Tasks

The following tasks are to assess your data science, statistics and programming skills. All the code should be written in Python (or R if needed). Please provide the solution to the tasks the source code for exercises.

Please note, that since is quite lengthy task, you are not required to solve every task. But the solution of at least one of 1st and 2nd task and at least one of 3rd and 4th task is required.

1. Theory I. Classical statistics

Data generating process of *y* is:

$$y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, \dots, n,$$

where ϵ_i are i.i.d. $N(0, \sigma^2)$, ϵ_i and x_i are uncorrelated.
Please find the data in file data_classical_statistics.csv.

a) Write the data generating process in matrix notation.
b) Write the formula for ordinary least squares estimator of α and β. Calculate the estimates using this formula.
c) Fit the model using some ready made function in python (or R). Check if you get the same estimates of α and β. What is the estimate of σ?
d) Test the statistical hypothesis that $\beta = 2$ against the alternative that $\beta \neq 2$ with the significance level of 5%.

2. Theory II. Linear algebra

1. Given points $p_1 = (x_1, y_1) = (2, 3)$ and $p_2 = (x_2, y_2) = (3, 0)$, find line connecting two points $y = w_0 + w_1 x$ and write linear regression in the matrix form of $A\vec{w} = \vec{y}$, with coefficient matrix A, parameter vector $\vec{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$ and dependent variable vector $\vec{y}$. Find w_0 and w_1.
2. What is the unit vector in the same direction as $(3, 2, 2, 2, 2)$?
3. What is the projection of the vector $(3, 5, -9)$ onto the direction $(0.6, -0.8, 0)$?
4. A three-dimensional dataset has a covariance matrix $\Sigma = \begin{pmatrix} 4 & 2 & -3 \\ 2 & 9 & 0 \\ -3 & 0 & 9 \end{pmatrix}$, what is the variance in the direction of $(1, 1, 0)$?
5. For a particular four-dimensional data set, the top two eigenvectors of the covariance matrix are

$$\frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \frac{1}{2}\begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \end{pmatrix}.$$

What is the PCA projection of point $(2, 4, 2, 6)$ into two dimensions? Write it in the form (a, b).

3. Chemometrics

Chemometrics is the science of extracting information from chemical systems by data-driven means. Chemometrics is inherently interdisciplinary, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, biochemistry, medicine, biology and chemical engineering.

In Brolis Sensor Technology, ultra-compact laser-based integrated sensor technology molecule-sensing is based on relation between near infrared absorption spectra measurements and target molecule concentration. In this task, we provide you a database of artificial near infrared absorption spectra of aqueous glucose solution with corresponding concentrations and two reference spectra of glucose $\varepsilon_{glucose}$ and water ε_{water} separately. Assume that at every wavelength absorption's A relationship on reference spectra and concentration c is described by function

$$A = a + d\,(c\varepsilon_{glucose} + (1-c)\varepsilon_{water}) + u,$$

where u is random noise and a, d, c and c are real numbers.

Fit a model to the 10 spectra Sample_Gluc-xxx.h5 given in Data_prepared_syntetic.zip and estimate parameters a, d, c. Numbers in the file names are real concentrations. Evaluate concentration prediction accuracy using RMSE (root mean squared error) measure.

4. Classification problem

In data_for_classification.csv you will find the data, where target variable is y, which should be categorized into two categories: 0 if y < 244, 1 if y >= 244. Your task is to make a classification model which predicts the category of y. Variables x_.. are possible predictors, but not all of them must be used in the model. You can freely choose the model you want to use, briefly explain why are you using this model, evaluate the prediction accuracy by calculating the percentage of True positive (both real and predicted category is 1) and True negative (both real and predicted category is 0) predictions. Find which predictors should be used in order to get the best prediction accuracy. This is real world data, the target variable is hard to predict, so do not expect very good model performance.

5. Code efficiency

In file code_efficiency.py you will find a inefficiently (in terms of calculation time) written code which imports data and calculates maximum value of column z1_32 at every 5 seconds period. Rewrite a code in such a way that it would do the same thing but would run much faster. What is the running time ratio between slow code given and fast code you wrote?