Biomedical Data Science & AI

Exercise sheet 7

SoSe2025

Prof. Dr. Holger Fröhlich, Shammi More, Jiajun Qiu, Anna Weller

Due date: May 27th

Questions

Exercise 1 - NMF Clustering (17 points)

1. Investigate the given gene expression (GE) dataset gedata.csv and carry out NMF consensus clustering to cluster patients/samples based on their GE profile.
    a. Perform NMF consensus clustering for different number of clusters $k = 2, \ldots, 7$ and visualize the consensus matrices. **(3 points)**
    b. Consider the distribution of values in the consensus matrix. Use e.g. seaborn.ecdfplot to plot the empirical cdf of this distribution for different k. How can you use these plots to assess the stability of the clustering? **(2 points)**
    c. Identify the optimal number of clusters and briefly explain your approach. **(1 point)**
    d. Repeat the consensus clustering with k-means clustering and compare the results. **(2 points)**

2. Inform yourself about sparse NMF (sNMF) and Non-Negative Matrix Tri-Factorization (NMTF).
    a. What is the primary difference between NMF, sNMF and NMTF? **(2 points)**
    b. Which of the three NMF would be most suited for the dataset mentioned in question 1? Briefly explain the reasoning behind your suggestion. No programming needed! **(3 points)**

3. PCA and NMF are both matrix factorization methods. Write down the corresponding formulas and compare them. What is similar, what is mathematically different? Describe a situation where NMF is favored over PCA. **(2 points)**

4. Have a look at this paper: https://arxiv.org/abs/1512.07548 and briefly explain why k-means clustering can be understood as a matrix factorization problem as well. **(2 points)**

Exercise 2 - Machine Learning (8 points, no programming needed!)

The type of machine learning (e.g. supervised learning, unsupervised learning, etc.) depends on the problem at hand. Assume that we have an Alzheimer's disease (AD) dataset where rows represent participants and columns represent different collected measurements (such as patient characteristics, MRI measurements, and cognitive tests). We are provided with diagnoses (healthy and AD) of participants.

1) You are asked as a data scientist to predict the diagnosis status of 20 participants based on a model trained with data from 500 participants.
    a) What type of machine learning (ML) would you choose for this task and why? **(1 point)**
    b) What are the steps you should consider before training your ML algorithm? **Hint:** List the potential preprocessing steps you would carry out. **(2 points)**

2) Assume we do not have information about the diagnosis (i.e. no label) of participants. Answer the following questions.
    a) What type of machine learning would you use to group the participants based on the collected measurements? **(1 point)**
    b) Based on your answer in part (2a), suggest a model that you would use to continue this task. Please explain how you can determine the number of groups that separate your participants. **(2 points)**

3) You are asked to investigate the age distribution of healthy versus AD participants, name a visualization plot that can be used in this case, and explain why you think this plot would work best. **(2 points)**