Exercises for the lecture “Data Mining in Bioinformatics”

Organizer: Peter Meinicke
Submission Date: 28.06.2025, 12 p.m.

Sheet 4: Application of PCA to mass spectrometry data in metabolomics

Task 1: Literature and data set

a) The mass spectrometry data to be analyzed are from a metabolic profiling experiment of *Arabidopsis thaliana*. Study the corresponding article “Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps” (e.g. from https://almob.biomedcentral.com/articles/10.1186/1748-7188-3-9) and load the corresponding .csv file (additional file 3) of the measurements.

b) Find out how the experiment is set up and how the data set is structured. What information can be obtained from the data set?

Task 2: Application of principal component analysis

For all subtasks, it is a good idea to write reusable functions so that you can use them also in subsequent tasks.

a) Perform a sample-based principal component analysis (PCA) of the metabolite-specific intensities. Consider columns 6 to 77 as 837-dimensional data vectors and store these last 72 columns transposed as rows in the actual data matrix `x_data_mat` for further analysis. First, show the corresponding eigenvalue spectrum (function: `plt.bar`).

b) Make a scatterplot (`plt.plot`) of the first two principal components, representing the conditions in different colors/symbols including a legend. Can the samples (measurements) from different conditions be distinguished well in this plot?

c) Visualize the loadings (eigenvector coordinates) of the first two principal components as a scatterplot. Can this loadings-plot be used to identify specific groups (clusters) of marker candidates or “outliers”?

Task 3: Normalization

a) Perform a normalization of the intensity values before PCA. This can be done by normalizing the original columns (samples) or the rows (metabolite candidates) according to Euclidean unit norm. Test the two different variants! What effect does the respective normalization have on the results of the PCA, compare these according to task 2 parts a) to c).

b) Visualize the intensity profiles of the corresponding metabolite candidates (function `plt.imshow`) for some promising clusters that you select in the loadings plot. For this purpose it is necessary to select the corresponding candidates over the range of values of the loadings (interval selection, Boolean indexing). For the visualization of the intensity profiles it is useful to summarize the 9 repetitions per condition by the mean (function `mean`) or the median (function `median`) and to normalize the resulting 8-dimensional vectors according to Euclidean unit norm. Interpret the result and compare it with the results from the original publication.

Submission instructions

Please submit your solution by **28th June, 12 p.m.** using the corresponding homework folder in StudIP. Upload a **zip** compressed file containing all documents and data that you use in your solution. The file name should indicate the author, for example *Peter_Meinicke_Sheet4.zip*.
With your solution you should provide the Python source code together with the documentation within a **Jupyter Notebook** answering all the questions above. Use Markdown cells for your answers and explanations well separated from the Python code.
For the programming in Python/numpy do not use any code or functions from special bioinformatics libraries or toolboxes! Make sure that your code is readable and understandable for others, i.e. use descriptive names for the variable definitions according to the conventions on **Sheet 0**, do not use implicit variables (hard-coded values without a variable) and make use of comments!
For reasons explained in the lecture, we do not recommend the use of **A.I. tools** (such as ChatGPT) but, in principle, it is not prohibited. However, if parts of the solutions or code were generated with A.I. tools, it is **mandatory to specify** to what extent (percentage) and for what purpose it was used. Not indicating the actual use of A.I. tools can lead to disqualification from the exercises. If you are unsure about the specification(s) please ask us and do not make any assumptions!