به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس شناسایی آماری الگو

استاد رحمتی

تمرین چهارم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

سوال ۱

a و b و c) با درنظرگرفتن مرکزهای اولیه، داده‌های هر خوشه عبارت است از:

بر همین مبنا مرکز جدید سه خوشه تغییر خواهد کرد:

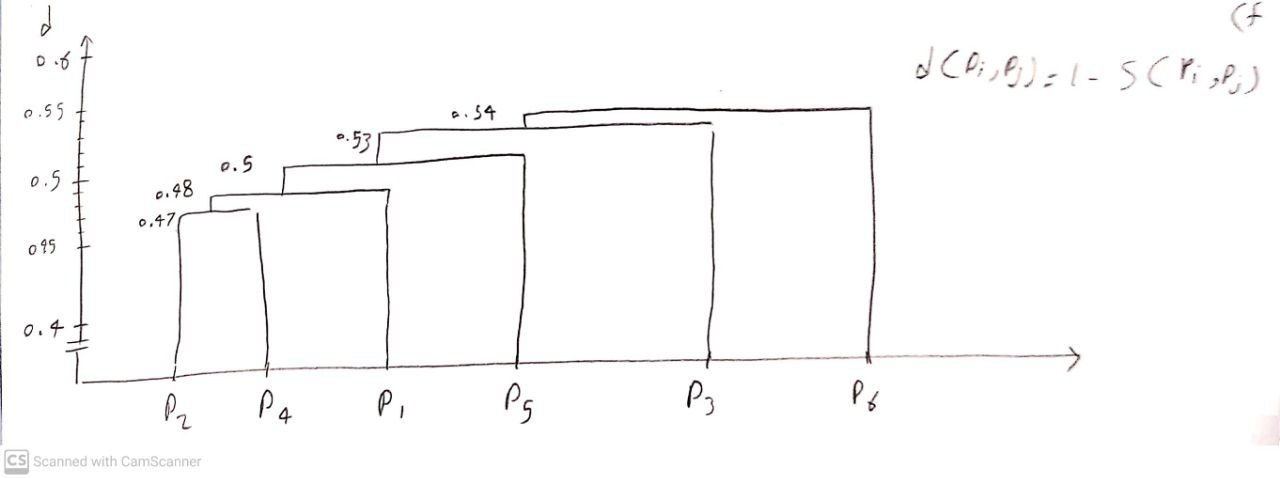
d) باتوجه به مراکز جدید خوشه‌ها، داده‌های خوشه‌ها در دومین گام عبارت است از:

باتوجه به اینکه جای داده‌ها تغییر نکرده است پس مراکز خوشه‌ها هم تغییری نخواهد کرد و لذا الگوریتم در پایان گام دوم خاتمه می‌یابد.

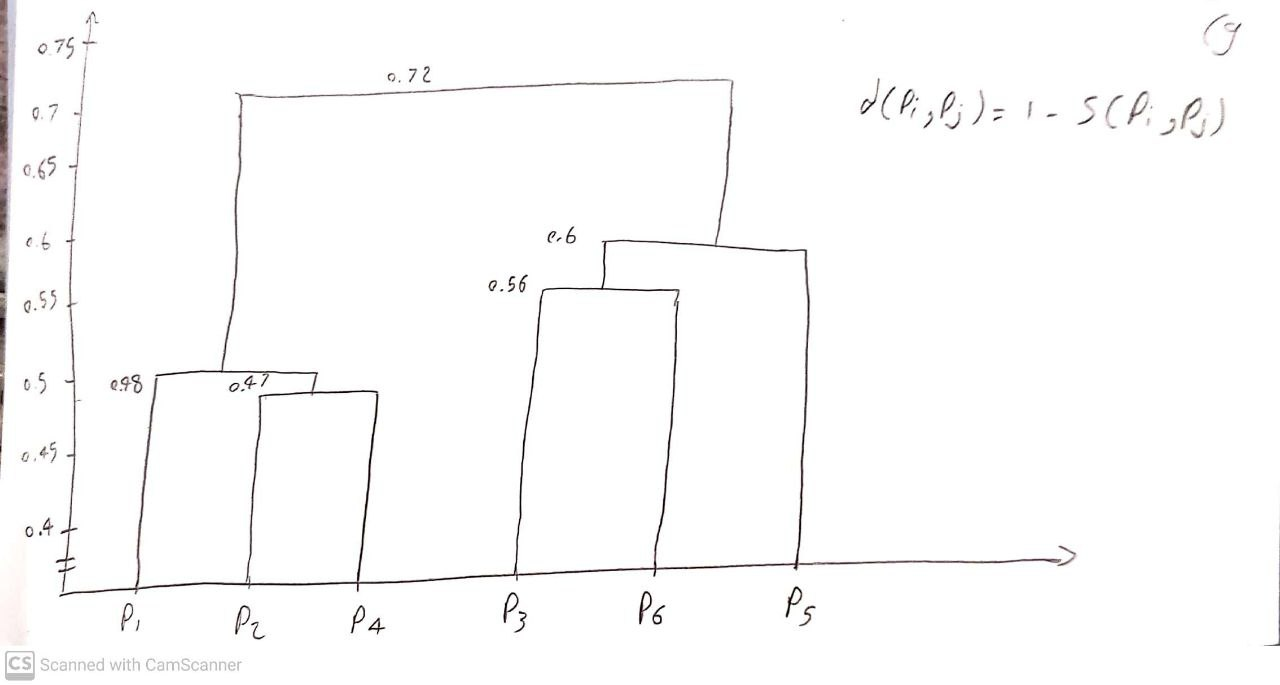
e) حالت بهینه سراسری زمانی رخ می‌دهد که یک خوشه دارای داده‌ی ۳- و ۱ باشد و خوشه دیگر دارای داده ۸. اگر شرایط به گونه‌ای پیش برود که یک خوشه بدون داده بماند و یا اینکه یک خوشه دارای ۸ و ۱ و دیگری دارای ۳- باشد به بهینه سراسری دست پیدا نخواهیم کرد. لذا می‌توان این شرایط را در نظر گرفت:

با اشتراک‌گیری از شرایط بالا به بازه‌های زیر می‌رسیم:

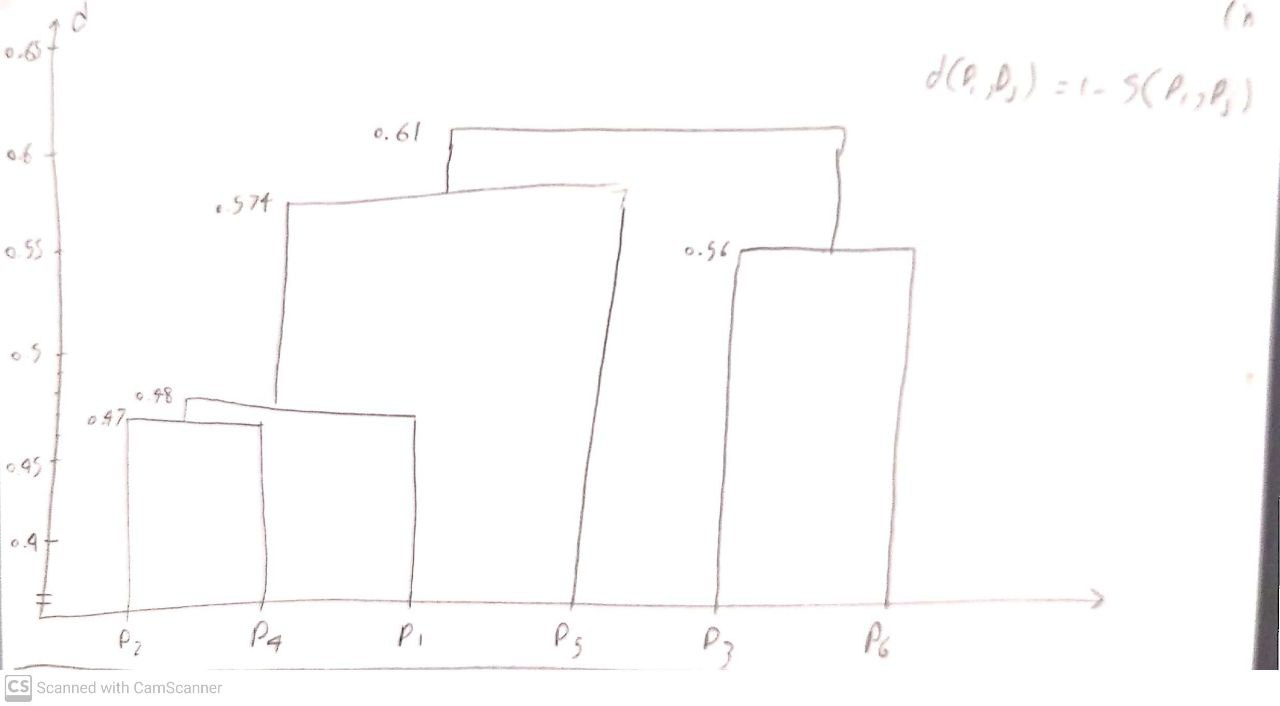
f)



g)



h)



سوال ۲

b)

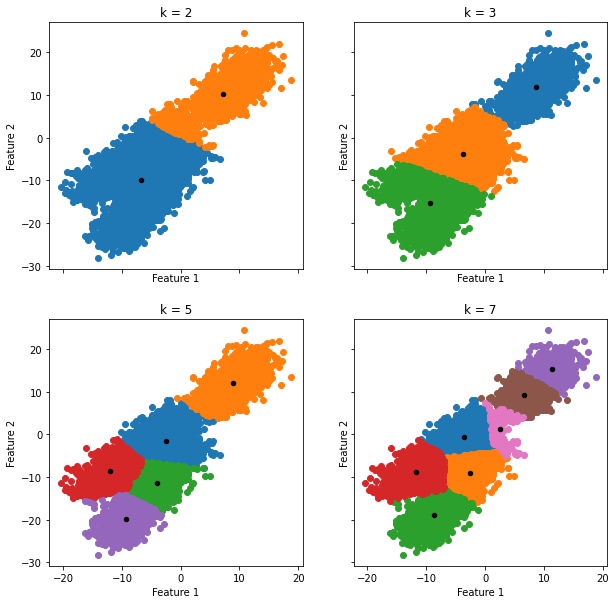
c)

d) به صورت عادی این مراکز، مراکزی هستند که همگرا شده‌اند و اگر الگوریتم خوشه‌بندی با این مراکز شروع شود حد آستانه همگرایی شکسته می‌شود و در همان گام اول الگوریتم به پایان می‌رسد. یک احتمال دیگر هم که وجود دارد آن است که تعداد گام درنظر گرفته شده برای خوشه‌بندی اول کافی نباشد. در این صورت ممکن است در دفعه دوم خوشه‌بندی مراکز کمی تغییر کند.

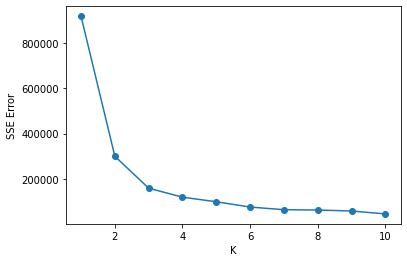
e) اگر داده‌های بابرچسب داشته باشیم، استفاده از یک الگوریتم نظارت‌شده معمولا به نتایج بهتری منجر می‌شود؛ چراکه در این حالت دانش بیشتری وجود دارد که یک الگوریتم نظارت‌شده از آن استفاده می‌کند ولی یک الگوریتم بدون نظارت آن را نادیده می‌گیرد. در مقایسه یک الگوریتم غیرنظارت‌شده دیگر با k-means هم به قطعیت نمی‌توان نظری داد و بسته به نوع داده‌ها و پارامتر‌های هر الگوریتم ممکن است k-means بهتر باشد و ممکن است الگوریتم دیگر عملکرد بهتری داشته باشد.

سوال ۳

a)



b)

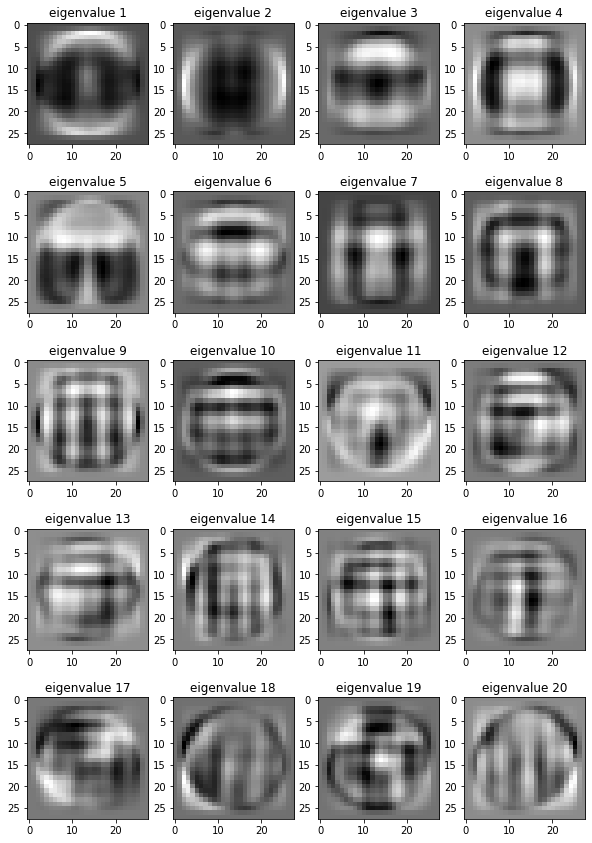


در این نمودار انتظار می‌رود که با افزایش مقدار k خطای خوشه‌بندی هم کاهش یابد اما مقدار کاهش روند ثابتی را ندارد؛ در ابتدا کاهش شدید است تا به نقطه‌ای برسیم که تعداد خوشه‌های متناسب با داده‌ها بدست آید و از آنجا به بعد کاهش کمتری رخ خواهد داد. لذا یک شکستگی در نمودار دیده می‌شود که تعداد خوشه مناسب را نمایان می‌کند. در این نمودار هم ۳ خوشه به نظر مناسب می‌آیند.

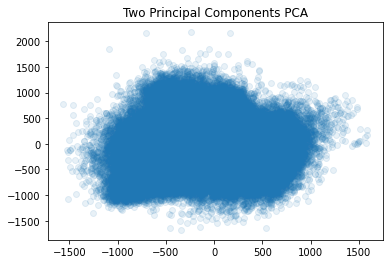
c)

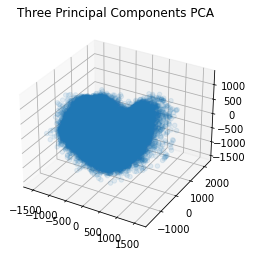
سوال ۴

a) بردارهای مربوط به مهم‌ترین مولفه‌ها عبارت‌اند از:

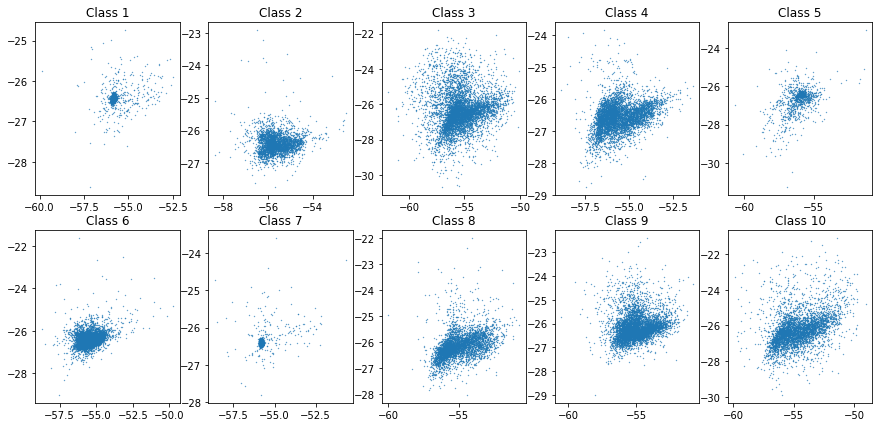


خروجی PCA به ازای یک، دو و سه مولفه اصلی:

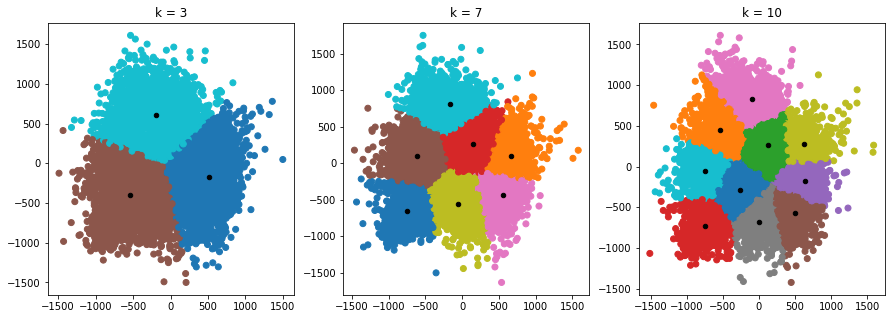




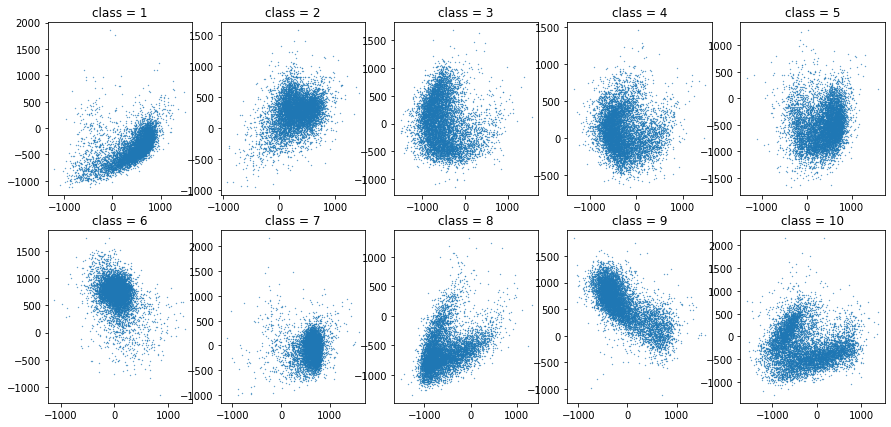
b) باتوجه به اینکه LDA‌ به برچسب کلاس هم توجه می‌کند، در نمودار زیر هر کلاس در یک زیرنمودار جداگانه آورده شده است:



c) نتایج خوشه‌بندی به شرح زیر است:



اگرچه خروجی PCA‌ ذاتا دارای اجزای جدا از هم نبوده ولی خوشه‌بندی روی خروجی‌ بدست آمده از PCA، خوشه‌بندی معقول است و فضا به به خوشه‌های تقریبا مساوی شکسته شده است. اما یک نکته‌ای که باید به آن توجه کرد این است که از آنجا که PCA‌ توجهی به کلاس‌های هر داده نداشته است، لزومی هم وجود ندارد که هر خوشه بتواند نماینده یک یا چند کلاس مشخص باشد و حتی این امکان وجود دارد که PCA‌ تفاوتی که در کلاس‌ها وجود داشته است را از بین برده باشد. لذا برای تحلیل بهتر نمونه‌های کلاس‌های مختلف را ترسیم می‌کنیم:



اگرچه برخی از کلاس‌ها تا حدی از سایر کلاس‌ها جدا شده است ولی نسبت به حالت LDA‌ داده‌های کلاس‌ها بیشتر در هم فرو رفته‌اند و نتایج خوشه‌بندی برای جداکردن داده‌ها مطلوب به نظر نمی‌رسد.

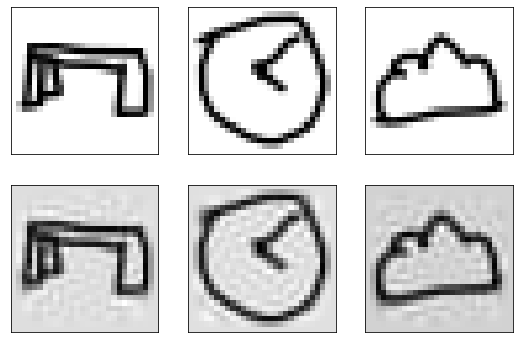
d) نتایج زیر حاصل می‌شود:



نتایج این قسمت با قسمت قبل چندان تفاوت ندارد و این بار هم یک خوشه نمی‌تواند نماینده داده‌های یک کلاس باشد. علت اینکه نقطه شروع متفاوت در اینجا کم تاثیر است احتمالا بدلیل آن است که داده‌ها خروجی PCA‌ هستند و در یک فضا به صورت نسبتا یک دست پخش شده‌اند.

e) اهمیت و واریانس داده‌ها برای هر مولفه متناسب با مقدار ویژه متناسب با آن مولفه است. چنانچه مجموعه‌ای از k مولفه داشته باشیم. حاصل تقسیم مجموع مقادیر ویژه آن‌ها بر مجموع مقادیر ویژه کل مولفه‌ها نشان می‌دهد که چه میزان از واریانس حفظ شده است. با بررسی مقادیر ویژه دریافتم که باید حداقل 271 مولفه اول را حفظ کرد تا این مقدار از واریانس باقی بماند.

در تصویر زیر سه نمونه اصلی (سطر بالا) به همراه خروجی حاصل شده از 271 مولفه اول (سطر پایین) را می‌توانید مشاهده کنید:



سوال ۵