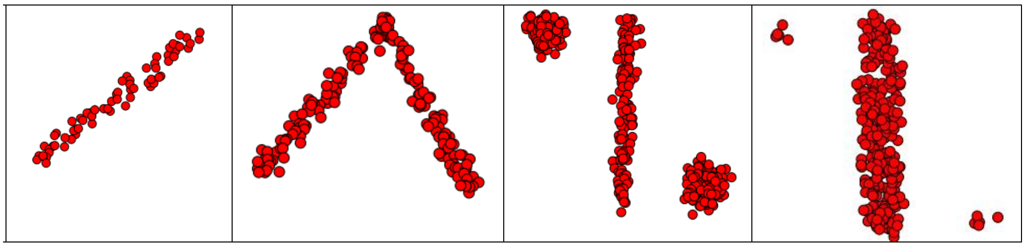
Assignment 6

Machine Learning, Summer term 2015, Ulrike von Luxburg

To be discussed in exercise groups on June 1/3

Exercise 1 (Direction of principal components, 1 point) Guess and plot the direction of eigenvectors on each image. Draw roughly the data projected on these eigenvectors.



Exercise 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 their 1- age, 2- time spent playing with computer, 3- time spent in facebook and 4- time spent doing sport. Then they run a PCA on the data.

- What does it mean if a single eigenvector covers 90% of the data variance?
- How would you interpret the results if the eigenvector $v_1 = [0, 1, 1, 1]^T$ covers 85% of the data variance.

Exercise 3 (PCA, 4 points) Implement PCA in matlab. Do it in a three line matlab code: Subtract the mean of your data, calculate the covariance matrix C, and find its eigenvalues and eigenvectors using the matlab command `[V,D] = eig(C)`.
To test your code, generate 500 samples from a Gaussian distribution with mean $\mu = [1, 1]$ and covariance $\Sigma = [2, -1; -1, 2]$. For generating the points you can use the matlab command `normrnd`. Apply your PCA code on this data and compare the result with the eigenvectors of the covariance matrix Σ.

Exercise 4 (PCA on USPS data, 2 points) Apply the PCA method on images of digits 5 from USPS dataset. The dataset is available in the course website from Assignment 1. Plot the image of the first and the second principal components as 16x16 grayscale images. You can either use your PCA implementation from Exercise 3, or use the matlab command `princomp`.

Exercise 5 (MapReduce with Matlab, 10 points (5+5)) For an introduction to MapReduce framework, please refer to `MapReduce.pdf`.
We want to analyze a part of shakespeare corpus stored in `shakespeareaa.csv`. You can read the file by using

```
ds = datastore('shakespeareaa.csv', 'Delimiter',';','TextscanFormats','%d%s%s%s%s%s');
```

Write the Map and the Reduce functions for the following tasks:

- Find the name of different rules in the play (use the `speaker` column).
- Find the words that appear more than 5 times in the play (use the `text_entry` column). Output the words and their frequency (helpful commands: `strsplit` for finding words and `addmulti` for improving performance).

Web page: `http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/teaching/2013-ss-vorlesung-ml/`
Login: “machine”, Password: “learning”