Florian Rösel, Frauke Liers Winter semester 2021/2022

Mathematics of Learning – Worksheet 1

Basics. [Solving linear equation systems.][1]
Solve the linear equation system $Ax = b$. A and b are given as

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 4 & 9 \\ 1 & 8 & 27 \end{pmatrix}, \quad b = \begin{pmatrix} 3 \\ 2 \\ 7 \end{pmatrix}.$$

Control yourself, if your solution is right. If you need some practice, generate some random linear equation systems and solve them.

Basics. [Norms.]
A mapping $\|\cdot\|$ from any (real) vector space V to the real numbers $\mathbb{R}$ is called a norm, whenever

$$\|v+w\| \leq \|v\| + \|w\|, \quad \|v\| = 0 \implies v = 0_V, \quad \|\lambda v\| = |\lambda|\|v\| \text{ for all } \lambda \in \mathbb{R}, v, w \in V.$$

Proof for the following statements if they are true or false.

1. Let $V = \mathbb{R}^n$ for some $n \in \mathbb{N}$. The euclidean norm

   $$\|v\|_2 := \sqrt{\sum_{i=1}^{n} v_i^2}$$

   is a norm.

2. Let $V = \mathbb{R}^n$ for some $n \in \mathbb{N}$. The mapping

   $$\|v\|_{\frac{1}{2}} := (\sum_{i=1}^{n} \sqrt{|v_i|})^2$$

   is a norm.

3. Let V be the space of convergent sequences. The mapping

   $$\|v\|_{lim} := \lim_{n\to\infty} v_n$$

   is a norm.

Exercise 1 [Python, Pandas, K-Means].
Install Python 3 on your computer and make sure you are able to import the following packages: `NumPy`, `Matplotlib`, `Pandas`. If you are new to Python you should first watch any Python introduction you find on your favorite video platform - or you look for written tutorials using your favorite search engine.

[1] There are lots of nice tutorial books for linear algebra and analysis available in our library. For a less formal introduction, you can, e.g., also consult wikipedia ;)

a) Download the dataset `faithful.csv` [2] from StudOn and load it into Python using the Pandas package.[3] Explore the dataset and visualize it as a two-dimensional plot using `Matplotlib`. Save the plot to a `png` file.

b) From plotting the data you should see two distinct clusters. Implement the K-means algorithm in Python (by completing the code `K-means_incomplete.py`) and test it (by running `python3 -i K-means.py` in a terminal). Apply K-means to `faithful.csv`.

Exercise 2 [Implementing EM for Clustering].
Implement the EM clustering algorithm for Gaussian mixtures as described on the slides. You can use the code `EM_incomplete.py`. Apply EM to `faithful.csv`.

Bonus [Experiments with K-Means and EM].
Generate own data sets. For example, take a few pictures of different objects (10 apples, 10 classrooms, 10 desks) with your smartphone camera (I propose to choose relatively low resolution), transform them to gray-scale matrices and apply the K-Means/EM Algorithm to the data set. Describe, visualize, and interpret your results.

Exercise 3 [Theory of K-means].
Letting $X \subset \mathbb{R}^M$ denote a finite set of N points, the i-th iteration of the K-means algorithm can be compactly written as ($\|\cdot\|$ is the euclidean norm)

$$\begin{cases} k_n^{(i)} \in \operatorname{argmin}_{k=1}^{K} \|x_n - m_k^{(i)}\|, & \forall n = 1, \dots, N, \\ C_k^{(i)} := \{n \in \{1, \dots, N\} \,:\, k_n^{(i)} = k\}, & \forall k = 1, \dots, K, \\ m_k^{(i)} := \frac{1}{|C_k^{(i)}|} \sum_{x \in C_k^{(i)}} x, & \forall k = 1, \dots, K, \end{cases}$$

where the first line means that *exactly one* element in the argmin is selected.

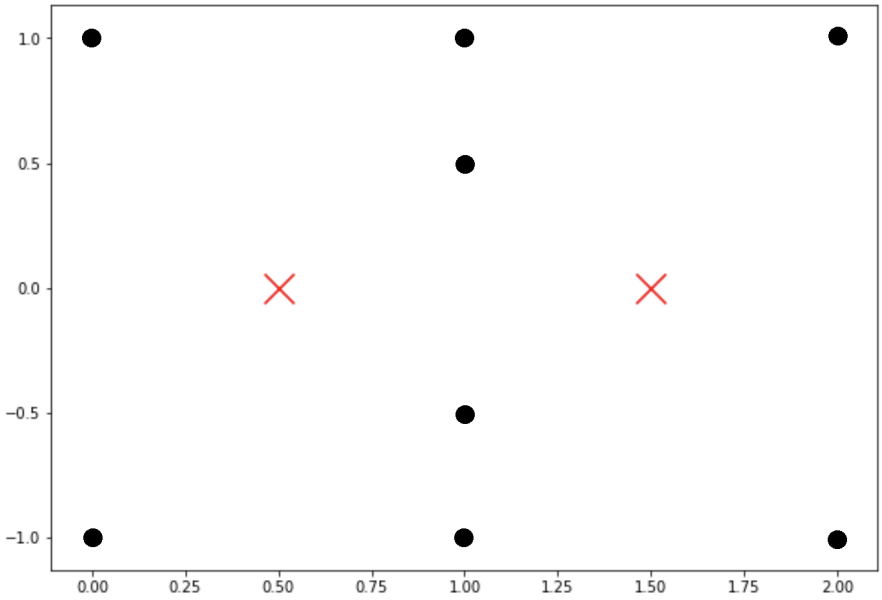
- Show that the iterates of the algorithm satisfy

$$\frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k^{(i)}} \|x - m_k^{(i)}\|^2 \leq \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k^{(i-1)}} \|x - m_k^{(i-1)}\|^2.$$

- Why is it important for this that every data point x_n is assigned to precisely one class?
- Try to extend the result to an arbitrary norm $\|\cdot\|$.
- Construct explicit solutions of K-means in the following situation, where the two red crosses correspond to the initialization $m_k^{(0)}$ of the means. How does this depend on the choice of assignment in the first line of K-means?

[2] See `https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat`

[3] You can learn how to use Pandas here: `https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html`.



Acknowledgement, for generating a previous version of this sheet, to Leon Bungert and Daniel Tenbrinck.