آمار و احتمال مهندسی

اساتید: دکتر توسلی‌پور، دکتر وهابی
دانشکده مهندسی برق و کامپیوتر، دانشکدگان فنی، دانشگاه تهران

تمرین کامپیوتری دوم ـ تخمین بیزی، ضریب همبستگی، توزیع مشترک

طراح: شهریار عطار
سوپروایزر: مهدی جمال‌خواه

تاریخ تحویل: ۱۶ آذر ۱۴۰۳

نکات

- هدف تمرین درک عمیق‌تر مفاهیم درس می‌باشد، در نتیجه زمان کافی برای تحلیل کردن نتایج اختصاص دهید.
- در ابتدای همه‌ی سوالات seed را سه رقم آخر شماره دانشجویی‌تان قرار دهید.
- پاسخ تمرین باید به صورت یک فایل زیپ با نام CA2_[Last-Name]_[Student-Id].zip بارگذاری شود. پاسخ سوالات تئوری و تحلیل نتایج‌ها باید به صورت Markdown در فایل Notebook یا در یک فایل pdf که شامل نمودارها و نتایج نیز هست، باشد.

بیشتر بدانیم: پارادوکس دو پاکت

به شما دو پاکت داده شده و هر کدام شامل مقداری پول می‌باشند؛ یک پاکت دو برابر دیگری پول در آن قرار دارد. شما باید یکی از این دو پاکت را انتخاب کنید. در ابتدا شما هیچ دانشی ندارید و دید یکسانی نسبت به هر دو پاکت دارید و یکی را انتخاب می‌کنید، حال به شما این شانس را می‌دهند که قبل از بازکردن پاکت انتخاب شده آن را عوض کنید، آیا پاکت را عوض می‌کنید؟

بیایید با دیدگاه بیزی که با آن آشنا شده‌اید این مساله را بررسی کنیم یعنی پس از انتخاب پاکت اول (وقوع یک رخداد) دیدگاهمان درباره مقدار پول هر پاکت را به‌روزرسانی کنیم:

- فرض کنید در پاکت انتخاب شده A دلار پول باشد. (ما آن را باز نکرده‌ایم و A مجهول است)
- پاکت دیگر به احتمال $\frac{1}{2}$ شامل $2A$ دلار است. و به احتمال $\frac{1}{2}$ شامل $\frac{A}{2}$ دلار است؛ بسته به اینکه پاکت انتخاب شده مقدار کم‌تر یا بیش‌تر را داشته باشد.
- در نتیجه امیدریاضی مقدار پول پاکت دیگر برابر است با:

$$E[M] = \frac{1}{2}(2A) + \frac{1}{2}\frac{A}{2} = \frac{5}{4}A$$

- این مقدار بیش‌تر از A است پس پاکت را عوض می‌کنیم.

همین روند را دوباره برای پاکت جدید می‌توانیم پیش بگیریم و دیدگاه‌مان را به‌روزرسانی کنیم و دوباره به این نتیجه می‌رسیم که پاکت قبلی را انتخاب کنیم. به این ترتیب تا ابد به تعویض نامه‌ها باید ادامه دهیم!؟

در سوال ۲ با دیدگاه بیزی بیش‌تر آشنا می‌شوید.

۱. توزیع شرطی، توزیع مشترک ۴۰ نمره

می‌خواهیم یک Queue System را بررسی کنیم. می‌دانیم که رسیدن Customer و Service دهی هر دو از توزیع پواسون پیروی می‌کنند، پس می‌توانیم نتیجه بگیریم که زمان بین آن‌ها توزیع نمایی دارد. کدی در اختیار شما قرار گرفته است که این سیستم را با پارامترهای مشخص شده (می‌توانید پارامترها را تغییر دهید و تاثیر آن را ببینید) شبیه‌سازی می‌کند و شما نیازی ندارید این بخش را خودتان پیاده‌سازی کنید، اما مطالعه کد به شما توصیه می‌شود. حال با توجه به داده‌های تولید شده به سوال‌های بعدی جواب دهید.

۱ـ توزیع‌های wait_times و arrival_times و service_times را بررسی کنید و نمودارهای مربوط به توزیع آن‌ها را رسم کنید. برای این کار می‌توانید از تابع histplot استفاده کنید. در زمان رسم نمودار با این تابع شما می‌توانید مقدار آرگومان KDE را تغییر دهید، در مورد کاربردهای این آرگومان و مفهوم آن توضیح کوتاهی دهید و نتیجه رسم تابع زمانی که مقدار این آرگومان را تغییر می‌دهید را با هم مقایسه کنید.

۲ـ با کمک تابع jointplot از کتابخانه seaborn توزیع مشترک arrival_times و service_times را بررسی کنید.

۳ـ مراحل بالا برا برای دو متغیر تصادفی wait_times و arrival_times نیز تکرار کنید. آیا می‌توان از این نودار نتیجه گرفت که این دو متغیر با همدیگر ضریب همبستگی بالایی دارند؟ توضیح دهید.

۴ـ برای دو متغیر wait_times و arrival_times یک scatter plot رسم کنید و ضریب همبستگی بین این دو متغیر را حساب کنید.

۵ـ حال با کمک متغیر های موجود، متغیر total_times را حساب کنید و توزیع آن را بدست آورید. توضیح دهید که چرا توزیع به این شکل است.

۶ـ توزیع total_times و wait_times را به شرط اینکه arrive_time کمتر از ۵۰ باشد را بدست آورید.

۲. تخمین بیزی ۳۰ نمره

می خواهیم با کمک توزیع بتا و تخمین بیزی و میزان head آمدن یک سکه (که داده آن در اختیار شما قرار گرفته) را تخمین بزنیم. ابتدا، چون اطلاعات اولیه نداریم فرض می‌کنیم که توزیع یکنواخت دارد و سپس پس از دیدن هر داده باید توزیع پسین (posterior) را با کمک داده جدید (likelihood) و توزیع پیشین (prior) بدست آوریم. فرض های اولیه خود و مراحل انجام را به طور کامل توضیح دهید. در انتها نیز تخمین خود را از به طور کامل گزارش کنید، بار دیگر همین کار را با در نظر گرفتن $Beta(۴, ۱۰)$ به عنوان توزیع پیشین مراحل بالا را تکرار کنید. از مقایسه این دو حالت چه نتیجه‌ای می‌گیرید؟

۳. کوواریانس و ضریب همبستگی ۴۰ نمره

دیتاست energy.csv مجموع مصرف برق در تعدادی از ایالت های آمریکا در هر ساعت بین سالهای ۲۰۰۴ تا ۲۰۱۸ را نشان می دهد.

۱ـ به کمک pandas.to_datetime ، نوع ستون Datetime را به تاریخ تغییر دهید.

۲ـ به کمک توابع pandas از ستون Datetime سال، ماه و ساعت هر داده را استخراج کنید و در ستونی مجزا ذخیره کنید.

۳ـ نمودار boxplot یکی از نمودارهای پرکاربرد برای نمایش بازه تغییرات یک متغیر می باشد. به وسیله این تابع که در کتابخانه seaborn وجود دارد، مصرف برق برحسب سال را نشان دهید.(برای این کار باید محور افقی را سال، و محور عمودی را مصرف برق قرار دهید.)

۴ـ با توجه به شکل نمودار، به نظر شما واریانس مصرف برق در روزهای سال ۲۰۰۴ بیشتر است یا سال ۲۰۰۵؟ درستی ادعا خود را نشان دهید.

۵ـ نمودار boxplot ساعت‌ـانرژی را رسم کنید. و آن را تحلیل کنید

۶ـ نمودار boxplot ماهـانرژی را رسم کنید و آن را تحلیل کنید.

۷ـ ضریب همبستگی مصرف برق و ساعت را در بین ساعت ها چهار صبح تا یک بعد از ظهر را نشان دهید. آیا با نمودار بالا مطابقت دارد؟

۸ـ ضریب همبستگی مصرف برق و ماه را در بین ماه های دوم تا چهارم و همچنین بین ماه های دهم تا دوازدهم را هر کدام جداگانه به دست آورید. اعداد به دست آمده را با توجه به فصول سال در تقویم میلادی تعبیر کنید.

حال میخواهیم ببینیم که آیا می توان از ضریب همبستگی رابطه علیتی نتیجه گرفت یا خیر.
۹ـ برای این کار محتوای فایل TV_LE_Physician را بخوانید. سپس ضریب همبستگی بین متغیر های عددی را بررسی کنید. به نظر شما آیا رابطه علیتی بین این متغیر ها وجود دارد؟ اگر نه به نظر شما چه چیزی باعث این ضریب همبستگی شده؟