کمک آن نماد، یک معادله تشکیل داد:

```
equation = x ** 2 + 5 * x
```

می‌توان نماد را با هر عدد یا نماد دلخواهی جایگزین کرد:

```
equation.subs({x:10})
```

همچنین می‌توان عملیات های ریاضی را به صورت نمادین انجام داد، برای مثال می‌توان از یک معادله مشتق گرفت:

```
deriv = sympy.diff(equation, x)
```

و همچنین می‌توان مقدار آن را به ازای عدد یا نماد مورد نظر نمایش داد:

```
deriv.subs({x:10})
```

[۱] coupon collector's problem
[۲] Monte Carlo
[۳] symbol

یکی از روش‌های حل مسئله جمع‌کننده کالابرگ کاهش مسئله به n مسئله مستقل است. متغیر تصادفی X_i را تعداد کالابرگ‌هایی که باید مشاهده کنیم تا کالابرگ نوع i ام برای اولین بار مشاهده شود تعریف می‌کنیم، این متغیر دارای چه توزیعی است؟ با توجه به اینکه مسئله دارای ماهیت برنولی (موفقیت — شکست) بوده، آیا احتمال مشاهده کالابرگ نوع i برای اولین بار برای همه i های ۱ تا n یکسان است؟

$$X = X_۱ + X_۲ + X_۳ + ...X_n$$

۳ـ با استفاده از $sympy$، متغیر s را به عنوان symbol تعریف کنید. تابع مولد گشتاور متغیر تصادفی X_i را به ازای $n = ۱۰$ و i از ۱ تا n تعریف کنید.

۴ـ با استفاده از خواص تابع مولد گشتاور، تابع مولد گشتاور X را بدست آورید.

۵ـ با استفاده از خواص تابع مولد گشتاور، میانگین متغیر تصادفی X را بدست آورید. پاسخ شما باید با پاسخ قسمت ۲ یکسان شود.

۳. تخمین بیزی و استنباط بیزی (۲۰+۲۰) نمره

دیتاست digits.csv دارای ۲۰۲ عدد دست نویس (۱۰۱ عدد ۸ و ۱۰۱ عدد ۹) به اندازه ۲۸ × ۲۸ می باشد، به طوریکه مقدار هر پیکسل(که عددی بین ۰ و ۲۵۵ می‌باشد) آمده است. همچنین عدد متناظر نیز در ستون label آمده است.

۱. اعداد ۲۰۱ام و ۲۰۲ام را در متغیرهایی ذخیره کنید و سپس از دیتافریم حذف کنید.

۲. برای ساده سازی، با قرار دادن threshold به اندازه ۱۲۸ اعداد را binary کنید به طوری ۱ به معنی روشن بودن و ۰ به معنی خاموش بودن هر پیکسل باشد.(توجه داشته باشید، label ها باید بدون تغییر باقی بمانند.)

۳. ۳ـ یک عدد به طور دلخواه انتخاب کنید و با استفاده از reshape آن را به آرایه دو بعدی با اندازه۲۸ × ۲۸ تبدیل کرده و سپس با استفاده از تابع $matplotlib.pyplot.imshow$ آن را نمایش دهید.

۴. متغیر تصادفی Y را احتمال روشن بودن پیکسل ۴۰۴ام به شرط ۸ = label و متغیر تصادفی برنولی N را روشن بودن یا نبودن پیکسل ۴۰۴ام تعریف میکنیم. حال می‌خواهیم به‌وسیله تخمین بیزی، تغییرات توزیع Y را پس از مشاهده هر نمونه مشاهده کنیم. ابتدا به دلیل نداشتن اطلاعات اولیه، Y دارای توزیع یکنواخت(توزیع بتا با $a = ۱$ و $b = ۱$) می‌باشد. حال با مشاهده هر داده با توجه به توزیع پیشین(prior) متغیر تصادفی Y و همچنین مشاهدات خود از نمونه جدید(likelihood) ، توزیع پسین (posterior) متغیر تصادفی Y را به‌دست آوریم. سپس در مشاهده داده بعدی، توزیع پسین به‌دست آمده، خود توزیع پیشین می‌شود و این روند تا دیدن آخرین داده ادامه دارد. بدین منظور شما کافیست کد الحاق شده را کامل کنید و سپس آن را اجرا کنید تا تغییرات توزیع متغیر تصادفی Y را مشاهده کنید. روابط مورد نیاز، در پایین آمده است. پس از اتمام، اگر بخواهیم یک عدد را به‌عنوان احتمال روشن بودن این پیکسل اعلام کنید، چه عددی را بیان می‌کنید؟ این عدد همان مد توزیع بتا پسین یا همان میانگین این پیکسل است.

$$f(Y|N) = \frac{P(N = n|Y = p)f(Y)}{\int_{۰}^{۱} P(N = n|Y = p)f(Y)}$$

$$\int_a^b f(x)dx \approx \frac{b-a}{N}\sum_{i=۰}^{N} f(x_i); x_i = a + i\frac{b-a}{N}$$

امتیازی:

۵. متغیر تصادفی X را به صورت برداری از متغیر های مستقل تعریف میکنیم به طوری که هر x_i احتمال روشن بودن پیکسل i ام را مشخص می کند. حال می‌خواهیم $P(X|label)$ را به دست آوریم. برای این کار می توانید، میانگین پیکسل ها را برای هر پیکسل

منحصر بفرد، یکبار برای داده ها با label = ۸ و یکبار برای داده ها با label = ۹ به‌دست آورید. در انتها شما باید دو آرایه با اندازه ۷۸۴ داشته باشید.

۶. حال میخواهیم با استفاده قانون بیز، $P(label|X)$ را برای اعداد ۲۰۱ام و ۲۰۲ام به‌دست آوریم. چهار مقدار $p(label = ۸|X^{۲۰۲})$ و $p(label = ۸|X^{۲۰۲})$ و $p(label = ۹|X^{۲۰۱})$ و $p(label = ۹|X^{۲۰۱})$ را محاسبه کنید. با توجه به این که پیکسل‌ها را مستقل از هم درنظر گرفتیم و همچنین برنولی بودن توزیع هر x_i می‌توان از فرمول زیر استفاده کرد:

$$p(label|X) = \frac{p(X|label)P(label)}{\sum_{label_i} p(X|label_i)P(label_i)}$$

$$p(X|label) = \prod_{i=۱}^{۷۸۴} p(x_i|label)^{x_i}(۱ - p(x_i|label))^{(۱-x_i)}$$