



بسمه تعالی

دانشکده ی مهندسی برق و کامپیوتر

درس آمار و احتمال مهندسی

تمرین کامپیوتری ۱

استاد: دکتر ربیعی

مهلت تحویل: ۳۰ آبان ماه ۹۹

طراح: امیرحسین ناظری



مقدمه:

هدف از این تمرین آشنایی با توزیع های آماری از جمله توزیع نرمال، توابع CDF, PDF, PMF و تحلیل داده، در زبان برنامه نویسی پایتون می باشد. که جهت بررسی موضوع در بخش اول ابتدا به توزیع نرمال و تحلیل آن پرداخته می شود و سپس در بخش دوم به مقدمه ای بر تحلیل داده و ارائه ی مثال هایی به کاربرد زبان برنامه نویسی پایتون در تحلیل داده ها اشاره می گردد.

بخش اول) توزیع نرمال:

همانطور که در درس آمار و احتمالات مهندسی خواندید توزیع نرمال، یکی از مهمترین توزیع ها در نظریه احتمال است. و کاربردهای بسیاری در علم فیزیک و مهندسی دارد. این توزیع توسط کارل فریدریش گاوس در رابطه با کاربرد روش کمترین مربعات در آمارگیری کشف شد. فرمول آن بر حسب دو پارامتر امید ریاضی μ و واریانس σ بیان میشود.

همچنین تابع توزیع نرمال یا گاوس از مهمترین توابعی است که در مباحث آمار و احتمالات مورد بررسی قرار می گیرد چرا که به تجربه ثابت شده است که در دنیای اطراف ما توزیع بسیاری از متغیرهای طبیعی از همین تابع پیروی می کنند. مانند: قد افراد، نويز در سیگنال های مخابراتی، خطاهای اندازه گیری، فشار خون، نمرات یک امتحان و... .

که تابع چگالی احتمال آن به صورت زیر است.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{(x-\mu)^2}{2\sigma^2})}$$

رابطه ی (1)

که اگر $\mu = 0$ و $\sigma = 1$ باشد به تابع توزیع نرمال استاندارد می گویند.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{x^2}{2})}$$

-آمار ها نشان می دهد ، میزان بارش سالیانه بر حسب میلی متر در استان تهران از توزیع نرمال با $X \sim N(200, 25)$ پیروی می کند، در این بخش قصد داریم این توزیع بارش را مورد تحلیل قرار دهیم.

انجام دهید:

۱- یک دنباله تصادفی $\underline{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ به ازای $n=100$ با توزیع نرمال $X_i \sim N(\mu = 200, \sigma^2 = 25)$ ایجاد کنید و سپس هیستوگرام این نمونه ها را بدست آورید و CDF آن ها را رسم کنید.

۲- با استفاده رابطه ی (۱) و خروجی bins (که از تابع هیستوگرام گرفته اید) ، CDF توزیع نرمال را رسم کنید.

۳- به ازای $n=100, 500, 1000$ در subplot قسمت های ۱ و ۲ را روی یکدیگر رسم کنید و علت اختلاف و همچنین علت روند بهبود این اختلاف را به صورت کامل توضیح دهید.

۴- احتمال اینکه میزان بارش بین $\mu - \sigma$ و $\mu + \sigma$ باشد را از روی CDF قسمت ۱ بیابید.

سپس به ازای $(\mu = 100, \sigma^2 = 10)$ و $(\mu = 10, \sigma^2 = 2.5)$ نیز احتمال فوق را بیابید. علت برابر بودن این مقادیر را تحلیل کنید.

***نکته:

۱- قبل از هر بار استفاده از توابع random خود جهت یکسان بودن مقادیر رندوم از تابع `np.random.seed(12345679)` استفاده کنید.

۲- در این قسمت مجاز به استفاده از `plt.hist()` نمی باشد و صرفا می توانید از `numpy.histogram()` استفاده کنید.

بخش دوم) مقدمه ای بر تحلیل داده:

در این بخش قصد داریم با استفاده از زبان برنامه نویسی پایتون ، به تحلیل داده های آماری واقعی بپردازیم.
سه dataset ضمیمه شده که به شرح زیر می باشند:

الف)tesla-stock-price : داده های مربوط به ارزش سهام بورس شرکت تسلا از سال ۲۰۱۵ تا ۲۰۱۸ که شامل ستون های date(تاریخ) ، close (ارزش سهام در هنگام اتمام بازه خرید و فروش) ، open(ارزش سهام در هنگام شروع بازه خرید و فروش) ، low (کمترین ارزش هر سهم) ، high(بیشترین ارزش هر سهم)

ب)monthly-milk : داده های مربوط به میزان تولید شیر یک مزرعه برحسب پوند بر هر راس گاو در سال های مختلف می باشد.

ج)migrants : داده های مربوط به تعداد مهاجرین در کشور های مختلف که ستون number1000 تعداد مهاجرین در آن کشور با واحد ۱۰۰۰ نفر است و ستون under18 درصد مهاجرین زیر ۱۸ سال در آن کشور است.

انجام دهید:

تمرین ۱: با استفاده از دیتاست tesla-stock-price.csv قطعه کدی بنویسید که به ازای ورودی ، تعداد دفعاتی که ارزش سهام شرکت در شروع بازه ی خرید و فروش بیشتر از آن بوده را برگرداند.

تمرین ۲: با استفاده از دیتاست monthly-milk.csv

الف)قطعه کدی بنویسید که نمودار میزان تولید شیر مزرعه را بر حسب زمان رسم کند .

*بدلیل تعداد زیاد داده ها و xtick ها فقط تعداد کمی از xtick ها را نمایش دهید

ب)از نمودار تولید شیر بر حسب زمان چه نتیجه ای میگیرید؟

ج) قطعه کدی بنویسید که ماه بیشترین تولید شیر را در یک تناوب پیدا کند؟

تمرین ۳: با استفاده از دیتاست migrants.csv قطعه کدی بنویسید که ستون های "percentarea" و "ccode" را dropکند و سپس نمودار میله ای ۱۰ کشور اول با بیشترین مهاجر را رسم کند.

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 کدهای 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 هر بخش 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 **CA#1_[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

موفق باشید...