به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس سیستم‌های هوشمند

تمرین شماره ۴

نام و نام خانوادگی محمدمهدی رحیمی

شماره دانشجویی ۸۱۰۱۹۷۵۱۰

دی ۱۴۰۰

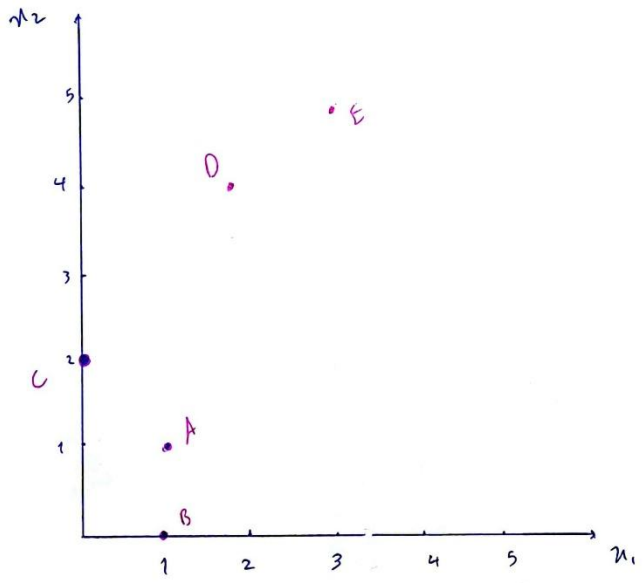
فهرست سوالات

سوال ۱: تحلیلی

در این سوال الگوریتم خوشه بندی را به دو روش به صورت دستی پیاده سازی می کنیم.

الف : خوشه بندی با روش کا-میانگین

برای حدس اولیه ابتدا نقاط را در نمودار رسم کرده و داریم:

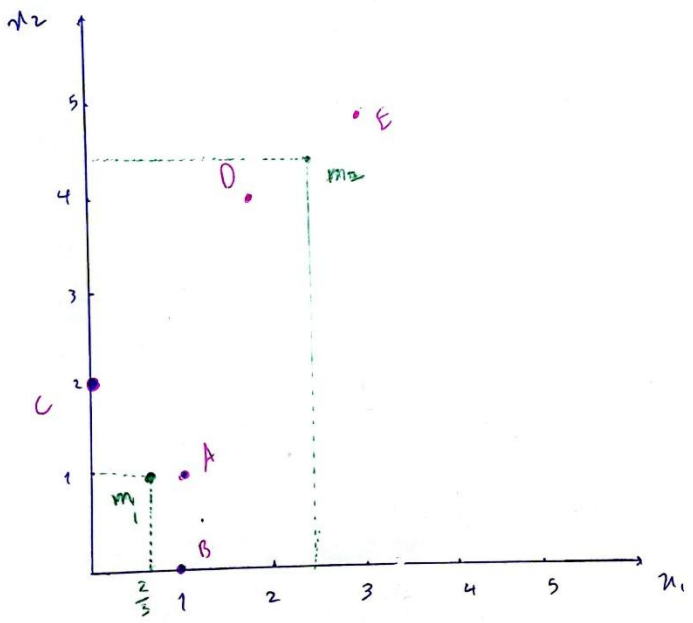


شکل ۱-۱: نمودار داده ها

همانطور که مشاهده می شود با توجه به اینکه دو خوشه داریم می توان حدس اولیه را بر اساس نمودار انتخاب کنیم که سریع تر همگرا شود. نقاط اولیه باید رندوم انتخاب شود اما در اینجا چون بعد مسئله و تعداد داده ها محدود می باشد می توان به صورت چشمی نقاط اولیه را انتخاب کنیم تا محاسبات کمتر شود. که به نظر می رسد A و B و C با هم در یک خوشه و باقی داده ها در خوشه دیگر باشد. برای نقطه اولیه برای افزایش همگرایی میانگین هر خوشه را به عنوان مرکز خوشه انتخاب می کنیم.

$$m_1 = \frac{A+B+C}{3} = \left(\frac{2}{3}, 1\right) \qquad m_2 = \frac{D+E}{2} = \left(\frac{5}{2}, \frac{9}{2}\right)$$

شکل ۱-۲ : محاسبه میانگین داده ها به عنوان مرکز خوشه



شکل ۱-۳ : محل قرار گیری مرکز خوشه ها و داده ها

حال فاصله داده ها را تا مراکز خوشه محاسبه می کنیم تا ببینیم که به کدام مرکز خوشه نزدیک تر می باشند و با توجه به نزدیک ترین مرکز خوشه آن ها را به آن خوشه انتقال می دهیم و در انتها با توجه به داده های هر خوشه مرکز خوشه را به روزرسانی می کنیم.

$d(A, m_1) = \frac{1}{3}$ $\quad d(A, m_2) = 3.8 \longrightarrow A \Rightarrow m_1$

$d(B, m_1) = 1.05$ $\quad d(B, m_2) = 4.74 \longrightarrow B \Rightarrow m_1$

$d(C, m_1) = 1.20$ $\quad d(C, m_2) = 3.53 \longrightarrow C \Rightarrow m_1$

$$m_1 = \frac{A+B+C}{3} = \left(\frac{2}{3}, 1\right)$$

$d(D, m_1) = 3.28$ $\quad d(D, m_2) = 0.70 \longrightarrow D \Rightarrow m_2$

$d(E, m_1) = 4.46$ $\quad d(E, m_2) = 0.70 \longrightarrow E \Rightarrow m_2$

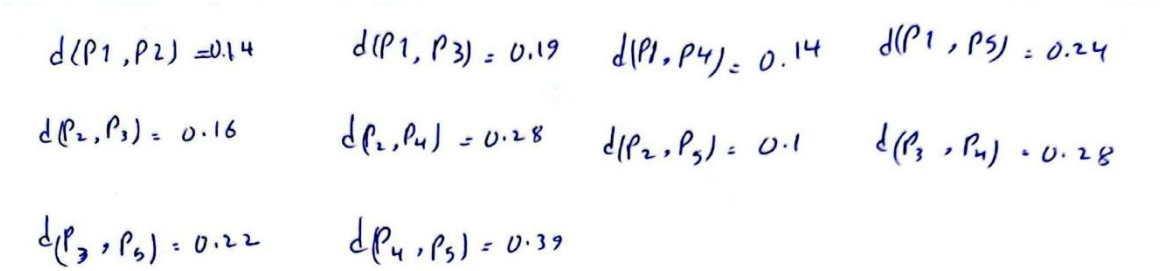
$$m_2 = \frac{D+E}{2} = \left(\frac{5}{2}, \frac{9}{2}\right)$$

شکل۱-۴ : محاسبات فاصله داده ها از مراکز خوشه ها

همانطور که مشاهده شد حدس اولیه درست بوده و داده ها به درستی خوشه بندی شده بودند و حتی مراکز خوشه نیز به درستی انتخاب شده بودند پس همان مراکز باقی ماندند.

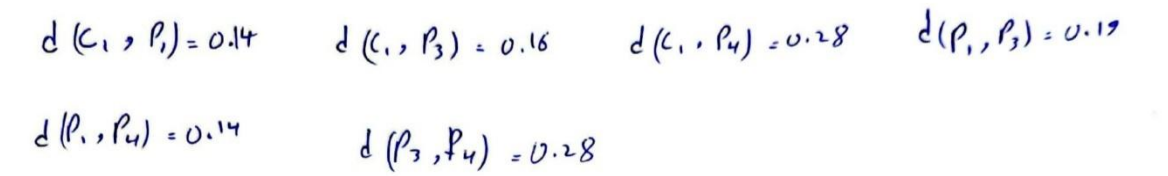
ب : خوشه بندی سلسله مراتبی

در این روش در هر مرحله فاصله داده ها را با نزدیک ترین داده ممکن می سنجیم و با توجه به آن تصمیم گیری میکنیم. در ابتدا که خوشه ای نداریم فاصله تمام داده ها از یکدیگر را بدست می آوریم.



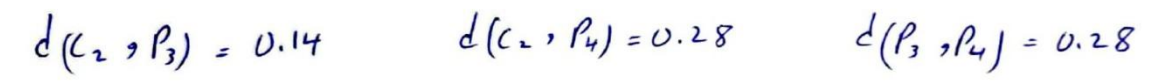
شکل ۱-۵ مرحله اول محاسبات ، بدون خوشه

همانطور که مشاهده کردید فاصله تمام داده ها با یکدیگر محاسبه شد و متوجه می شویم که P2 و P5 نزدیک ترین فاصله را دارند پس آن ها را در یک خوشه قرار داده و آن را خوشهC1 می نامیم و در مراحل بعد با توجه به این خوشه و اعضای آن فاصله را محاسبه می کنیم.



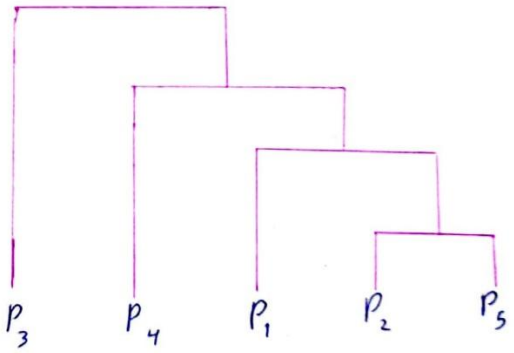
شکل ۱-۶ : مرحله دوم محاسبات ، یک خوشه

همانطور که مشاهده کردید P1 و خوشه C1 به یکدیگر نزدیک تر از بقیه موارد می باشند پس خوشه جدید که شامل P1 و P2 و P5 می باشد داریم که آن را C2 می نامیم و در مرحله بعد از آن استفاده می کنیم.



شکل ۱-۷ : مرحله سوم محاسبات

همانطور که مشاهده کردید P3 کمترین فاصله را دارد پس آن هم به خوشه ما اضافه می شود. و در انتها چون تنها گزینه باقی مانده P4 می باشد آن هم اضافه می شود. و نمودار درختی آن به شکل زیر می شود.



شکل ۱- ۸ : نمودار درختی در روش سلسله مراتبی

سوال ۲ : پیاده سازی الگوریتم خوشه بندی

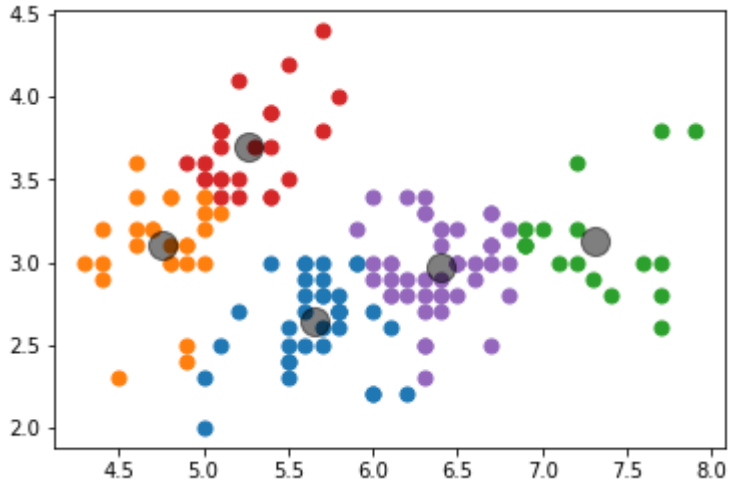
در این سوال خوشه بندی به روش کا-میانگین را انجام می دهیم که شامل دو بخش تخصیص دادن داده به خوشه و محاسبه مرکز خوشه می باشد. و برای حالت های مختلف انجام داده و نتایج را بررسی می کنیم. منظور از متغیر evaluation همان نسبتی است که در صورت سوال ذکر شده است.

در این سوال اشتباها از دو ویژگی از داده ها استفاده کرده ام. زیرا قسمت لود کردن داده ها را از دستور کار تمرین اول برداشتم که در آن جا فقط دو تا از ویژگی ها را استفاده می کرد و بنده نیز اشتباها تنها از همان دو داده استفاده کرده ام. اما با تغییر آن کد همچنان نیز جواب می دهد تنها نمودار ها بر اساس دو ویژگی رسم شده است و به دلیل کمبود وقت بنده به همین شکل ضمیمه کرده ام.

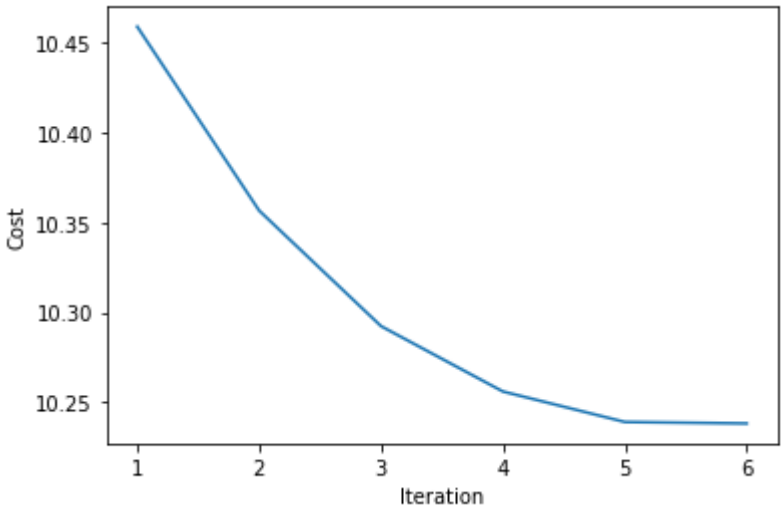
الف : تاثیر تعداد خوشه ها

برای این بخش کدی داریم که از توابع مختلف تشکیل شده است. تابعی برای تخصیص داده ها به یک خوشه و تابعی برای محاسبه مرکز خوشه و تابع کلی که از دو تابع ذکر شده استفاده می کند و الگوریتم را تکرار می کند تا همگرا شود. تابعی برای محاسبه هزینه و رسم کردن نمودار و همچنین ارزیابی الگوریتم با توجه به روشی که در صورت سوال ذکر شده است نیز داریم.

ابتدا برای زمانی که تعداد خوشه ها ۵ باشد داریم:



شکل۲-۱ : شمایی از خوشه ها و مراکز و داده های هر خوشه

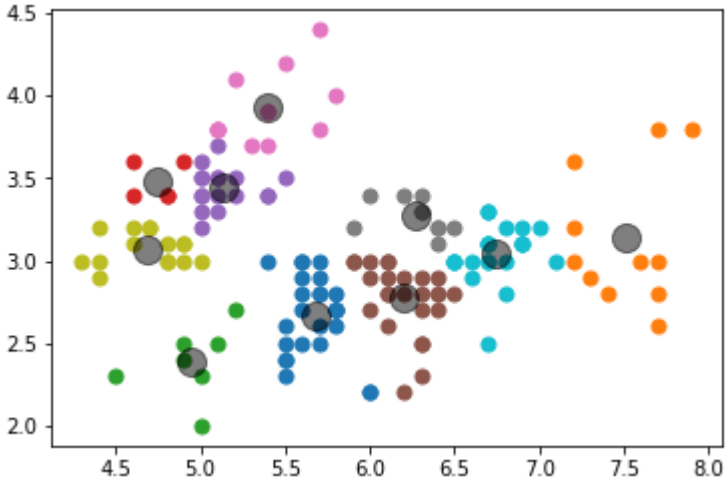


شکل۲-۲ : تابع هزینه به ازای هر بار اجرای الگوریتم

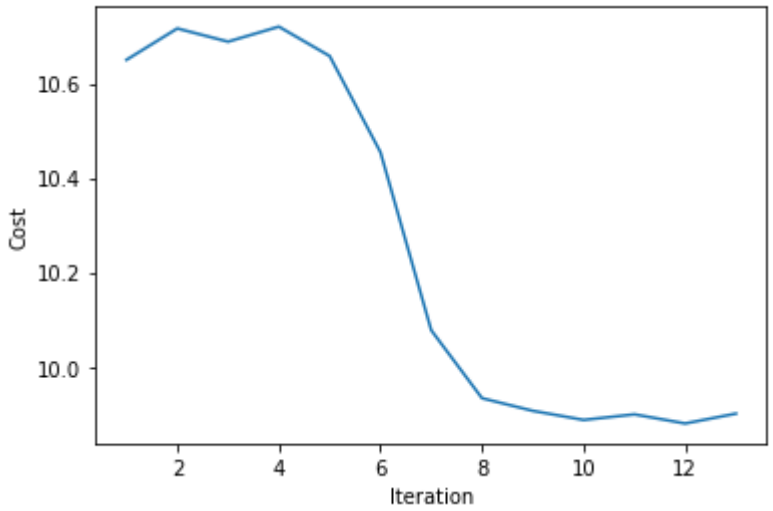
If k = 5 in 6 iteration evaluation is 0.0662747171489313

شکل ۲-۳ : خروجی کد برای تعداد خوشه ۵

همانطور که مشاهده کردید برای تعداد خوشه ۵ موارد خواسته شده آورده شد حال برای تعداد خوشه ۱۰ مراحل را تکرار می کنیم و خروجی ها را نمایش می دهیم.



شکل۲-۴ : شمایی از خوشه ها و مراکز و داده های هر خوشه

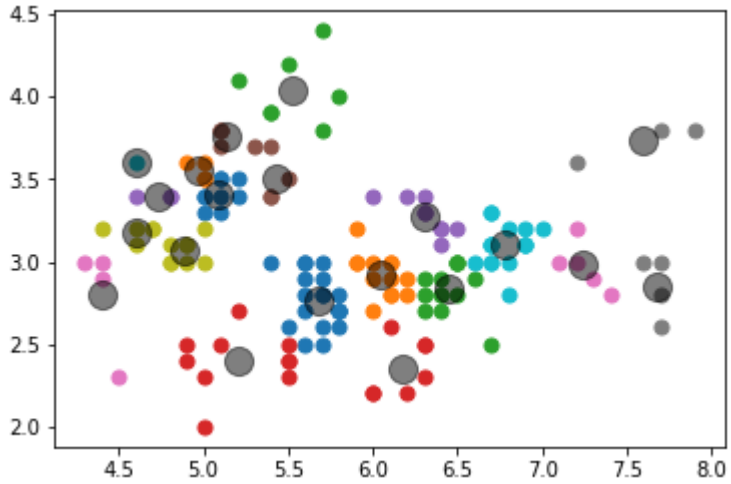


شکل۲-۵ : تابع هزینه به ازای هر بار اجرای الگوریتم

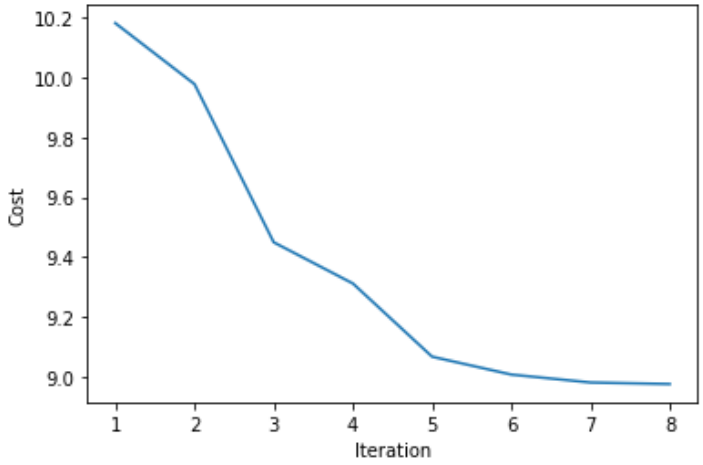
```
If k =  10 in  13   iteration evaluation is  0.021963931899853906
```

شکل ۲-۶ : خروجی کد برای تعداد خوشه ۱۰

حال برای ۲۰ خوشه داریم:



شکل۲-۷ : شمایی از خوشه ها و مراکز و داده های هر خوشه



شکل۲-۸ : تابع هزینه به ازای هر بار اجرای الگوریتم

If k = 20 in 8 iteration evaluation is 0.006692058776211913

شکل ۲-۹ : خروجی کد برای تعداد خوشه ۲۰

همانطور که مشاهده کردی دبرای حالت های مختلف موارد خواسته شده را نمایش دادیم. و اما از نظر عملکرد ۲۰ خوشه عملکرد بهتری داشت. پس تعداد خوشه مناسب ۲۰ می باشد زیرا نسبتی که از ما خواسته شده بود تا محاسبه شود برای مقدار ۲۰ کمترین میزان خود رسید.

ب : تاثیر تعداد آزمایش

در این بخش برای هر تعداد خوشه به تعداد ۲۰۰ بار الگوریتم را اجرا کرده سپس مقدار هزینه و نسبت گفته شده را محاسبه کرده و در نهایت از داده ایی که به دست آوردیم واریانس و میانگین می گیریم.

برای تعداد خوشه ۵ داریم:

If k = 5 :
Mean of evaluation is 0.0687715712960851 and variance of evaluation is 9.890071345204933e-06
Mean of cost is 10.339654956607237 and variance of cost is 0.022587861348128863

شکل ۲- ۱۰ : میانگین و واریانس ، نسبت گفته شده و هزینه برای ۵ خوشه

If k = 10 :
Mean of evaluation is 0.02228354790075541 and variance of evaluation is 1.1551150326198643e-06
Mean of cost is 9.937342335259515 and variance of cost is 0.0648504344936697

شکل ۲- ۱۱ : میانگین و واریانس ، نسبت گفته شده و هزینه برای ۱۰ خوشه

If k = 20 :
Mean of evaluation is 0.007281798908908153 and variance of evaluation is 2.3517778410437826e-07
Mean of cost is 9.116999355598137 and variance of cost is 0.08777458955570426

شکل ۲- ۱۲ : میانگین و واریانس ، نسبت گفته شده و هزینه برای ۲۰ خوشه

همانطور که مشاهده کردید در تعداد زیادی تکرار آزمایش نیز باز هم تعداد خوشه ۲۰ عملکرد بهتری داشت و مقدار نسبت گفته شده و تابع هزینه بسیار کمتر می باشد که تنها موردی که در آن از بقیه بیشتر می باشد واریانس تابع هزینه می باشد.

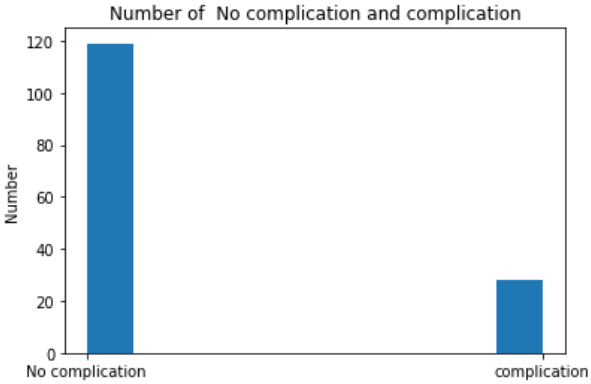
سوال ۳ : یادگیری نیمه نظارت شده

در این سوال ابتدا طبقه بند رگرسیون لاجستیک را ایجاد می کنیم سپس عملکرد آن را بررسی می کنیم سپس از داده های بدون برچسب استفاده کرده و طبقه بند را بهبود می دهیم.

الف : رگرسیون لاجستیک

رگرسیون لاجستیک یک طبقه بند باینری می باشد که به شکل s می باشد و خروجی آن به صورت احتمال است که برای مثال اگر احتمال بالای ۵۰ درصد باشد آن را به طبقه یک نسبت می دهد. این طبقه بند کاربرد زیادی دارد.

در این بخش تابعی را ایجاد کردیم که ورودی ما کل داده ها می باشد و به نسبت گفته شده داده ها را تقسیم می کند سپس نمودار هیستوگرام را برای داده هایی که برچسب دارند را رسم یم کنیم که داریم:



شکل ۳-۱ : نمودار هیتوگرام داده های برچسب دار

همانطور که مشاهده می شود داده ها متعادل نمی باشند و تعداد داده هایی که عوارض ندارند بیشتر می باشد.

ب : ارزیابی طبقه بند

کد این قسمت همراه قسمت قبل بوده و در این بخش با استفاده از کتابخانه ها و دستورات آماده آن ها یک رگرسیون لاجستیک را آموزش داده و بر روی داده های تست آزمایش می کنیم. که بعد از پیش بینی موارد خواسته شده را نمایش می دهیم.

```
F1 score =  0.8992974238875878
Accuracy =  0.9529926209346816
col_0          0   1
complication
0           2719   7
1            165  768
```

شکل ۳-۲ : دقت و ماتریس آشفتگی و معیار F1

ج : یادگیری نیمه نظارت شده

در این بخش با توجه به داده هایی که برچسب دارند ابتدا یک طبقه بند رگرسیون لالجستیک را آموزش می دهیم. سپس در مرحله بعد از این طبقه بند استفاده کرده و داده های آموزش بدون برچسب را بررسی می کنیم. آن هایی که با احتمالی بزرگ تر از احتمال آستانه در دسته ای قرار بگیرند به آن ها برچسب اختصاص می دهیم. حال روند قبل را تکرار میکنیم تا مرحله ای که یا داده های بدون برچسبی باقی نماند یا داده ای با احتمال حداقلی وجود نداشته باشد.در ادامه برای احتمال ۰,۷ این آزمایش را انجام می دهیم. و داریم.

```
Iteration 0
9955  data added
Iteration 1
640  data added
Iteration 2
110  data added
Iteration 3
33  data added
Iteration 4
14  data added
Iteration 5
6  data added
Iteration 6
2  data added
Iteration 7
0  data added
```

شکل۳-۳: تعداد داده های اضافه شده در هر مرحله

```
F1 score =  0.9547738693467336
Accuracy =  0.977862804044821
col_0          0   1
complication
0           2723   3
1             78  855
```

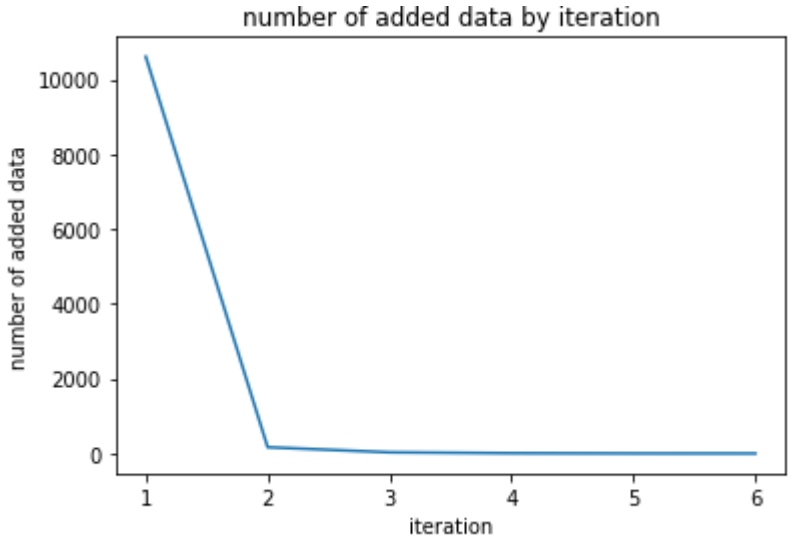
شکل ۳-۴: دقت و ماتریس آشفتگی و معیار F1

همانطور که مشاهده می کنید عملکرد طبقه بند بهبود یافته است.

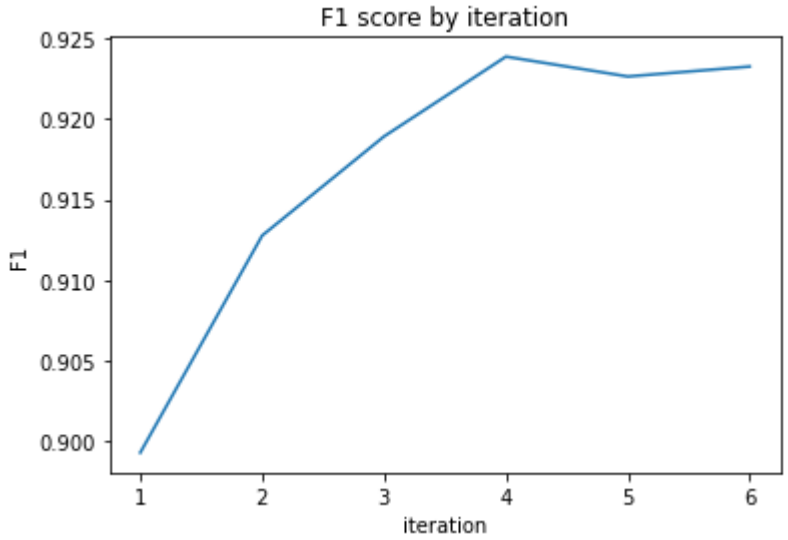
د: شرایط استفاده

در این بخش آن احتمالی که داده ها را به طبقه خاصی تخصیص می داد را عوض کرده و نمودار های خواسته شده را رسم می کنیم.

ابتدا برای حد استانه ۰,۵۵ داریم:

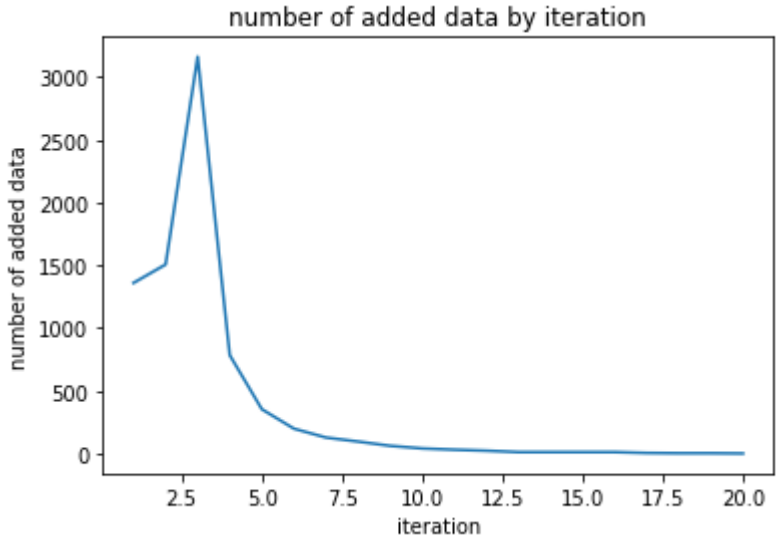


شکل۳-۵ : نمودار داده های اضافه شده بر حسب تکرار با حد استانه ۰,۵۵

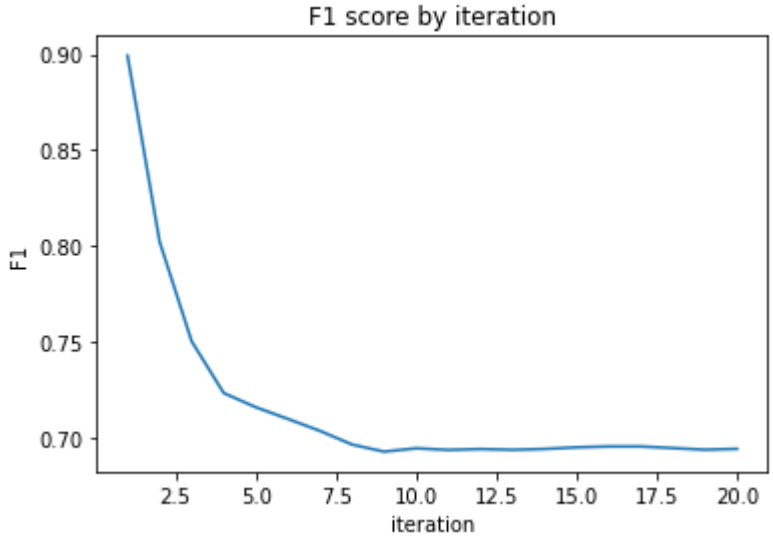


شکل ۳-۶ : نمودار معیار F1 بر حسب تعداد تکراربا حد استانه ۰,۵۵

حال برای حد استانه ۰,۹۹ داریم:

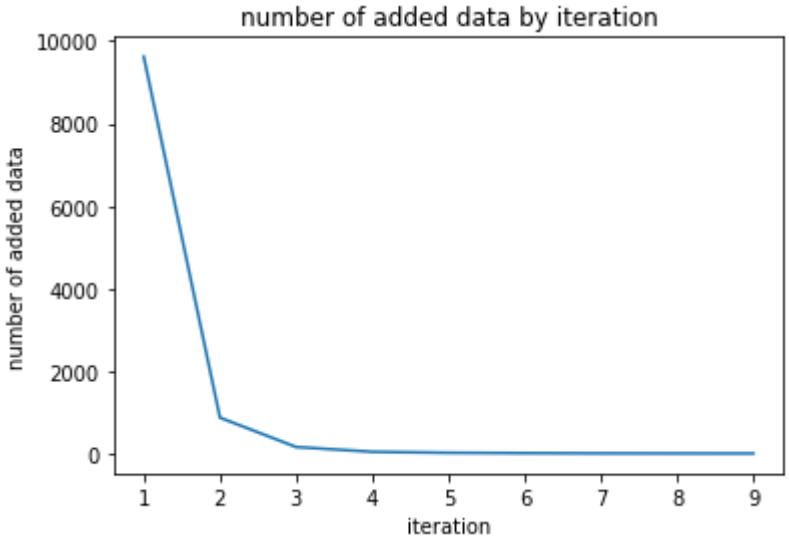


شکل۳-۷ : نمودار داده های اضافه شده بر حسب تکرار با حد استانه ۰,۹۹

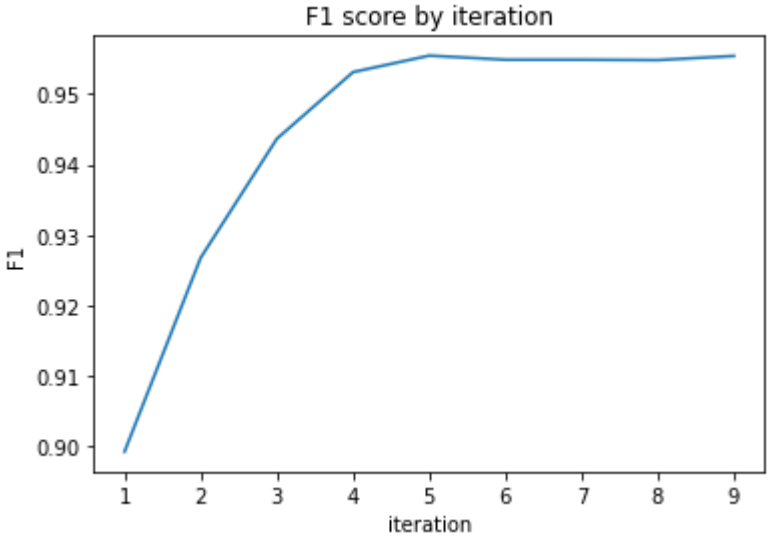


شکل ۳-۸: نمودار معیار F1 بر حسب تعداد تکراربا حد استانه ۰,۹۹

در این حالت معیار F1 کمتر شده است. حال با جستوجو باینری سعی می کنیم حد.د عدد مناسب را پیدا کنیم پس الان برای حد استانه ۰,۷۵ تست می کنیم.



شکل۳-۹ : نمودار داده های اضافه شده بر حسب تکرار با حد استانه ۰,۷۵



شکل ۳-۱۰: نمودار معیار F1 بر حسب تعداد تکراربا حد استانه ۰,۷۵

برای این حد استانه دقت و معیار F1 بسیار بهبود یافتن و با همین روش می توان بهترین احتمال را پیدا کرد.

سوال ۴ : مقدمات احتمال

الف : سوالات تحلیلی

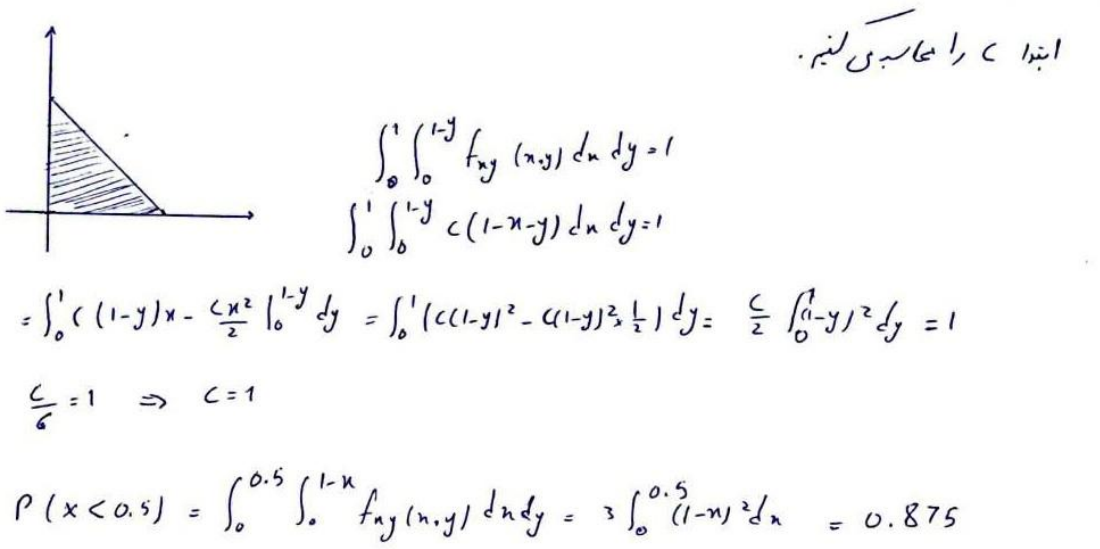
الف-۱:

چون سکه خراب می باشد نمی توان با یک بار انداختن آن مسئله را حل کرد. فرض میکنیم که اگر اول شیر و بعد از آن خط آمد نفر اول برنده شده است و اگر دفعه اول خط آمد و بعد از آن شیر آمد نفر دوم بنده می باشد. البته برعکس حالت ذکر شده نیز می شود زیرا احتمال دو حالت یکسان است و اما اگر حالت دیگری به غیر از حالات نام برده پیش آمد مانند دو شیر پشت سر هم عملیات را باید تکرار کنیم. و اما به سراغ این می رویم که چرا احتمال حالت ها ذکر شده یکی می باشد . با توجه به فرض سوال احتمال شیر و بعد از آن خط به شکل p(1-p) می باشد و برای حالت خط و بعد از آن شیر داریم (1-p)p است که مشاهده می کنیم احتمال ها برابر می باشند.

الف-۲:

الف-۳:

ابتدا خواسته اول را محاسبه می کنیم که داریم:



شکل ۴-۱:محاسبات خواسته اول

احتمال محاسبه شده نشان می دهد که یکی از افراد زمان کمتری از نصف را کار کرده است.

و برای خواسته دوم داریم:

$$E[X+Y] = \int_0^1 \int_0^{1-x} (x+y) f_{xy}(x,y)\, dx\, dy = 6 \int_0^1 \int_0^{1-x} (x+y)(1-x-y)\, dy\, dx$$

$$= \int_0^1 (1+2x^3-3x^2)\, dx = (-x^3 + \frac{x^4}{2} + x)\Big|_0^1 = 0.5$$

شکل ۴-۲ : محاسبات خواسته دوم

این خواسته نشان دهنده میانگین زمان صرف شده هر دو فرد می باشد.

ب : سوالات شبیه سازی

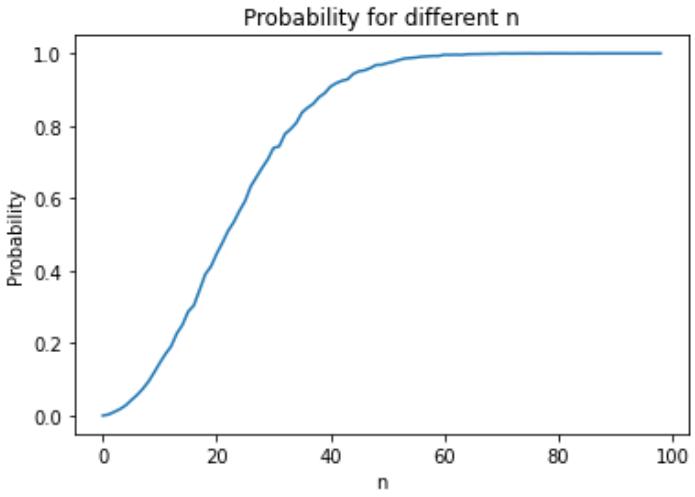
ب-۱:

در این بخش اول تابعی را ایجاد کرده که با توجه به تعداد n دلخواه به صورت تصادفی روز هایی تولید کرده و در لیستی قرار می دهیم و یک لیست که در آن عوض های تکراری نباشد از لیست قبلی می سازیم. حال اگر طول این دو لیست برابر نباشد یعنی عضو تکراری در لیست اولیه بوده است. از همین روش استفاده کرده و برای تعداد ۵۰ نفر و تکرار ۱۰۰۰۰ احتمال زیر را داریم:

```
Probability when n =  50  is equal :  0.97
```

شکل : محاسبه احتمال برای جمعیت ۵۰

حال برای خواسته قبلی حلقه ای ایجاد کرده تا مرحله قبل را برای جمعیت یک تا ۱۰۰ اجرا کرده و نمودار احتمال را بر حسب جمعیت رسم می کنیم.



شکل : نمودار احتمال بر حسب تعداد جمعیت

ب-۲:

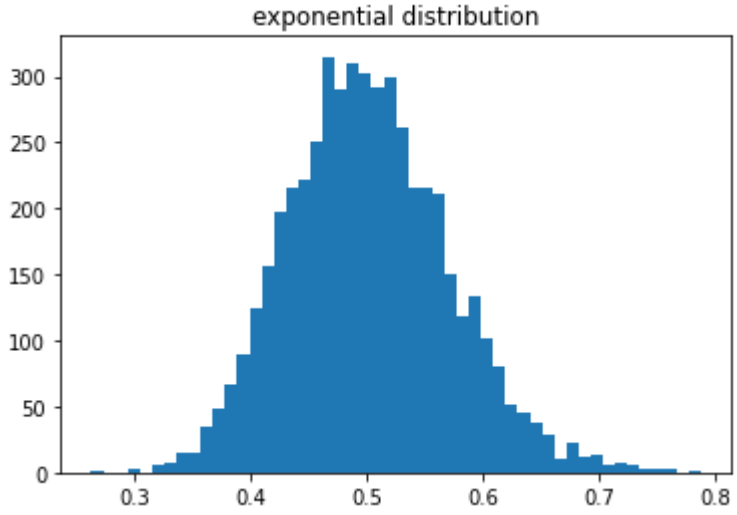
ابتدا داده ای با توزیع نمایی با نرخ دو به اندازه ۱۰۰۰۰ ایجاد کرده سپس نمونه ای ۵۰ تایی از داده ها را انتخاب کرده و میانگین آن را محاسبه می کنیم. این عمل را ۵۰۰۰ بار تکرار می کنیم. با توجه به اطلاعات به دست آمده می توان نمودار را رسم کرد و همچنین اگر میانگین و واریانس را برای تمام داده ها حساب کنیم می توان آن را با مقادیری که از قضیه حد مرکزی هست مقایسه کرد که با توجه به قضیه حد مرکزی و نرخ داده شده داریم:

$$\lambda = 2\ \mu = \frac{1}{\lambda}\ \sigma = \frac{1}{\lambda} \ \Rightarrow \mu = 0.5\ \sigma = 0.5$$

حال اگر کد را اجرا کنیم ابتدا میانگین و واریانس را محاسبه کرده و داریم:

```
Mean of exponential distribution =  0.5014169375475495
Standard deviation of exponential distribution =  0.5016470387856562
```

شکل : میانگین و واریانس برای توزیع نمایی



شکل : توزیع میانگین نمونه ها از توزیع نمایی

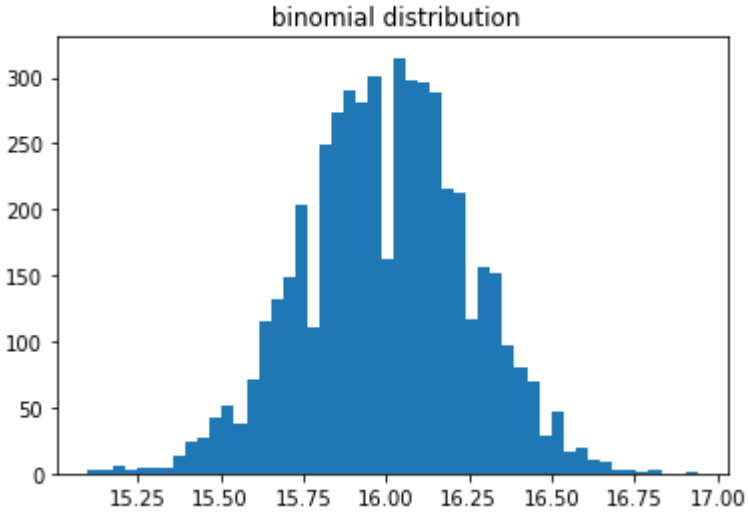
حال مراحل قبل را تکرار کرده اما این دفعه با توزیع دو جمله ای انجام می دهیم. در ابتدا مقادیر میانگین و واریانس با استفاده از قضیه حد مرکزی بدست می آوریم.

$$n = 20\ p = 0.8\ \mu = np\ \sigma = \sqrt{np(1-np)} \Rightarrow \mu = 16\ \sigma = 1.7888$$

و حال برای این قسمت نیز داریم.

```
Mean of binomial distribution =  15.999108
Standard deviation of binomial distribution =  1.7758364801794109
```

شکل : میانگین و واریانس برای توزیع دو جمله ای



شکل : توزیع میانگین نمونه ها از توزیع دو جمله ای

به طور کلی این قضیه ارتباط بین متغیر های تصادفی و توزیع نرمال را بیان می کند که نمونه ای از یک توزیع را با استفاده از امید ریاضی و واریانس قابل بیان است که این را به صورت عملی نمایش دادیم.