دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر

گزارش تکلیف پنجم درس شناسائی آماری الگو

دانشجو:
سید احمد نقوی نوزاد
ش-د: ۹۴۱۳۱۰۶۰

استاد:
دکتر رحمتی

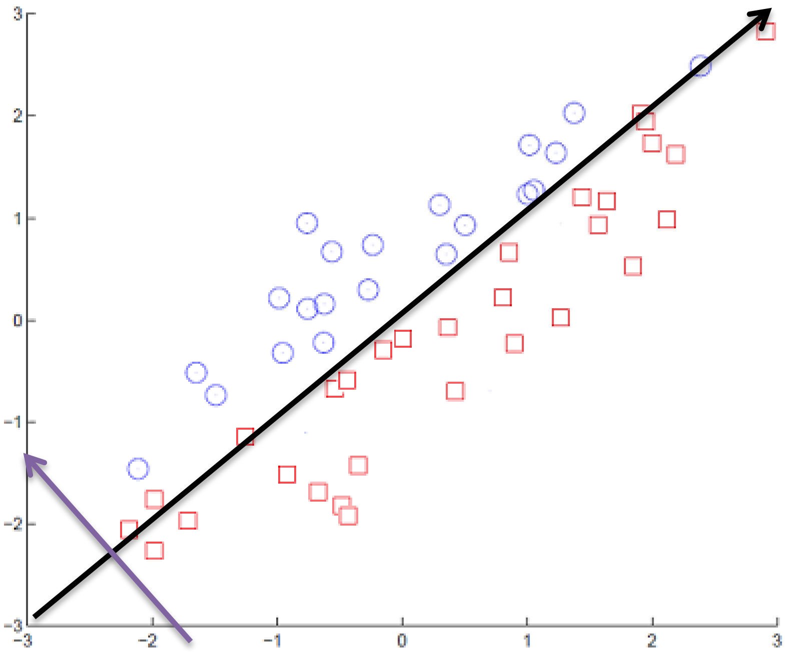
پائیز ۹۵

جواب سوال ۱

با توجه به صورت سؤال، در صورتی که در هنگام اجرای PCA، تعداد ابعاد در فضای جدید را برابر با تعداد ابعاد در فضای اصلی قرار دهیم، تنها اتفاقی که می‌افتد این است که دستگاه مختصات در راستای پراکندگی داده‌ها دچار چرخش شده است و از نظر بازنمائی داده‌ها اتفاق خاصی نیفتاده است. چرا که در کاربرد اصلی PCA، تسهیل بازنمائی داده‌ها با ابعاد زیاد می‌باشد.

جواب سوال ۲

توضیحات مکفی در ذیل تصویر قید شده است:



شکل ۱. توزیع داده‌های دو کلاس مثبت (نقاط دایره‌ای‌شکل) و منفی (نقاط مربعی‌شکل)؛ پیکان مشکی: جهت اولین جزء اصلی (**First Principle Component**)؛ پیکان بنفش: جهت بردار تصویرسازی با استفاده از ضابطه‌ی فیشر (**Fisher Linear Discriminant Method Projection Line**).

جواب سوال ۳

داده‌های مسئله به قرار زیر می‌باشند:

$$(-1,1), (0,0), (1,1)$$

قسمت الف)

$$X=\begin{pmatrix}-1&0&1\\1&0&1\end{pmatrix}$$

$$mean=\begin{pmatrix}\frac{-1+0+1}{3}\\\frac{1+0+1}{3}\end{pmatrix}=\begin{pmatrix}0\\\frac{2}{3}\end{pmatrix}$$

$$Y=X-mean=\begin{pmatrix}-1-0&0-0&1-0\\1-\frac{2}{3}&0-\frac{2}{3}&1-\frac{2}{3}\end{pmatrix}=\begin{pmatrix}-1&0&1\\\frac{1}{3}&-\frac{2}{3}&\frac{1}{3}\end{pmatrix}$$

$$S=YY^t=\begin{pmatrix}-1&0&1\\\frac{1}{3}&-\frac{2}{3}&\frac{1}{3}\end{pmatrix}\begin{pmatrix}-1&\frac{1}{3}\\0&-\frac{2}{3}\\1&\frac{1}{3}\end{pmatrix}=\begin{pmatrix}2&0\\0&\frac{2}{3}\end{pmatrix}$$

$$|S-\lambda I|=0\Rightarrow\begin{vmatrix}2-\lambda&0\\0&\frac{2}{3}-\lambda\end{vmatrix}=0\Rightarrow(2-\lambda)\left(\frac{2}{3}-\lambda\right)=0\Rightarrow\lambda_1=2,\lambda_2=\frac{2}{3}$$

The first principle component is the eigenvalue corresponding to the largest eigenvector λ_1=2.

Then we have:

$$Sv_1=\lambda_1 v_1\Rightarrow\begin{pmatrix}2&0\\0&\frac{2}{3}\end{pmatrix}\begin{pmatrix}v_{11}\\v_{12}\end{pmatrix}=2\begin{pmatrix}v_{11}\\v_{12}\end{pmatrix}\Rightarrow\begin{cases}2v_{11}=2v_{11}\\\frac{2}{3}v_{12}=2v_{12}\end{cases}\Rightarrow v_{11}=1,v_{12}=0\Rightarrow v_1=\begin{pmatrix}1\\0\end{pmatrix}$$

$$Sv_2=\lambda_2 v_2\Rightarrow\begin{pmatrix}2&0\\0&\frac{2}{3}\end{pmatrix}\begin{pmatrix}v_{21}\\v_{22}\end{pmatrix}=\frac{2}{3}\begin{pmatrix}v_{21}\\v_{22}\end{pmatrix}\Rightarrow\begin{cases}2v_{21}=\frac{2}{3}v_{21}\\\frac{2}{3}v_{22}=\frac{2}{3}v_{12}\end{cases}\Rightarrow v_{21}=0,v_{22}=1\Rightarrow v_2=\begin{pmatrix}0\\1\end{pmatrix}$$

قسمت ب)

$$X_{new}=v_1^tX=\begin{pmatrix}1&0\end{pmatrix}\begin{pmatrix}-1&0&1\\1&0&1\end{pmatrix}\Rightarrow X_{new}=\begin{pmatrix}-1&0&1\end{pmatrix}$$

$$E(X_{new})=\frac{-1+0+1}{3}=0,\quad E(X_{new}^2)=\frac{1+0+1}{3}=\frac{2}{3}\Rightarrow$$

$$V(X_{new})=E(X_{new}^2)-E^2(X_{new})\Rightarrow=V(X_{new})=\frac{2}{3}$$

قسمت ج)

$$X_{new}=\begin{pmatrix}-1&0&1\end{pmatrix}\to X_{rec}=\begin{pmatrix}v_1^t\\v_2^t\end{pmatrix}X_{new}^{aug}=\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}-1&0&1\\0&0&0\end{pmatrix}\Rightarrow X_{rec}=\begin{pmatrix}-1&0&1\\0&0&0\end{pmatrix}$$

Here we write the MATLAB code!

$$\mathrm{Rec\_}Err=X-X_{rec}=MSE=mean\left(sum\left(\left(\left(\begin{pmatrix}-1&0&1\\1&0&1\end{pmatrix}-\begin{pmatrix}-1&0&1\\0&0&0\end{pmatrix}\right).\wedge 2\right)\right)\right)=mean\begin{pmatrix}1&0&1\end{pmatrix}$$

$$\Rightarrow\mathrm{Rec\_}Err=\frac{2}{3}$$

جواب سوال ۴

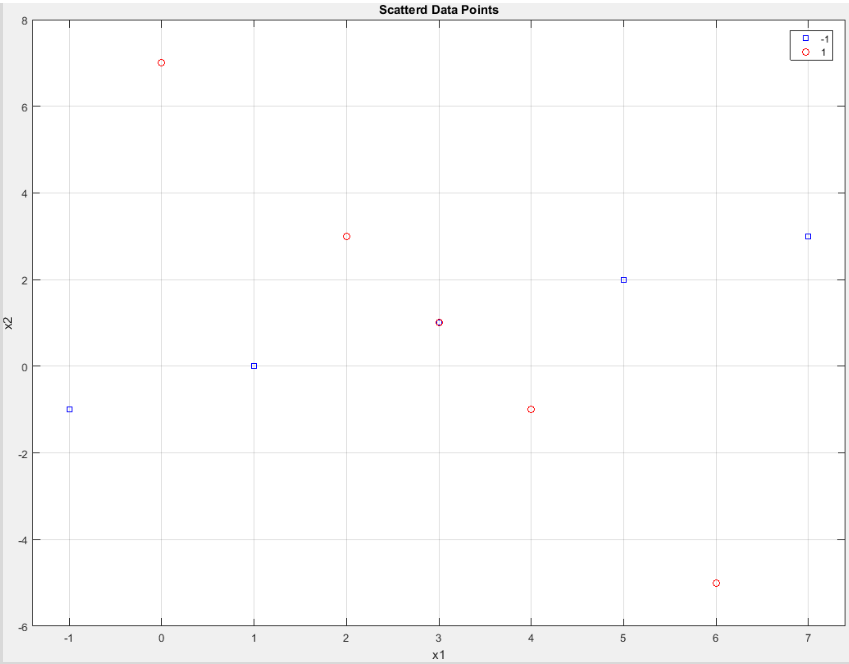
کدهای مربوط به این سؤال در فایل ex04 قرار دارد.

داده‌های مسئله به قرار زیر می‌باشند:

$$X=\begin{bmatrix}3 & 2 & 4 & 0 & 6 & 3 & 1 & 5 & -1 & 7\\ 1 & 3 & -1 & 7 & -5 & 1 & 0 & 2 & -1 & 3\end{bmatrix}$$

به طوری که هر سطر معادل یک بُعد و هر ستون نیز معادل یک نمونه‌داده می‌باشد.

قسمت الف)



شکل ۲. توزیع داده‌های دو کلاس مثبت (نقاط دایره‌ای‌شکل) و منفی (نقاط مربعی‌شکل).

قسمت ب)

$$Y=X-E(X)=\begin{bmatrix}3 & 2 & 4 & 0 & 6 & 3 & 1 & 5 & -1 & 7\\ 1 & 3 & -1 & 7 & -5 & 1 & 0 & 2 & -1 & 3\end{bmatrix}-\begin{bmatrix}3\\ 1\end{bmatrix}\Rightarrow$$

$$Y=\begin{bmatrix}0 & -1 & 1 & -3 & 3 & 0 & -2 & 2 & -4 & 4\\ 0 & 2 & -2 & 6 & -6 & 0 & -1 & 1 & -2 & 2\end{bmatrix}$$

قسمت ج)

$$E(Y)=zeros(2,10);$$

$$C=\frac{1}{10-1}YY^{t}\Rightarrow C=\begin{bmatrix}6.67 & -2.22\\ -2.22 & 10\end{bmatrix}$$

$$C=\begin{bmatrix} 6.67 & -2.22 \\ -2.22 & 10 \end{bmatrix} \rightarrow$$

$$|C-\lambda I|=0 \Rightarrow \begin{vmatrix} 6.67-\lambda & -2.22 \\ -2.22 & 10-\lambda \end{vmatrix}=0 \Rightarrow (6.67-\lambda)(10-\lambda)-(-2.22)^2=0 \Rightarrow \lambda_1=11.11, \lambda_2=5.56$$

The first principle component is the eigenvalue corresponding to the largest eigenvector $\lambda_1=11.11$. Then we have:

$$Cv_1=\lambda_1 v_1 \Rightarrow \begin{bmatrix} 6.67 & -2.22 \\ -2.22 & 10 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix}=11.11\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} \Rightarrow \begin{cases} 6.67v_{11}-2.22v_{12}=11.11v_{11} \\ -2.22v_{11}+10v_{12}=11.11v_{12} \end{cases} \Rightarrow v_1=\begin{pmatrix} -.8944 \\ -.4472 \end{pmatrix}$$

$$Cv_2=\lambda_2 v_2 \Rightarrow \begin{bmatrix} 6.67 & -2.22 \\ -2.22 & 10 \end{bmatrix}\begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix}=5.56\begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} \Rightarrow \begin{cases} 6.67v_{21}-2.22v_{22}=5.56v_{21} \\ -2.22v_{21}+10v_{22}=5.56v_{22} \end{cases} \Rightarrow v_2=\begin{pmatrix} -.4472 \\ .8944 \end{pmatrix}$$

قسمت د)

در این قسمت می‌بایست واریانس داده‌های تصویرشده بر روی اولین جزء اصلی را محاسبه نمائیم. داریم:

$$X_{new}=v_1^t X=[-.8944 \quad -.4472]\begin{bmatrix} 3 & 2 & 4 & 0 & 6 & 3 & 1 & 5 & -1 & 7 \\ 1 & 3 & -1 & 7 & -5 & 1 & 0 & 2 & -1 & 3 \end{bmatrix} \Rightarrow$$

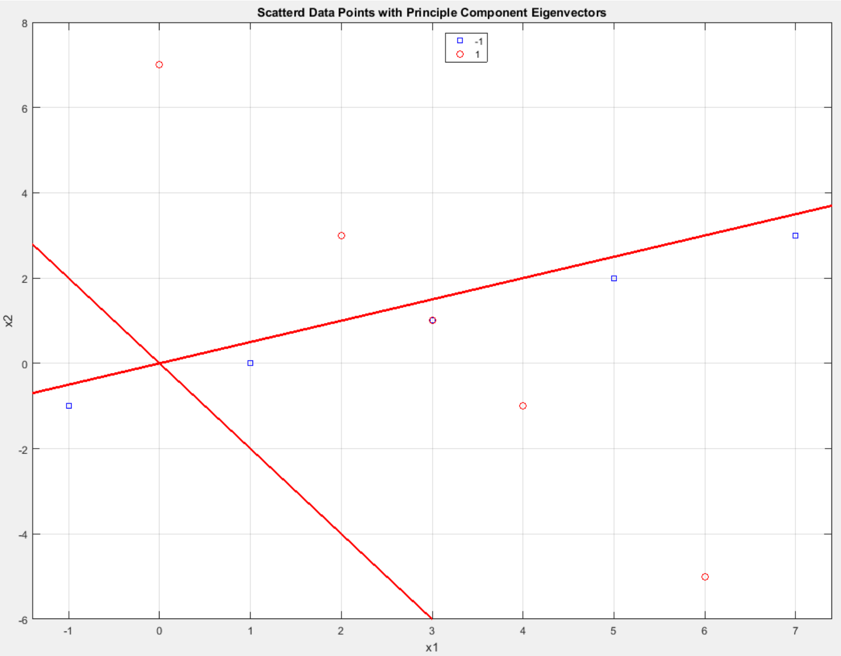
$$X_{new}=[0 \quad 2.2361 \quad -2.2361 \quad 6.7082 \quad -6.7082 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

$$Var(X_{new})=11.11$$

قسمت ه)

The principle components are as follows:

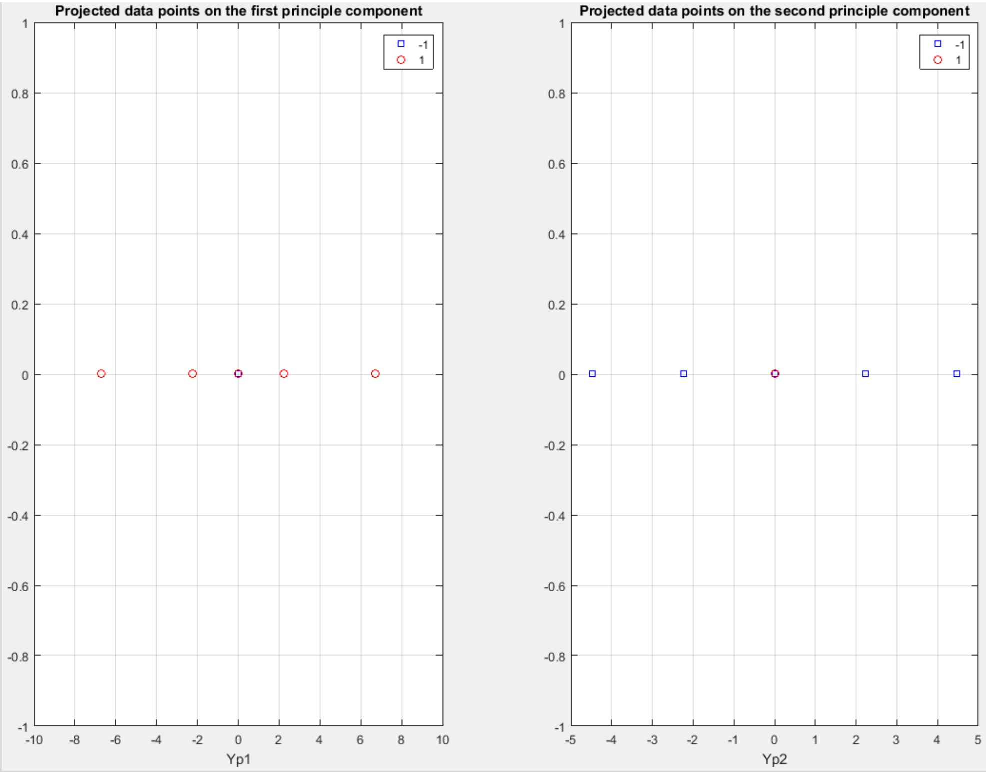
$$v_1=\begin{pmatrix} -.8944 \\ -.4472 \end{pmatrix}, \quad v_2=\begin{pmatrix} -.4472 \\ .8944 \end{pmatrix}$$



شکل ۳. توزیع داده‌های دو کلاس مثبت (نقاط دایره‌ای‌شکل) و منفی (نقاط مربعی‌شکل) به همراه اجزاء اصلی رسم‌شده (که به دلیل تنظیمات درونی نرم‌افزار متلب ظاهراً بر یکدیگر عمود نمی‌باشند).

قسمت و)

در این قسمت مجموعه داده‌ی Y را بر روی تک‌تک Principle Componentها تصویر می‌نمائیم. نتایج به شرح ذیل می‌باشند:

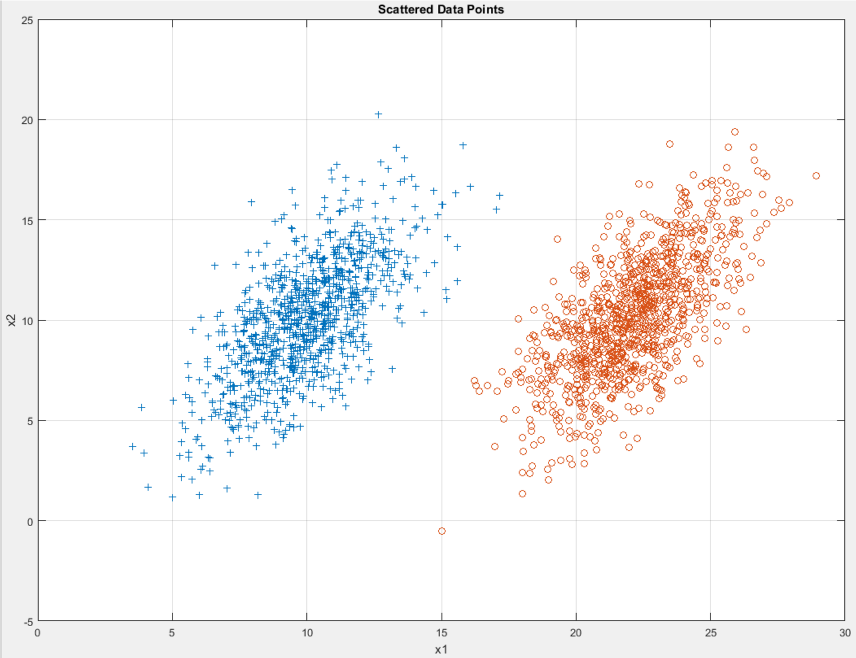


شکل ۴. سمت چپ: تصویر داده‌های ماتریس Y بر روی اولین جزء اصلی؛ سمت راست: تصویر داده‌های ماتریس Y بر روی دومین جزء اصلی؛ همانطور که پیداست از آن‌جا که در راستای اولین جزء اصلی واریانس داده‌ها بیشینه می‌باشد، لذا تصویرسازی آن‌ها بر روی اولین جزء اصلی سبب می‌گردد تا شاهد تفکیک‌پذیری مناسبی نباشیم، ولی در مورد دومین جزء اصلی چون میزان واریانس داده‌های تصویرشده کمتر می‌باشد، لذا شاهد تفکیک‌پذیری بیشتری در داده‌های تصویرشده می‌باشیم.

جواب سوال ۵

کدهای مربوط به این سؤال در فایل ex05 قرار دارد.

$$\mu_1 = [10 \quad 10]^T , \quad \mu_2 = [22 \quad 10]^T , \quad \Sigma_1 = \Sigma_2 = \begin{bmatrix} 4 & 4 \\ 4 & 9 \end{bmatrix}$$



شکل ۵. داده‌های ترسیم‌شده مربوط به دو کلاس با مشخصات قیدشده در بالا.

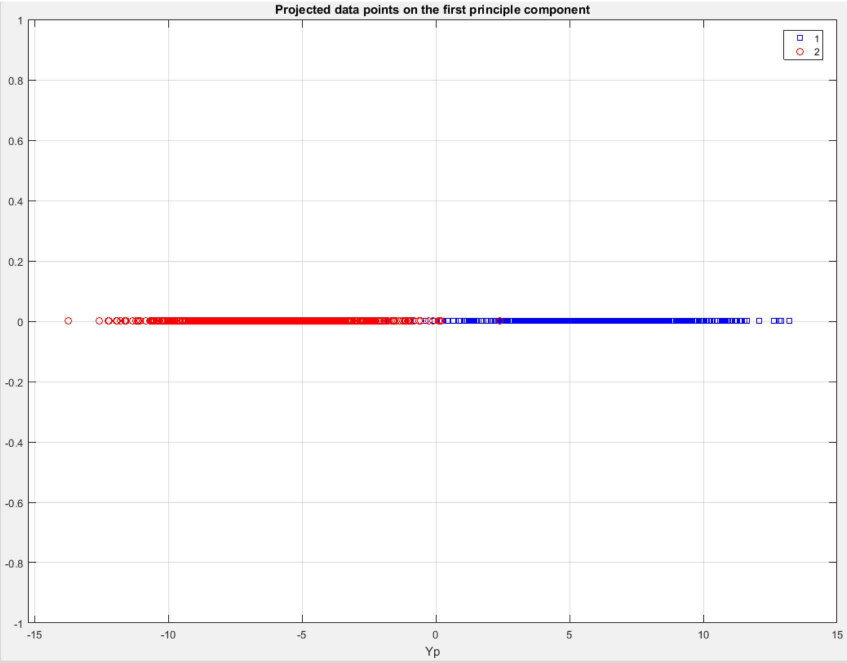
قسمت الف)

The equation of first principle component is as follows:

$$-.9911x_1 - .1334x_2 = 0$$

شکل ۶. توزیع داده‌های دو کلاس به همراه معادله‌ی اولین جزء اصلی مربوط به کل داده‌ها که PCA داده‌ها را بر روی آن تصویر می‌کند؛ همان‌طور که قابل مشاهده است واریانس داده‌ها در راستای این خط بیشینه می‌باشد که این همان خاصیت اصلی PCA می‌باشد.

قسمت ب)

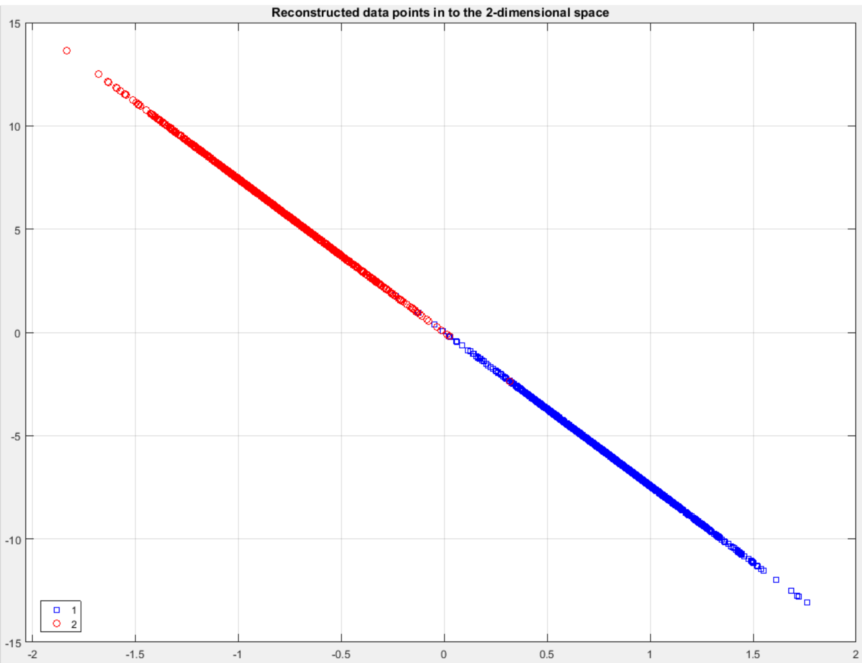


شکل ۷. تصویر داده‌های دو کلاس بر روی اولین جزء اصلی؛ همان‌طور که پیداست داده‌های دو کلاس پس از تصویرسازی تا حدی دارای هم‌پوشانی می‌باشند که علت این مسئله نیز همان واریانس بالاتر داده‌ها در راستای اولین جزء اصلی است. لازم به ذکر است که میزان این هم‌پوشانی در راستای دومین جزء اصلی کمتر می‌باشد.

قسمت ج)

همان‌طور که پیداست داده‌های دو کلاس پس از تصویرسازی تا حدی دارای هم‌پوشانی می‌باشند که علت این مسئله نیز همان واریانس بالاتر داده‌ها در راستای اولین جزء اصلی است. لازم به ذکر است که میزان این هم‌پوشانی در راستای دومین جزء اصلی کمتر می‌باشد.

قسمت د)



شکل ۸. داده‌های بازسازی‌شده از فضای یک‌بُعدی نگاشت‌شده به فضای دوبعدی اصلی.

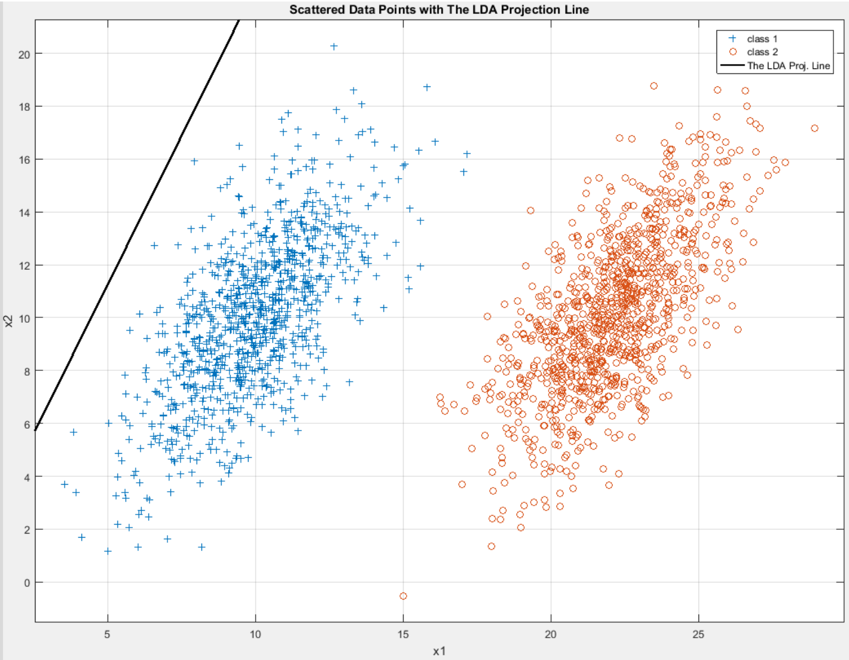
میزان خطای بازسازی در این‌جا مانند قسمت (ج) سؤال سوم محاسبه می‌شود. طبق محاسبات کامپیوتری داریم:

The reconstruction error is: 91.29523

قسمت ه)

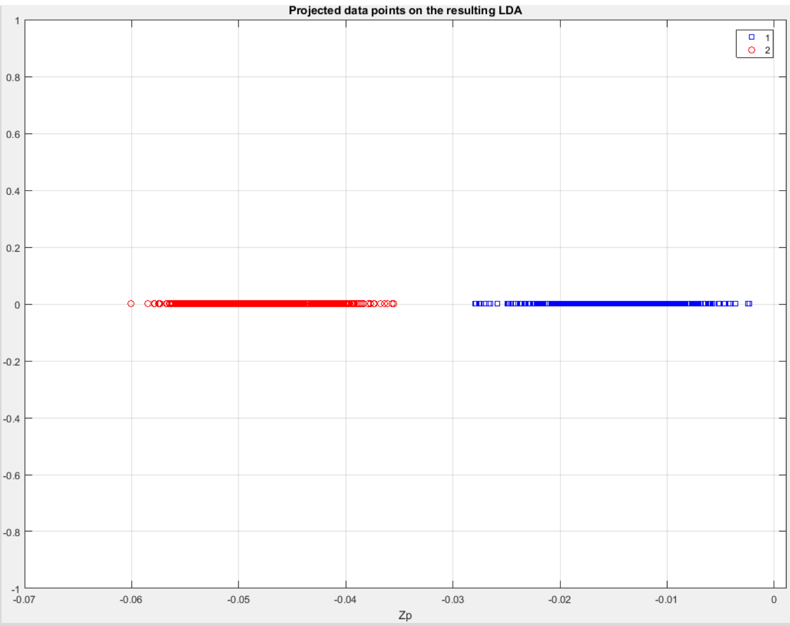
در این‌جا جهت به دست‌آوردن خط تصویرسازی حاصله از روش جداسازی خطی فیشر[1] (LDA[2]) از ذکر محاسبات خودداری کرده و به ذکر کد متلب بسنده می‌نمائیم. داریم:

```
mean1 = mean(C1); mean2 = mean(C2);
S1 = (C1-repmat(mean1,m1,1))'*(C1-repmat(mean1,m1,1));
S2 = (C2-repmat(mean2,m2,1))'*(C2-repmat(mean2,m2,1));
Sw = S1 + S2;
W = Sw\(mean1-mean2)';
```



شکل ۹. توزیع داده‌های دو کلاس به همراه خط تصویرسازی حاصله از روش جداسازی خطی فیشر (**Fisher Linear Discriminant Method Projection Line**) که یکی از روش‌های **LDA** می‌باشد.

قسمت و)



شکل ۱۰. تصویر داده‌های دو کلاس بر روی بردار تصویرسازی **LDA**؛ همان‌طور که کاملا پیداست، داده‌های نگاشت‌شده کاملا از یکدیگر به صورت خطی قابل تفکیک می‌باشند.

[1] Fisher Linear Discriminant Method Projection Line
[2] Linear Discriminant Analysis (LDA)

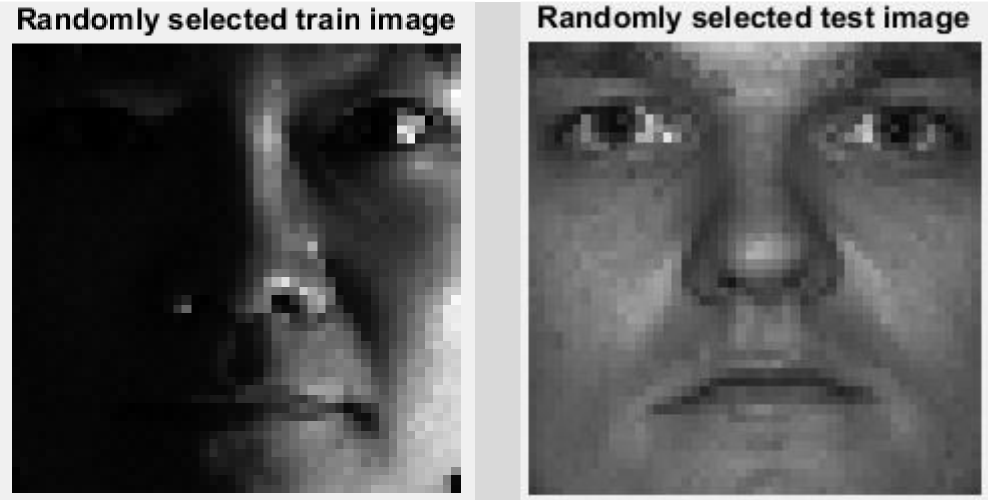
قسمت ز)

همان‌طور که از قسمت (و) کاملا پیداست، داده‌های نگاشت‌شده کاملا از یکدیگر به صورت خطی قابل تفکیک می‌باشند. چرا که روش LDA رسماً به دنبال برداری است که پس از نگاشت داده‌ها بر روی آن بتوان از هر کدام از روش‌های جداساز خطی استفاده نمود و البته در مورد داده‌های این سوال، از آن‌جا که ماتریس کوواریانس داده‌های دو کلاس کاملا با یکدیگر برابرند، معادله‌ی مرز تصمیم‌گیری قطعاً یک خط راست می‌باشد.

جواب سوال ۶

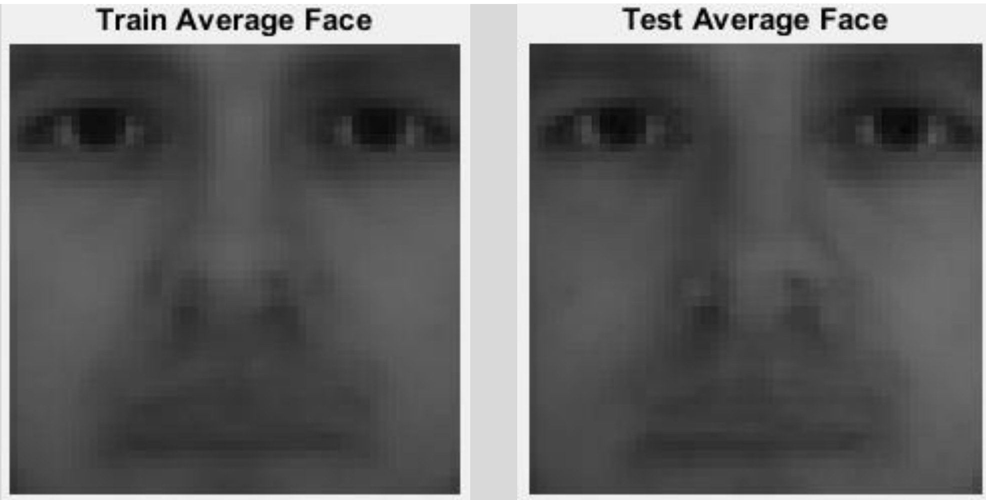
کدهای مربوط به این سؤال در فایل‌های ex06 و eigFaceFeat قرار دارند.

قسمت ب)



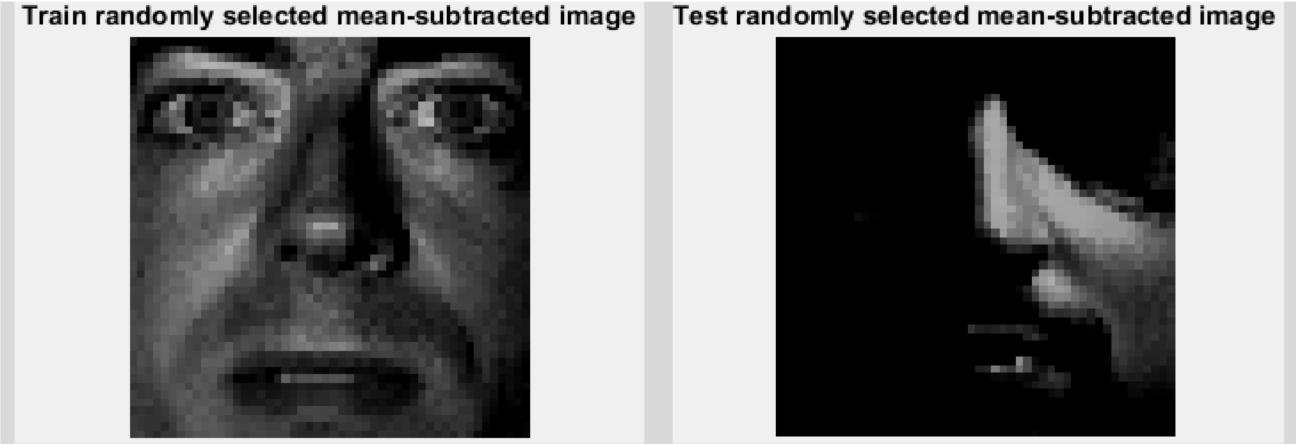
شکل ۱۱. سمت چپ: تصویر به صورت تصادفی انتخاب‌شده از مجموعه‌داده‌ی آموزشی؛ سمت راست: تصویر بـه صـورت تصادفی انتخاب‌شده از مجموعه‌داده‌ی آزمایشی.

قسمت ج)



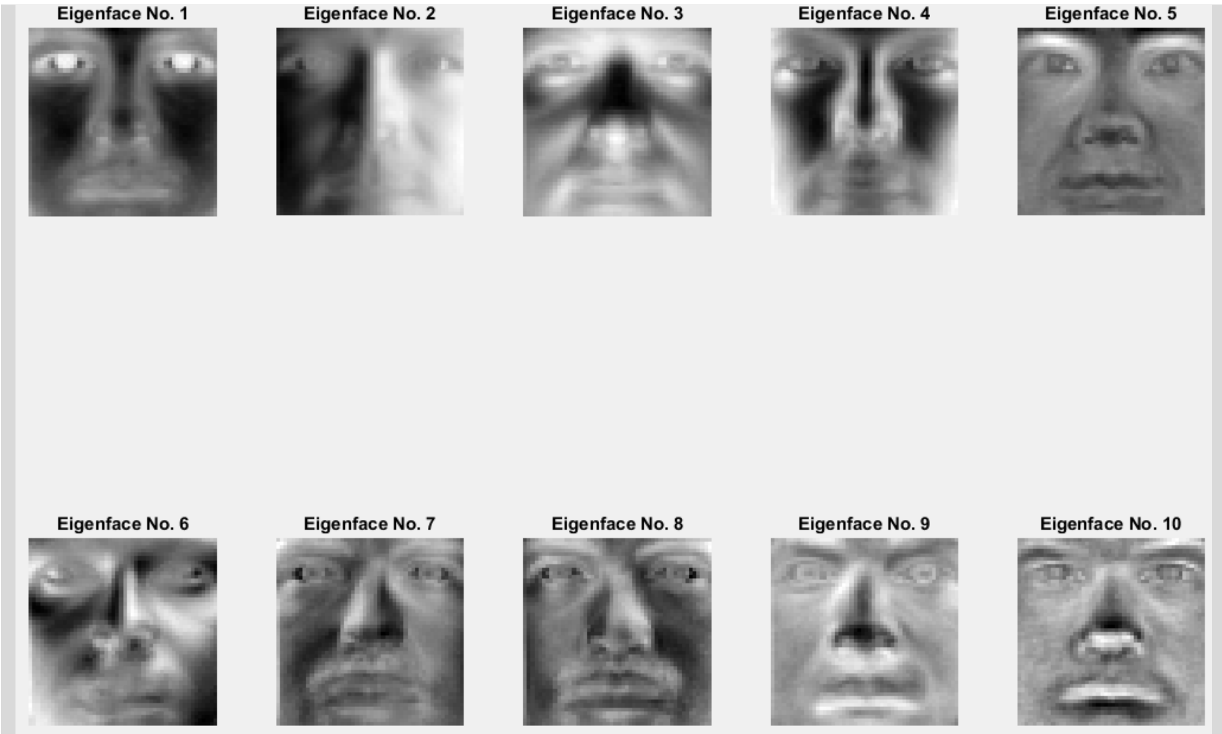
شکل ۱۲. سمت چپ: تصویر میانگین مجموعه‌داده‌ی آموزشی؛ سمت راست: تصویر میانگین مجموعه‌داده‌ی آزمایشی.

قسمت د)



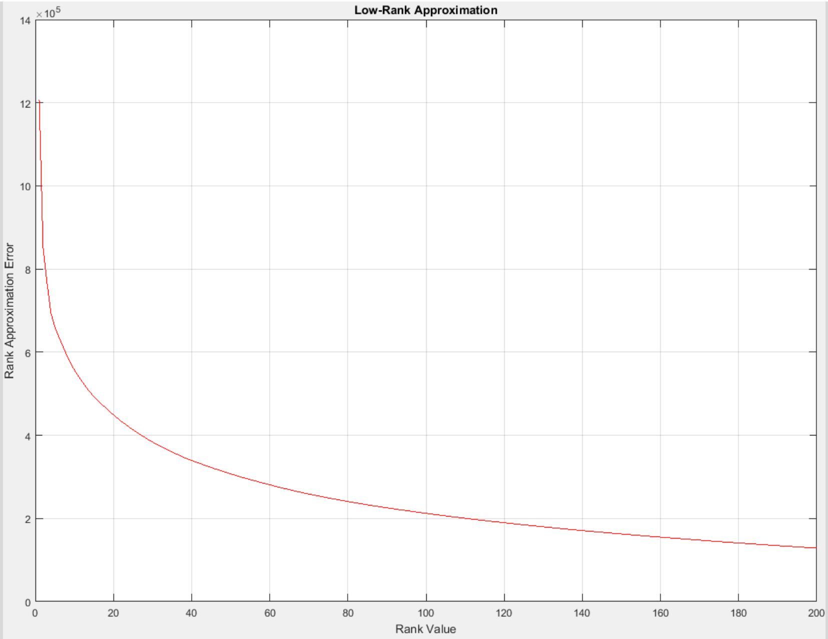
شکل ۱۳. سمت چپ: تصویر انتخاب‌شده به صورت تصادفی و از تصویر میانگین‌کسرشده از مجموعه‌داده‌ی آموزشی؛ سمت راست: تصویر انتخاب‌شده به صورت تصادفی و از تصویر میانگین‌کسرشده از مجموعه‌داده‌ی آزمایشی.

قسمت ه)



شکل ۱۴. تعداد ۱۰ به اصطلاح **eigenFace** اول ماتریس **V** حاصل از تجزیه‌ی ماتریس داده‌های آموزشی **X** به مقادیر منفرد (**Singular Values**) آن به صورت $X = U\Sigma V^T$ *where* $[U, \Sigma, V] = svd(X)$.

قسمت و)



شکل ۱۵. نمودار خطای تخمین رتبه‌ی r-اُم برای ماتریس داده‌های آموزشی X و به ازای r=1,...200.

قسمت ز)

کد این قسمت در فایل eigFaceFeat.m قرار دارد.

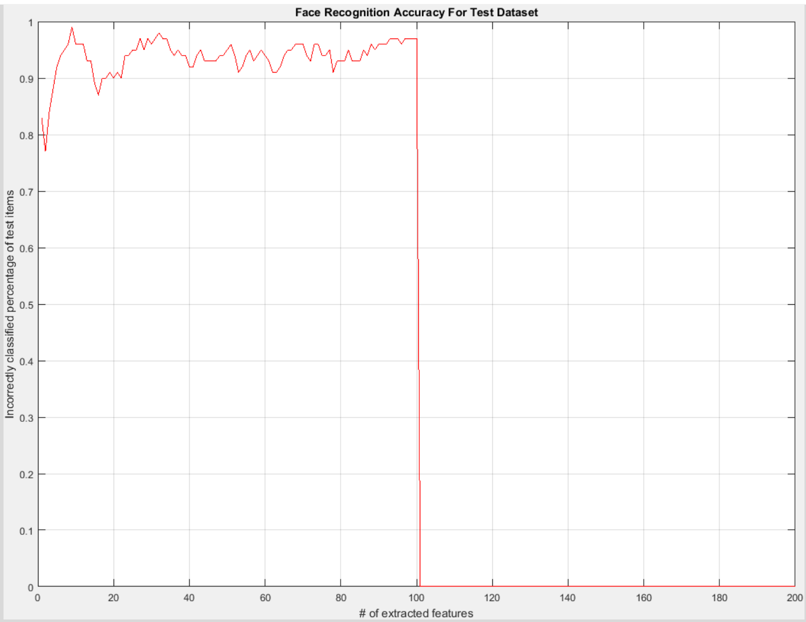
قسمت ح)

از آن‌جا که داده‌های این سؤال دارای بیشتر از یک برچسب می‌باشند (برچسب‌ها اعداد ۱ تا ۱۰ می‌باشند)، لذا ناچاریم تا از Multinomial Logistic Regression جهت پیش‌بینی کلاس داده‌ها استفاده نمائیم. به همین منظور از تولباکس متلب با نام mnrfit() و mnrval() استفاده خواهیم کرد.

نتیجه‌ی قسمت اول سؤال به صورت زیر می‌باشد:

Classification accuracy on the test set with low-rank = 10 is:
# of incorrectly classified items: 97
# of correctly classified items: 3

نتیجه‌ی قسمت دوم سوال نیز به قرار زیر می‌باشد:



شکل ۱۶. نمودار درصد خطای دسته‌بندی با استفاده از مدل Multinomial Logistic Regression بر روی داده‌های تست پس از آموزش مدل بر روی داده‌های آموزشی و به ازای r=1,...200.

در مورد قسمت اخیر باید گفت که به دلیل شدیداً طولانی‌بودن رویه‌ی آموزش مدل Multinomial Logistic Regression و پس از حدود دو روز در حال اجرابودن برنامه، به دلیل ضیق وقت از کسب نتیجه‌ی نهائی منصرف شدیم! و به تقریبا نصف نتیجه‌ی نهائی بسنده نمودیم. اما آن‌طور که از نمودار حاصله پیداست، حتی به ازای نصف داده‌ها شاهد تغییرات مفهوم و قابل اتکائی از صحت دسته‌بندی به ازای افزایش رتبه‌ی ماتریس نمی‌باشیم. چرا که انتظار ما این بود که با افزایش مقدار رتبه‌ی ماتریس، طبعاً مدل ما بهتر آموزش دیده و در مورد داده‌های آزمایشی نیز بهتر تصمیم‌گیری نماید.