Exercises for “Data Mining in Bioinformatics”

Organizer: Peter Meinicke
Submission date: 17.05.2025, 12 p.m.

Sheet 1: Math- and Python(`numpy`)-Basics

Task 1: Eigenvalues and eigenvectors

For an eigenvalue λ_i of the matrix $\mathbf{A}$ and the associated eigenvector $\mathbf{u}_i$ holds:

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

a) Determine the eigenvalues λ_i of the matrix

$$\mathbf{A} = \begin{pmatrix} 7 & 4 \\ 4 & 13 \end{pmatrix}$$

and the corresponding eigenvectors $\mathbf{u}_i$ first analytically (in written form) then numerically in Python with `numpy`.

b) Represent the vector $\mathbf{x} = [3, 1]^T$ as linear combination of the eigenvectors (in writing). In general, what are the eigenvalues and eigenvectors of a diagonal matrix? Consider as an example

$$\mathbf{D} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$$

Task 2: Random data

Generate an artificial 2D data set, first with 100 points according to a uniform distribution using `x_data_mat = np.random.rand(n_points, n_dims)`, where a data point corresponds to a row vector in the data matrix. Visualize the points using the function `plot` (or `scatter`). Import the functions with
`from matplotlib.pyplot import plot (or scatter)`.
Move all points so that the mean vector of the resulting points is is the zero vector (hint: see function `np.mean`). Rotate these points using a rotation matrix (variable name: `r_rot_mat`)

$$\mathbf{R} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

and visualize the rotated points. Based on the representation of the data points by row vectors, how to implement the rotation? Estimate the *variance* of the resulting x_1 component, i.e., the first column dimension of the rotated data matrix. For which α (variable name: `alpha_angle`) does this variance become maximum? Test this with an angle resolution of one degree (angle measure!) and plot the variance as a function of the angle with `plot`.
Repeat the experiment for another 10 random data sets and then increase the number of data points step by step to $n = 1000, 10000, 100000$. What do you

observe about the results when you sample 10 random data sets for each n? What would you expect in terms of the variance maximizing angle, what do the corresponding function plots show? ☹

What would be *theoretically* the angles for which the variance is maximized? Calculate the solution analytically in writing the expected value of the squared x_1-component, which results from the rotation around angle α (see definition of variance!). Hint: the probability density function $p(\mathbf{x})$ of the underlying uniform distribution of the data generation process has to be used.

Task 3: Covariance

What exactly does the function `c_cov_mat = np.cov(x_data_mat, bias=True, rowvar=False)` do? Show the corresponding formula and explain the additional arguments. Determine for the data sets of the previous task and the corresponding matrix $\mathbf{C}$ (`c_cov_mat`) the vector $\mathbf{v} = [\cos\alpha, \sin\alpha]^T$ (variable name: `dir_vec`) with angle α so that

$$\mathbf{v}^T\mathbf{C}\mathbf{v}$$

becomes maximum. Explain (in writing) why this task is equivalent to the previous one.

Submission instructions

Please submit your solution by **17th May, 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_Sheet1.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!