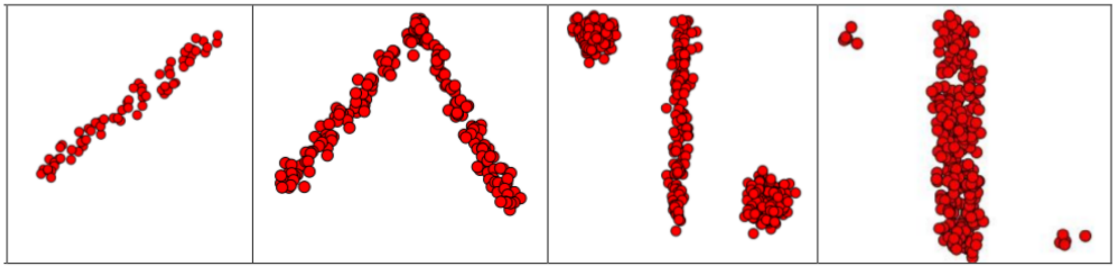
Assignment 07

Machine Learning, Summer term 2018
Norman Hendrich, Marc Bestmann, Philipp Ruppel
May 30, 2018

Solutions due by June 03

Assignment 07.1 (Direction of principal components, 2 points)

Guess and plot the direction of eigenvectors on each image. Draw roughly the data projected on these eigenvectors.



Assignment 07.2 (Expressing the variance, 2 points)

A computer game company runs a survey among visitors of their website. Around 1000 people participate in this survey, and they provide: 1) their age, 2) the time spent playing with the computer, 3) the time spent in facebook, and 4) the time spend doing sport. Then they run a PCA on the data.

a. What does it mean if a single eigenvector covers 90% of the data variance?

b. How would you interpret the results if the eigenvector $v_1 = [0, 1, 1, 1]^T$ covers 85% of the data variance?

Assignment 07.3 (Eigenfaces in sklearn, 1+1+1+2+2 points)

For this exercise, we start with the Eigenface demo in scikit-learn, `http://scikit-learn.org/0.15/_downloads/face_recognition.py`.

a. Download the script, then try to run. This first downloads the example faces dataset. Install any missing dependencies, e.g. *Pillow* (a forked version of the Python Imaging Library). Next, fix the small bugs and deprecation warnings to get the script running with your version of *scikit-learn*.

b. Describe the dataset: total number of images, size of the input images, value ranges of the pixel values, number of classes (aka persons), etc. For each of the classes (persons), randomly select a couple of images and plot them.

c. After splitting the dataset into training and test sets, the demo runs a PCA on the training set. Is the initial choice of $nc = 150$ principal components a good choice? Plot a histogram of the corresponding eigenvalues and estimate the resulting projection error.

d. The demo next uses the *GridSearchCV* module to repeatedly train a SVM with radial-basis functions, in order to find good choices of C and γ for the SVM. It then also runs classification using the selected SVM and prints the results (including the confusion matrix).

Repeat the experiment with smaller and higher numbers of principal components (e.g. $nc = 50$ and $nc = 250$). Which is the smallest number of principal components that still results in acceptable recognition results?

e. Do you think that the Eigenfaces algorithm could be used for automatic passport control (checking a photo of a person against the biometric photo in her passport)? Explain.