Exercise 1. ***Dimensionality reduction.***

(a) What are the main motivations for reducing a dataset's dimensionality? What are the main drawbacks?

(b) Describe the Johnson–Lindenstrauss lemma.

(c) What is the curse of dimensionality?

Exercise 2. ***Principal component analysis.***

Suppose we have a dataset with 4 points:

$$\mathcal{D} = \{(1,5),(0,6),(-7,0),(-6,-1)\}$$

(a) Plot the dataset and try to guess two principal components ($k = 2$).

(b) Compute the empirical covariance matrix, its eigenvalues and eigenvectors. Do the eigenvectors correspond to your guess of principal components? Please do not forget the assumptions of PCA. (The dataset should be centered and we want unit eigenvectors.)

Exercise 3. ***Implementing dimensionality-reduction methods.***

(a) Implement a random-projection method in `python`. The random projection $X^{\mathrm{RP}} \in \mathbb{R}^{k\times N}$ of N d-dimensional samples $X \in \mathbb{R}^{d\times N}$ is given by

$$X^{\mathrm{RP}} = RX\,. \tag{1}$$

A computationally efficient way to generate $R \in \mathbb{R}^{k\times d}$ has been proposed by Achlioptas (see lecture). Generate the components of R according to

$$R_{ij} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}. \tag{2}$$

(b) Estimate the reduced dimension k according to Theorem 2 in Achlioptas (2001).[1] Try out different values of β. Discuss and visualize the connection of the estimation of k to the Johnson–Lindenstrauss lemma.

(c) Download the Reuters Corpus Volume I (RCV1) dataset. It is a multilabel dataset in which each data point can belong to multiple classes. The total number of classes is 103. Each data point has a dimensionality of 47,236.

Hint. The Reuters dataset can be directly accessed in ***sklearn****:* *https://scikit-learn.org/0.18/datasets/rcv1.html*.

[1] https://users.math.msu.edu/users/iwenmark/Teaching/MTH995/Papers/JL_Database_Subgaussian.pdf

(d) Select 500 data points from RCV1 that belong to at least one of the first three classes. Use Achlioptas' method to project the selected data points to a lower-dimensional space. Then implement and visualize a matrix that stores the differences between distances of data points in the original space and in the lower-dimensional projection.