Exercise Sheet 3: Data Science Methods

Technische Universität München

Winter semester 2024-2025

Instructions

You can hand in your solutions during the next tutorial on 12/12/24 (note the 5/12 is DIES ACADEMICUS, so no tutorial) or submit them earlier at the group office 1269 in the Physics Department. For code, you can send it to `alan.zander@tum.de`.

1 Exercise: Nelder-Mead Simplex Algorithm

1.1 Implement the Nelder-Mead Simplex Algorithm presented in the lecture, i.e. write a function in Python that minimizes a function passed as an argument. This function should return at least the minimum, the value of the minimized function at the minimum and the number of iterations needed to find it.

1.2 Minimize with your self-made function the Rosenbrock function

$$f(x, y) = (a - x)^2 + b\left(y - x^2\right)^2, \tag{1}$$

for $a = 1$ and $b = 100$. Furthermore, minimize

$$g(x, y) = |x - 10| + |y + 1|. \tag{2}$$

1.3 Now minimize the functions above with the built-in `scipy.optimize.minimize` with the option `method='Nelder-Mead'`. Measure the time it takes for both your minimizer and `minimize` from `scipy` to find the minimum of the Rosenbrock function (1). Also, how many iterations were needed?

2 Exercise: Hypothesis Testing

Consider two hypotheses H_0, H_1 Consider two hypotheses H_0 and H_1:
Under H_0, the random variable $\mathbf{X} = (x_1, x_2)$ is distributed as a multivariate normal:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0. \\ 0. \end{bmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Under H_1, the random variable $\mathbf{X} = (x_1, x_2)$ is distributed as a multivariate normal:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 2. \\ 0. \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.2 & 0.2 \\ 0.2 & 1 \end{bmatrix}$$

Furthermore, consider the statistics $T_1(x_1, x_2) = x_1$, $T_2(x_1, x_2) = x_2$, $T_3(x_1, x_2) = x_1^2 + x_2^2$ and $T_4(x_1, x_2) = e^{x_1/2} - x_2$.

2.1 Determine the distributions of these statistics via Monte Carlo sampling with respect to the null hypothesis $P(T_i|H_0)$ and to the alternative hypothesis $P(T_i|H_1)$, for example, by making a histogram.

2.2 Construct a ROC-curve for all four test statistics.

2.3 For a fixed size of $\alpha = 0.05$, what is the best statistic? For this statistic, what would be the cutoff value t_0 delimiting the rejection and acceptance region?

2.4 Write a function in Python that computes the p-value for an observation $X = (x_1, x_2)$. What is the p-value of $X = (2, 1)$ for each statistic T_1, T_2, T_3, T_4?

2.5 Choose your favorite statistic and plot the distribution of the p-value under both H_0 and H_1.

3 Exercise: Pearson's χ^2-Test (Bonus)

Pearson's χ^2-test is a statistical method widely used to determine whether observed data fit a theoretical distribution. It compares the frequencies of observed data $\mathcal{O}_i$ in predefined categories (bins, $i = 1, \ldots, k$) to the expected frequencies $\mathcal{E}_i$ calculated with the theoretical distribution. One of the key advantages of this test is that the Pearson statistic,

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{k} \frac{(\mathcal{O}_i - \mathcal{E}_i)^2}{\mathcal{E}_i}, \tag{3}$$

follows a known distribution under the null hypothesis, namely the χ^2-distribution (surprise!), given $\mathcal{E}_i \gtrsim 5$ and that the observations are independent of one another.

In this exercise, we will investigate whether or not the daily returns of the stock market are normally distributed (more specifically, the daily returns of SPY, the largest ETF in the U.S.).

3.1 Load the file `SPY_close_price.csv` into your working file (Python, Jupyter Notebook). Each row corresponds to the closing price of a day in the stock market for SPY. Calculate the relative change in price with respect to the previous day and make a histogram to visualize the data (the daily change or return of SPY).

3.2 Fit a Gaussian distribution to the data, i.e. estimate the parameters μ, σ^2 (e.g. via MLE) assuming the data is normally distributed. This will be our null-hypothesis H_0 we want to reject or not. Plot this Gaussian in the histogram from before for comparison.

3.3 We need to bin the data into $\{\mathcal{O}_i\}$. For that, divide the data into the following 22 bins: `[-100., -3.5, -3.15, -2.8, -2.45, -2.1, -1.75, -1.4, -1.05, -0.7, -0.35, 0., 0.35, 0.7, 1.05, 1.4, 1.75, 2.1, 2.45, 2.8, 3.15, 3.5, 100.]`. This choice ensures that $\mathcal{E}_i \gtrsim 5$.

3.4 Now determine the probabilities p_i for an observation (the daily change of SPY) to be in the i^{th}-bin, assuming the null-hypothesis H_0. The expected frequency in then $\mathcal{E}_i = n_{\text{data}} p_i$, where $n_{\text{data}} = \sum_{i=1}^{k} \mathcal{O}_i$.

3.5 We have all ingredients to compute Pearson's χ^2_{Pearson}-statistic, but first we choose a significance level of $\alpha = 0.05$, i.e. we reject H_0 if the p-value of the data is smaller than α. Now compute χ^2_{Pearson} and the p-value, given that our statistic is distributed according to χ^2_{k-1} under H_0, i.e. $P(\chi^2_{\text{Pearson}}|H_0) \sim \chi^2_{k-1}$. Can we reject H_0?

3.6 Verify via MC sampling (as we did in Exercise 2) that indeed $P(\chi^2_{\text{Pearson}}|H_0) \sim \chi^2_{k-1}$.