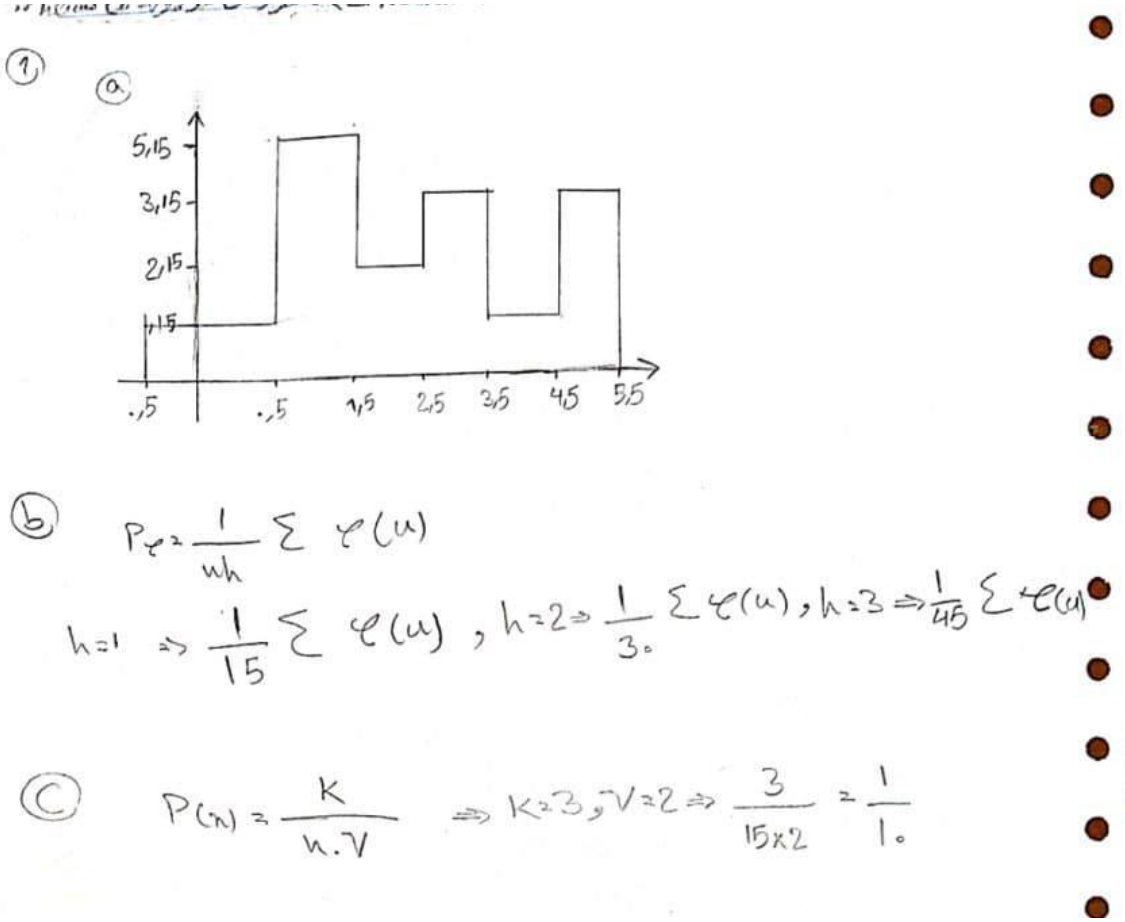
مهدی فیروزبخت

۴۰۰۱۳۱۰۲۷

تمرین سری سوم شناسایی الگو

۱)



(D) $P_{\varphi}(0)=\frac{1}{30}$, $P_{\varphi}(1)=\frac{5}{30}$, $P_{\varphi}(2)=\frac{2}{30}$, $P_{\varphi}(3)=\frac{3}{30}$, $P_{\varphi}(4)=\frac{1}{30}$

$P_{\varphi}(5)=\frac{3}{30}$

(E) $h=4$ $P_{\varphi}(4)=\frac{1}{6.}\times 5$, $P_{\varphi}(1.)=\frac{1}{6.}$, $P_{\varphi}(16)=\frac{4}{6.}$

(F) ؟

(٢

(2) (A) $\mu=\begin{bmatrix}6,4\\4,6\\5,7\end{bmatrix}$

$b=\begin{bmatrix}-5,4 & -3,4 & -2,4 & -.,4 & .,6 & 1,6 & 1,6 & 1,6 & 2,6 & 3,6\\ -4,6 & -.,6 & -1,6 & 2,4 & -3,4 & 3,4 & -2,6 & 5,4 & .,4 & 1,4\\ -2,7 & .,3 & 1,3 & -2,7 & -.,7 & 4,3 & -1,7 & -3,7 & 2,3 & 3,3\end{bmatrix}$

$Cov=\begin{bmatrix}74,4 & 43,6 & 27,2\\ 43,6 & 92,4 & 10,8\\ 27,2 & 10,8 & 68,1\end{bmatrix}$

$d = |cov - \lambda I| \Rightarrow \begin{cases} \lambda_1 = 31{,}9 \\ \lambda_2 = 65{,}3 \\ \lambda_3 = 137{,}6 \end{cases} \Rightarrow v_1 = \begin{bmatrix} -1{,}7 \\ 1{,}09 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} -0{,}16 \\ -0{,}165 \\ 1 \end{bmatrix}$

$v_3 = \begin{bmatrix} 1{,}7 \\ 1{,}9 \\ 1 \end{bmatrix}$

e ⇒ است λ_3 بزرگترین مقدار ویژه eigen vector = v_3

$F \Rightarrow y = e^T x \Rightarrow \begin{bmatrix} -2{,}6 & -6{,}62 & -5{,}8 & 1{,}1 & -6{,}5 & 13{,}4 \\ -3{,}9 & 9{,}2 & 7{,}4 & 12{,}08 \end{bmatrix}$

(g) $\Rightarrow \mu_1 = \begin{bmatrix} 4{,}9 \\ 3{,}2 \\ 5 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 6 \\ 8{,}1 \\ 4{,}8 \end{bmatrix}$

(h) $\Sigma_1 = \begin{bmatrix} 9{,}29 & 3{,}02 & -3{,}9 \\ 3{,}02 & 2{,}76 & -2{,}5 \\ -3{,}9 & -2{,}5 & 7{,}8 \end{bmatrix}$

$\Sigma_2 = \begin{bmatrix} 6{,}4 & 1{,}5 & 1{,}7 \\ 1{,}5 & 1{,}69 & 1{,}12 \\ 1{,}7 & 1{,}12 & 11{,}56 \end{bmatrix}$

(i) $S_w = S_1 + S_2 \Rightarrow$

$$\begin{bmatrix} 156,9 & 45,2 & -22 \\ 45,2 & 44,5 & -13,8 \\ -22 & -13,8 & 193,6 \end{bmatrix}$$

(j) $S_B =$

$$\begin{bmatrix} 1,21 & 5,39 & -0,22 \\ 5,39 & 24,01 & -0,98 \\ -0,22 & -0,98 & 0,04 \end{bmatrix}$$

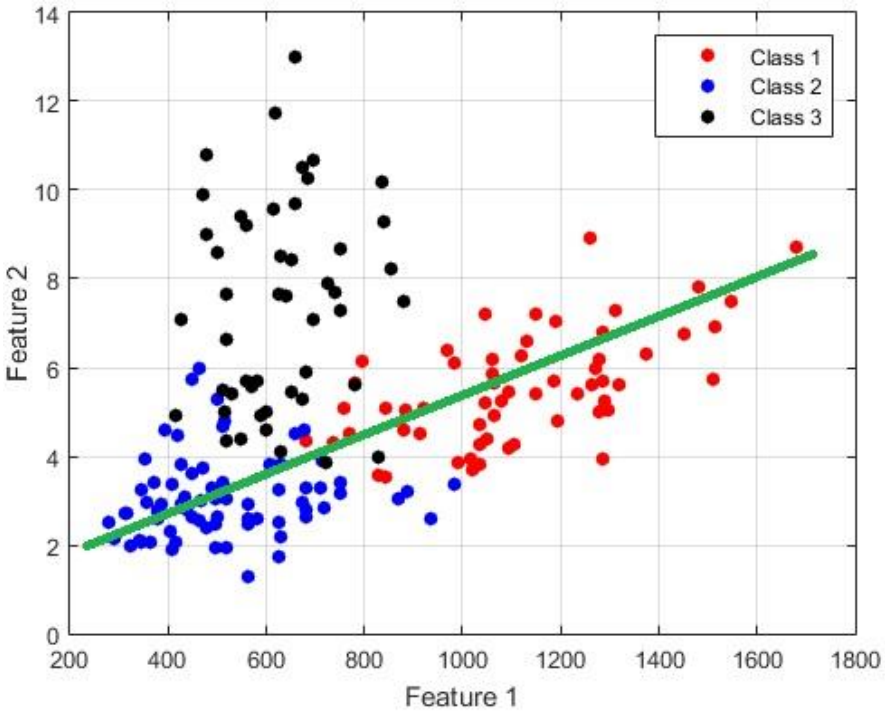
(K) $v = S_w^{-1}(\mu_1 - \mu_2)$

$$S_w^{-1} = \begin{bmatrix} 0,009 & -0,009 & 0,003 \\ -0,009 & 0,023 & 0,001 \\ 0,003 & 0,001 & 0,005 \end{bmatrix} \times \begin{bmatrix} -1,1 \\ -4,9 \\ -,2 \end{bmatrix} = \begin{bmatrix} 0,3 \\ -0,13 \\ -0,07 \end{bmatrix}$$

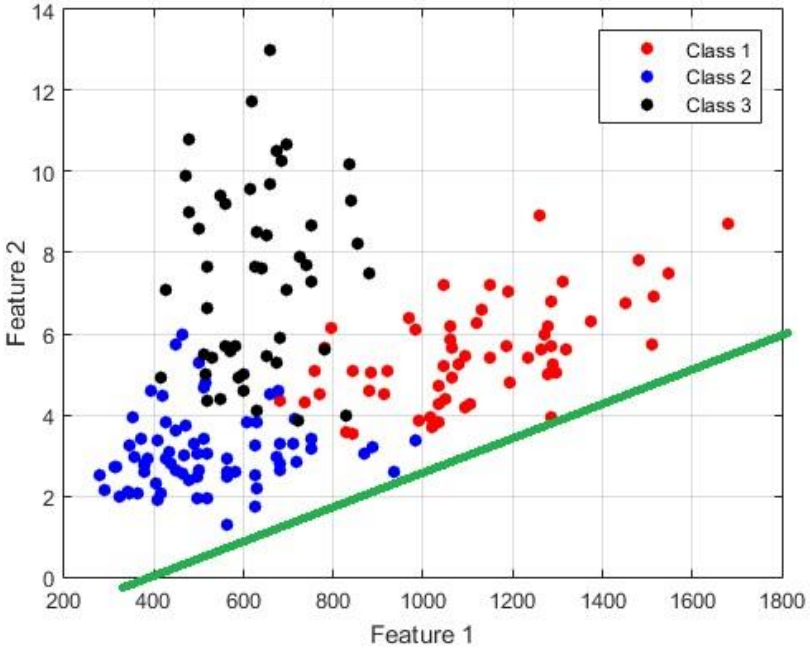
(L) $w_1 \Rightarrow [-,43 \quad -0,11 \quad -0,27 \quad -0,09 \quad -0,141 \quad -0,5 \quad -0,07 \quad -0,17 \quad -0,141 \quad -0,154]$

$w_2 = [-0,85 \quad -1,03 \quad -0,63 \quad -1,04 \quad -1,1 \quad -0,74 \quad -0,79 \quad -0,87 \quad -1,04 \quad -0,8]$

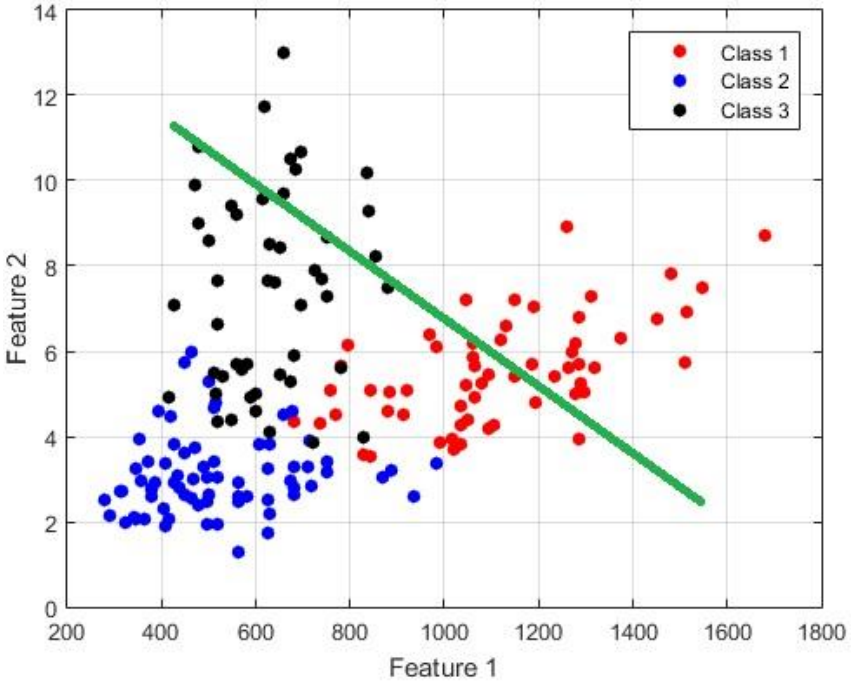
M)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

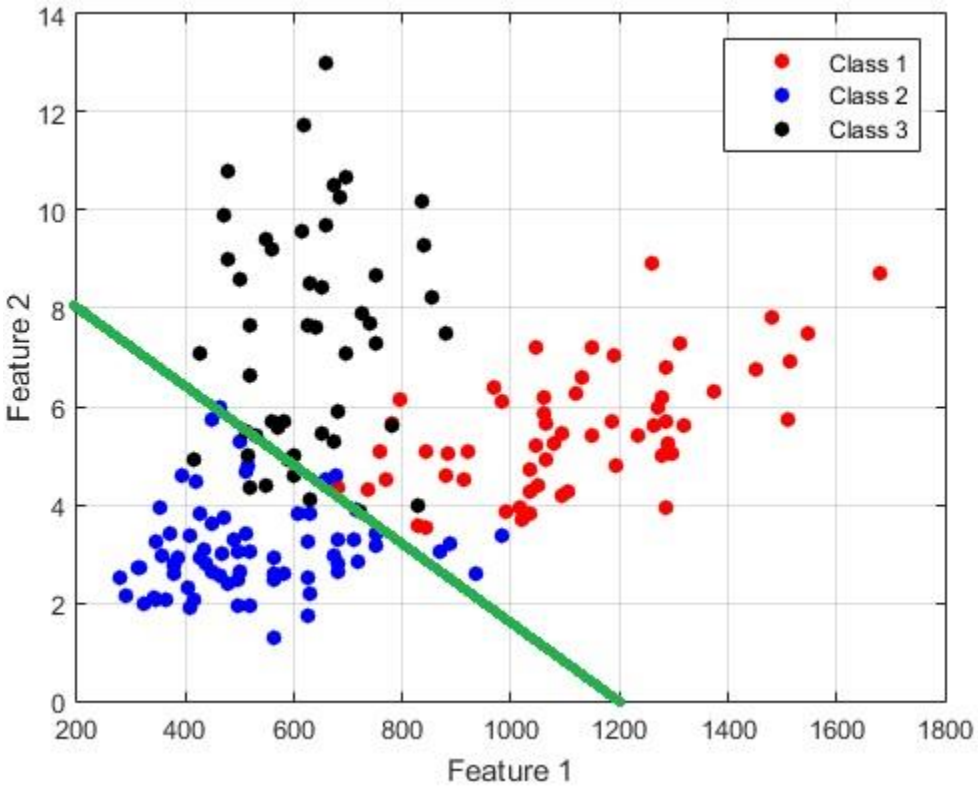
n)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

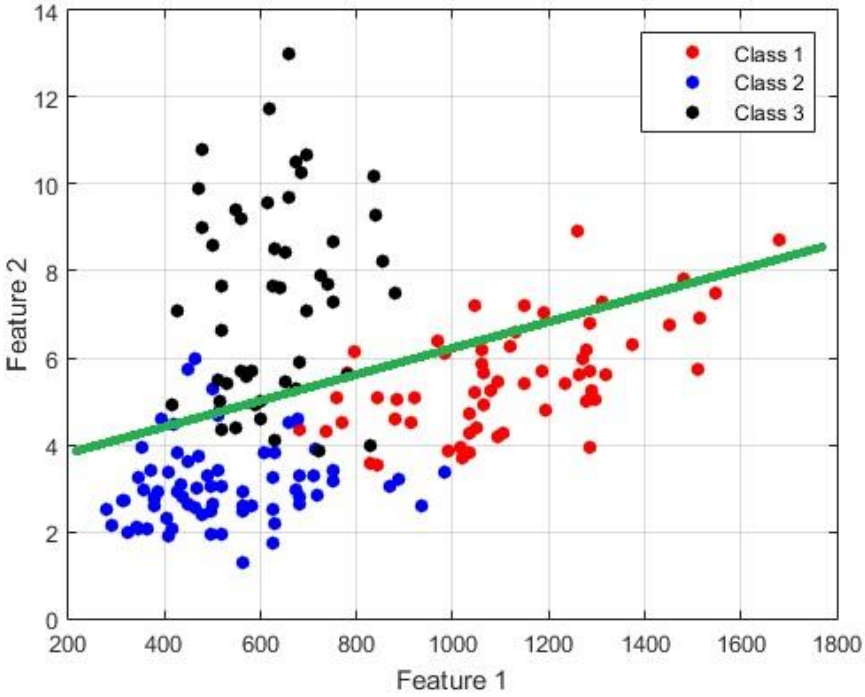
o)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

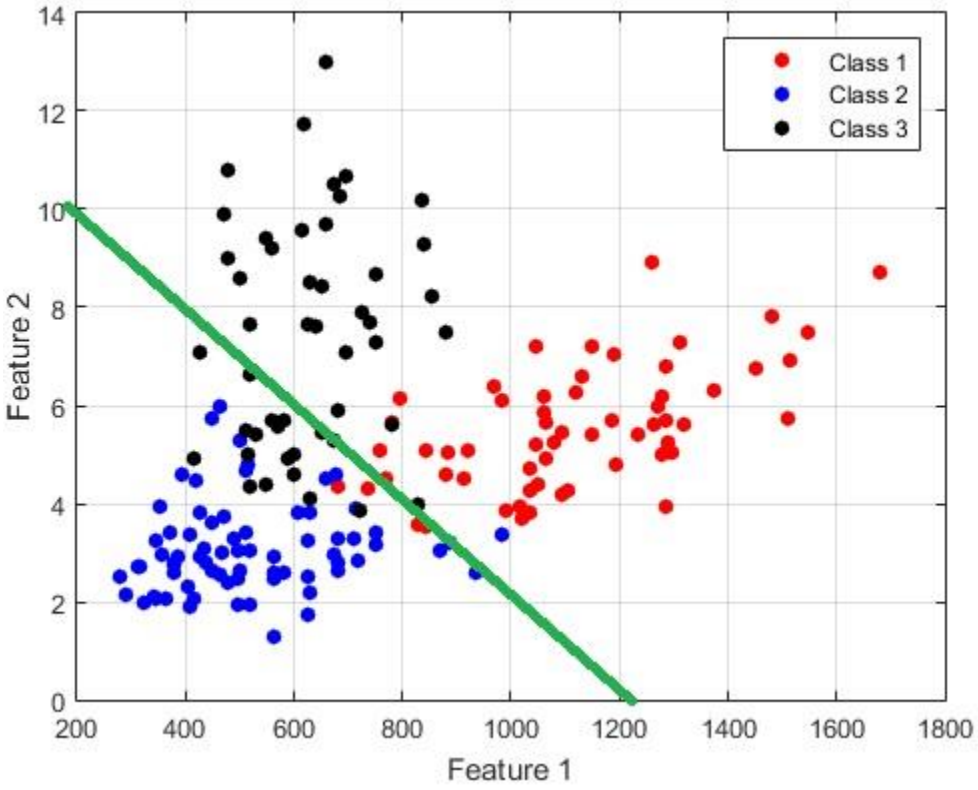
p)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

q)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

r)


Class 1
Class 2
Class 3
Feature 2
Feature 1
0
2
4
6
8
10
12
14
200
400
600
800
1000
1200
1400
1600
1800

۳)

در ابتدا داده ها را برای پردازش میخوانیم.

سپس در مرحله قبل از پردازش ، همانگونه که در متن سوال خواسته شده بود، داده هایی که مقدار آن ها در هدف بی مقدار بوده اند ، حذف کرده ایم. سپس در مرحله بعدی یک ستون با نام ماه ساخته ایم که با فرمولی از داده های date استخراج میشود و برای پردازش داده های دیگر نیاز داریم. در مرحله بعدی داده های باران امروز و فردا را از حالت کاتگوری به عددی تبدیل کرده ایم.

در مرحله بعدی قبل از پردازش ، یک حلقه بر روی ماه ها نوشته ایم که بر اساس ستون ماه ، داده های مربوط به آن ماه را استخراج میکنیم ، سپس همانگونه که در سوال خواسته شده است ، داده هایی که مقادیر گمشده دارند را با میانگین آن ماه محاسبه میکنیم. در این مرحله ، داده های جدید ما بدون هیچ مقدار گمشده ای هستند.

در مرحله بعدی برای سوالات ، ابتدا ماتریس همسبتگی را برای ویژگی ها استخراج میکنیم و بر اساس این ماتریس و داده های مورد نیاز چند ویژگی را حذف میکنیم.

سپس برای سوالات بعدی ابتدا داده ها را جدا میکنیم و ۴۰ درصد داده ها را برای تست کنار میگذاریم و باقی برای آموزش هستند. با سپس یک طبقه بندی کننده نزدیک ترین همسایه را برای این مسئله آموزش میدهیم که با تست های مختلف k = 20 را برای این مسئله انتخاب کرده ام. سپس بر روی داده های تست محاسبه کرده و مقدار صحت آن را محاسبه میکنیم.

در مرحله آخر یک بردار از ویژگی های هوای شهر رشت را ساخته و به این طبقه بندی کننده میدهیم تا معین کند که فردا باران می آید یا خیر.

۴)-

۵)

ابتدا داده ها را برای این سوال فراخوانی میکنیم. سپس با استفاده از کتابخانه pca و با مقدار و داده ها این pca را محاسبه میکنیم و با آن یک دیتافریم جدید میسازیم و بردار های ویژه همانطوری که در سوال خواسته شده بود نمایش داده میشود.

در این مرحله برای پیدا کردن بهترین مقدار پارامتر برای این سوال ، برای مقدار پهنا که در kernelDensity قرار دارد از کتابخانه آن و همچنین برای پیدا کردن بهترین مقدار از Grid Search استفاده کرده ایم و برای تست و محاسبه آن بازه ۰ تا ۱۰ را به آن داده ایم که در این بازه ۲۰ داده را تست کند و بهترین آن را به عنوان بهترین پهنا به خروجی بدهد.

۶)

در این سوال ابتدا یک تابع برای خواندن تمام تصاویر یک فولدر با لیبل آن نوشته ایم. سپس داده ها را فراخوانی کرده و از این تابع استفاده میکنیم و داده ها را دریافت میکنیم. سپس برای محاسبه eigenfaces یک تابع نوشته ایم که در آن از pca استفاده کرده و با داده های آموزش آن را آموزش میدهیم و مقدار بردار های ویژه را در W و میانگین ها را در mu ذخیره میکنیم. سپس از این تابع استفاده کرده و با ۵۰ بردار ویژه آن را میسازیم و مقادیر W و mu را نمایش میدهیم.

سپس همانطوری که در سوال خواسته شده تصاویر را به صورت گرید ۵ در ۱۰ و اندازه هر تصویر به صورت ۱۶۰ در ۳۰ نمایش میدهیم.

سپس در قسمت بعدی ۱۰ تصویر به صورت رندوم انتخاب کرده ایم و برای مقادیر ویژه ۱ تا ۱۰ تمام eigenface ها را محاسبه کرده و هر تصویر را به صورت ۸۰ در ۶۰ در هر مرحله نمایش میدهد. برای نمایش این تصاویر ، ابتدا مقدار وزن ها را محاسبه کرده تصاویر را با استفاده از آن میسازیم.

سپس برای تست کردن داده های تست ، تابع eigenfaces_test را میسازیم که در آن مقدار داده های جدید را با استفاده از داده های گذشته میسازیم ، به این گونه که ابتدا از میانگین آنها را کم میکنیم سپس در مقادیر ویژه کرده تا مقادیر جدید ساخته شود .
سپس یک طبقه بندی کنند knn را با k = 1 ساخته و داده های آموزش را به آن میدهیم. سپس داده های تست را با آن کلاس بندی کرده و میزان خطای آن را با استفاده از نمودار نمایش میدهیم. این کار را برای ۱ مقدار ویژه تا ۵۰ انجام میدهیم.

برای سوال بعدی همین کار را دوباره امتحان کرده بدون در نظر گرفتن ۵ مقدار ویژه اولی و نمودار آن را رسم میکنیم.

در حالت دوم ، در ابتدا مقدار خطا زیاد بوده است اما هر چه تعداد بردار ویژه زیاد میشود برای حالت دوم زودتر مقدار خطا کاهش میابد و در انتها مقدار خطای آن از خطای حالت قبل کمتر است.

۷)

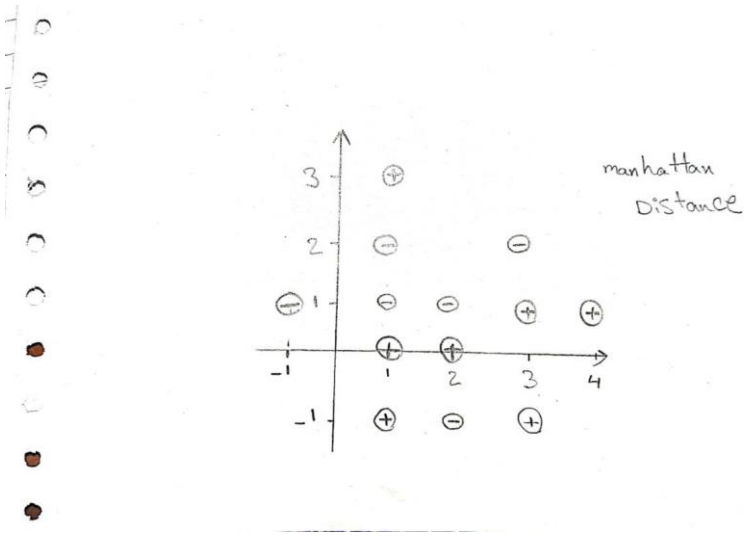
الف)

روش knn نسبت به نویز مقاوم نیست. زیرا اگر یک داده پرتی داشته باشیم پنجره برای رسیدن به آن داده بزرگ میشود که اگر داده های زیادی در یک جا جمع شده باشند و به این داده پرت نیاز دارند و پنجره بزرگ شده و باعث میشود که مقدار غیر واقعی برای بعضی نقاط میانی برای این توزیع ایجاد شود.

ب)

برای روش $k = 1$ ، مقدار کلاس داده برابر با نزدیک ترین داده میشود که در این حالت خطایی رخ نمیدهد. اما برای k های بزرگتر از ۱ چون نیاز به مقایسه با داده های اطراف است تا بتواند نزدیک ترین همسایه ها را بیابد ، مقدار خطا برابر فاصله داده تا داده های دیگر میشود.

پ)



ت)

در زمانی که داده ها به صورت کاملا یکنواخت پراکنده شده باشند در این حالت تخمین توزیع پارزن و نزدیک ترین همسایه برابر خواهند بود.

ث)

در pca ما به دنبال ترکیب ابعاد هستیم تا بتوانیم ابعاد را در تعداد کمتر و با کیفیت بهتری داشته باشیم و مولفه هایی که ما در این مسئله استفاده میکنیم ، مولفه هایی هستند که این ابعاد با آن ها درگیر هستند و اگر از مولفه های بیشتری استفاده کنیم ، ابعاد با این مولفه ها درگیری ندارند و از آنها استفاده ای نمیکنند و فقط یک سربار برای ما ایجاد میکنند پس مفید نیستند.

ج)

هنگامی که از pca استفاده میکنیم ، بردار های ویژه ای که مقادیر ویژه زیادی دارند را انتخاب میکنیم ، به طور کلی به گونه ای بردار هایی که مقادیر ویژه زیادی دارند از کیفیت بهتر و ارزش بالاتری برخوردار هستند. در تصویر ، ابتدا ماتریس را از d*d به ماتریس d^2 1 * تبدیل میکنیم ، یعنی هر پیکسل تصویر برای ما به صورت یک بعد در نظر گرفته میشود ، پس در تصاویر وقتی تعدادی بردار ویژه با مقادیر عالی ترین انتخاب میشوند به گونه ای مقادیر نویز که مقادیر ویژه کمی دارند در نظر گرفته نمیشوند و اینگونه در سیستم محاسبات حذف میشوند.

د)

برای محاسبه بردار ویژه لازم است scatter matrix محاسبه شوند که از ضرب ماتریس تصویر منهای میانگین در خودش به دست می آید که در تصاویر به علت ابعاد بالا (مثلا یک تصویر ۱۰۰ * ۱۰۰ تعداد ابعاد آن ۱۰۰۰ است و این مقدار برای تصاویر بزرگتر خیلی زیاد میشود) که به علت ابعاد زیاد آن ، محاسبه مقدار و بردار ویژه و مراحل آن بسیار دشوار بوده و قابل حل نخواهد بود. برای حل این مشکل به جای ضرب ماتریس تصاویر در ترانهاده خودش ، به صورت برعکس ، ترانهاده تصاویر در ماتریس تصاویر ضرب میشود که در این حالت ابعاد ماتریس scatter بدست آمده از n * n خواهد بود که n برابر تعداد داده های مسئله میباشد.