תרגיל בית מספר 4

מבוא ללמידה חישובית

בנימין אסתרליס ותומר סמרה

315636761 321164923
tomersamara@mail.tau.ac.il esterlis@mail.tau.ac.il

הכתובת של קבצי הקוד + התמונות בצבע:specific/a/home/cc/students/math/esterlis/Downloads/ML-HW3

1 חלק תיאורטי

1. (א) נרצה להראות זאת על ידי דוגמה בה מתקיים: $\forall i\, error(h_i) = 2\epsilon$ אך $\mathbb{P}\left(h(x) \neq y\right) = error(h) = 3\epsilon$, $\epsilon \in (0, \frac{1}{3})$.
נגדיר את xבאופן אחיד בקטע $(0,1)$
לשם כך נגדיר מחלקת השערות הכוללת את 3 ההשערות הבאות:

$$h_i(x) = \begin{cases} 1 & x \in \left(\frac{i-1}{3}\epsilon, \frac{i}{3}\epsilon\right) \\ 0 & otherwise \end{cases}$$

ולכן מתקיים כי:

$$h(x) = majority(h_1(x), h_2(x), h_3(x)) = 0$$

ולכן אם נקח את cפונקצית המטרה להיות $c = \begin{cases} 1 & x \in (0, \epsilon) \\ 0 & otherwise \end{cases}$, נקבל כי $error(h) = 3\epsilon$, אך $\forall i\, error(h_i) = 2\epsilon$.

(ב) נקח $\mathcal{H}$מחלקת השערות וניקח $h_i, \dots, h_{2k+1} \in \mathcal{H}$, נניח שמתקיים $error(h_i) \leq \epsilon$ נרצה להראות שמתקיים $error(h) \leq 2\epsilon$.
נגדיר:

$$\chi_i = \begin{cases} 1 & h_i(x) \neq c(x) \\ 0 & otherwise \end{cases}$$

ניתן לראות כי הינו המציין של השגיאה ב h_i, ולכן $\mathbb{E}\left[\chi_i\right] = error(h_i) = \epsilon$.
על מנת שתהיה שגיאה בh צריך כי תהיה שגיאה בלפחות $k+1$ השערות ולכן צריך כי:

$$\mathbb{P}\left(\sum_{i=1}^{2k+1} \chi_i \geq k+1\right) \overset{Markov}{\leq} \frac{\mathbb{E}\left[\sum_{i=1}^{2k+1} \chi_i\right]}{k+1} = \frac{\sum_{i=1}^{2k+1} \mathbb{E}\left[\chi_i\right]}{k+1} =$$

$$= \frac{\sum_{i=1}^{2k+1} \epsilon}{k+1} = \frac{(2k+1)\epsilon}{k+1} \leq 2\epsilon$$

2. (א) ניווכחנו בעובדה זו בכיתה.

(ב) נרצה להראות כי השגיאה ביחס להעשרה הנוכחית בשלב הבא היא בדיוק חצי.
לשם כך נבחין

$$\begin{aligned}
\underset{x \sim \mathcal{D}_{t+1}}{\mathbb{P}}\left(h_t(x) \neq y\right) =& \sum_{i=1}^{m} \mathcal{D}_{t+1}(x_i) \cdot \chi_{(h(x_i) \neq y_I)} = \\
=& \sum_{i=1}^{m} \frac{\mathcal{D}_t\left(x_i\right) \cdot e^{-\alpha_t \cdot y_i \cdot h_t(x_i)}}{Z_t} \cdot \chi_{(h(x_i) \neq y_i)} \overset{sgn(y_i) \neq sgn((h(x_i))}{=} \\
=& \sum_{\{i : h_t(x_i) \neq y_i\}} \frac{\mathcal{D}_t(x_i)}{Z_t} \cdot e^{\alpha_t} \overset{(a)}{=} \sum_{\{i : h_t(x_i) \neq y_i\}} \frac{\mathcal{D}_t(x_i)}{2\sqrt{\epsilon_t \cdot (1-\epsilon_t)}} \cdot \frac{\sqrt{\epsilon_t \cdot (1-\epsilon_t)}}{\epsilon_t} = \\
=& \sum_{\{i : h_t(x_i) \neq y_i\}} \frac{\mathcal{D}_t(x_i)}{2 \cdot \epsilon_t} = \frac{1}{2\epsilon_t} \cdot \underset{x \sim \mathcal{D}_t}{\mathbb{P}}\left(h_t(x) \neq y\right) \frac{1}{2\epsilon_t} \cdot \epsilon_t = \frac{1}{2}
\end{aligned}$$

(ג) על מנת להראות זאת נניח בשלילה כי $h_t(x)=h_{t+1}(x)$,
לפי הגדרה נקבל כי:$h_t(x)=arg\min_w \mathbb{P}\left(h_t(x)\neq y\right)$, אך לפי סעיף (ב) ניתן לראות כי $\mathbb{P}\left(h_t(x)\neq y\right)=\frac{1}{2}$
ולכן $min\mathbb{P}\left(h(x)\neq y\right)=\frac{1}{2}$, משמע מצאנו הפלגות D_{t+1} אשר עבורה $\forall h\in\mathcal{H}\, error(h)\geq\frac{1}{2}$, ולכן קבילנו כי $\mathcal{H}$ הנ"ל אינה $weak-learnable$ וזה סתירה.
ולכן $h_t(x)\neq h_{t+1}(x)$

3. (א) נרצה לחשב את $\bar{K'}$לשם כך נחשב $\bar{K'_{i,j}}=\langle y_i,y_j\rangle$ $\forall i,j\in[m]$ באופן הבא:

$$\begin{aligned}
\langle y_i,y_j\rangle &= \left\langle \phi(x_i)-\frac{1}{m}\sum_{k=1}^{m}\phi(x_k),\phi(x_j)-\frac{1}{m}\sum_{k=1}^{m}\phi(x_j)\right\rangle = \\
&= \langle\phi(x_i),\phi(x_j)\rangle+\left\langle\phi(x_i),-\frac{1}{m}\sum_{k=1}^{m}\phi(x_k)\right\rangle+\left\langle-\frac{1}{m}\sum_{k=1}^{m}\phi(x_k),\phi(x_j)\right\rangle+\left\langle-\frac{1}{m}\sum_{k=1}^{m}\phi(x_k),-\frac{1}{m}\sum_{k=1}^{m}\phi(x_k)\right\rangle = \\
&= \bar{K}_{i,j}-\frac{1}{m}\sum_{k=1}^{m}\langle\phi(x_k),\phi(x_j)\rangle-\frac{1}{m}\sum_{k=1}^{m}\langle\phi(x_i),\phi(x_k)\rangle+\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}\langle\phi(x_k),\phi(x_l)\rangle = \\
&= \bar{K}_{i,j}-\frac{1}{m}\sum_{k=1}^{m}\bar{K}_{k,j}-\frac{1}{m}\sum_{k=1}^{m}\bar{K}_{i,k}+\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}\bar{K}_{k,l}
\end{aligned}$$

ניתן לראות כי סיבוכיות זמן הריצה היא:
משך זמן חישוב פעולת ה$\bar{K}_{i,j}$ הינו $O(1)$ כי הם נתונים לנו, ומספר פעולות החישוב המתבצע הוא: $O(m^2)$ ולכן בסה"כ: $O(m^2)$.

(ב) נרצה להראות כי u_j הינו צירוף ליניארי של $\phi(x_i)$.
בנוסף אנו ראינו בכיתה כי:

$$\Sigma=\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\cdot\phi(x_i)^T$$

בנוסף אנו ידועים כי u_j הינם ו"ע של Σ ולכן:

$$\begin{aligned}
&\Sigma u_j=\lambda_j u_j\Rightarrow\frac{1}{m}\sum_{i=1}^{m}\phi(x_i)\cdot\phi(x_i)^T\cdot u_j=\lambda_j\cdot u_j\Rightarrow \\
&\Rightarrow u_j=\frac{1}{\lambda_j\cdot m}\sum_{i=1}^{m}\phi(x_i)\cdot\phi(x_i)^T\cdot u_j\Rightarrow \\
&\overset{\phi(x_i)^T\cdot u_j\, is\, scalar}{\Rightarrow} u_j=\frac{1}{\lambda_j\cdot m}\sum_{i=1}^{m}\left(\phi(x_i)^T\cdot u_j\right)\cdot\phi(x_i)
\end{aligned}$$

וכאשר נציב $a_i^j=\left(\phi(x_i)^T\cdot u_j\right)\cdot\frac{1}{\lambda_j\cdot m}$

כעת נראה כיצד מחשבים את הו"ע:

$$
\begin{aligned}
&Au_k = \lambda_k u_k \Rightarrow A\sum_{i=1}^{m} \alpha_i^k \phi(x_i) = \lambda_k \sum_{i=1}^{m} \alpha_i^k \phi(x_i) \Rightarrow \\
&\quad \Rightarrow \left(\frac{1}{m} \sum_{j=1}^{m} \phi(x_j)\phi(x_j)^T \right) \cdot \left(\sum_{i=1}^{m} \alpha_i^k \phi(x_i) \right) = \lambda_k \sum_{i=1}^{m} \alpha_i^k \phi(x_i) \Rightarrow \\
&\quad \Rightarrow \frac{1}{m} \sum_{j=1}^{m} \phi(x_j)^T \sum_{i=1} \alpha_i^k \cdot K(x_j, x_i) = \lambda_k \sum_{i=1}^{m} \alpha_i^k \phi(x_i) \Rightarrow \\
&\quad \overset{mutliple\, by\, \phi(x_k)^T\, from\, left}{\Rightarrow} \frac{\phi(x_k)^T}{m} \sum_{j=1}^{m} \phi(x_j)^T \sum_{i=1}^{m} \alpha_i^k \cdot K(x_j, x_i) = \phi(x_k)^T \cdot \lambda_k \sum_{i=1}^{m} \alpha_i^k \phi(x_i) \Rightarrow \\
&\quad \Rightarrow \frac{1}{m} \sum_{j=1}^{m} K(\phi(x_j), \phi(x_k)) \sum_{i=1}^{m} \alpha_i^k \cdot \overline{K_{j,i}} = \lambda_k \sum_{i=1}^{m} \alpha_i^k \overline{K_{j,i}} \Rightarrow \\
&\quad \Rightarrow \frac{1}{m} \sum_{j=1}^{m} \overline{K_{j,k}} \alpha^k \cdot \overline{K_{k,k}} = \lambda_k \sum_{i=1}^{m} \alpha_i^k \overline{K_{j,i}} \Rightarrow \frac{1}{m} \overline{K}^2 \alpha^k = \lambda_j \overline{K} \alpha^k
\end{aligned}
$$

ולכן $\bar{K}\alpha^k$ הינו ו"ע של $\bar{K}$.
לכן α^kהינו ו"ע עם ע"ע 0 או שמתקיים: $K\alpha^k = m\lambda_k\alpha^k$.
בנוסף מכוון ש$\|u_j\|^2 = 1$נקבל כי:

$$
\begin{aligned}
&1 = \left\langle \sum \alpha^i \phi(x_i), \sum \alpha^j \phi(x_j) \right\rangle \Rightarrow \\
&\quad \Rightarrow 1 = \sum \alpha^i \alpha^j K_{i,j} \Rightarrow \\
&\quad \Rightarrow {\alpha^j}^T \left(K\alpha^j \right) = 1 \Rightarrow \\
&\quad \Rightarrow {\alpha^j}^T (m \cdot \lambda_j \cdot \alpha^j) = 1 \Rightarrow \\
&\quad \Rightarrow \|\alpha^j\|^2 = 1
\end{aligned}
$$

לכן, כדי לחשב את המקדים α^jשל u_j, נחשב את הו"ע ה-i של K(אחרי נירמול מסעיף (א)) ונחלק את הוקטור המקבל בשורש הע"ע שלו.

$$
K\alpha^j = \theta\alpha^j \Rightarrow \alpha^j_{normalized} = \frac{\alpha^j}{\sqrt{\theta}} \Rightarrow \|\alpha^j_{normalized}\|^2 = \frac{1}{\theta} \Rightarrow \|v\|^2 = 1
$$

(ג) נחשב את $\langle u_j, \phi(x) \rangle$באופן הבא: (את u_j חישבנו בסעיף ב)

$$
\langle u_j, \phi(x) \rangle = \left\langle \sum_{i=1}^{m} \alpha_i^j \phi(x_i), \phi(x) \right\rangle = \sum_{i=1}^{m} \alpha_i^j \left\langle \phi(x_i), \phi(x) \right\rangle = \sum_{i=1}^{m} \alpha_i^j K(x_i, x)
$$

צריך לחשב את כל α_j^iוניתן לחשב את כל ה-α-ות בתור מטריצה בזמן $O(m^2)$ באמצעות קסמי המתמטיקה, ולכן בסה"כ נקבל כי הסיבוכיות עבור $\langle u_j, \phi(x) \rangle$ הינה $O(m^3)$.
ולכן ניתן לחשב את כל המכפלות $\langle u_j, \phi(x) \rangle$ $\forall j \in [m]$הינו ב$O(k \cdot m^3)$.
($\star$)הסיבוכיות הינה $O(k \cdot m^2)$ כאשר Kנתון וידוע מראש, אם לא ניתן לחשוב ב$O(d)$ולכן בסה"כ: $O(k \cdot d \cdot m^3)$.

4. נרצה לפתור את הבעיה הבאה:

$$\underset{w}{argmin}\|w\|$$
$$s.tX^TXw=X^Ty$$

על מנת לבצע זאת ננסה לחלץ את w מהתנאי.
ראשית, נעזר בפירוק SVD האומר כי $X=U\Sigma V^T$ עבור $V_{n\times n},U_{m\times m}$ מטריצות אורתונורמליות, ו$\Sigma_{m\times n}$מטריצה עם ערכים סינגולריים באיברים $(\Sigma)_{i,i}=\sigma_{i,i}$ $\forall i\in[min\{m,n\}]$, ושאר אפסים (#סמיון).
לפי הנתון אנו רואים כי עמודות X אינן ב"ת ולכן קיימים ערכים סינגולריים שערכם הוא 0, ולכן Σ יכולה להיות עם 0 -ים באלכסון.
ולכן:

$$X^TXw=X^Ty\Rightarrow V\Sigma U^TU\Sigma^TV^Tw=UV\Sigma U^Ty\Rightarrow$$
$$\overset{U\,is\,orthonormal\,matrix}{\Rightarrow}V\Sigma^T\Sigma V^Tw=V\Sigma^TU^Ty\Rightarrow$$
$$\overset{multiple\,by\,V^T\,from\,left}{\Rightarrow}V^TV\Sigma^T\Sigma V^Tw=V^TV\Sigma^TU^Ty\Rightarrow$$
$$\overset{V\,is\,orthonormal\,matrix}{\Rightarrow}\Sigma^T\Sigma V^Tw=\Sigma^TU^Ty$$

כעת, ניווכח במספר עובדות:

(א) $\Sigma^T\Sigma$ הינה מטריצה אלכסונית מגדול $n\times n$ אשר האברים באלכסון שלה הינם הערכים הסינגולריים בריבוע,σ_i^2, למעט מספר כלשהו של איברים באלכסון אשר הינם 0 כי ישנה אפשרות שהערכים הסינגרולריים הינם 0 או כי עמודות Xתלויים אחת בשנייה, ראה ציור:

$$\Sigma^T\Sigma=\begin{pmatrix}\sigma_1^2 & 0 & 0 & \dots & \dots & \dots & 0\\ 0 & \sigma_2^2 & 0 & \dots & \dots & \dots & \vdots\\ \vdots & 0 & \ddots & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \sigma_k^2 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & 0 & 0 & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & \dots & 0 & \dots & \dots & 0\end{pmatrix}$$

(ב) בנוסף, ניווכח בעובדה כי הפתרון של המשוואה $Ax=b$ עבור A אלכסונית, הינו:

$$x=\begin{pmatrix}\frac{b_1}{A_{11}}\\ \vdots\\ \frac{b_n}{A_{nn}}\end{pmatrix}=A^\star b$$

כאשר $A^\star$הינה המטריציה ההופכית של A, אך אם יש איברים אפס על האלכסון הם נשארים אפס, ז"א היא תראה מהצורה הבאה:

$$A^\star=\begin{pmatrix}A_{11}^{-1} & 0 & \dots & 0\\ 0 & \ddots & \dots & \vdots\\ \vdots & \ddots & A_{kk}^{-1} & 0\\ 0 & \dots & 0 & 0\end{pmatrix}$$

ולכן בעקבות 2 העובדות שציינתי נקבל כי הפתרון למשוואה $\Sigma^T\Sigma V^Tw=\Sigma^TU^Ty$ הינו:

$$V^Tw=\left(\Sigma^T\Sigma\right)^\star\Sigma^TU^Ty$$

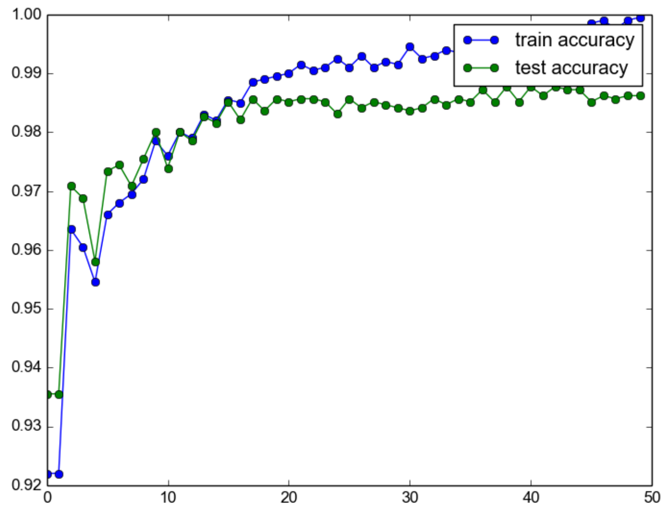
ולכן על מנת לבודד את wנכפיל ב Vמשמאל ונקבל:

$$w = V\left(\Sigma^T\Sigma\right)^{\star}\Sigma^T U^T y$$

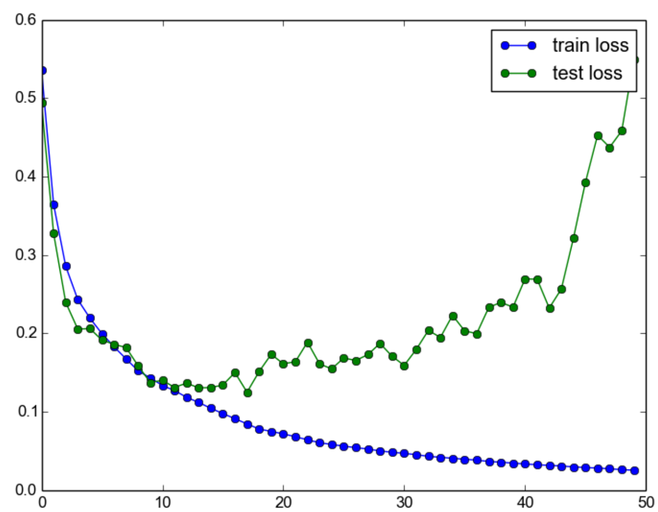
ולכן, צריך לקבל את ה w המינימאלי מהצורה $w = V\left(\Sigma^T\Sigma\right)^{\star}\Sigma^T U^T y$ כאשר יש משתנים חופשיים.
כעת, ניתן לראות כי ב $\left(\Sigma^T\Sigma\right)^{\star}$ יש איברים אשר ערכם הוא אפס, ולכן למען הנוחיות נניח כי הם השורות האחרונות של המטרציה.
ולכן נקבל כי הערכים אשר בשורות האחרונות של w הינם איברים חופשיים, ולכן על מנת להגיע ל $\|w\|$ מינימאלי נציב בהם 0 ונקבל הדרוש.
ולכן wעבורו $\|w\|$מינימאלי הינו $V\left(\Sigma^T\Sigma\right)^{\star}\Sigma^T U^T y$.

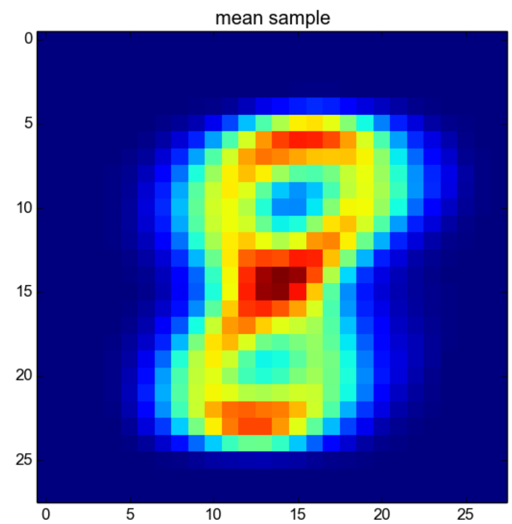
2 חלק מעשי

1. (א) ברור כי השגיאה צריכה לרדת על ה $train\, data$ בכל איטרציה, זה נובע לפי שאלה 2, כי אנו יוצרים את ההשערה לפי ה$train\, data$.
עברו ה $test\, data$ ניתן לראות כי בהתחלה אחוז השגיאה יורד וזה מעיד על נכונות ההאלגוריתם, אך בהמשך ניתן לראות כי אחוז השגיאות גדל כי כנראה אחרי מספר איטריות אנו מתאימים את עצנו יותר מידי לדגימות,
אך כנראה הדגימות אינן תואמות ברמה מדוייקת למיצאות, ולכן קיבלנו כי ההעשרה יוצרת שיגאות לאחר מספר רב של איטרציות.



(ב) ראינו בכיתה כי התהליך $Adaboost$ מביא למינימום את השגיאה האקספוננציאלית, וכן זה מה שקרה לפי הגרף גם על $train\, data$
אך לעומת זאת ב $test\, data$ אחרי שמריצים על מספר מסוים של איטרציות אני מקבלים כי השגיאה הולכת וגדלה.





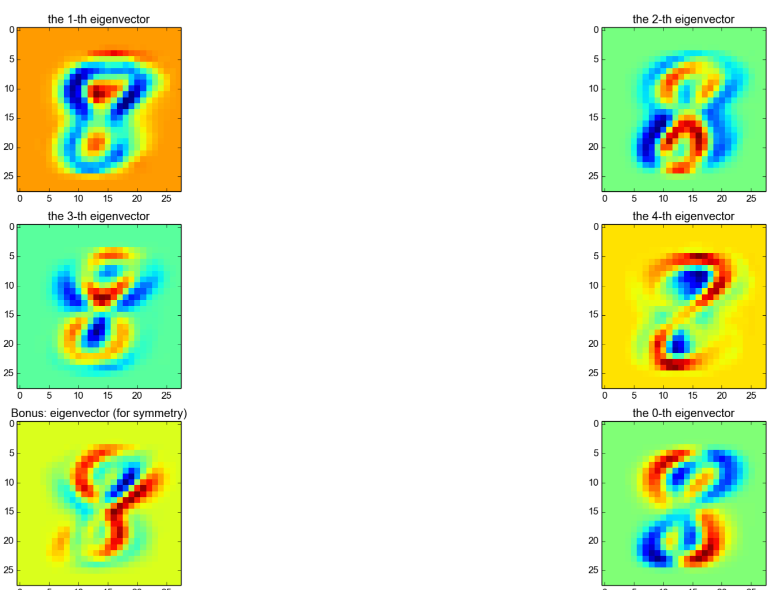
2. (א)

זה הגיוני, נסביר למה:

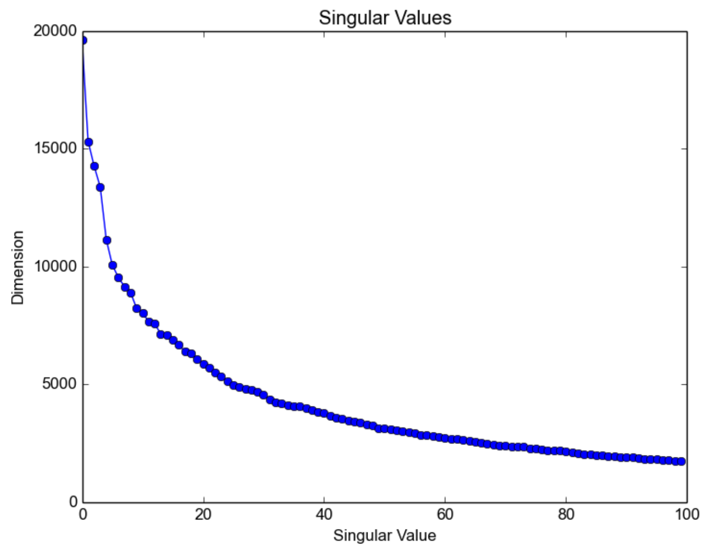
ראשית, זה ניתן לראות כי התמונה נראת כמו המספר 8.

שנית, מספר נכבד מן השמיניות נבדים בחלקים העגולים שלהם ובאוריינטציה, אבל לכולם יש באמצע את החיתוך בין שני העיגולים וזה בא לידי ביטוי בתמונה כאשר באמצע יש אדום בוהק.

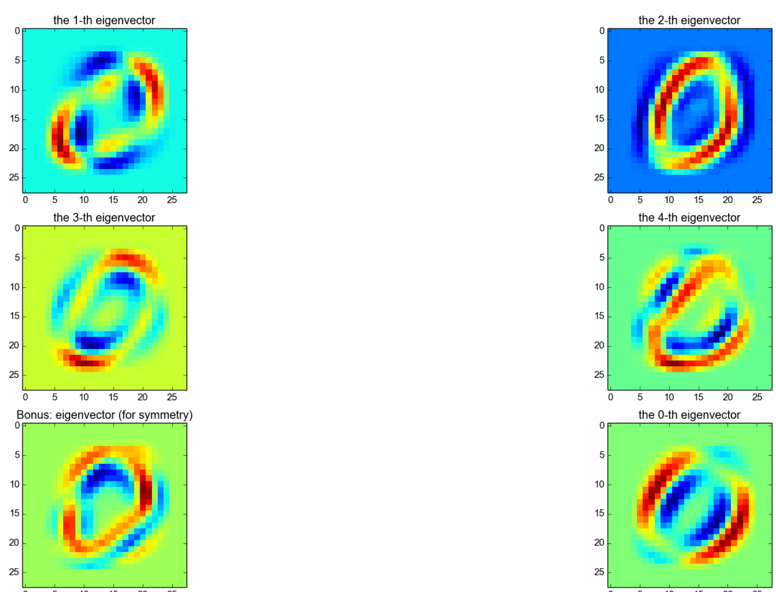
mean:



ע”ע:

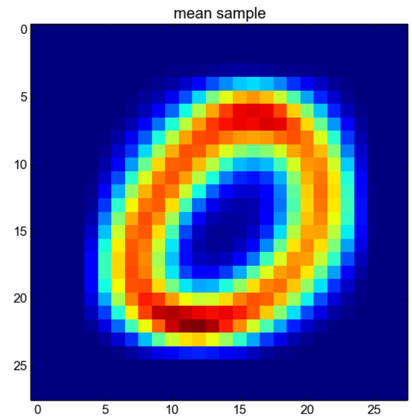


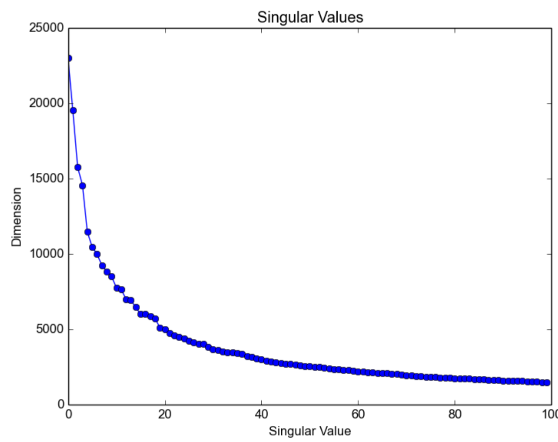
(ב) ע"ע:



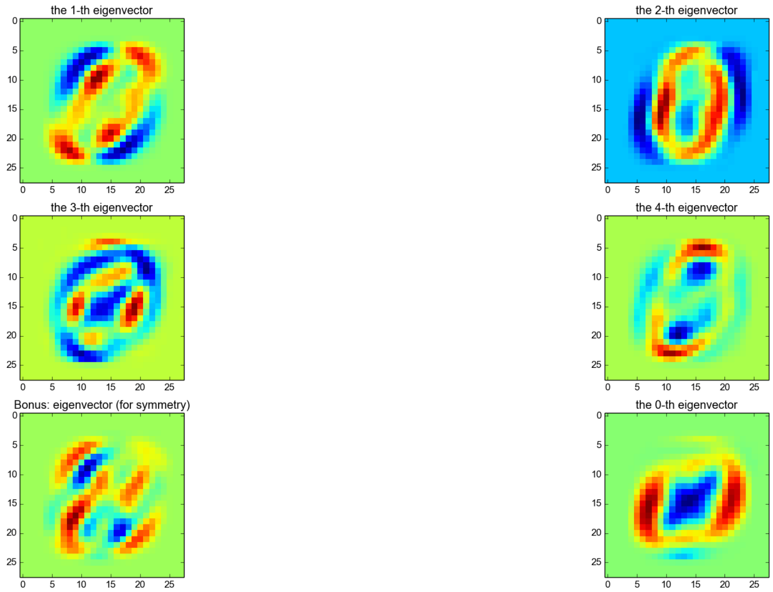
זה הגיוני מאותם נימוקים של הסעיף הקודם.

$mean$:



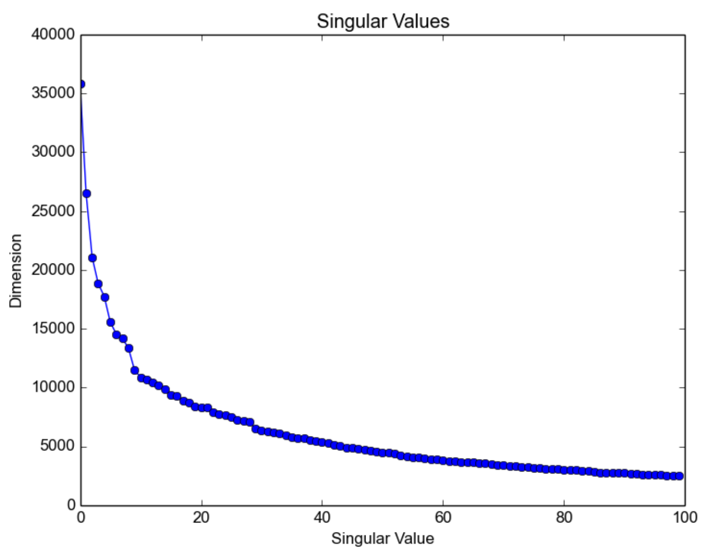


(ג) כעת נציג את התוצאות עבור 8 ו0 ביחד.
ברור כי התוצאות צריכה להראות כקומבינציה של 8 ו0, כי 8 ו0 בעלי אותם חלקים עגולים בראש ובתחתית.
$:mean$

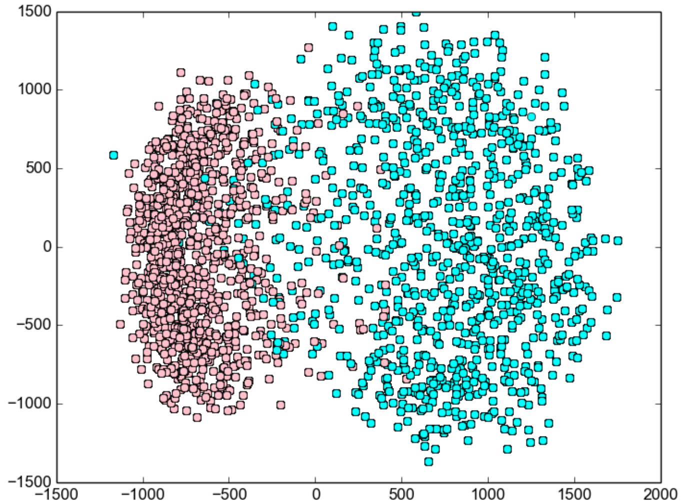


כצפוי. הערכים העצמיים הם קומבינציות שונות של $0's$ ו $8's$, בניגוד לסעיפים א וב, בהם בבירור הע"ע ייצגו $principal\ components$ של אפסים ושמיניות.

ע"ע:

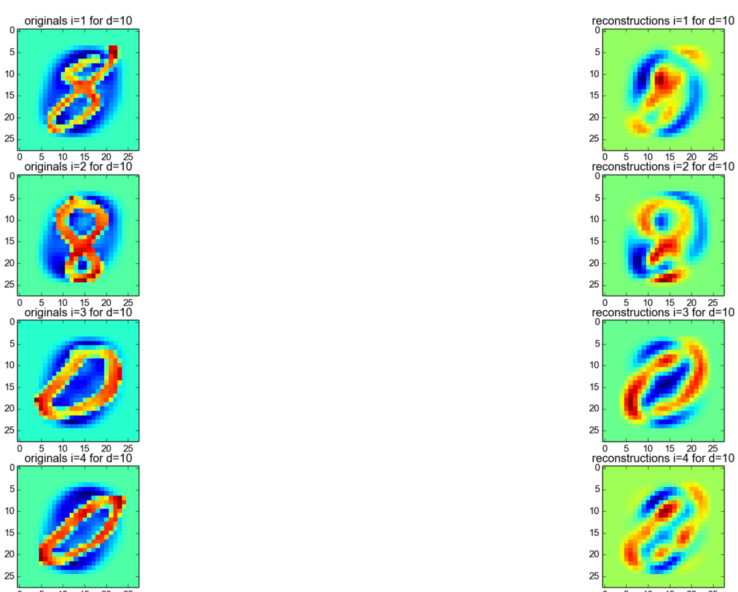


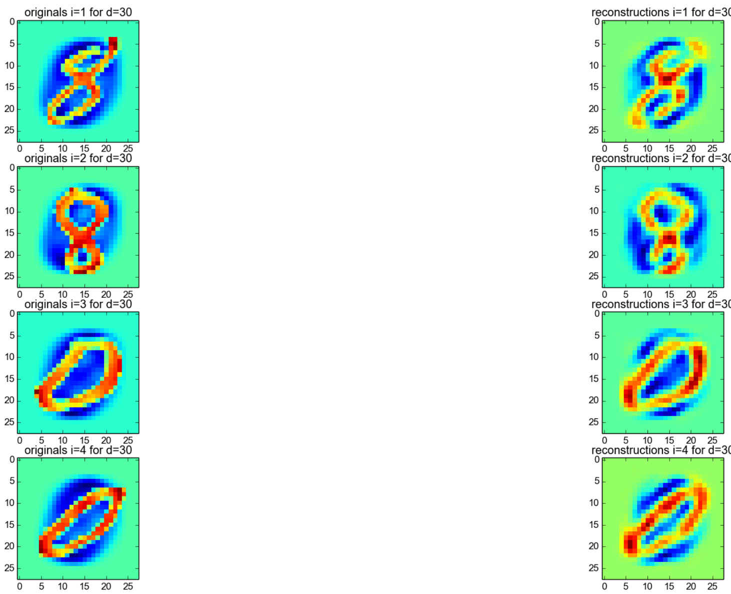
(ד) לקחנו $k = 2$ בתצוגה דו-מימדית בה ביצענו חיזוי של התמונות לתוך שני צירים עיקריים ונקבל:

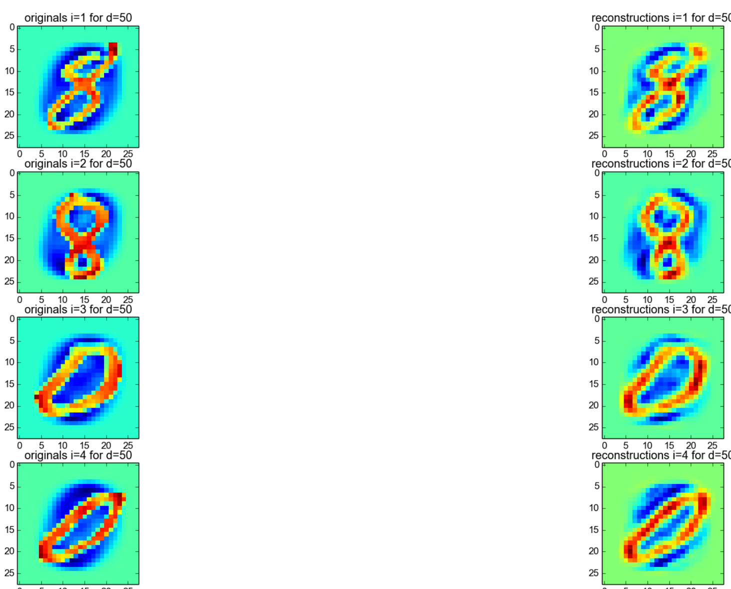


אנו יכולים לראות כי הדגימות שיש לנו מופרדות באופן כמעט מושלם,
מה שקרה הוא שהפקדנו את הדגימות שיש לנו לפי שני צרים עיקריים שמואנכים אחד לשני כך שהכיוונים שלהם נותנים לנו את השונות הגבוהה לפי הדגימות.

(ה) הנה כמה דוגמאות של שיחזור בעזרת PCA עם $k = 10, 30, 50$ עם 2 תמונות מכל מחלקה.
כמו שנבחין כעת, השיחוק מתקרב לתמונה האמיתית ככל שk הולך וגדל.
זה צפוי וברור מאליו, ככל שk הולך וקטן יותר מידע אנחנו עוזבים, ומשארים בצד.




originals i=1 for d=30
originals i=2 for d=30
originals i=3 for d=30
originals i=4 for d=30
reconstructions i=1 for d=30
reconstructions i=2 for d=30
reconstructions i=3 for d=30
reconstructions i=4 for d=30


originals i=1 for d=50
originals i=2 for d=50
originals i=3 for d=50
originals i=4 for d=50
reconstructions i=1 for d=50
reconstructions i=2 for d=50
reconstructions i=3 for d=50
reconstructions i=4 for d=50