تمرین چهارم درس یادگیری ماشین
دکتر باباعلی

سیدعلیرضا مولوی

فهرست مطالب

۱ PCA[۱]

داده هایی که در دسترس داریم:

$$D = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)} | \forall i (x^{(i)} \in R^d \wedge x^{(i)} \sim i.i.d)\} \quad (۱)$$

هدف ما این است که داده ها را از فضای d به فضای d' منتقل کنیم به صورتی که تا حد امکان اطلاعات مربوط به داده حفظ شوند.

۱.۱ مدل محاسباتی

مدل محاسباتی برای کاهش بعد از فضای d به d' به صورت زیر در نظر گرفته ایم.

$$z = f(x; W) = W^T x \in R^{(d',1)} \quad \text{Data projection} \quad (۲)$$

$$x \in R^{(d,1)}, \quad E[x] = (0, 0, \cdots, 0)^T \quad (۳)$$

$$W = (\omega_1, \omega_2, \cdots, \omega_{d'}) \in R^{(d',d)}, \quad \omega_i^T \omega_j = \delta_{i,j} \quad (۴)$$

$$x^* = Wz \in R^{(d,1)} \quad \text{Data reconstruction} \quad (۵)$$

هر $w^i \in R^{(d,1)}$ یک مولفه اصلی[۲] است. x^* بازسازی داده بر روی فضای اولیه است.
فرض گرفتیم داده ها که میانگین صفر دارند، اگر هم میانگین صفر نبود می توانیم از طریق معادله زیر میانگین داده را صفر می کنیم:

$$x = x - E[x]$$

توجه: در عمل توصیه می شود علاوه بر صفر کردن میانگین، انحراف از معیار داده ها هم ۱ کرد:

$$x = \frac{x - E[x]}{\sigma(x)}$$

با این کار خطا به طور قابل توجهی کاهش می یابد. (در پیاده سازی کد، من از این روش استفاده کردم)
توجه: زمانی که میانگین داده را صفر می کنیم و یا داده را استاندارد می کنیم، باید در هنگام بازسازی داده، معکوس این تغییرات را اعمال کنیم:

$$x^* = (zW\sigma) + \mu \quad (۶)$$

۲.۱ مدل محاسباتی در حالت برداری

اکنون مدل برای حالت **برداری** باز نویسی می کنیم:

$$X = \begin{pmatrix} \cdots & x^{(1)} & \cdots \\ \cdots & x^{(1)} & \cdots \\ & \vdots & \\ \cdots & x^{(n)} & \cdots \end{pmatrix} \in R^{(N,d)} \quad (۷)$$

$$Z = f(X; W) = XW = \begin{pmatrix} \cdots & z^{(1)} & \cdots \\ \cdots & z^{(1)} & \cdots \\ & \vdots & \\ \cdots & z^{(n)} & \cdots \end{pmatrix} \in R^{(N,d')} \quad (۸)$$

(۹)

$$X^* = ZW^T \in R^{(N,d)} \quad (۱۰)$$

$$X^* = (ZW^T\sigma) + \mu \quad \text{If we standarized our data} \quad (۱۱)$$

[۱] Principal Component Analysis
[۲] Principal Component

که N تعداد کل داده ها است.

۳.۱ معیار بهینه سازی

هدف: مولفه اصلی $\omega^i \in R^{(d,1)}$ را به گونه ای بیابیم که پراکندگی[۳] داده ها در فضای کوچکتر d' بیشینه شوند.

$$W^* = argmax_W Var(Z) \quad (۱۲)$$
$$s.t : \omega_i^T \omega^j = \delta_{ij} \quad (۱۳)$$

$$z_1 = \omega_1^T x \in R \quad E[z_1] = \omega_1^T E[x] = 0 \quad (۱۴)$$
$$V(z_1) = \frac{1}{N}\sum_n^N \left(z_1^{(n)}\right)^2 = \frac{1}{N}\sum_n^N \left(\omega_1^T x^{(n)}\right)^2 \quad (۱۵)$$
$$= \frac{1}{N}\sum_n^N (\omega_1^T x^{(n)})(x^{(n)^T}\omega_1) \quad (۱۶)$$
$$= \omega_1^T \left(\frac{1}{N}\sum_n^N x^{(n)} x^{(n)^T}\right)\omega_1 \quad (۱۷)$$
$$= \omega_1^T S \omega_1 \quad (۱۸)$$

که در واقع S در معادله ۱۸ معادله ماتریس پراکندگی (ماتریس کوواریانس[۴]) داده های X است. اکنون تابع معیار را می توان به صورت معادله ۱۹ بازنویسی کرد.

$$\begin{aligned} \omega_1^* &= argmax_\omega\ \omega_1^T S \omega_1 \\ s.t &: \omega_1^T \omega_1 = 1 \end{aligned} \quad (۱۹)$$

نکته:

$$S = \frac{1}{N} X^T X$$

۴.۱ محاسبه W بهینه

از طریق بهینه سازی لاگرانژ معادله ۱۹ را حل می کنیم.

$$\mathcal{L}(\omega_1, \lambda) = \omega_1^T S \omega_1 + \lambda(1 - \omega_1^T \omega_1) \quad (۲۰)$$
$$\frac{\partial L}{\partial \omega_1} = 0 \implies 2S\omega_1 - 2\lambda\omega_1 = 0 \quad (۲۱)$$
$$\implies S\omega_1 = \lambda\omega_1 \implies X^T X \omega_1 = N\lambda\omega_1 \quad (۲۲)$$
$$V(z_1) = \omega_1^T S \omega_1 = \omega_1^T \lambda \omega_1 = ||\omega_1||^2 \lambda = \lambda \quad (۲۳)$$
$$\implies \lambda = V(z_1) \quad (۲۴)$$

در فرآیند حل معادله بهینه سازی ۱۹ به معادله ۲۲ رسیدیم که معادله مقدار ویژه است. هدف ما بیشینه کردن $\omega_1^T S \omega_1$ که این معادل λ است (معادلات ۲۳ و ۲۴)، بنابراین مقدار بهینه ω_1 برابر با بردار ویژه متناظر با مقدار ویژه بیشینه ماتریس پراکندگی X است.

بنابراین برای یافتن $W = (\omega_1, \omega_2, \cdots, \omega_{d'})$ کافی است بردار ویژه متناظر با d' بزرگترین مقادیر ویژه ماتریس پراکندگی داده را در نظر بگیریم.

توجه: ما در نظر گرفتیم که X دارای میانگین صفر است.

[۳] Variance
[۴] Covariance

۵.۱ پیاده سازی بهینه

۱.۵.۱ استفاده از تجزیه مقدار منفرد[۵]

هدف ما پیدا کردن d' بردار ویژه متناظر با بزرگترین مقدار ویژه های ماتریس X^TX است. محاسبه مقدار ویژه ها برای X^TX زمان بر است، پس ما از تجزیه مقدار منفرد که زمان کمتری می گیرد استفاده می کنیم.

$$X = U\Sigma V^T \tag{۲۵}$$

$$X^TX = (V\Sigma^T U^T)(U\Sigma V^T) = V\Lambda V^T \tag{۲۶}$$

بنابراین طبق معادله ۲۶ میتوانیم برای یافتن بردار ویژه X^TX از تجزیه مقدار منفرد استفاده کرد.
بنابراین برای یافتن $W = (\omega_1, \omega_2, \cdots, \omega_{d'})$ کافی است d' بردار V متناظر با d' بزرگترین مقادیر منفرد X را در نظر بگیریم.

۲.۵.۱ استفاده از خاصیت هرمیتی

ماتریس پراکندگی یک ماتریس هرمیتی است.

$$(X^TX)^T = X^TX$$

بنابراین:

$$X^TX = V\Lambda V^T = V\Lambda V^{-1}$$

بنابراین می توان میتوان معادله $V\Lambda V^T$ را حل کرد که سریعتر از $V\Lambda V^{-1}$ است. برای حل $V\Lambda V^T$ باید از تابع $numpy.linalg.eigh$ استفاده کرد.

---

[۵] Singular Value Decomposition (SVD)

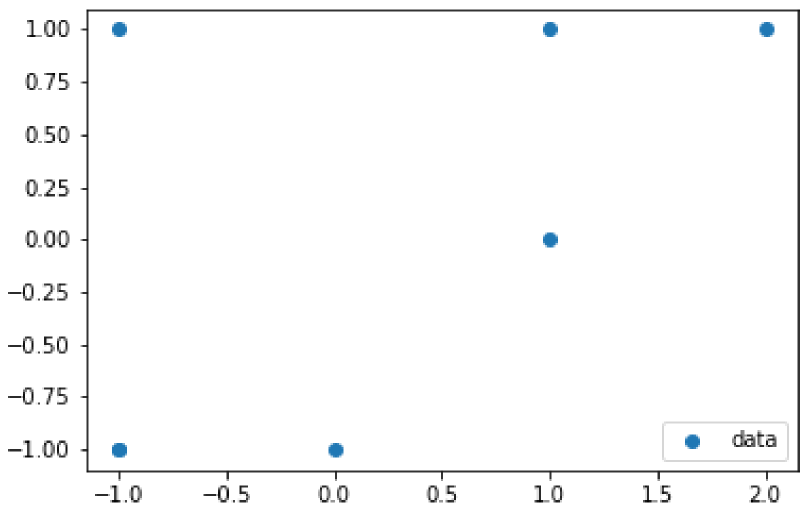
۲ سوال دوم

۱.۲ داده ها

$$Data = \left\{ \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\} \quad (۲۷)$$

$$X = \begin{pmatrix} 0 & -1 \\ 1 & 0 \\ 2 & 1 \\ 1 & 1 \\ -1 & 1 \\ -1 & -1 \\ -1 & -1 \end{pmatrix} \quad (۲۸)$$

شکل ۱: داده ها در فضای ۲ بعدی



۲.۲ حل مسئله

در ابتدا میانگین هر **بعد** را بدست می آوریم و X را 0 میانگین می کنیم:

$$\mu = \frac{1}{7}\sum_{i=1}^{7} x^{(i)} = \frac{1}{7}\begin{pmatrix}1\\0\end{pmatrix} \quad (۲۹)$$

$$\hat{X} = X - \mu = \frac{1}{7}\begin{pmatrix}-1 & -7\\ 6 & 0\\ 13 & 7\\ 6 & 7\\ -8 & 7\\ -8 & -7\\ -8 & -7\end{pmatrix} \quad (۳۰)$$

اکنون ماتریس کوواریانس $\hat{X}$ را محاسبه می کنیم:

$$Cov = \frac{1}{7}\hat{X}^T\hat{X} = \frac{1}{49}\begin{pmatrix}435 & 196\\ 196 & 294\end{pmatrix} = \frac{1}{7}\begin{pmatrix}62 & 28\\ 28 & 42\end{pmatrix} \quad (۳۱)$$

اکنون بردار ویژه و مقادیر ویژه ماتریس کوواریانس ۳۱ را بدست می آوریم. چون مابه دنبال راستای بردار های ویژه هستیم میتوانیم که مسئله بردار ویژه و مقدار ویژه را برای مضربی از ماتریس کوواریانس حل کنیم:

$$S = \begin{pmatrix}31 & 14\\ 14 & 21\end{pmatrix} = \frac{7}{2}Cov \quad (۳۲)$$

اکنون مسئله مقدار ویژه و بردار ویژه را برای S حل می کنیم:

$$\begin{Vmatrix}31-\lambda & 14\\ 14 & 21-\lambda\end{Vmatrix} = \lambda^2 - 52\lambda + 445 = 0 \quad (۳۳)$$

$$\implies \lambda_1 = 26 + \sqrt{221} \quad \lambda_2 = 26 - \sqrt{221} \quad (۳۴)$$

اکنون بردار های ویژه متناظر را بدست می آوریم:

$$(S - \lambda I_2)v = 0 \quad (۳۵)$$

توجه: راستای بردار های ویژه برای ما اهمیت دارند بنابراین نرم بردار های ویژه رو به عنوان خروجی نهایی لحاظ می کنیم. (در واقع طبق تعریف لازم است نرم ۱ باشد)

برای λ_1:

$$\left(\begin{pmatrix}31 & 14\\ 14 & 21\end{pmatrix} - (26+\sqrt{221})I_2\right)\begin{pmatrix}v_1\\v_2\end{pmatrix} = 0 \quad (۳۶)$$

$$\left(\begin{pmatrix}5-\sqrt{221} & 14\\ 14 & -5-\sqrt{221}\end{pmatrix}\right)\begin{pmatrix}v_1\\v_2\end{pmatrix} = 0 \quad (۳۷)$$

$$\begin{pmatrix}v_1\\v_2\end{pmatrix} = \begin{pmatrix}\frac{14y}{-5+sqrt(221)}\\ y\end{pmatrix} \quad (۳۸)$$

$$(y = 1, normalize, numeric\,eval) \implies \begin{pmatrix}0.81741556\\0.57604844\end{pmatrix} \quad (۳۹)$$

برای λ_2:

$$\left(\begin{pmatrix}31 & 14\\ 14 & 21\end{pmatrix} - (26-\sqrt{221})I_2\right)\begin{pmatrix}v_1\\ v_2\end{pmatrix} = 0 \tag{۴۰}$$

$$\left(\begin{pmatrix}5+\sqrt{221} & 14\\ 14 & -5+\sqrt{221}\end{pmatrix}\right)\begin{pmatrix}v_1\\ v_2\end{pmatrix} = 0 \tag{۴۱}$$

$$\begin{pmatrix}v_1\\ v_2\end{pmatrix} = \begin{pmatrix}\frac{-14y}{-5+sqrt(221)}\\ y\end{pmatrix} \tag{۴۲}$$

$$(y=1, normalize, numeric\, eval) \implies \begin{pmatrix}-0.81741556\\ 0.57604844\end{pmatrix} \tag{۴۳}$$

بنابراین:

$$\left(\lambda_1 = 26+\sqrt{221}, V_1 = \begin{pmatrix}0.81741556\\ 0.57604844\end{pmatrix}\right) \tag{۴۴}$$

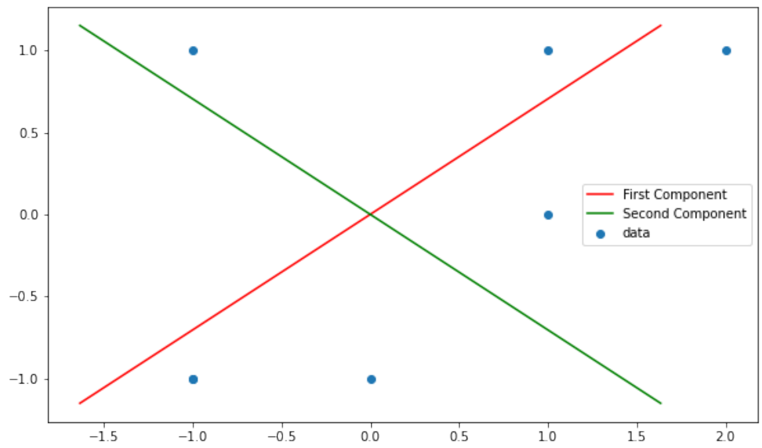
$$\left(\lambda_2 = 26-\sqrt{221}, V_2 = \begin{pmatrix}-0.81741556\\ 0.57604844\end{pmatrix}\right) \tag{۴۵}$$

که V_1 و V_2 مولفه های اصلی اند.

۳.۲ مولفه های اصلی بر روی داده

در شکل ۲ مولفه ها و داده ها رسم شده اند.

شکل ۲: مولفه های اصلی بر روی داده ها در فضای ۲ بعدی



۴.۲ نمایش داده ها در فضای ۱ بعدی

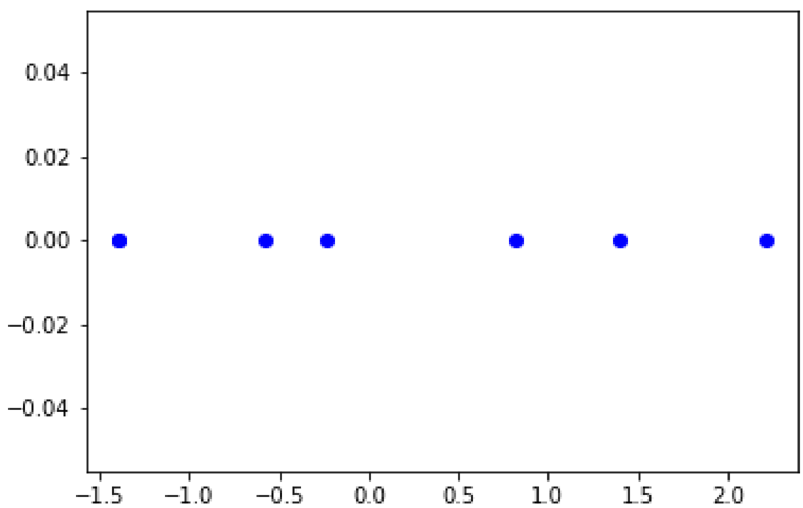
ما مولفه اصلی اول (متناظر با مقدار ویژه بزرگتر) را در نظر می گیرم:

$$W = \begin{pmatrix} 0.81741556 \\ 0.57604844 \end{pmatrix} \tag{۴۶}$$

$$Z = \hat{X}W = \begin{pmatrix} -0.57604844 \\ 0.81741556 \\ 2.21087956 \\ 1.393464 \\ -0.24136712 \\ -1.393464 \\ -1.393464 \end{pmatrix} \tag{۴۷}$$

Z(۴۷) داده ها در فضای ۱ بعدی اند. در شکل ۳ نمایش داده ها بر روی مولفه اول رسم شده است.

شکل ۳: نمایش داده ها در بر روی مولفه اول



۵.۲ بازسازی داده ها

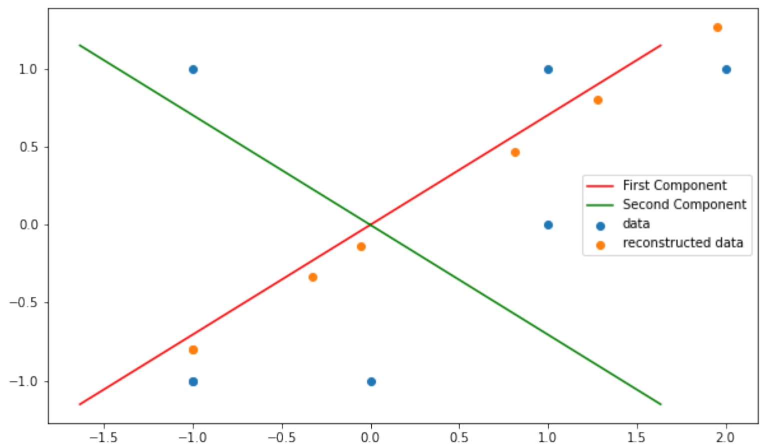
$$X_{reconstructed} = ZW^T + \mu = \tag{۴۸}$$

$$\begin{pmatrix} -0.57604844 \\ 0.81741556 \\ 2.21087956 \\ 1.393464 \\ -0.24136712 \\ -1.393464 \\ -1.393464 \end{pmatrix} \begin{pmatrix} 0.81741556 & 0.57604844 \end{pmatrix} + \frac{1}{7}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \tag{۴۹}$$

$$\begin{pmatrix} -0.32801381 & -0.3318318 \\ 0.81102534 & 0.47087096 \\ 1.9500645 & 1.27357371 \\ 1.2818963 & 0.80270276 \\ -0.0544401 & -0.13903915 \\ -0.99618201 & -0.80270276 \\ -0.99618201 & -0.80270276 \end{pmatrix} = X_{reconstructed} \tag{۵۰}$$

در شکل ۴ داده ها بازسازی شده به رنگ نارنجی مشخص شده اند. بازسازی داده ها به طور دقیق منطبق بر داده های اصلی نیست، زیرا درصدی اطلاعات از دست رفته است. خطای $RMSE$[۶] بین داده اصلی و داده بازسازی شده برابر است با: 0.3787

شکل ۴: بازسازی داده از روی مولفه اصلی اول به فضای ۲ بعدی



نکته

اگر علاوه بر اینکه میانگین داده ها را صفر کردیم، انحراف از معیار را ۱ کنیم (داده را استاندارد کنیم) نتیجه بهتر خواهد شد.

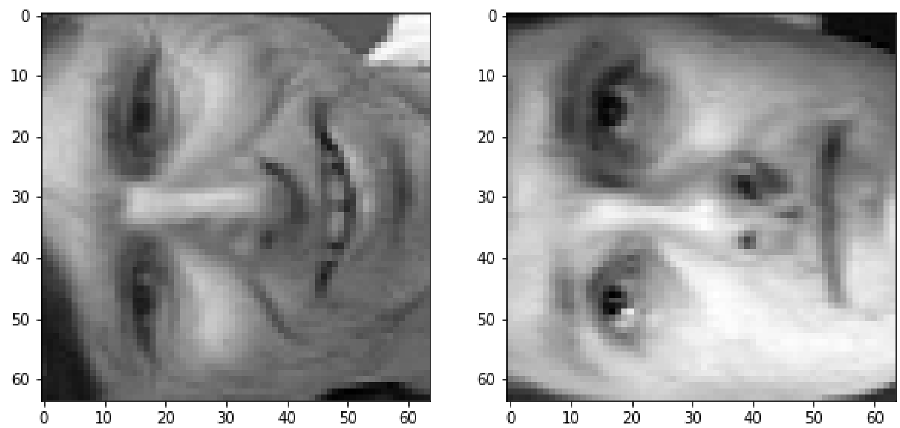
[۶] Root Mean Squared Error

۳ سوال سوم

۱.۳ داده ها

داده ها متشکل از ۴۰۰ عدد عکس سیاه و سفید چهره با وضوح (۶۴،۶۴) است. ما داده ها را به ۲ بخش آموزشی و تست تقسیم می کنیم به طوریکه داده های تست 0.3 کل داده ها است. در شکل ۵ ۲ نمونه تصویر را مشاهده می کنید.

شکل ۵: نمونه تصاویر



۲.۳ مولفه های اصلی

پس از آموزش بر روی داده آموزشی PCA تعداد مقادیر منفرد ناصفر (تعداد مقادیر ویژه ناصفر) ۲۸۰ است، بنابراین سایر مولفه ها به قدری ناچیز اند که محاسبات به دلیل underflow در محاسبات عددی برابر صفر شده اند. می توانیم داده ها به بعد ۲۸۰ نظیر کنیم به طوریکه **تقریبا** اطلاعاتی از دست ندهیم.

خطای RMSE بر روی زمانی که داده ها را به بعد ۲۸۰ منتقل می کنیم:

$$\text{RMSE on train} = 6.7286 \times 10^{-14} \quad \text{RMSE on test} = 10.3370$$

۳.۳ بررسی بر روی داده ها

برای بررسی ابعاد زیر را در نظر گرفتم: (درصد پراکندگی مربوط به هر مجموعه مولفه اصلی مشخص شده است)

$$\begin{pmatrix} \text{dims} \\ \text{vars} \end{pmatrix} = \begin{pmatrix} 280 & 192 & 138 & 98 & 68 & 45 & 28 & 16 & 8 & 3 \\ 1 & 0.9 & 0.8 & 0.7 & 0.6 & 0.5 & 0.4 & 0.3 & 0.2 & 0.1 \end{pmatrix}$$

در جدول ۱ شکل ۶ خطای بازسازی با ابعاد ذکر شده، نمایش داده شده است. با توجه به شکل ۶ متوجه می شویم که مدل overfit کرده است.

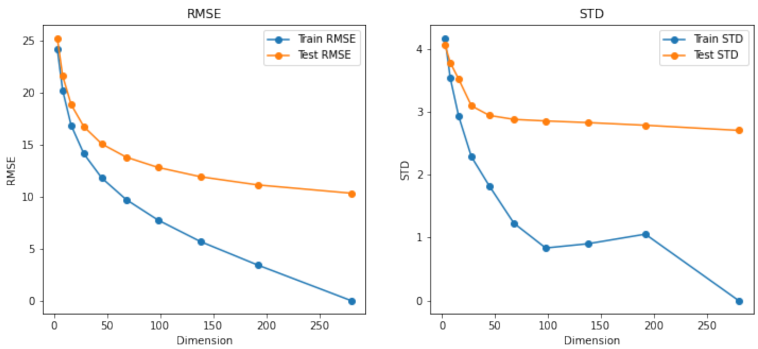
در شکل ۷ و ۸ ما دو تصویر را از داده آموزشی و داده تست انتخاب کرده ایم و برای ابعاد ذکر شده رسم کرده ایم. در شکل های ۱۰ و ۹ ما ۱۰ داده آموزشی و تست انتخاب کرده ایم و بازسازی آن ها را برای ابعاد ۳۰۰۰ و ۲۸۰ رسم کرده ایم. باتوجه به اشکال واضح است که مدل بر روی داده آموزشی overfit کرده است، زیرا با اینکه داده آموزشی حتی با بعد ۲۸۰ تقریبا اطلاعاتی را از دست نداده است، ولی داده های تست حتی با استفاده بعد ۳۰۰۰ چندان مشابه داده اصلی نیستند.

تصاویر ضمیمه شده اند.

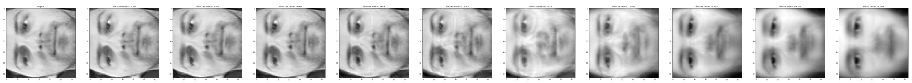
جدول ۱: خطای RMSE مرتبط با ابعاد

Dim	280	192	138	98	68	45	28	16	8	3
RMSE on train	0.000	3.4213	5.6816	7.7505	9.7212	11.7917	14.139	16.8854	20.2609	24.2394
RMSE on test	10.337	11.133	11.9219	12.8165	13.8075	15.075	16.7452	18.8807	21.6531	25.2744

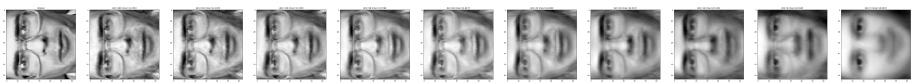
شکل ۶: خطای RMSE و مقدار انحراف از معیار (std) برای داده های آموزشی و تست



شکل ۷: داده آموزشی



شکل ۸: داده تست



شکل ۹: بررسی بر روی داده های آموزش: ستون سمت چپ داده های اصلی، ستون وسط داده های بازسازی شده از بعد ۳۰۰۰، ستون سمت راست داده های بازسازی شده از بعد ۲۸۰

شکل ۱۰: بررسی بر روی داده های تست: ستون سمت چپ داده های اصلی، ستون وسط داده های بازسازی شده از بعد ۳۰۰۰، ستون سمت راست داده های بازسازی شده از بعد ۲۸۰