Machine Learning A

2023-2024

Home Assignment 7

Sadegh Talebi **Christian Igel**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **23 October 2023, 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your full source code in the PDF file, only selected lines if you are asked to do so.
- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF. The programming language of the course is Python.
- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in *SpeedGrader*. Zipped PDF submissions will not be graded.
- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.
- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Handwritten solutions will not be accepted. Please use the provided latex template to write your report.

1 Clustering (25 points)

Consider a dataset consisting of 6 data points A, B, C, D, E, F as shown in Figure 1. The pair next to each point shows the x and y coordinates of each data point. We wish to group the datapoints into k clusters according to the K-means criterion, and using the `k-means++` algorithm (Arthur and Vassilvitskii, 2007).

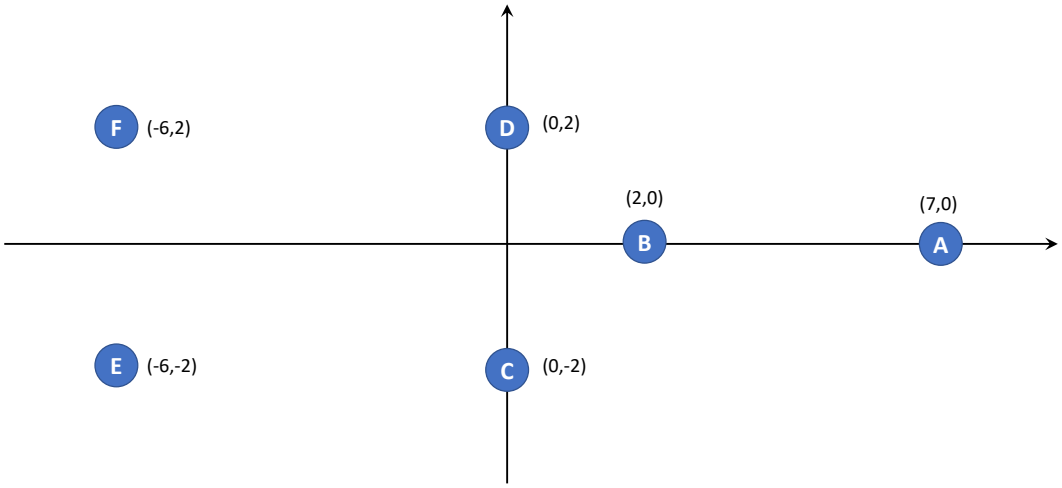

Figure 1: Dataset for Question 1

1. Consider $k = 1$. What is the cluster center? (I.e., what is the *centroid* of the entire dataset?)

Now consider $k = 3$. We would like to study how `k-means++` performs clustering on this dataset.

2. How does the algorithm choose the first initial cluster center c_1? (In other words, determine the probability of each data point being chosen as c_1).
3. Conditioned on A being chosen as c_1, how does the algorithm choose c_2?
4. Conditioned on $c_2 = F$ and $c_1 = A$, how does the algorithm choose c_3?
5. (Optional) Suppose that the algorithm chooses $c_3 = B$ —hence, $c_1 = A, c_2 = F, c_3 = B$. Assume we stop here. Determine the clustering C based on the cluster centers c_1, c_2, c_3 and compute the corresponding cost ϕ.

   Recall that for the K-means problem, the cost ϕ of a dataset $\mathcal{X}$ and a set of cluster centers $\mathcal{C} = \{c_1, \ldots, c_k\}$ is

$$\phi(\mathcal{X}, \mathcal{C}) = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|_2^2$$

6. Suppose that the `k-means++` algorithm chooses $c_3 = B$ —hence, $c_1 = A, c_2 = F, c_3 = B$. Determine the *final* clustering C returned by `k-means++` based on the cluster centers c_1, c_2, c_3. Also argue how you determined that the algorithm has converged.

Deliverables: Report all necessary computations in tasks above

2 Sleep Well Revisited (25 points)

Sleep is one of the most fundamental physiological processes, and abnormal sleeping patterns are associated with poor health. They may, for example, indicate brain- & heart diseases, obesity and diabetes. During sleep our brain goes through a series of changes between different *sleep stages*, which are characterized by specific brain and body activity patterns. *Sleep staging* refers to the process of mapping these transitions over a night of sleep. This is of fundamental importance in sleep medicine, because the sleep patterns combined with other variables provide the basis for diagnosing many sleep related disorders (Kales and Rechtschaffen, 1968, Iber and AASM, 2007). The stages can be determined by measuring the neuronal activity in the cerebral cortex (via electroencephalography, EEG), eye movements (via electrooculography, EOG), and/or the activity of facial muscles (via electromyography, EMG) in a *polysomnography* (PSG) study. The classification into stages is done manually. This is a difficult and time-consuming process, in which expert clinicians inspect and segment the typically 8–24 hours long multi-channel signals. Contiguous, fixed-length intervals of 30 seconds are considered, and each of these *segments* is classified individually. Algorithmic sleep staging aims at automating this process. The state-of-the-art in algorithmic sleep staging is marked by deep neural networks, which can be highly accurate and robust, even compared to human performance, see the recent work by Perslev et al. (2019) and references therein.

This assignment considers algorithmic sleep staging. The data is based on a single EEG channel from the Sleep-EDF-15 data set (Kemp et al., 2000, Goldberger et al., 2000). The input is given by an intermediate representation from the U-Time neural network architecture (Perslev et al., 2019), the targets are sleep stages. We created a training and test set, the inputs and the corresponding labels can be found in `X_train.csv` and `y_train.csv` and `X_test.csv` and `y_test.csv`, respectively. Download and extract the data from `https://github.com/christian-igel/ML/blob/main/data/Sleep-EDF-15_U-Time/`.

2.1 Principal component analysis

Perform a principal component analysis of the training data `X_train.csv`. Plot the eigenspectrum (see the plot on slide 28 of the *PCA* slides for an example). How many components are necessary to "explain 90 % of the variance"? Visualize the data by a scatter plot of the data projected on the first two principal components. Use different colors for the different classes in the plot.

Deliverables: description of software used; plot of the eigenspectrum; number of components necessary to explain 90 % of variance; scatter plot of the data projected on the first two principal components with different colors indicating the 5 different classes

2.2 Clustering using k-means

Perform 5-means clustering of `X_train.csv`. After that, project the cluster centers to the first two principal components of the training data. Then visualize the clusters by adding the cluster centers to the plot from the previous exercise.

Briefly discuss the results.[1]

Deliverables: description of software used; one plot with cluster centers and data points; short discussion of results

2.3 Clustering using k-means++

Repeat the last part using 5-means++ and compare the resulting clusters with the ones obtained with regular 5-means. Are the clusters similar or different, and why? Provide argumentation, considering factors like the initialization of cluster centers, convergence behavior, and the impact on the clustering results.

Deliverables: description of software used; one plot with cluster centers and data points; discussion of results and discussion on comparison

References

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Symposium of Discrete Algorithms (SODA'07)*, pages 1027–1035, 2007.

[1] In this example, you do not get 5 nicely separated clusters in 2D. A better looking projection of the data can be achieved using a non-linear dimensionality reduction technique, for example *non-linear t-Distributed Stochastic Neighbor Embedding* (t-SNE) (**?**). Although not part of the exam, we recommend that you try it out.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

C. Iber and AASM. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*. American Academy of Sleep Medicine, Westchester, I. L., 2007.

A. Kales and A. Rechtschaffen. *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Allan Rechtschaffen and Anthony Kales, editors*. U. S. National Institute of Neurological Diseases and Blindness, Neurological Information Network Bethesda, Md, 1968.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9): 1185–1194, 2000.

M. Perslev, M. Hejselbak Jensen, S. Darkner, P. J. Jennum, and C. Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.