Mathematics of Learning – Worksheet 2

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Basics [Get some literature sources].
Get the two books recommended for reading in this course (see module manual) -

- Goodfellow et al., Deep learning. e.g. `https://www.deeplearningbook.org/` (if you find a better source, let me know)
- Hastie et al., The Elements of Statistical Learining (available as full text pdf in our library)

Read one (you choose which) subsection dealing with "unsupervised learning" (in the Hastie book). Explain it to a fellow student.

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.

a) Download the dataset `faithful.csv` [1] from StudOn and load it into Python using the Pandas package.[2] 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`.

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

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

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 $X \subset \mathbb{R}^D$ 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)

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

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

- Show that the iterates of the algorithm satisfy

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

- 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 crosses correspond to the initialization $m_k^{(0)}$ of the means, and the dots represent the data points. How does this depend on the choice of assignment in the first line of K-means?