מערכות לומדות תשע"ז - תרגיל 3

למידה לא מפוקחת

בתרגיל זה תשתמשו בשיטות של למידה לא מפוקחת לחילוץ מאפיינים ולהפחתת מימדים.

עידכון הספריה

ודאו שמותקנת אצלכם הגרסה היציבה האחרונה של scikit-learn (0.18.1). הנה תקציר ההוראות להתקנה:

א. פיתחו console (cmd.exe) והקלידו
ב. python -m pip install --upgrade pip
ג. pip install -U scikit-learn

נושא 1 – חילוץ מאפיינים אוטומטי

בתרגיל הקודם פיתחתם ומימשתם מאפיינים של ספרות הכתובות בכתב יד, המרתם את הקלט למרחב המאפיינים וביצעתם בו סיווג. כעת תשתמשו במודל למידה לא מפוקחת למציאה אוטמטית של מאפיינים מן הקלט הלא מסווג, כשלב מקדים לסיווג.

תחילה הורידו והריצו את הקוד הבא:
http://scikit-learn.org/stable/auto_examples/neural_networks/plot_rbm_logistic_classification.html#sphx-glr-auto-examples-neural-networks-plot-rbm-logistic-classification-py

בקוד זה, מודל הנקרא Bernoulli Restricted Boltzmann machine. זהו סוג של רשת נוירונים מלאכותית המהווה מודל גנרטיבי הלומד את ההתפלגות של הקלט. הקוד משתמש במודל לביצוע חילוץ מאפיינים אוטומטי. אחר כך הנתונים מומרים למרחב המאפיינים ועליו מתבצע סיווג בעזרת logistic regression. לצורך השוואה מתבצע גם סיווג (בצורה בלתי תלויה) בעזרת logistic regression על הקלט המקורי. בחנו את הקוד והבינו את פעולתו. שימו לב במיוחד לנקודות הבאות:

- הגדלת סט הנתונים על ידי הזזות של פיקסל יחיד לארבעה כיוונים. פעולה זו גם מגדילה את כמות הנתונים המתויגים וגם מאפשרת "אדישות" להזזה (לפחות ברמה בסיסית).
  הבינו כיצד מתבצעת ההזה: השימוש בפונקצית למבדא, השימוש בקונבולוציה עם גרעיני ההזזה, שכפול התיוג הנכון לכל דוגמה מוזזת. השימוש במשתנה "_" ליצירת ערכי ה Y (התיוג) המשוכפלים. ראו גם ההסבר כאן
  https://shahriar.svbtle.com/underscores-in-python

- נורמליזצית הקלט.
- חלוקה לסט אימון ולסט מבחן.
- השימוש ב Pipeline הכולל שני שלבים: הראשון BernoulliRBM, והשני LogisticRegression. קיראו על המחלקה Pipeline בתיעוד.
- האימון של המודל המשולב. האימון של מודל logisticRegression נפרד.
- הצגת ביצועי המסווגים על סט המבחן, והתוצאות עצמן.
- הצגת המרכיבים (_components) שנמצאו על ידי BernoulliRBM.

בשימוש עצמאי ב BernoulliRBM ההמרה למרחב המאפיינים מתבצעת על ידי המתודה transform של BernoulliRBM. אך כאשר המודל הזה נמצא כחלק מ Pipeline יש שימוש במתודה fit_transform, המבצעת את ה fit המחשב את המרכיבים, ואחר כך transform הממיר את הקלט למרחב המאפיינים. התוצאה היא טרנספורמציה לא לינארית (הפונקציה הלוגיסטית) של הקלט בעזרת המרכיבים.
במהלך הלמידה, הרשת מנסה לשחזר את הקלט וכך מהווה מודל הלומד את התפלגות הקלט. המסגרת הכללית בו משחזרים את הקלט נקרא autoencoder ו Restricted Boltzmann machine מהווה אחת השיטות למימוש מסגרת זו.

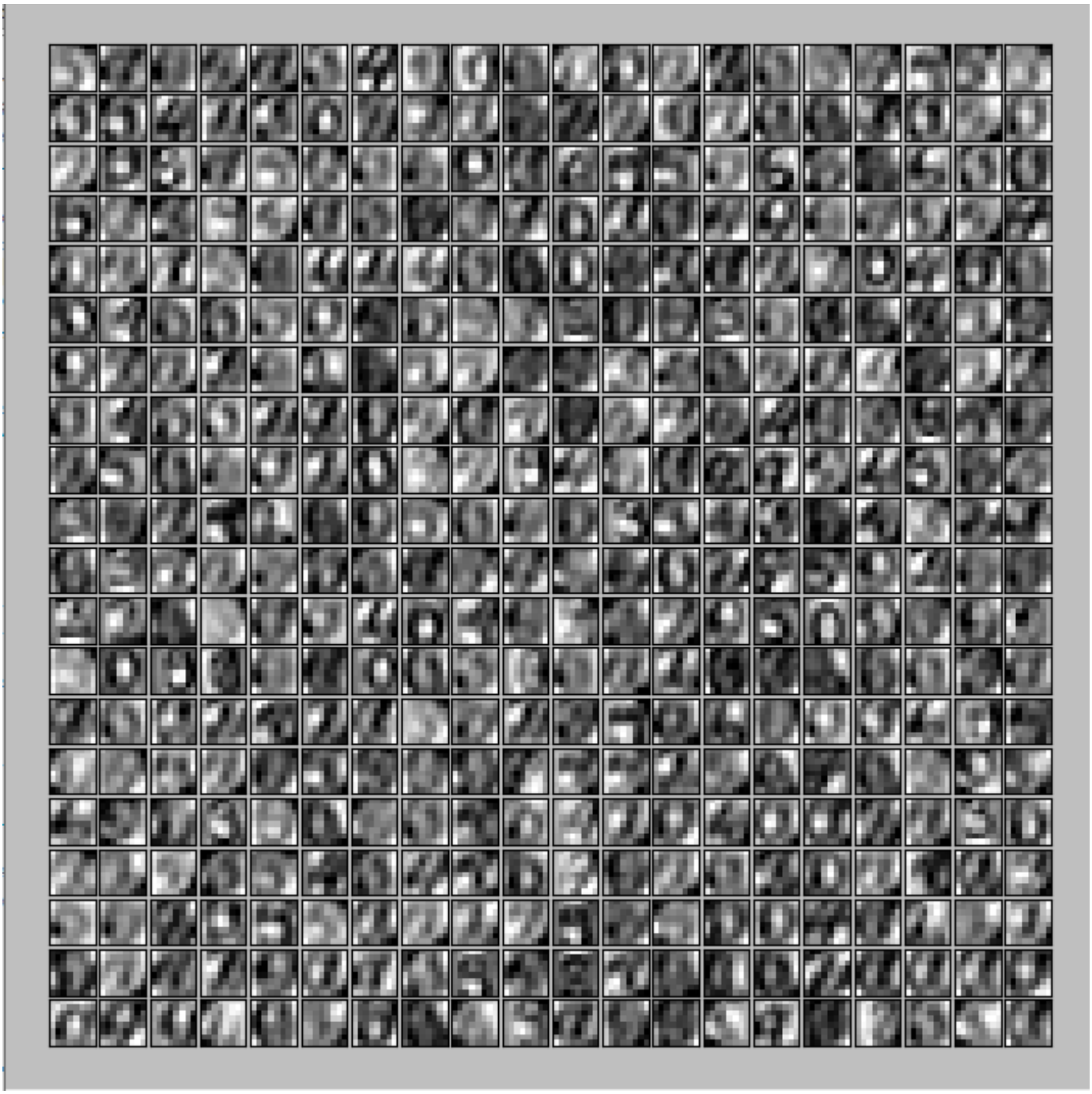
למתעניינים:
קיראו כאן הסבר פשוט על סוג הרשת הזו ושימושיה
https://deeplearning4j.org/restrictedboltzmannmachine

וזה מאמר המשתמש ב Restricted Boltzmann machine כשלב עיבוד מקדים ללימוד דמיון בין קטעי מוזיקה
http://mirg.city.ac.uk/blog/wp-content/uploads/2013/09/rbm-features-for-music-similarity.pdf

(שאלה 1)
פרמטר חשוב של המודל הוא מספר היחידות הנסתרות ברשת, או במונחי הקוד שלנו מספר הרכיבים (_components) הנלמדים. מספר זה הוא המימד של מרחב המאפיינים החדש, ובשאלה זו נחקור את השפעתו על תוצאות הסיווג.
נשים לב שמימד הקלט המקורי הוא (8x8=64), כך שאם נבחר מספר קטן מזה נבצע הפחתת מימדים.
הריצו את הקוד מספר פעמים עבור ערכי _rbm.n_components שהם ריבועי המספרים מ 2 ועד 20 (כלומר 400,... ,4,9,16).

עבור כל הרצה שימרו את ערך ה precision הממוצע, ואת הזמן בשניות שלקח בכל ההרצה **חלק האימון של ה pipeline בלבד**. העזרו ב ()time.clock.

עבור כל הרצה הציגו את _rbm.n_components בצורה גרפית, על ידי שינוי הקוד המקורי כך שיוצגו ... ,2x2, 3x3, 4x4 (בעזרת subplot). הנה התוצאה הרצויה עבור 20x20:



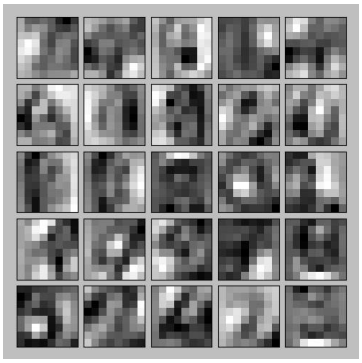
בסיום ההרצות הציגו שני גרפים:

- ה precision הממוצע כנגד מספר הרכיבים. הציגו גם את ערך ה precision הממוצע עבור מסווג הlogistic regression (0.77) בתור קו אופקי, כך שאפשר יהיה להשוות אליו.
- הזמן בשניות לכל הרצה כנגד מספר הרכיבים.

להגשה בחלק זה:

א. הקוד המלא בקובץ ex3_1.py.

ב. הוסיפו לקובץ התשובות (Word) את כל התצוגות הגרפיות של הרכיבים, החל מ 4 רכיבים ועד 400. שנו את גודלן של התמונות (אחרי ההעתקן אל קובץ ה Word) כך שגודל כל תת תמונה של רכיב יהיה בקרוב זהה לזה שבדוגמה למעלה. למשל כך עבור 5X5:



ג. הוסיפו לקובץ ה Word את שני הגרפים. הקפידו על כותרות וטקסט לצירים.

ד. הוסיפו לקובץ ה Word תשובות:

1. כיתבו את המימדים של המשתנים הללו (העזרו ב shape):

X_train, X_test, rbm.transform(X_train), rbm.intercept_hidden_
_mean_hiddens

2. הסתכלו בקוד של rbm.py במתודה transform הקוראת ל המחשבת בפועל את הטרנספורמציה.

כיתבו את החישוב של הטרנספורמציה במונחים של המשתנים שבסעיף הקודם. הסבירו כיצד המימדים של המשתנים מתאימים לחישוב.

נושא 2 – הפחתת מימדים

בחלק זה תכירו ותשתמשו ב PCA (Principal Component Analysis). שיטה זו מורידה את המימד של הקלט על ידי מציאת סט צירים המותאמים לקלט, ובחירה k מתוכם כאשר k קטן (בהרבה) מן המימד המקורי n.

תחילה הורידו והריצו את הקוד הבא:
http://scikit-learn.org/stable/auto_examples/applications/face_recognition.html#sphx-glr-auto-examples-applications-face-recognition-py

הקוד הזה מסווג תמונות פנים של אנשים מפורסמים ל 7 מחלקות (7 מפורסמים). הדוגמה משתמשת ב PCA כשלב של מציאת מאפיינים. הקוד מבצע את השלבים האלה:

- מוצא 150 צירי PCA שהם 150 הצירים המתאימים ביותר לקלט. מימד המרחב המקורי הוא (50x37 = 1850), כך שזו הפחתה משמעותית. כל ציר הוא "תמונת בסיס" או "פרצוף בסיס" ומספר האיברים בו כמימד המרחב המקורי.

- ממיר את הנתונים (סט האימון וסט הבחינה) לצירים אלה. כל תמונה מיוצגת על ידי וקטור קואורדינטות שלה בצירים החדשים (כל תמונה מהווה צירוף לינארי של הצירים, כאשר מקדמי הצירוף הם הקואורדינטות שלה). אורך כל וקטור הוא 150. וקטורי הקואורדינטות מהווים את מרחב המאפיינים להמשך.

- מבצע סיווג על ידי SVM על וקטורי הקואורדינטות של התמונות. ראשית אימון על סט האימון, ולאחריו בחינה על סט הבחינה.

- מציג את ביצועי המסווג על סט המבחן.

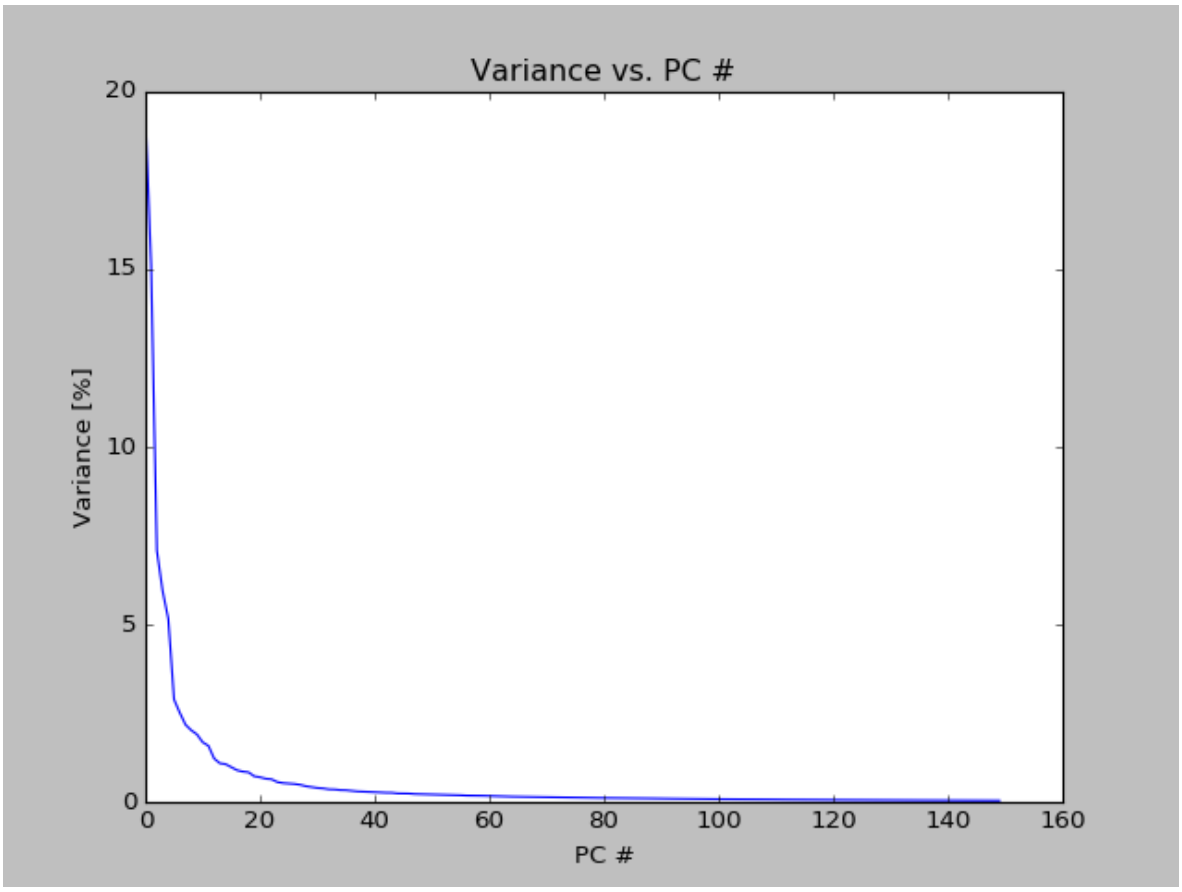
- ולבסוף מציג שני חלונות. בראשון דוגמאות של התמונות וסיווגן, ובשני אוסף של פרצופי בסיס (Eigen faces).

בחנו את הקוד היטב וראו כי אתם מבינים הכל.
שימו לב לשימוש ב GridSearchCV לחיפוש הפרמטרים המתאימים למסווג SVM.

(שאלה 2)

השתמשו בקוד והוסיפו לו \ שנו אותו כך:

א. אחרי השורה המחשבת את ה eigenfaces הוסיפו קטע קוד להצגת השונות המוסברת על ידי וקטורי הבסיס החדשים. השתמשו ב pca.explained_variance_ratio_. הציגו את ערכו כפול 100 בגרף. הוסיפו כותרות וכיתוב לצירים. התוצאה צריכה להראות כך:

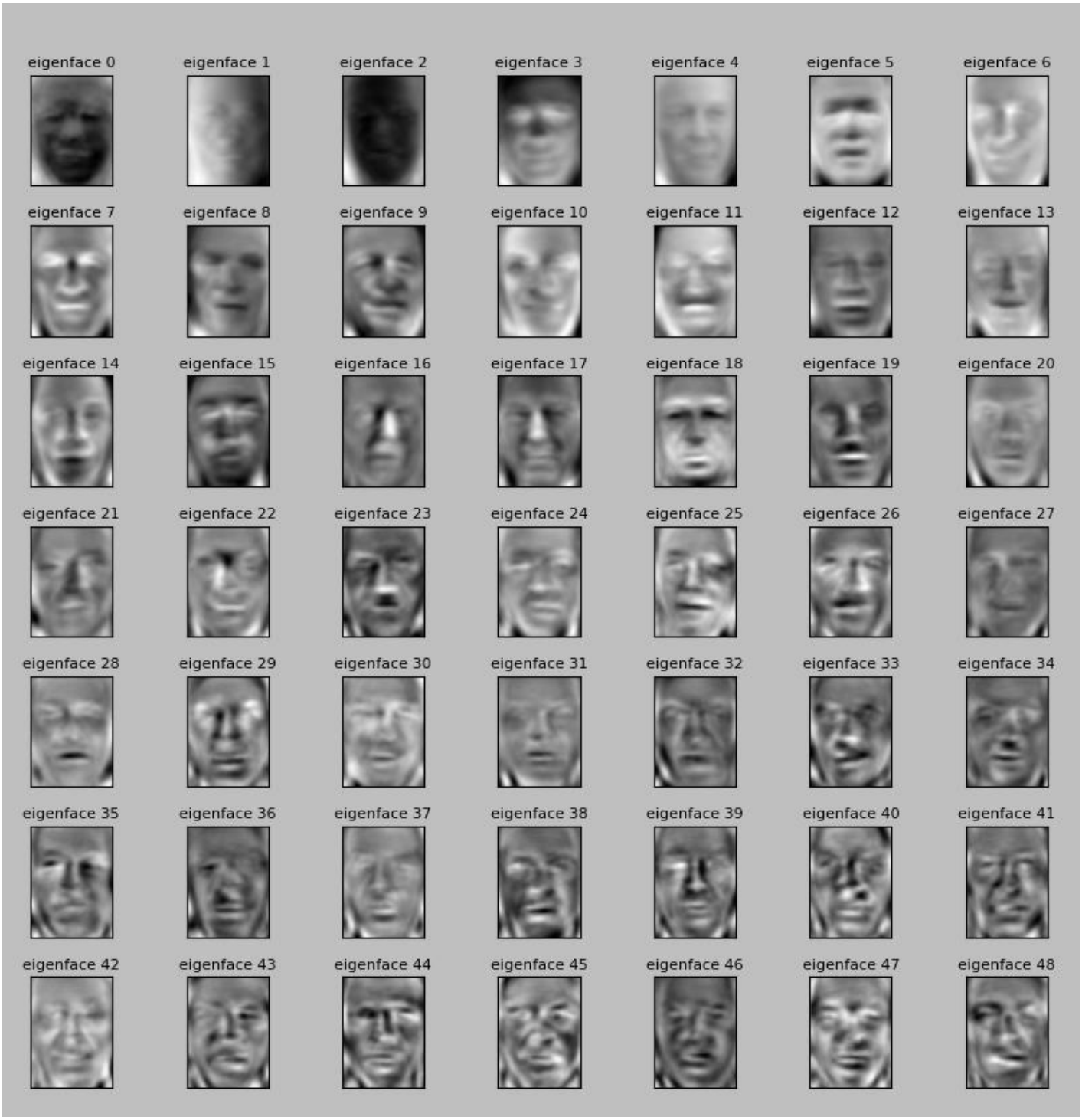


ב. מדדו כמה זמן לוקח האימון בעזרת GridSearchCV. דווחו על איכות הסיווג ועל הזמן בקובץ התשובות. אם תרצו להריץ שוב כדאי לכם להשתמש בערכים שהשיטה מוצאת (עבור המסווג הטוב ביותר) ולהכניס אותם ישירות לקוד שלכם.

ג. שנו את התצוגה של תמונות הפנים ושל פרצופי הבסיס כך שתכיל 7 x 7 תמונות. השתמשו בקוד הזה במקום ההגדרות הקיימות

```
plt.figure(figsize=(1.45 * n_col, 1.5 * n_row))
plt.subplots_adjust(bottom=0.03, left=.01, right=.99, top=.93,
hspace=.36)
```

ד. התוצאות צריכות להראות כך:


eigenface 0
eigenface 1
eigenface 2
eigenface 3
eigenface 4
eigenface 5
eigenface 6
eigenface 7
eigenface 8
eigenface 9
eigenface 10
eigenface 11
eigenface 12
eigenface 13
eigenface 14
eigenface 15
eigenface 16
eigenface 17
eigenface 18
eigenface 19
eigenface 20
eigenface 21
eigenface 22
eigenface 23
eigenface 24
eigenface 25
eigenface 26
eigenface 27
eigenface 28
eigenface 29
eigenface 30
eigenface 31
eigenface 32
eigenface 33
eigenface 34
eigenface 35
eigenface 36
eigenface 37
eigenface 38
eigenface 39
eigenface 40
eigenface 41
eigenface 42
eigenface 43
eigenface 44
eigenface 45
eigenface 46
eigenface 47
eigenface 48

ה. הוסיפו מסווג SVM שיאומן על סט האימון המקורי לפני המרתו למרחב PCA, ויבחן על סט המבחן המקורי לפני המרתו למרחב PCA. בצעו חיפוש בעזרת GridSearchCV אחר המסווג הטוב ביותר לנתונים אלה. מדדו כמה זמן לוקחת הפעלת GridSearchCV. דווחו על איכות הסיווג ועל הזמן בקובץ התשובות.

ו. ענו בקובץ התשובות: האם הפחתת המימדים הועילה לסיווג סט הנתונים הזה? מדוע? האם בשאלה הראשונה הועילה הפחתת מימדים או לא? הסבירו תוך התייחסות למימד הקלט ולכמות הנתונים.

להגשה בחלק זה:

- הקוד המלא בקובץ ex3_2.py.
- התשובות לשאלות 2ב, 2ה, 2ו.

הגשת התרגיל

א. תאריך הגשה: עד יום ראשון, 29.2.17, בשעת חצות הלילה.
ב. ניתן להגיש בזוגות.
ג. יש לכתוב **שם \ שמות + תז** בראשית כל מסמך מוגש (**כולל בקבצי הקוד**).
ד. כל מגיש (ביחיד או בזוג) צריך לדעת להסביר כל מה שנעשה בפתרון המוגש. חלק מן המגישים ידרשו להסביר את הפתרון שלהם למרצה.
ה. יש להגיש מסמך Word המכיל את כל התשובות לתרגיל. שם מסמך זה יהיה **ex3.docx**.
הקפידו שמיספור סעיפי התשובות שלכם יהיה זהה למיספור סעיפי השאלות.
ו. לכל פונקציה צריך להיות תיעוד.
ז. יש להגיש את כל הקוד לתרגיל בשני קבצים לפי ההנחיות למעלה. בתחילת כל קובץ יבואו הגדרות כל הפונקציות. בהמשך הקובץ יבוא חלק ההרצה. חלק זה **יופרד על ידי הערות לכל אחד מסעיפי השאלות.**
ח. שלושת הקבצים ישכנו בתוך תיקייה הכוללת את שמכם.
שם התיקיה למגיש יחיד:
EX3FamilyName
שם התיקיה לשני מגישים:
EX3Family1Family2
התיקיה תארז לקובץ **zip** בעל אותו שם כשל התיקיה (לא rar).