باسمه تعالی

تمرین سری ششم درس یادگیری ماشین

استاد درس: جناب آقای دکتر باباعلی

نام دانشجو: ایمان کیانیان

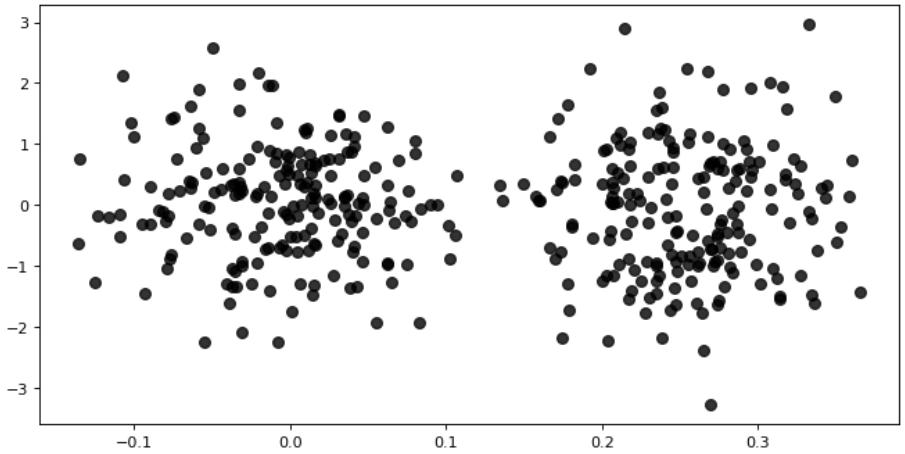
شماره دانشجویی: ۶۱۰۳۰۰۲۰۳

۱- مقدمه

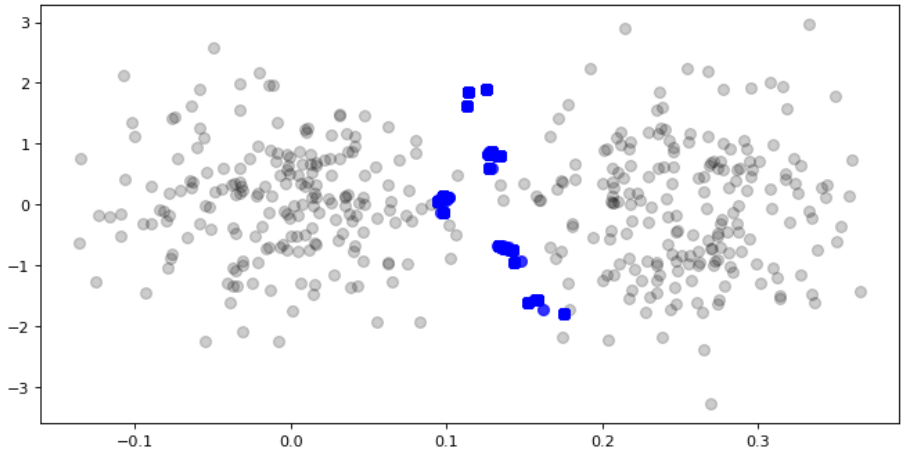
در این تمرین قرار است با استفاده از ۳ دیتاست داده شده، عمل خوشه بندی یا clustering را انجام دهیم. دیتاست اول شامل ۴۰۰ سمپل و دو فیچر است. دیتاست دوم شامل ۴۵۰ سمپل و ۲ فیچر و دیتاست سوم ، شامل ۵۰۰ سمپل و ۲ فیچر است. در قسمت های بعد خواسته های صورت سوال را مطرح کرده ، نتیجه را ذکر میکنیم و تحلیل لازم را برای هر نتیجه ارائه میکنیم.

۲- قسمت a

در این سوال ابتدا از ما خواسته شده داده هایی داده های درون فایل اول مروبط به کلاسترینگ را نمایش دهیم. داده ها در فضای دو بعدی به صورت زیر است:



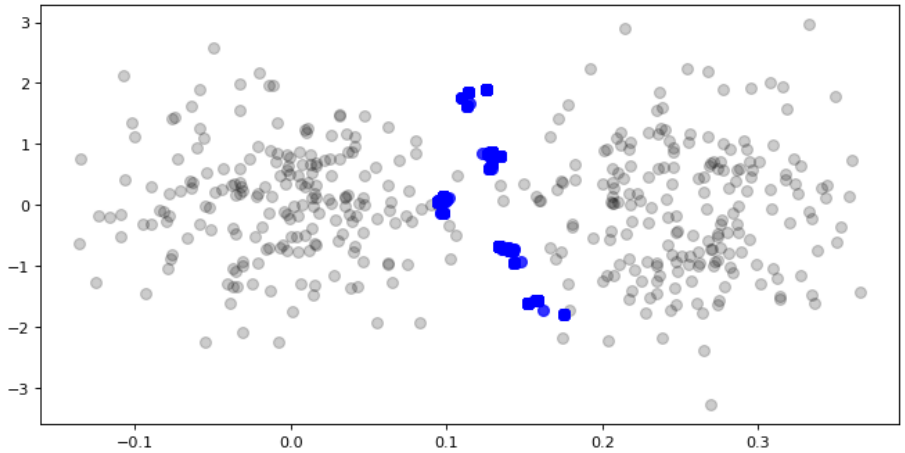
که به صورت چشمی دیده میشود که احتمالا دارای دو خوشه هستیم . یعنی اگر از تعریف density based استفاده کنیم دو خوشه خواهیم داشت. اما چون به ما در سوال گفته شده نه K=5 را در نظر بگیریم احتمالا مراکز به هم نزدیک خواهند شد. در این قسمت از ما خواسته شده ۲۰۰ بار بصورت مستقل kmeans را بصورت مقدار دهی اولیه random اجرا کنیم و هر بار centroid ها را در کنار داده ها نمایش دهیم که شکل به صورت زیر خواهد بود:



همانطور که مشاهده می شود این کلاستر نتوانسته حالت یا شکل مطلوبی داشته باشد چون داده های ما تقریبا ۲ کلاستر دارد ولی ما خواستیم ۵ کلاستر فیت کنیم. از طرفی خاصیت تصادفی بودن انتخاب ها کمی ما را دچار اشتباه میکند.

۳- قسمت b

در این قسمت از ما خواسته شده است با استفاده از همان روش بالا فقط استفاده از روش `k-means++` یک مقدار دهی اولیه بهتری برای `centroid` ها در الگوریتم `kmeans` پیدا کنیم. اما چون داده های ما ظرفیت ۵ کلاستر را ندارد ، نمیتوانیم `centroid` های اولیه خوبی پیدا کنیم.

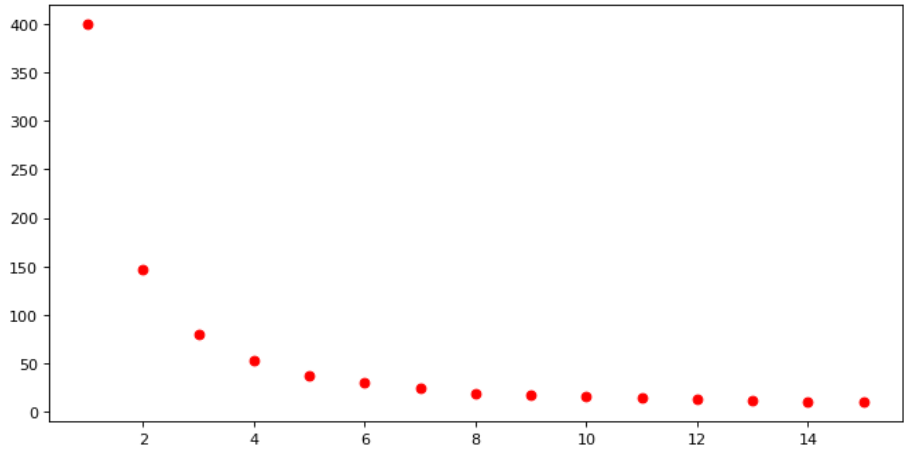


فرق این روش با روش قبل در این است که کمی از حالت random خارج میشویم و انتخاب های ما بر اساس معیار زیر شکل میگیرد.
اولین centroid در هر مرحله random اختیار میشود و انتخاب های بعدی برای centroid های بعدی (۴ تا) به صورت زیر است:

$$\Pr(\boldsymbol{c}_j = \boldsymbol{x}^{(i)}) \propto \min_{k<j} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{c}_k \right\|^2$$

۴- قسمت c

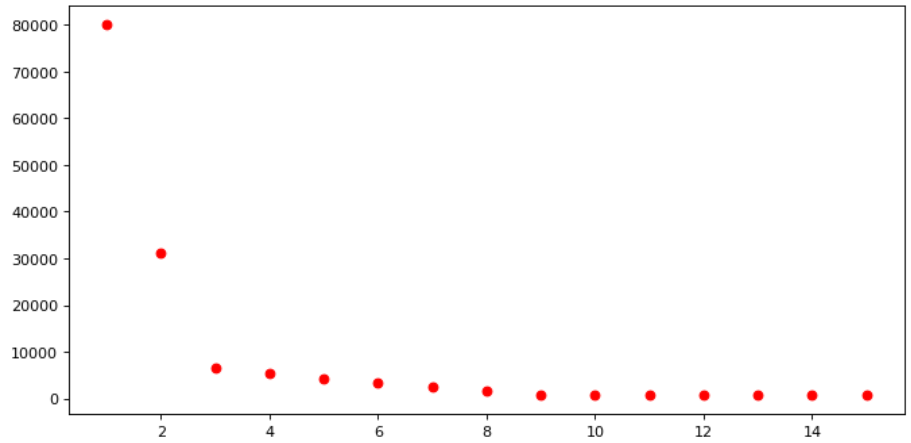
با استفاده از داده های سری اول، که شامل ۴۰۰ داده (سمپل) بودند نمودار خطا ، تعداد کلاستر ها را به ازای ۲۰۰ بار مقداردهی به کلاستر ها و انتخاب بهترین آنها و به ازای k = 1, 2 ... , 15 ، را رسم کردیم. نمودار مذکور به شکل زیر است:



از ما خواسته شده است با استفاده از این شکل نقطه knee را مشخص کنیم. به وضوح در نقطه ی مربوط به K=2 **یا** K=3 ، محل knee یا elbow است یعنی در این نقطه میتوانیم K قابل قبولی داشته باشیم که مطابق انتظارات ما هم بود.

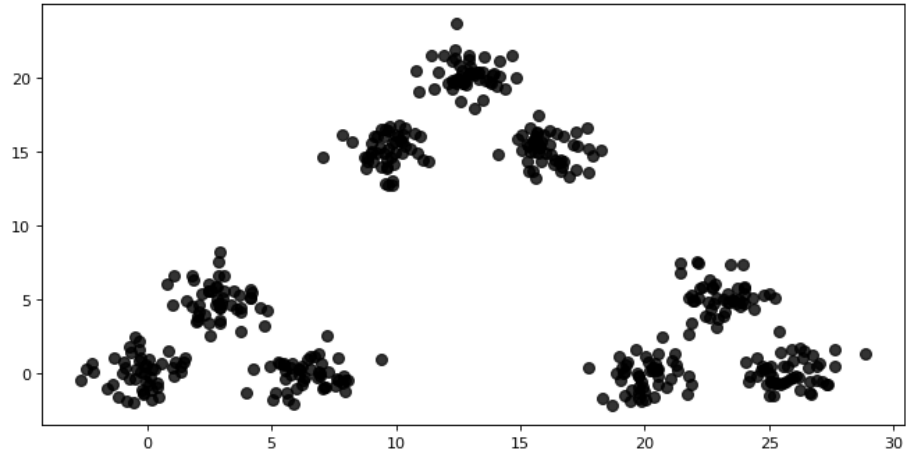
۵- قسمت d

حال با استفاده از دیتاست دوم باید قسمت c را تکرار کنیم. شکل بدست آمده به صورت زیر است:



که مشاهده میشود در نقطه K=3 به نقطه مورد نظر یعنی *knee* یا *elbow* میرسیم که مورد انتظار ما نیز هست چون دیتاست به صورت چشمی یا با استفاده از تعریف *density* نیز شامل ۳ کلاستر است.

همچنین مشاهده میشود در نقطه *K=9* یک *knee* یا *elbow* دیگر داریم که علت آن این است که دیتاست ما به صورت زیر است (به صورت اختیاری رسم شده است) :

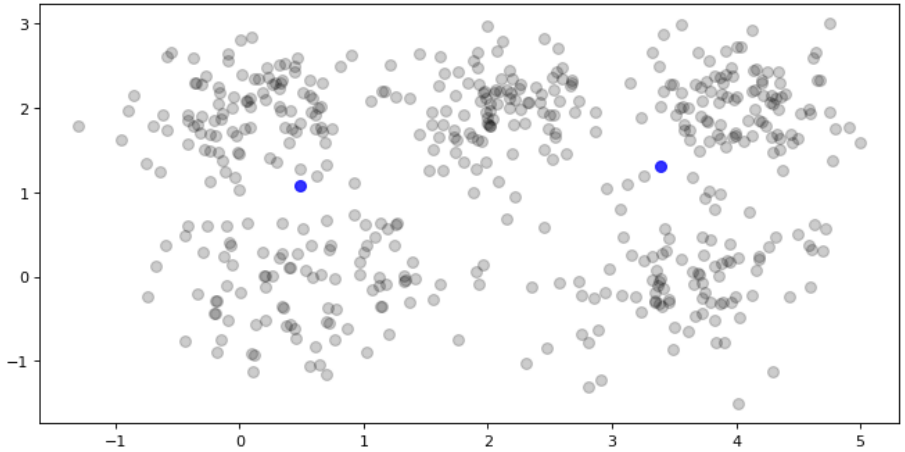


واضح است که به صورت کلی ۳ کلاستر در تعریف *density* داریم و هر کلاستر به ۳ کلاستر دیگر تبدیل میشوند و در مجموع با *K=9* کلاستر هم میتوانیم نتیجه خوبی داشته باشیم.

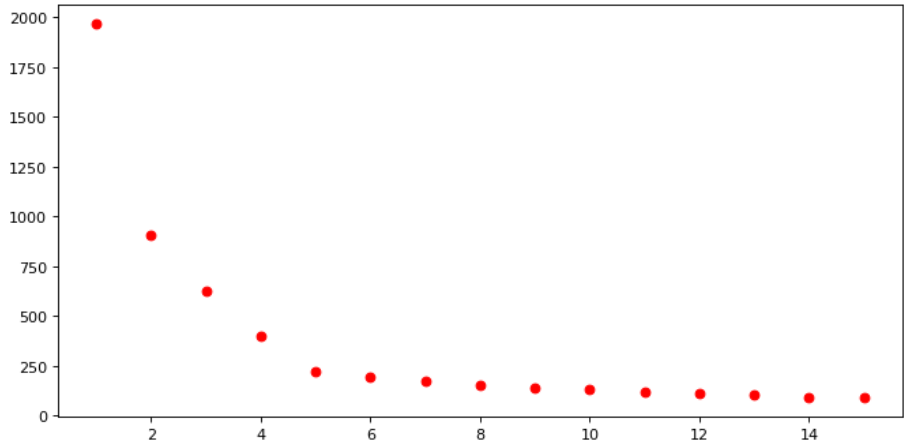
پس نتیجه در این قسمت $K=3$ **و** $K=9$ شد که قابل انتظار بود.

۶- قسمت e

حال برای دیتاست سوم و برای مقدار $K=2$ ، و با تعداد تکرار ۵۰۰ بار به دنبال مراکز کلاستر های دیتاست سوم میگردیم. بهترین کلاستر های کسب شده با استفاده از روش $k-means++$ و ۵۰۰ بار تکرار مقداردهی اولیه، به صورت شکل زیر است. در شکل زیر داده های اصلی با رنگ مشکی و $centroid$ ها به صورت نقطه های آبی نشان داده شده اند.



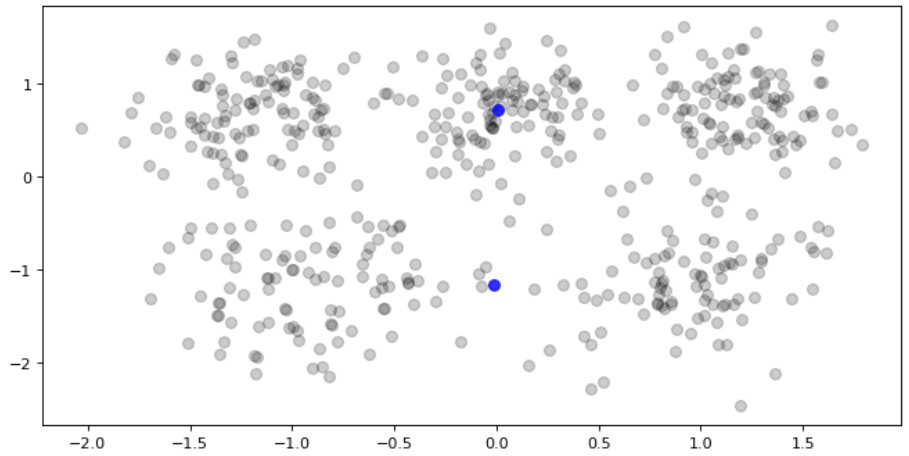
نتیجه با استفاده از ۲ کلاستر قابل قبول است. چون K کم است . برای اینکه K مطلوب را برای این دیتاست هم پیدا کنیم، به صورت اختیاری نمودار مشابه با دو قسمت قبل رسم می کنیم.



که همانطور که انتظار میرود *knee* یا *elbow* در نقطه *K=5* رخ می دهد و علت اینکه دقت خوبی برای *K=2* در این دیتا نداشتیم به علت کوچک بودن *K* است.

۷- قسمت f

در این قسمت میخواهیم تاثیر نرمال سازی روی داده های سوم را بسنجیم. ابتدا نرمال سازی را همانطور که سوال خواسته انجام میدهیم و مشابه با حالت قسمت e با این تفاوت که مقدار دهی اولیه ۵۰۰ بار انجام میشود و بهترین انتخاب میشود، خوشه بندی را اعمال می کنیم . نتیجه به صورت زیر است:



که مشاهده میشود به عمل `kmeans` کمک شده است چون شکل بهتری نسبت به حالت e داریم. البته لازم به ذکر است همچنان K مناسب نیست و برای اینکه بتوانیم نتیجه خوبی داشته باشیم که بتوانیم کار های انجام گرفته را مقایسه کنیم نیاز داریم که K مناسبی را اختیار کنیم که برای دیتاست اول $K=2$ ، دیتاست دوم $K=3$ و دیتاست سوم $K=5$ مقادیر مناسبی هستند (همانطور که ذکر شد).

بنابراین نتیجه گیری به این صورت است که یکی از مشکلات الگوریتم `kmeans` انتخاب اولیه K است. اگر این انتخاب K بد باشد (مانند نتایج موجود در این تمرین) نمیتوانیم کلاسترینگ خوبی را داشته باشیم بنابراین باید تا جایی که ممکن است روی انتخاب K دقت کنیم تا نتایج خوشه بندی مناسبی را بگیریم.