بسمه تعالی

دانشکده ی مهندسی برق و کامپیوتر

درس آمار و احتمال مهندسی

پروژه نهایی + هندزآن

طراح: امیرحسین ناظری | مهلت تحویل: ۱۲ بهمن ماه۹۹ | استاد: دکتر ربیعی

Hands On: (۲۰ نمره)

مطابق خواسته های مطرح شده در فایل هندزآن (Hands On.ipynb) کد آن را تکمیل کنید. همچنین حتما به توضیحاتی که در جلسه ی توجیهی داده می شود توجه کنید.

پروژه: (۸۰ نمره)

مقدمه:

هدف از این پروژه آشنایی با تحلیل داده های واقعی و پیاده سازی دانسته های شما از تئوری آمار و احتمال می باشد.

تحلیل داده:

در این بخش قصد داریم با استفاده از زبان برنامه نویسی پایتون ، به تحلیل داده های آماری واقعی بپردازیم.

dataset (FIFA2020.csv) ضمیمه شده اطلاعات مربوط به بهترین بازیکنان تاریخ فوتبال جهان تا سال ۲۰۲۰ می باشد که شامل ستون هایی مانند: ملیت (nationality)، امتیاز (overall)، وزن(weight)، قد(height)، توانایی شوت زدن (shooting) ، توانایی دریبل زدن(dribbling)، سرعت(pace) و... می باشد. درواقع هر یک از ستون ها یک متغیر تصادفی می باشد.

انجام دهید:

۱- همانطور که در فایل Dataset مشاهده می کنید، تعداد از داده های کمّی N/A (Not A Number) هستند و همچنین تعدادی از داده های کیفی Icons هستند که نشان دهنده نامعلوم بودن این مقادیر می باشد.

برای جایگزین کردن داده های کمّی نامعلوم چه راهکاری پیشنهاد می کنید؟ راهکار خود را برای داده های ستون (pace) و ستون (dribbling) پیاده کنید و دیتاست جدید را جایگزین دیتاست قبل کنید.

۲- متغییر تصادفی age را در نظر بگیرید:

الف)نمودار جعبه ای آن را رسم کنید و مقادیر (min, Q_1, Q_2, Q_3, max) و توضیح دهید هر کدام از این مقادیر به چه معنا هستند.

ب)هیستوگرام این داده را رسم کنید و سپس با scale کردن مناسب این نمودار ، PMF سنین مختلف را نیز رسم کنید.

ج) همانند قسمت ۲ بخش ۱تمرین کامپیوتری ۱ ،CDF این متغییر تصادفی را از روی هسیتوگرام رسم کنید.

د)احتمال اینکه سن بازیکنان بین ۲۰ تا ۲۴ سال باشد را با توجه به رابطه ی زیر از روی CDF آن، بیابید.

$$P(20 < x < 24) = F_x(24) - F_x(20)$$

۳- متغییر تصادفی weight را در نظر بگیرید و به صورت تصادفی n=۱۰۰ نمونه از این متغییر انتخاب کنید:

الف)میانگین ، واریانس و انحراف معیار این نمونه ها را بیابید.

ب)یکی از ابزارهایی که برای مقایسه شهودی دو توزیع به می رود ، نمودار Q-Q(quantile-quantile) می باشد.

یک نمونه ی n=۱۰۰ تایی از توزیع نرمال با σ و μ (میانگین و واریانس n نمونه) بدست آمده در قسمت "الف" ایجاد کنید سپس با استفاده این دو مجموعه nتایی و نمودار Q-Q، توزیع آماری وزن بازیکنان را با توزیع نرمال مقایسه کنید و نتیجه را تحلیل کنید.

ج)سپس قسمت "ب" به ازای n=500,2000 تکرار کنید، چه نتیجه ای می گیرید؟

۴- دو متغییر تصادفی weight و height را در نظر بگیرید:

الف)یکی از نمودارهایی که برای نمایش وابستگی دو متغیر تصادفی استفاده می شود ، نمودار پراکندگی نقطه ای (scatter plot) می باشد.نمودار پراکندگی نقطه ای این دو متغییر را رسم کنید،چه نتیجه ای می گیرید؟

ب)توزیع های این دو متغییر های تصادفی را با استفاده از نمودار Q-Q مقایسه کنید.

۵- یکی از توزیع های آماری توزیع پواسون(Poisson) است: $P(X=x)=\frac{\lambda^x e^{-\lambda}}{x!} \quad x \in \mathbb{Z}$

الف)به ازای $\lambda = 3$ تعداد نمونه ی n=۵۰۰۰ از آن برداشته و هیستوگرام آن را رسم کنید.

ب)به ازای ۵و۵۰و۵۰۰۰=n و $\lambda = 3$ با استفاده از نمودار Q-Q توزیع این نمونه ها را با توزیع نرمال مقایسه کنید و نتایج را بر اساس قضیه ی حد مرکزی (CLT) توجیه کنید.

۶- دو متغیر تصادفی weight را مستقل و pace را وابسته در نظر بگیرید:

سپس نمودار رگرسیون خطی (براساس معیار LMSE) را با استفاده از رابطه ی زیر رسم کنید و نتیجه را تحلیل کنید.

$$y = ax + b \quad a = \frac{r_{xy}\sigma_x}{\sigma_y} \quad b = \mu_y - a\mu_x$$

**نکته:

۱- برای خواندن دیتاست از تابع روبرو استفاده کنید:df = pd.read_csv('FIFA2020.csv',encoding = "ISO-8859-1")

۲-در ابتدای کد های خود جهت یکسان بودن مقادیر رندوم از تابع np.random.seed(12345679) استفاده کنید.

(امتیازی)مفهوم و شبیه سازی فرایند زنجیره ی مارکوف: (۵۰ نمره امتیازی)

۱-مفهوم زنجیره مارکوف:

زنجیره مارکوف مدلی تصادفی برای توصیف یک توالی از رویدادهای احتمالی است که در آن احتمال هر رویداد فقط به حالت رویداد قبلی بستگی دارد. زنجیره مارکوف که به افتخار آندری مارکوف ریاضی‌دان اهل روسیه این گونه نام‌گذاری شده یک سیستم ریاضی است که در آن انتقال میان حالات شمارا، از حالتی به حالت دیگر صورت می‌گیرد. زنجیره مارکوف یک فرایند تصادفی بدون حافظه‌است بدین معنی که احتمال حالت(یا حالات) بعد تنها به حالت فعلی بستگی دارد و مستقل از گذشته‌ی آن است. این نوع بدون حافظه بودن خاصیت مارکوف نام دارد. زنجیره مارکوف در مدل‌سازی دنیای واقعی کاربردهای زیادی دارد.

یکی از معروف‌ترین زنجیره‌های مارکوف که موسوم به «drunk random walk » است یک پیاده‌روی تصادفی است که در آن در هر قدم موقعیت با احتمال برابر به اندازه ۱+ یا ۱- تغییر می‌کند. در هر مکان دو انتقال ممکن وجود دارد یکی به عدد صحیح بعدی(۱+) و یکی به عدد صحیح قبلی(۱-). احتمال هر انتقال فقط به حالت کنونی بستگی دارد. برای مثال احتمال انتقال از ۵ به ۶ برابر با احتمال انتقال از ۵ به ۴ است و هر دوی این احتمالات برابر با ۰/۵ هستند. این احتمالات مستقل از حالت قبلی (که یا ۴ بوده یا ۶) هستند.

مثالی دیگر عادات غذایی موجودی است که فقط انگور، پنیر و کاهو می‌خورد و عادات غذایی او از قوانین زیر پیروی می‌کند:

او فقط یک بار در روز غذا می‌خورد.

اگر امروز پنیر بخورد فردا انگور یا کاهو را با احتمال برابر خواهد خورد.

اگر امروز انگور بخورد فردا با احتمال ۰/۱ انگور، با احتمال ۰/۴ پنیر و با احتمال ۰/۵ کاهو خواهد خورد.

اگر امروز کاهو بخورد فردا با احتمال ۰/۴ انگور و با احتمال ۰/۶ پنیر خواهد خورد.

عادات غذایی این موجود را می‌توان با یک زنجیره مارکوف مدل‌سازی کرد به دلیل این که چیزی که فردا می‌خورد (حالت بعدی) تنها به چیزی که امروز خورده‌است (حالت فعلی) بستگی دارد. یکی از ویژگی‌های آماری که می‌توان در مورد این زنجیره محاسبه کرد امید ریاضی درصد روزهایی است که انگور خورده‌است (در یک دوره طولانی).

*جهت درک بهتر فرآیند و زنجیره ی مارکوف (Markov process and Markov chain) ، لینک های زیر را مشاهده کنید.

https://www.youtube.com/watch?v=o-jdJxXL_W4

http://setosa.io/ev/markov-chains/

۲- شبیه سازی فرایند زنجیره ی مارکوف:

-مسائل زیر را با استفاده از کتابخانه های Numpy وMatplotlib شبیه سازی کنید.

انجام دهید:

۱. فرض کنید در ظرفی ۱۰ عدد قرص کامل وجود دارد.یک بیمار در هر وعده نیمی از یک قرص کامل را مصرف می کند، بدین معنا که اگر قرصی که بر می دارد کامل باشد، نیمی از آنرا مصرف و نیمه دیگر را در ظرف برمیگرداند و اگر قرص نیمه باشد آنرا مصرف می کند. در هر مرحله ، قبل از برداشتن قرص احتمال اینکه فرد یک قرص نیمه بردارد را بدست آورید.

*خروجی این شبیه سازی باید شامل آرایه ای از احتمال خواسته شده در هر مرحله و نمودار آن باشد.

**راهنمایی: جهت شبیه سازی ظرف از یک آرایه ی ۲*۱ استفاده کنید.

۲. در یک شهر هر ساعت پیشامد آن وجود دارد که یک نفر متولد شود یا فوت کند، یا هر دو پیشامد در یک ساعت باهم رخ دهند.

پیشامد فوت و تولد نیز مستقل از هم هستند. و احتمال هر کدام در هر ساعت تابعی از جمعیت شهر در آن ساعت است.

$$\begin{cases} P_{birth}(i) = \dfrac{0.5}{X_1} \\ P_{death}(i) = \dfrac{0.5}{X_1} \end{cases} \quad ,where\ X_1\ is\ the\ initial\ population$$

الف) اگر جمعیت اولیه شهر ۱۰۰ نفر باشد، به ازای ۱۰۰۰ ساعت میزان جمعیت در هر ساعت را بدست آورید

* خروجی این شبیه سازی باید شامل آرایه ای از میزان جمعیت در هر ساعت و نمودار آن برحسب ساعت باشد.

ب) ۲۰۰ =n مرتبه آزمایش "الف" را انجام دهید و نتیجه ی هر آزمایش (میزان جمعیت نهایی در پایان ساعت ۱۰۰۰ام) را در یک آرایه ی ۲۰۰*۱ ذخیره کنید. سپس نمودار این آرایه و همچنین نمودار هیستوگرام این آرایه را رسم کنید.

پ) به ازای n=10,100,1000 هیستوگرام خروجی هر کدام از این مجموعه آزمایشات را رسم کنید.چه نتیجه ای میگیرد؟

موفق باشید...

نکات تحویل:

- **حتما قوانین درس را مطالعه فرمایید.**
- **از آنجا که این تمرین بیشتر شامل تحلیلات آماری است و تحویل حضوری ندارد، لذا ۶۰٪ نمره به گزارش کار، توضیحات و نتایج خواسته اختصاص دارد لذا به همراه کد های خود می بایست گزارش کار کاملی از توضیحات و نتایج خواسته ارائه دهید، همچنین برای راحتی میتوانید بجای گزارش کار در هر قسمت با استفاده از Markdown توضیحات خواسته شده را بنویسید.**
- **فایل های شما باید شامل دو فایل ipynb. برای این پروژه و هندزآن باشد.**
- **هر قسمت از هر بخش باید در سلولی جداگانه انجام شود و سلول ها با استفاده از Markdown های مناسب از یکدیگر جدا شوند.**
- **فایل های خود را به صورت زیپ شده با فرمت Project_[full name]_[student number] در صفحه ی درس آپلود کنید.**
- **هدف این تمرین یادگیری شماست. در صورت کشف تقلب مطابق قوانین درس با آن برخورد خواهد شد.**
- **سوالات خود را در خصوص این تمرین از طریق ایمیل زیر مطرح نمایید:**

ah.nazeri1@gmail.com