بنام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

درس آمار و احتمال مهندسی

تمرین کامپیوتری 1

مهلت تحویل : 14 آذر

استاد : دکتر ربیعی

مقدمه : در این پروژه قرار است با استفاده از ابزاری که در "آموزش پایتون قسمت_2" یاد گرفتید ، به کدزنی و تحلیل های آماری بپردازید .

- **سوال اول** (Hidden Markov models/HMM) :

شما در شهری عجیب زندگی می‌کنید که هوای آن یا یک روز کاملا آفتابی می‌باشد یا کاملا ابری . در روبروی خانه‌ی شما فردی عجیب به نام شایان زندگی میکند که هر روز تنها یکی از سه کار پیاده روی (Walking) ، خرید (Shopping) و تمیز کاری (Cleaning) را انجام می‌دهد . شما به عنوان یک تحلیل گر تصمیم می‌گیرید که رفتار آب و هوای این شهر و فرد مورد نظر را برای مدتی بررسی کنید و دو مشاهده زیر را جمع آوری می‌کنید :

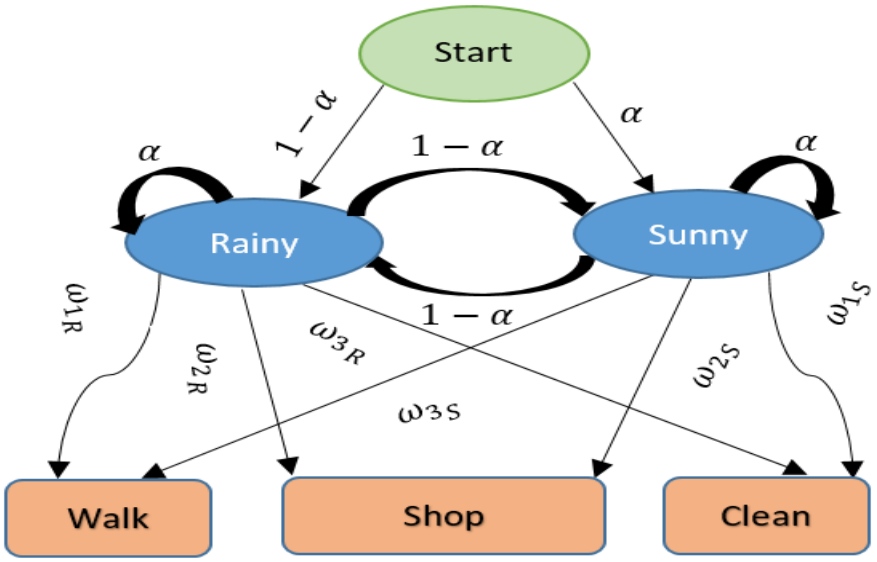
1. امروز هوا آفتابی یا ابری است ؟ (S_n : وضعیت هوا در روز n ام)
2. شایان امروز چی کاری انجام داده است ؟ (O_n : کار انجام شده در روز n ام)

بعد از مدتی به نتایج جالبی می‌رسید . بسته به این که هر روز هوای شهر چگونه باشد ، به احتمال های زیر وضعیت هوا تغییر خواهد کرد :

$$Sunny \xrightarrow{1-\alpha} Rainy \quad , \quad Rainy \xrightarrow{1-\alpha} Sunny$$

همچنین بسته به این که هر روز هوا آفتابی یا ابری باشد ، شایان به احتمال های زیر یکی از سه کار ذکر شده را انجام می‌دهد :

$$Sunny\ day \rightarrow \begin{cases} Walking \rightarrow \omega_{1_S} \\ Shopping \rightarrow \omega_{2_S} \\ Cleaning \rightarrow \omega_{3_S} \end{cases} \qquad Rainy\ day \rightarrow \begin{cases} Walking \rightarrow \omega_{1_R} \\ Shopping \rightarrow \omega_{1_R} \\ Cleaning \rightarrow \omega_{1_R} \end{cases}$$



شکل 1-1 : HMM Diagram

خلاصه‌ی داده های سوال در شکل 1-1 آورده شده است .

با استفاده از اطلاعات بدست آمده ، قصد دارید از امروز که هوا آفتابی است پیش بینی از رفتارهای شایان در روز های آینده داشته باشید .

A. بخش اول

به این منظور قرار است با زبان پایتون تابع هایی بنویسید که احتمال ها یا خروجی های مطلوب خواسته شده در زیر را به شما برگرداند :

$$P\left(S_n = \begin{cases} Sunny \\ Rainy \end{cases}\right) \quad .1$$

$$P\left(O_n = \begin{cases} Walk \\ Shop \\ Clean \end{cases}\right) \quad .2$$

$$m > n\ , P\left(S_n = Sunny \middle| O_m = \begin{cases} Walk \\ Shop \\ Clean \end{cases}\right) \quad .3$$

4. محتمل ترین وضعیت هوا در 9 روز آینده اگر :

$$O_{i=1:9} = \{\text{'Shop','Shop','Walk','Walk','Clean','Clean','Walk','Shop','Shop'}\}$$

B. بخش دوم

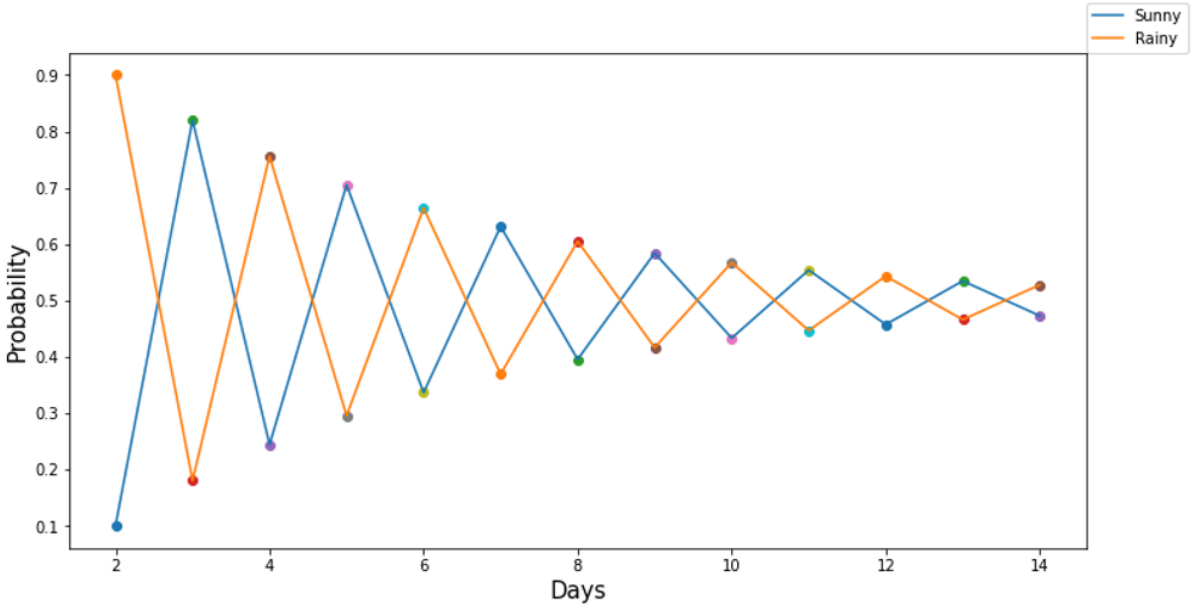
1. برای $\alpha \in np.linspace(0.1,0.9,9)$ و برای $n \in np.linspace(2,14,13)$ ، احتمال های $P(S_n = Sunny)$ و $P(S_n = Rainy)$ را در یک نمودار و بر حسب n رسم کنید .(برای هر α یک نمودار جدا خواهید داشت) و سپس به سوالات زیر پاسخ دهید :

a. با افزایش α چه تغییری در نمودار های بوجود آمده مشاهده می‌کنید ؟

b. مقدار مرزی α که از آنجا به بعد $for\ \forall n, P(S_n = Rainy) \leq P(S_n = Sunny)$ باشد ، کدام است ؟

c. با توجه به روند تغییرات احتمال های بالا با افزایش n ، $\lim_{n\to\infty} P\left(S_n = \begin{cases} Sunny \\ Rainy \end{cases}\right)$ را بدست آورید .

*راهنمایی : شکل حاصل شده برای یک α نامعلوم ، بصورت زیر می‌باشد :



شکل 2-1 : Resulted figure for an unknown α

2. برای $\alpha \in np.linspace(0.1,0.9,9)$ و برای $n \in np.linspace(2,14,13)$ ، احتمال های $P(O_n = Clean), P(O_n = Shop), P(O_n = Walk)$ را در یک نمودار و بر حسب n رسم کنید .(برای هر α یک نمودار جدا خواهید داشت) و سپس به سوالات زیر پاسخ دهید .

*از جدول زیر در محاسبات خود استفاده کنید :

	Walk	Shop	Clean
Sunny	0.3	0.6	0.1
Rainy	0.4	0.2	0.4

جدول 1-1 : Emission matrix

a. با افزایش α چه تغییری در نمودار های بوجود آمده مشاهده می‌کنید ؟

b. مقدار مرزی α که از آنجا به بعد، شرط زیر برقرار باشد :

$$For\ \forall n\ , P(O_n = Clean) \leq P(O_n = Walk) \leq P(O_n = Shop)$$

c. با توجه به روند تغییرات احتمال های بالا با افزایش n ، $\lim_{n\to\infty} P\left(O_n = \begin{cases} Walk \\ Shop \\ Clean \end{cases}\right)$ را بدست آورید .

d. تفاوت اصلی نمودار های بدست آمده در قسمت اول نسبت به این قسمت در چیست ؟دلیل را توجیه کنید .

3. برای $n \in [2,4]$ و $m \in np.linsapce(n+1, n+10, 10)$ ، احتمال های زیر را در یک نمودار و بر حسب مقادیر $\alpha \in [0.2,0.4,0.6]$ رسم کنید (یک Subplot 1*3 بر حسب مقدار α خواهید داشت) و سپس به سوالات زیر پاسخ دهید :

$$P(S_n = Sunny|O_m = Walk)\ , P(S_n = Sunny|O_m = Shop), P(S_n = Sunny|O_m = Clean)$$

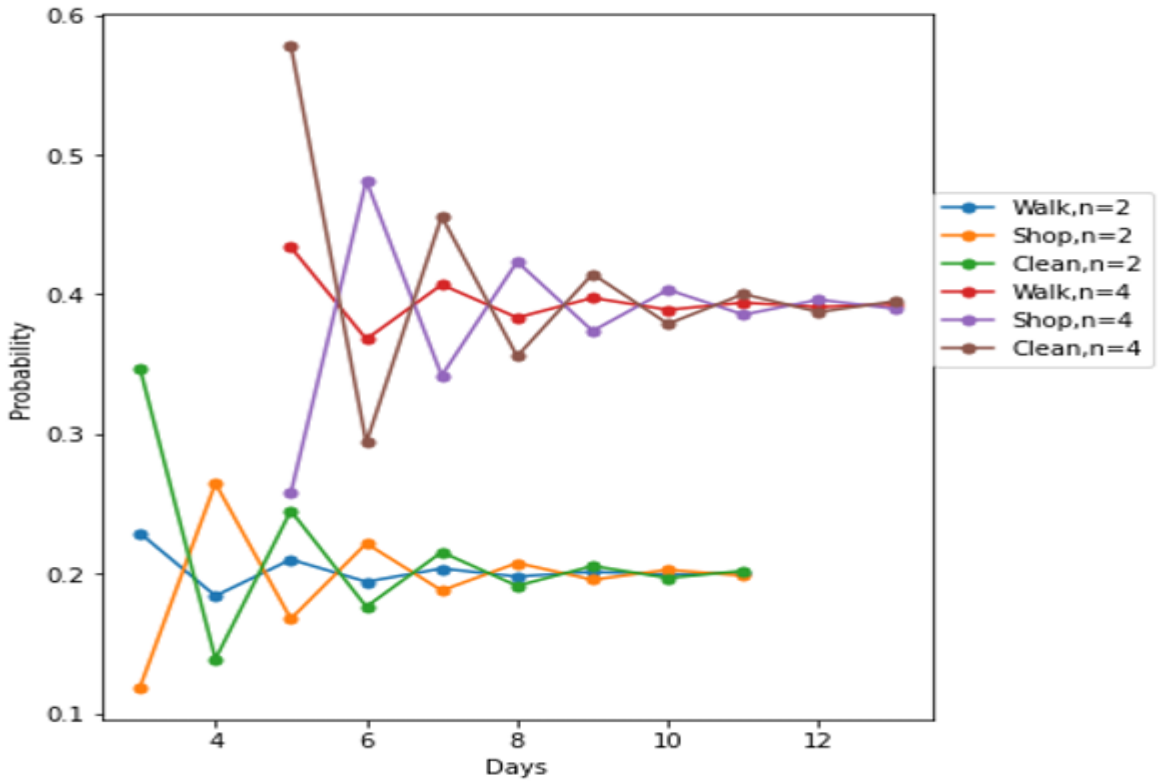
a. با افزایش α ، روند تغییرات نمودار هر یک از احتمال های بالا بر حسب روز چه تغییراتی می‌کند ؟

b. برای کدام مقدار α ، نمودار احتمال های بالا از هم قابل تفکیک هستند ؟

c. آیا در هر نمودار ، مقادیر احتمالات بدست آمده برای n=1 و n=2 رفتار یکسانی را نشان می‌دهد ؟ اگر خیر ، برای کدام مقدار α این رفتار عوض شده است ؟ دلیل را به کمک نتایج بدست آمده از قسمت اول و دوم توجیه کنید .

d. با توجه به روند تغییرات نمودار ها ، $\lim_{m\to\infty} P\left(S_n = Sunny \middle| O_m = \begin{cases} Walk \\ Shop \\ Clean \end{cases}\right)$ را بدست آورید . (در صورتی که برای هر α مقادیر تفاوت دارد ، بصورت جدا گزارش دهید)

*راهنمایی : یکی از 3 شکل حاصل شده برای α نامشخص بصورت زیر می‌باشد :



شکل 3-1 : Resulted figure for an unknown α

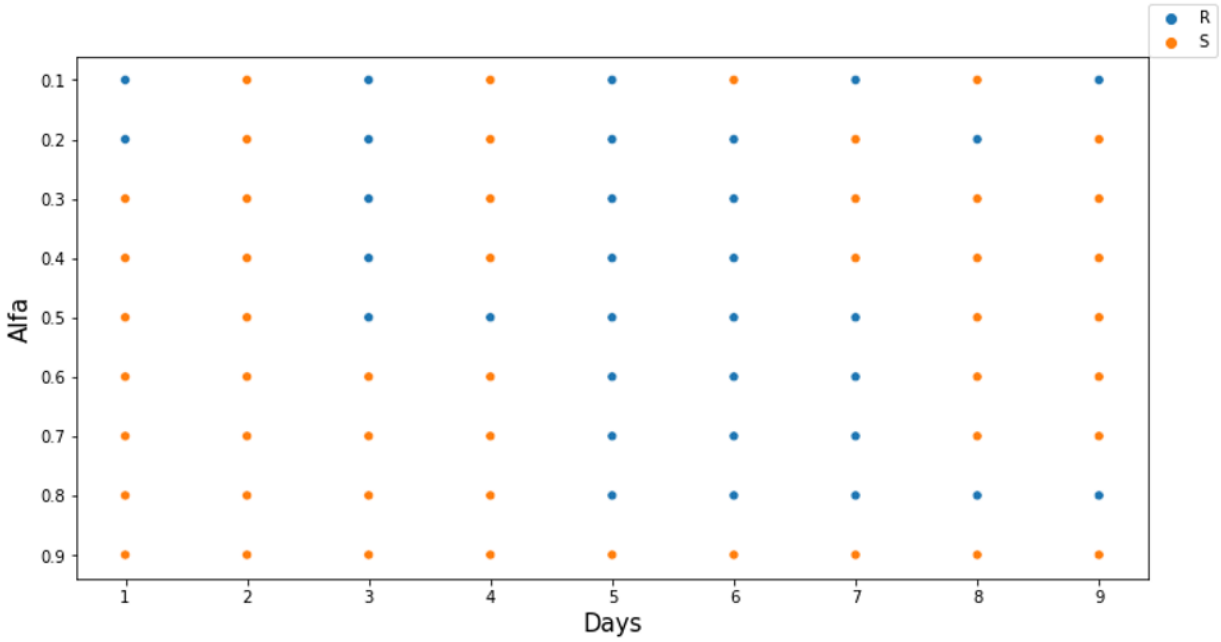
4. با کمک تابعی که در قسمت 4 بخش اول نوشتید ، می‌خواهیم تاثیر هر یک از پارامتر های α ، ω_{1_S} ، ω_{1_R} را در خروجی بدست آمده مشاهده کنیم. (در مجموع سه نمودار داریم) .
در هر نمودار پارامتر مجهول را بر حسب روز رسم کنید و "hue" در Scatterplot را وضعیت هوا در هر روز قرار دهید ("Rainy" یا "Sunny") .

همچنین پارامتر مجهول را $P_{unkown} = np.linspce(0.1,0.9,9)$ در نظر بگیرید و برای سایر پارامتر های معلوم از جدول 1-1 و $\alpha = 0.2$ استفاده کنید و سپس به سوالات زیر پاسخ دهید :

a. تغییر در کدام یک از پارامتر های مجهول بالا ، تاثیر بیشتری در تغییر وضعیت آب و هوا در روز های مختلف دارد ؟

b. برای هر کدام از پارامتر های مجهول ، مقداری که در آن بیشترین تغییر در وضعیت آب و هوا در 9 روز آینده رخ می‌دهد را اعلام کنید .

*راهنمایی : نمودار بدست آمده بر حسب پارامتر مجهول α به صورت زیر می‌باشد :



شکل 2-4-2-1: Resulted figure for setting α as an unknown parameter

➢ سوال دوم (Sites Rating Analysis)

فرض کنید که در سال 2015 هستید و امروز پنج شنبه است و شما می‌خواهید با دوستان خود به سینما بروید . از آنجا که احتمال دارد در هنگام خرید بلیط به صورت حضوری بیش از حد منتظر بمانید ، تصمیم می‌گیرید که به صورت اینترنتی از سایت Fandango فیلم دلخواه و جای خود را رزرو کنید .

در سایت Fandango ، برای هر فیلم امتیازی از 0 تا 5 (Rating) آورده شده و شما تصمیم می‌گیرید فیلم "Taken 3" که امتیاز آن برابر 4.5 است را انتخاب کرده و به تماشای فیلم بروید .

فرض کنید که بعد از دیدن فیلم ، نسبت به توقعی که از Rating آن داشتید راضی نبودید و تصمیم می‌گیرید که برسی کنید آیا سایت Fandango برای فروش بیشتر، Rating فیلم ها را بیشتر از مقدار واقعی آنها اعلام می‌کند ؟

برای این منظور از دیتاست Fandango_info.csv استفاده می‌کنید تا به تحلیل بپردازید .

در دیتاست داده شده ، دو ستون Rating و Stars آورده شده که به صورت زیر تعریف می‌شوند :

- Stars: امتیازی بین 0 و 5 که در سایت Fandango برای هر فیلم نمایش داده می‌شود.
- Rating: امتیاز واقعی موجود برای هر فیلم که از کاربران پنهان است.

A. بخش اول :

1. در ابتدا می‌خواهیم رابطه بین محبوبیت فیلم و امتیازی که به آن داده شده را پیدا کنیم . به این منظور ستون "votes" را برحسب ستون "Rating" با ابعاد و تنظیمات مناسب رسم کنید . نتیجه را گزارش دهید .
2. 10 تا فیلم اولی که بیشترین Vote را از طرف مردم پیدا کردند را گزارش دهید .
3. ابتدا فیلم هایی را که تعداد Vote برای آنها صفر است را از دیتاست حذف کنید . سپس pdf مربوط به ستون "Rating" و "Stars" را با تنظیمات مناسب در یک نمودار همراه با legend رسم کنید .
   (❋ x_axis در نمودار بدست آمده باید بین 0 تا 5 باشد) . چه نتیجه ای می‌گیرید ؟
4. در دیتاست موجود ، یک ستون جدید به نام "Stars_diff" ایجاد کنید که تفاوت بین ستون "Stars" و "Rating" می‌باشد . حال نموداری رسم کنید که تعداد تکرار هر کدام از مقادیر ستون "Stars_diff" را نشان دهد . بیشترین مقداری که سایت Fandango به Rating یک فیلم اضافه کرده است ، چقدر می‌باشد ؟

B. **بخش دوم :** در این بخش می‌خواهیم تفاوت امتیاز دهی سایت Fandango را با سایر سایت ها مقایسه کنیم و ببینیم آیا برای بقیه سایت ها هم می‌توان نتیجه گیری یکسانی داشت ؟
برای این منظور از دیتاست All_sites_info.csv برای این تحلیل استفاده می‌کنیم .

a. در ابتدا با Rotten Tomatoes شروع می‌کنیم که دو نوع امتیاز دهی در آن وجود دارد :

- Critic reviews: امتیازهایی که توسط منتقدان رسمی اعلام می‌شود.
- User reviews: امتیاز هایی که توسط کاربران داده می‌شود.

1. به کمک دو ستون مشخص شده در دیتاست جدید ، آنها را بر حسب هم با تنظیمات مناسب رسم کنید. به طور کلی چه نتیجه‌ای می‌گیرید ؟
2. برای آنکه تفاوت نظر کاربران با منتقدان را بهتر نمایش دهید ، ستون جدیدی بنام "Rotten_diff" ایجاد کنید که تفاوت بین ستون منتقدان با کاربران می‌باشد . به طور میانگین ، بین نظرات کاربران و منتقدان چند درصد تفاوت وجود دارد (* از قدر مطلق استفاده کنید)
3. هیستوگرام ستون "Rotten_diff"را یک بار بصورت عادی و بار دیگر بصورت قدر مطلق آن رسم کنید (Kde=True و bins=25 در نظر بگیرید). از روی نمودار های بدست آمده ، Outlier ها را مشخص کنید .
4. 5 فیلمی که امتیاز کاربران بیشترین تفاوت را با امتیاز منتقدان دارد را معرفی کنید و بر عکس آنرا نیز انجام دهید .

b. حال برای Metacritic و IMDB برسی می‌کنیم :

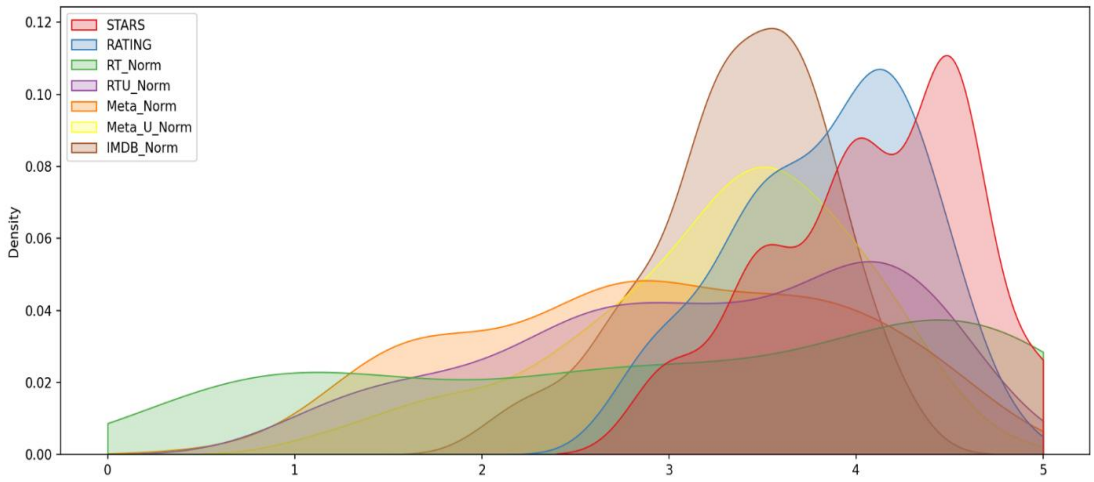
1. ستون IMDB_user_vote_count را بر حسب Metacritic_user_vote_count با تنظیمات مناسب رسم کنید . آیا رابطه مشخصی بین این دو برقرار است ؟
2. فیلم های مربوط به دو تا Outlier نمودار بدست آمده را معرفی کنید .

c. در نهایت می‌خواهیم با رسم pdf های مربوط به هر سایت ، نتیجه گیری کنیم .

1. ابتدا باید دیتاست جدید از ادغام دو دیتاست بخش اول و دوم بر حسب ستون "Film" داشته باشیم. (*راهنمایی : می‌توانید از تابع merg استفاده کنید ، در دیتاست جدید 13 ستون خواهید داشت)
2. سپس برای تمام ستون هایی که حاوی امتیاز یک فیلم می‌باشند(7 ستون) ، اعداد را بین 0 تا 5 اسکیل کنید.

3. سپس pdf های ستون های بدست آمده را در یک نمودار با تنظیمات مناسب رسم کنید (❋ دقت کنید که در نمودار بدست آمده بازه مجاز برای x-axis باید بین 0 تا 5 باشد) و نتیجه گیری کنید .

❋راهنمایی : نمونه ای از شکل نهایی بصورت زیر می‌باشد :



شکل 3-3-2-2 : Final pdfs in one plot

نکات تحویل :

1. کدهای نهایی تحویلی هر سوال را در یک فایل .ipynb در نهایت قرار دهید (هر بخش و زیر بخش ها را با فرمت Markdown از هم جدا کنید) . (در نهایت دو فایل .ipynb آماده برای آپلود خواهید داشت)
2. تمامی شکل های خروجی خواسته شده در هر زیر بخش را با زیرنویس مربوط به آن زیربخش (به شکل های در صورت پروژه دقت کنید) مشخص کرده و در گزارش خود قرار دهید. همچنین در هر زیر بخش ، متناسب با مقدار خواسته شده توضیح و پاسخ دهید . در نهایت گزارش و کد های خود را به در قالب فایل zip. و به فرمت CA_num-Last_name-std_num در صفحه درس آپلود کنید.
3. هدف از تمرین های کامپیوتری کمک به یادگیری شماست. بنابراین در صورت مشابهت بیش از حد در بخش های پروژه ، از شما نمره کم خواهد شد .
4. در صورتی که نسبت به پروژه سوال یا ابهامی داشتید ، از طریق ایمیل sh.vassef@ut.ac.ir یا در گروه تلگرامی با من در ارتباط باشید.

موفق باشید .