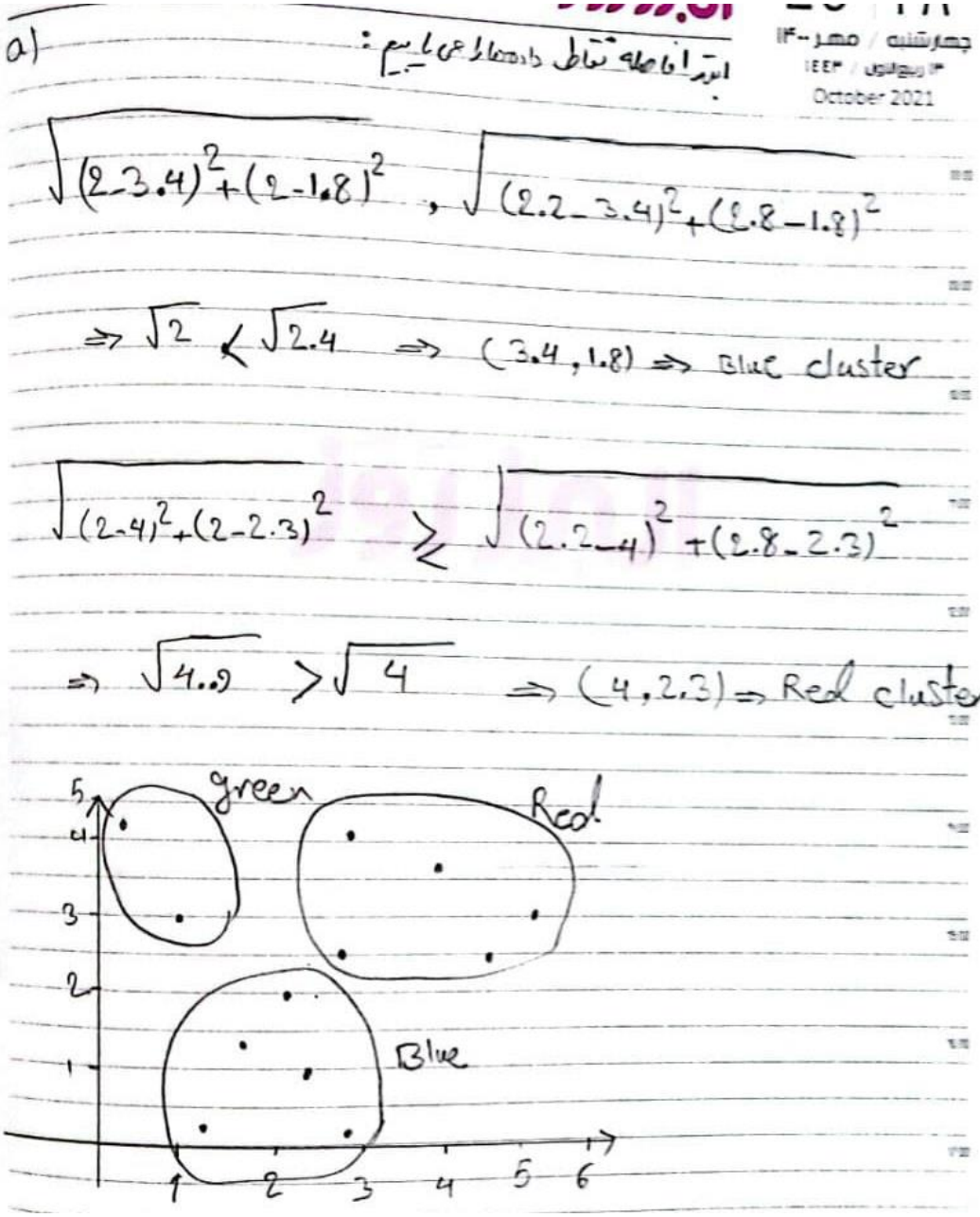
مهدی فیروزبخت

۴۰۰۱۳۱۰۲۷

تمرین سری چهارم شناسایی الگو

۱)



۲۹ | 21
پنجشنبه / مهر ۱۴۰۰
۱٤ ربیع‌الاول / ١٤٤٣
October 2021

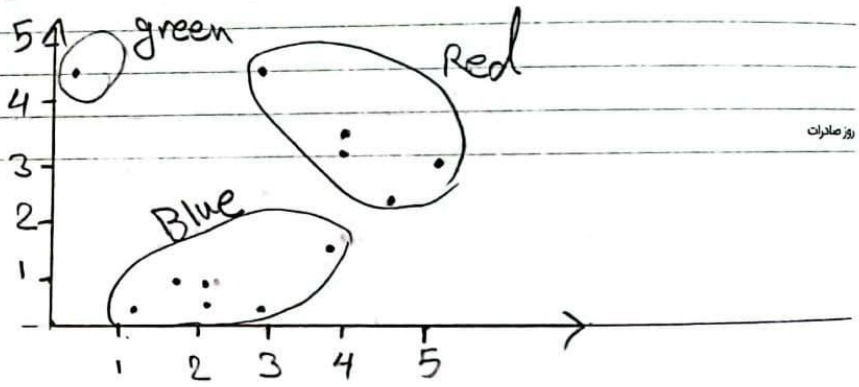
بی‌گدار به آب بزنید!

(B) $\mu_1 = (3.55, 3.25)$

$\mu_2 = (2.34, 1.04)$ $\mu_3 = (0.4, 4.5)$

$$\sqrt{(2.34-3.4)^2+(1.04-1.8)^2} \lesssim \sqrt{(3.55-3.4)^2+(3.25-1.8)^2}$$

$$\Rightarrow \sqrt{1.85} < \sqrt{2.3} \rightarrow \text{Blue cluster}$$



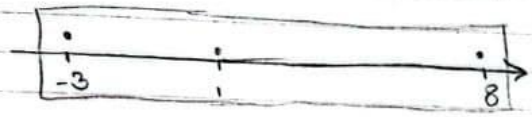
روز صادرات

۳۰ | 22
پنجشنبه / مهر ۱۴۰۰
۱۵ ربیع‌الاول / ١٤٤٣
October 2021

(C) چون ⇒ iteration صفر به Converge رسیده بودیم
و رده‌ها تغییر نکرده است پس گره سبز بدون تغییر بوده است.

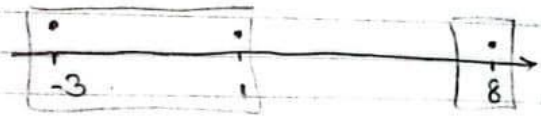
(D) چون ⇒ iteration = 2 تغییر ایجاد شده پس به همگرایی رسیده‌ایم

(E) $\mu_1 = -30$ $\mu_2 = 0$ ⇒ Conv in iteration = 2



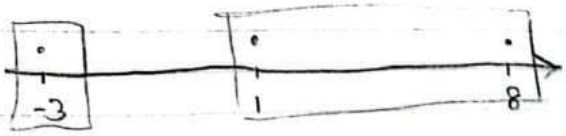
$\mu_1 = \infty$
$\mu_2 = 2$

$\mu_1 = -1$ $\mu_2 = 6$ ⇒ Conv in iteration = 2



$\mu_1 = -1$
$\mu_2 = 8$

$\mu_1 = -2$ $\mu_2 = 3$ ⇒ Conv in iteration = 2



$\mu_1 = -3$
$\mu_2 = 4.5$

۲۱ 24

یکشنبه / آبان ۱۴۰۰
۱۷ ربیع‌الاول / ۱۴۴۳
October 2021

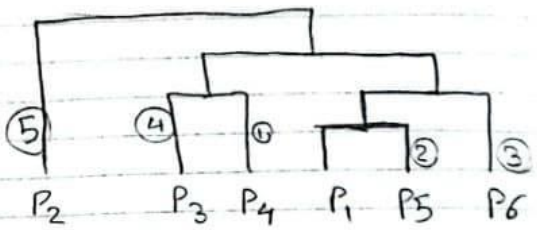
بی‌گدار به آب بزنید!

F

① .,28 ③ .,37

② .,31 ④ .,38

⑤ .,39

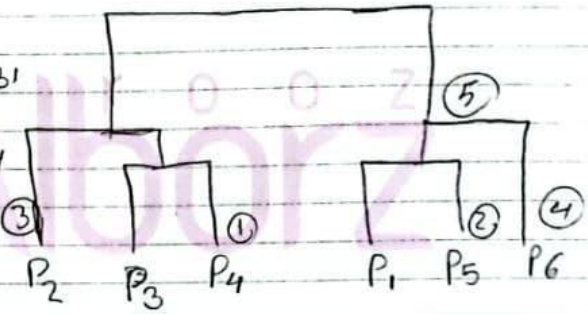


G

① .,28 ② .,31

③ .,39 ④ .,44

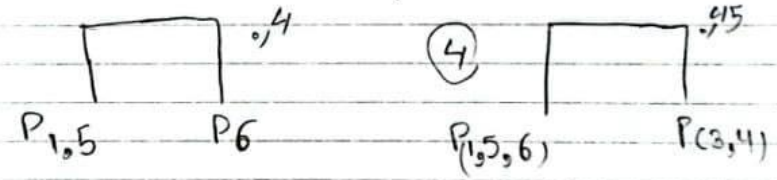
⑤ .,52



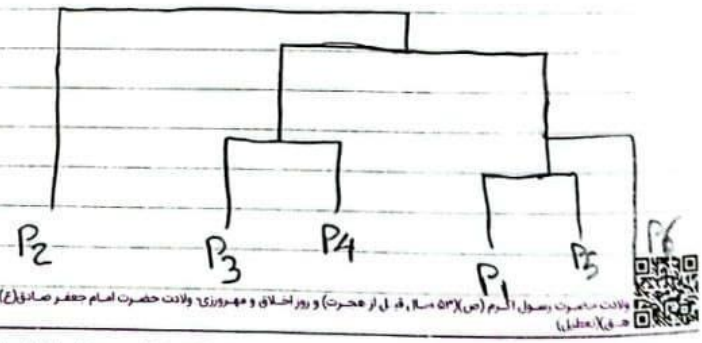
H

③ .,4

④ .,45



⑤ .,47



M	T	W	T	F	S	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

ولادت حضرت رسول اکرم (ص)(۵۳ سال قبل از هجرت) و روز اخلاق و مهرورزی؛ ولادت حضرت امام جعفر صادق(ع) مؤسس مذهب جعفری(۸۳ هـ.ق)(معتبر)

در حقیقت موفقیت حاصل یک درصد کار است که خود از ۹۹ درصد خطا حاصل شده است. (نوشیرو هوندا)

۲)

سوال دوم با عنوان Question 2 در فولدر IMP قرار داده شده است.

برای قسمت a ، ابتدا در تابع خواسته شده در نظر گرفته ایم که تعداد iter ها از تعداد ماکسیمم بیشتر نشده و همچنین مقدار sse همواره از حدی tol بیشتر بود ، کد ها تکرار شود.

در این کد ها ابتدا فاصله هر داده از میانگین های ورودی محاسبه شده و در ارایه ای ریخته شده است. سپس بر اساس این فاصله داده ها در k گروه ، لیبل بندی شده اند. سپس میانگین های جدید محاسبه شده، مقدار sse به دست امده و چرخه دوباره تکرار میشود تا به همگرایی برسد.

برای قسمت b ، ۵ بار و با میانگین های متفاوت تابع را فراخوانی کرده ایم که در هر بار تکرار مقدار دقت آن محاسبه شده است. همه حالات در تعداد تکرار های متفاوت توانسته اند به مقدار دقت ۰٫۸۵۴۱ همگرا شده اند. اما تعداد تکرار ها برای رسیدن به این دقت متفاوت بوده است.

برای قسمت c ، مقدار دقت در هر تکرار نمایش داده شده است که در نهایت به مقدار ۰٫۸۵۴۱ همگرا شده است.

اگر از میانگین های واقعی نیز برای این مثال استفاده شود نیز در نهایت باز هم به مقدار دقت ۰٫۸۵۴۱ همگرا شده و تفاوتی با میانگین های تصادفی داده شده در حالات گذشته ندارد و فقط توانسته در تکرار ۵ به این مقدار برسد.

در مورد قسمت آخر،

چون در این مثال داده ها به صورت ۲بعدی بوده و خطی هستند ، الگوریتم k-means بهترین خروجی را در این مثال خواهد داشت ، همانطوری که دیده شد در قسمت آخر نیز همانند الگوریتم k-medoid مقادیر اصلی را قرار داده ایم و به مقدار یکسانی رسیده ایم که در حالات قبلی k-means با تعداد تکرار کمتری به این مقدار رسیده بوده است.

۳)

سوال سوم در فولدر IMP و با نام Question 3 قرار داده شده است.

در این مثال از کد k-means سوال قبلی استفاده شده است.

برای قسمت الف داده های متفاوت تصادفی انتخاب شده و میانگین با ضربدر نمایش داده شده است.

برای قسمت b ، داده های تصادفی برای k های متفاوت اخذ شده که با اجرا و نمایش داده های متفاوت به مقادیر k و sse متفاوت دسترسی پیدا کرده ایم که بهترین k برای این سوال k = 4 خواهد بود.

برای قسمت c ، ابتدا برای تغییر تابع kmeans به kmedoid ابتدا همانند تابع قبلی فاصله را تا میانگین های ابتدایی محاسبه کرده (که این میانگین های تصادفی از بین داده های اولیه انتخاب شده اند) و سپس برای به دست آوردن medoid های جدید ، میانگین هر دسته را محاسبه کرده و فاصله هر داده ی دسته را از میانگین به دست آورده و داده ای که کمترین فاصله را تا میانگین دسته دارد به عنوان medoid آن دسته انتخاب میکنیم و باقی کار های تابع همانند تابع kmeans به محاسبه هر دسته میپردازد.

برای این مقادیر نیز در انتها برای هر k ، medoid های نهایی مشخص شده اند.

برای قسمت e ، کدها در قسمت GeneQuestion قرار داده شده است.

در این سوال ابتدا داده ها را با کتابخانه sklearn به صورت AgglomerativeClustering فیت شده است و دندوگرام آن نمایش داده شده است. در این سوال با کم کردن میانگین از داده ها ، میانگین داده های جدید را ۰ کرده ایم و روش correlation به روش کوسینوسی همگرا میشود.

بله این تابع توانسته که داده ها را به ۲ گروه دسته بندی کند.

برای سوال f ، کد آن در Compares قرار داده شده است. در این سوال ، ابتدا داده ها را با AgglomerativeClustering فیت شده است با هر دو روش و با purity_score و normalized_mutual_info_score مقادیر آن ها محاسبه شده است و تمام دقت ها در این کد قرار داده شده است.

۴)

برای قسمتهای مرتبط با pca ، کد آن در pca قرار داده شده است.

برای قسمت lda ، کد ان در LDA قرار داده شده است.

برای پیاده سازی هر دو از فرمول های آن استفاده شده است و در کد کامنت گذاری شده است.

قسمت a ، ابتدا با pca تمام ۲۰ بردار های ویژه به دست آورده ایم و همانطور که در سوال خواسته شده این ۲۰ بردار ویژه را نمایش داده ایم. سپس تصویر داده ها را بر روی ۲ بردار ویژه اصلی و سپس ۳ بردار اصلی نمایش داده شده است.

قسمت b ، ابتدا ماتریس هر تصویر را به صورت ستونی تبدیل میکنیم . سپس میانگین کلی را به دست می آوریم ، سپس برای هر کلاس میانگین آن را محاسبه میکنیم. با استفاده از این میانگین ها ماتریس های SB و SW را به دست می آوریم. و با استفاده از این ماتریس ها بردار های ویژه را محاسبه میکنیم. سپس تصویر هر داده را بر روی ۲ بردار اصلی و سپس بر روی ۳ بردار اصلی به دست می آوریم.

برای قسمت c ، از تابع kmeans که در سوالهای قبلی نوشته ایم استفاده میکنیم. ابتدا تبدیل هر تصویر را بر روی بردار های اول و دوم به دست می آوریم و به تابع kmeans میدهیم و این تابع بر اساس مقدار k این داده ها را دسته بندی میکند. همچنین مقدار میانگین نهایی در تصاویر مشخص شده است. در مورد مقایسه این دسته بندی با دسته بندی قسمت a ، مشخص است که در قسمت اول دسته بندی بسیار پیچیده صورت گرفته و مرز تصمیم مشخصی در بعضی از دسته ها دیده نمیشود در صورتی که در این مثال ها کاملا داده ها از یکدیگر جدا شده اند و مرز تصمیم مشخصی دارند.

برای قسمت d ، کاملا مشابه قسمت قبلی رفتار میکنیم تنها تفاوت مقدار میانگین های ابتدایی است که به صورت تصادفی نبوده و مشخص شده است. در مورد مقایسه این

داده مثال ها با قسمت اول ، همانند سوال قبل این مثال ها کاملا مرز های جدایی دارند و متفاوت با قسمت اول هستند که مرز تصمیم پیچیده ای دارند و مشخص نیست. اما در مقایسه مثال های این سوال با سوال قبلی ، از نظر ظاهری و مرز تصمیم شباهت بسیار زیاد با یکدیگر دارند با این تفاوت که مقدار صحت و دقت مثال هایی که مقدار اولیه آنها از ترکیب میانگین ها به دست می آید بیشتر است.

برای قسمت e ، ابتدا با ازمون و خطا تعداد بردار های ویژه مناسب برای اینکه مقدار variance_ratio از ۰/۹۵ بیشتر باشد را پیدا کرده ایم که تعداد آنها برابر ۲۷۳ است. سپس به صورت رندوم ۳ تصویر انتخاب شد و با استفاده از این تعداد بردار ویژه آنها را بازسازی کرده ایم و مشاهده کرده ایم که تصویر به دست آمده بسیار شبیه تصویر واقعی است. همچنین در ادامه این داده ها را با این تعداد ویژگی و با kmeans دسته بندی کرده ایم.

برای قسمت f ، ابتدا داده ها را دسته بندی کرده ایم به روش kmeans ، سپس به صورت رندوم از هر دسته ۱۰ داده را جدا کرده این و با بردار های ویژه آنها را بازسازی کرده ایم. مشاهده میشود که بعضی از داده های به دست آمده که اشتباه هستند در بعضی از موارد بسیار شبیه به یکدیگر بوده اند و در دسته های یکسان قرار گرفته اند. اما در کل با بررسی مقدار دقت و صحت مشاهده میشود که داده ها به خوبی دسته بندی نشده اند و در بعضی از دسته ها موارد اشتباه زیادی دیده میشود. هر چند در برخی دسته بندی ها مانند دسته بندی چتر مشاهده میشود که دسته بندی به خوبی صورت گرفته و مقدار خطا بسیار کم است.

برای قسمت g ، از تابع all_indices برای به دست آوردن ایندکس داده ها استفاده میشود تا جای هر داده مربوط به هر دسته را به دست آوریم. از این داده ها استفاده کرده و داده های مربوط به هر کلاس را در هر دسته استخراج میکنیم و از این داده های جدید برای نمایش bar graph استفاده میکنیم. مشاهده میشود همانند حرف گفته شده در سوال قبل ، در بعضی از کلاس ها مانند کلاس چتر تعداد داده های کلاس چتر زیاد بوده و خطا کم است در حال که در دسته بندی اول که مربوط به میز است ، تعداد خطا ها بسیار زیاد است.

برای قسمت H ، داده ها را بر روی ویژگی هایی که بردار ویژه بالاتری دارد تصویر میکنیم. ابتدا بر روی ۲ ویژگی و سپس برای ۳ ویژگی این کار را انجام داده و نمایش میدهیم.

۵)

a)

در الگوریتم های مبتنی بر نمونه تصمیم گیری ها و محاسبات همگی به نمونه های موجود در داده ها دارد و فاز آموزشی ندارد. در الگوریتم k-means برای محاسبه فاصله با مرکز خوشه ها و محاسبه فاصله برای داده های جدید همه در فاز آزمایش قرار دارد و اموزش خاصی ندارد و داده های جدید فاصله با این داده ها محاسبه شده و هر چه فاصله کمتر ، به آن گروه تعلق دارد.

b)

داده های پرت در نتیجه میانگین دسته ها اهمیت دارد و بر آن تاثیر منفی میگذارد.

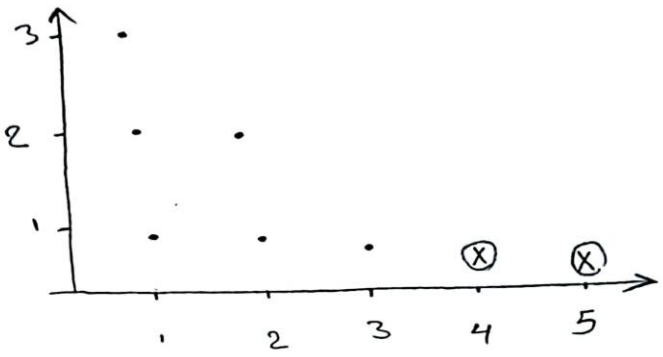
c)

بله، اگر بعد از انجام محاسبات داده های خوشه ها و خوشه ها تغییری نکنند و مشابه حالت قبل باشد به همگرایی رسیده ایم.

d)

حالت اول امکان پذیر نیست، اگر مقدار k مقداری در نظر گرفته شود به تعداد آن میانگین خواهیم داشت که فاصله هر داده تا این داده ها محاسبه میشود و حتما به k دسته تبدیل میشود و امکان ندارد تعداد دسته ها از k بیشتر شود. اما اگر میانگین ها به گونه ای انتخاب شوند که داده ها به تعدادی از این میانگین ها نزدیک تر از باقی باشد

ممکن است حالت دوم ایجاد شود ، یعنی ممکن است تعداد دسته ها از k کمتر شود. به مثال زیر توجه کنید :



خوشه ها میانگین هستند و نقطه ها داده هستند .

در این مثال همه داده ها در دسته میانگین (۱و۲) قرار می گیرند و تعداد دسته ها برابر ۱ است درحالی که K=2 بوده است.

e)

اگر تعداد خوشه ها مشخص باشد استفاده از k-means بهتر است.

اگر تمایز بین داده ها ساده باشد و به سادگی قابل جداسازی باشند اما تعداد دسته ها مشخص نباشد استفاده از hierarchical بهتر است.

اگر تعداد متغیر ها زیاد باشد ، k-means سریعتر است.

اگر نیاز به تفسیر بیشتر در نتایج دسته بندی باشد روش hierarchical بهتر خواهد بود.