Mathematics of Learning – Worksheet 2 - – Discussion on October 26/27th, 2023

- 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

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

  Solution. We give the proof in two steps; first we show that

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

  We see this easily by rearranging the sum a little bit - which works since we have at both sides exactly N terms:

$$\sum_{k=1}^{K} \sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i-1)}\|^2 = \sum_{n=1}^{N} \|x_n - m_{k_n^{(i)}}^{(i-1)}\|^2 \leq$$
$$\leq \sum_{n=1}^{N} \|x_n - m_{k_n^{(i-1)}}^{(i-1)}\|^2 = \sum_{k=1}^{K} \sum_{n \in C_k^{(i-1)}} \|x_n - m_k^{(i-1)}\|^2,$$

  this works because $k_n^{(i)}$ has been set to minimize the terms $\|x_n - m_k^{(i-1)}\|$.
  Next we show that

$$\sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i)}\|^2 \leq \sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i-1)}\|^2 \tag{2}$$

  for all $k \in [K]$. If $C_k^{(i)} = \emptyset$, it obviously holds true.
  Otherwise, we can show inequality (2) with one of the following two possibilities:

1. We consider the function

$$f_k^{(i)}(m) := \sum_{n \in C_k^{(i)}} \|x_n - m\|^2$$

A necessary optimality condition of this function would be the gradient being 0, hence, $\forall d = 1, \ldots, D$,

$$\begin{aligned}\frac{\partial}{\partial m_d} \sum_{n \in C_k^{(i)}} \sum_{j=1}^{D} ((x_n)_j - m_j)^2 &= \sum_{n \in C_k^{(i)}} \sum_{j=1}^{D} \frac{\partial}{\partial m_d} ((x_n)_j - m_j)^2 \\ &= \sum_{n \in C_k^{(i)}} 2 \cdot ((x_n)_d - m_d) \cdot (-1) = 0.\end{aligned}$$

This is equivalent to $|C_k^{(i)}| m = \sum_{n \in C_k^{(i)}} x_n$ and therefore $m_k^{(i)}$ is a critical point.

The Hessian matrix $H_k^{(i)}$ of the function is defined as

$$H_{k,dj}^{(i)} := \frac{\partial}{\partial m_d \partial m_j} f_k^{(i)}(m) = \frac{\partial}{\partial m_j} \sum_{n \in C_k^{(i)}} -2((x_n)_d - m_d) = \begin{cases} 0 \text{ for } d \neq j, \\ 2 \cdot |C_k^{(i)}| \text{ otherwise.} \end{cases},$$

This is a diagonal matrix with positive entries and therefore positive definite. Hence, the critical point $m_k^{(i)}$ is a minimizer of $f_k^{(i)}$ and inequality (2) follows.

2. We have:

$$\begin{aligned}&\sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i-1)}\|^2 = \sum_{n \in C_k^{(i)}} \left(\|(x_n - m_k^{(i)}) + (m_k^{(i)} - m_k^{(i-1)})\|^2 \right) = \\ &= \sum_{n \in C_k^{(i)}} \Big(\|x_n - m_k^{(i)}\|^2 + \|m_k^{(i)} - m_k^{(i-1)}\|^2 + \\ &\qquad + 2 \left(x_n m_k^{(i)} - x_n m_k^{(i-1)} - m_k^{(i)} m_k^{(i)} + m_k^{(i)} m_k^{(i-1)} \right) \Big) \\ &\overset{(*)}{=} \sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i)}\|^2 + |C_k^{(i)}| \|m_k^{(i)} - m_k^{(i-1)}\|^2 + \\ &\qquad + 2 \underbrace{\left(|C_k^{(i)}| m_k^{(i)} m_k^{(i)} - |C_k^{(i)}| m_k^{(i)} m_k^{(i-1)} - |C_k^{(i)}| m_k^{(i)} m_k^{(i)} + |C_k^{(i)}| m_k^{(i)} m_k^{(i-1)} \right)}_{=0} \\ &= \sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i)}\|^2 + \underbrace{|C_k^{(i)}| \|m_k^{(i)} - m_k^{(i-1)}\|^2}_{\geq 0} \\ &\geq \sum_{n \in C_k^{(i)}} \|x_n - m_k^{(i)}\|^2,\end{aligned}$$

where $(*)$ holds because of $m_k^{(i)} = \frac{1}{|C_k^{(i)}|} \sum_{n \in C_k^{(i)}} x_n$. This shows inequality (2).

Taking the two inequalities (1) and (2) together, we get the desired result.

- Why is it important for this that every data point x_n is assigned to precisely one class?
  Solution. Basically the answer here is a discussion. One possible reasoning would be that otherwise you would double-count some distances.

- Try to extend the result to an arbitrary norm $\|\cdot\|$.
  Solution. It does not work. In the proof of the first part we have seen that the arithmetic mean is the minimum if you minimize the function f_k^i when using the squared Euclidean norm. This does not work for other (squared) norms anymore.

  Consider the maximum norm ($||x||_\infty = \max_{i=1,\dots,D} |x_i|$), one cluster ($K = 1$) containing $N = 3$ points in dimension $D = 2$: $x_1 = (0,0), x_2 = (1,0), x_3 = (0,1)$ and initial cluster center $m_0 = (\frac{1}{2}, \frac{1}{2})$. The algorithm just skips step 1 (since we only have one cluster, so re-assigning points to clusters is non-sense), and recalculates the cluster center as the arithmethic mean of the three points which results in $m_1 = (\frac{1}{3}, \frac{1}{3})$. However, we get

$$||x_1 - m_0||_\infty + ||x_2 - m_0||_\infty + ||x_3 - m_0||_\infty = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5$$

  and

$$||x_1 - m_1||_\infty + ||x_2 - m_1||_\infty + ||x_3 - m_1||_\infty = \frac{1}{3} + \frac{2}{3} + \frac{2}{3} = \frac{5}{3} \approx 1.67,$$

  implying that the new center has worse total energy. For the squared maximum norm we get something similar:

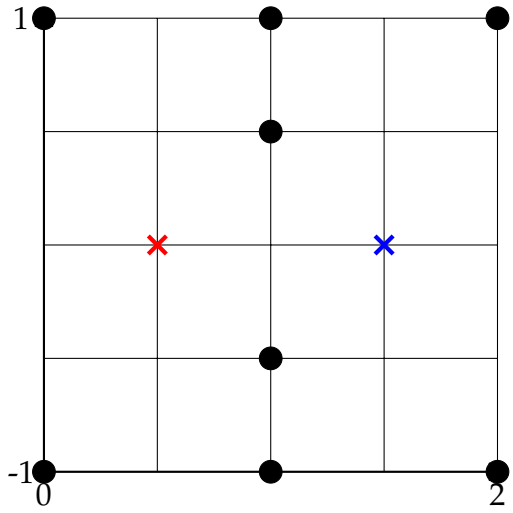
$$||x_1 - m_0||_\infty^2 + ||x_2 - m_0||_\infty^2 + ||x_3 - m_0||_\infty^2 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 0.75$$

  and

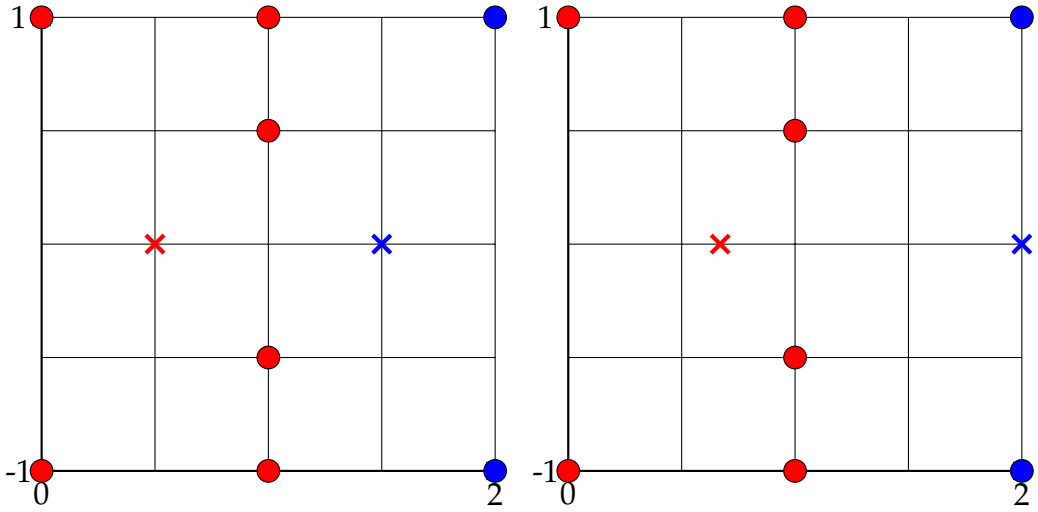
$$||x_1 - m_1||_\infty^2 + ||x_2 - m_1||_\infty^2 + ||x_3 - m_1||_\infty^2 = \frac{1}{9} + \frac{4}{9} + \frac{4}{9} = 1.$$

  Nevertheless, it is possible to adapt the update rule in step 3 that it fits for the maximum norm or for some arbitrary norms you could try (bonus exercise: think about how such an update step could look like and look it up in one of the proposed books for the lecture).

- 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?



Solution. A possible run of the algorithm...



Another possible run of the algorithm...