Machine Learning in ICT

TECHNISCHE UNIVERSITÄT DARMSTADT

Programming Exercise 1: Probability Theory and Statistics
SoSe 2022
Due date: May 11, 2022

Department of Electrical Engineering and Information Technology
Instructors: Profs A. Klein, H. Koeppl
Teaching Assistant: Y. Eich, yannick.eich@tu-darmstadt.de, G. Ekinci, gizem.ekinci@tu-darmstadt.de

General Information

In the practical exercises we work with Python 3.9. Make sure to submit .py files that print the solutions and plots the figures asked for in the exercises.

You can choose any integrated development environment (IDE) you like to work with from now on. I recommend PyCharm, but Jupyter Notebooks is another good choice here. You will find a lot of helpful tutorials on YouTube if you are completely new to Python.

As we do not want to write all functions from scratch, we can use different toolboxes. The most important ones for these exercises are NumPy for matrix operations, SciPy for algorithms and Matplotlib for plots. You can either install the toolboxes with programs like http://anaconda.com or add them directly in PyCharm.

Notice that not all exercises get evaluated. Please only submit solutions to the exercises indicated by the attainable points. Feel free to use the public discussion forum to ask for help, but do not provide solutions there.

Problem 1 (10 pts)

(a) A fair coin is tossed n times. Simulate a Bernoulli random variable with success probability $p = 0.5$. At each iteration k, compute a sample average of all k sampled elements. Afterwards, produce a plot of the average vs iterations. According to the law of large numbers, the sample average approaches the mathematical expectation, as $n \to \infty$. Take $n = 10^3$. **(5 pts)**

(b) Make n following experiments. At k-th experiment ($k = 1, \ldots, n$), draw k elements from a Bernoulli distribution m times, compute an average of k elements for each time, and save the result into a corresponding row of an n-by-m matrix (thus, you will have m averages in a k-th row). Afterwards, plot the results as the average vs the sample size using Matplotlib's plot and errorbar functions. According to the law of large numbers, the deviation of the average should decrease with the increase of the sample size. Take $n = 10^3, m = 10$. **(5 pts)**

(c) Assume that the coin is biased with the probability of successes $p \in \{0.3, 0.9\}$. Test the two previous programs and verify that it converges to a new mathematical expectation.

Problem 2 (10 pts)

(a) The sum of n independent Bernoulli random variables with success probability p has a binomial distribution with parameters n, p. Conduct the following procedure m times. Use the program from Problem 1 a.,

compute the number of successes for n tosses, and save this number into a corresponding row of an m-by-1 vector. After the procedure is done, draw a binomially-distributed random variable m times with parameters n, p, and save it into another m-by-1 vector. For both vectors, plot the resulting probability distributions with the help of a histogram and compute the mean-square error of the difference. Use the following parameter values: $n = 10^3, p = 0.3, m = \{10^3, 10^4, 10^5\}$. **(5 pts)**

(b) According to the Poisson limit theorem, as $n \to \infty$ and and $p \to 0$, the Binomial(n, p) distribution approaches Poisson(np) distribution. Assume $n = 10^3, p = \{10^{-1}, 10^{-2}, 10^{-3}\}, m = 10^5$ and modify the previous program in order to compare Binomial and Poisson distributions. **(5 pts)**

(c) According to the de Moivre-Laplace theorem, as $n \to \infty$ and p is fixed, the following approximation of a binomial by a normal distribution can be used

$$P\{X = k\} = \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}$$

Take $n = 10^3, p = 0.3, m = 10^5$. Modify the program again in order to compare binomial and normal distributions.

Problem 3

The central limit theorem states that the mean of a large number of independent random variables with finite mean and variance exhibits a normal distribution.

(a) Let $X_1, \ldots, X_n \sim \mathcal{U}(a, b)$. The mean and variance are given by $\mu = \frac{a+b}{2}$,$\sigma^2 = \frac{(b-a)^2}{12}$ Let $S_n = X_1 + \ldots X_n$ and $Z = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Make m iterations, at each iteration, generate $X_1, \ldots X_n$, compute Z, and save it into a vector. After all iterations, use histogram in order to find the distribution of Z and overlay it with $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ Use $a = 0, b = 1, n \in \{10, 25, 50, 100\}, m = 10^3$.

(b) Let $X_1, \ldots, X_n \sim$ Exp(λ). The mean and variance are $\mu = \frac{1}{\lambda}, \sigma^2 = \frac{1}{\lambda^2}$. Repeat the procedure from a. with $\lambda = 1, n \in \{10, 25, 50, 100\}, m = 10^3$.

(c) Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Repeat the procedure from a. with $\mu = 0, \sigma = 1, n \in \{10, 25, 50, 100\}, m = 10^3$.

(d) Let $X_1, \ldots, X_n \sim U(a, b), Y_1, \ldots, Y_n \sim$ Exp$(\lambda), Z_1, \ldots, Z_n \sim \mathcal{N}(\mu, \sigma^2)$.
Let $Z = \frac{\sum_{i=1}^{n}(X_i + Y_i + Z_i) - \sum_{i=1}^{n}(EX_i + EY_i + EZ_i)}{\sqrt{\sum_{i=1}^{n}(Var(X_i) + Var(Y_i) + Var(Z_i))}}$. Repeat the procedure from a. choosing the parameters as in a.-c.