Homework 2

Simulation and Performance Evaluation – University of Trento

You can solve the following assignments using any programming language. In doing so, make sure to implement the formulas explained in class, and try not to use functions made available by the languages to achieve the required tasks. However, you are allowed to use utility functions (e.g., for sorting, to compute quantiles, and of course to draw random numbers uniformly at random in the interval $[0, 1]$ when needed).

Exercise 1

Load the data from the CSV file `data_ex1.csv`. This data represents measurements of some quantity over a few days. In each line: the 1st value refers to the time of the measurement; the 2nd value is the measurement output. If you draw a scatter plot, you should see a clear trend in the data.

1. Use least squares to remove the trend. You only need polynomial functions for this.
2. After having verified that a good value for the maximum degree of the the polynomial is 5, remove the trend from the data and fit a Gaussian distribution to the resulting dataset.
3. Give the mean and variance of the distribution, and draw a QQ-plot to determine if the Gaussian approximation holds. Give a prediction interval for future samples from this Gaussian distribution.
4. Discuss what would happen if you fit a polynomial of degree different than 5 to the data.

Exercise 2

Load the data from the CSV file `data_ex2.csv`. These are samples from three different, independent Gaussian distributions, all mixed together.

1. Implement the Expectation-Maximization algorithm to fit a mixture of three Gaussian distributions to the data. Try both with and without the prior update step. Discuss the results.
2. Give the parameters of the distributions thus found, and plot the corresponding PDFs on top of the empirical PDFs of the data (e.g., the histogram).

Exercise 3

The binomial distribution describes the statistics of the number of successful events for a Bernoulli experiment repeated N times, where the probability of success of each experiment is p.

1. Execute $N_{\text{tr}} = 10000$ trials, each made of $N = 100$ Bernoulli experiments with probability of success $p = 0.05$. [*Hint*: to test whether a Bernoulli experiment is successful or not, draw a random number $u \sim \mathcal{U}(0, 1)$, and check if $u \leq p$.]
2. For each trial i, count the number of successes s_i, and draw the empirical probability mass function (PMF) of the number of successes throughout all trials. Compare against the theoretical binomial PMF.
3. Compare the empirical and the theoretical binomial distributions against a Poisson distribution of parameter $\lambda = Np$. Repeat the comparison for different values of N and p. When does the Poisson PMF accurately approximate the binomial PMF?

Exercise 4

In a popular board game, the players roll two dice at every turn. They want to test the fairness of the dice, so they note the number of occurrences of each possible result, from 2 to 12. The collected data are as follows

value	2	3	4	5	6	7	8	9	10	11	12
# occurrences	1	4	2	7	10	9	9	14	7	5	3

1. Find the probability mass function of the distribution
2. Run a chi-squared test on this data to check if it is in accordance with the discrete triangular distribution that characterizes a 2-dice roll:

$$P[X = k] = \frac{1}{36} \cdot \begin{cases} k - 1, & k = 2, \dots, 7 \\ 13 - k, & k = 8, 9, \dots, 12 \end{cases} \tag{1}$$

   Discuss the result of the test.

Exercise 5

Compute an approximate value for π by using Monte-Carlo simulation to approximate the ratio between the area of a circle of radius 1 to the area of the square circumscribed to it (which has side length equal to 2).

1. Set a stopping rule in terms of the confidence interval for the success probability (where a success occurs if a point falls within the circle). Make an algorithm that keeps drawing additional points until the stopping rule is satisfied.

Exercise 6

Verify that the CDF inversion formula for the generation of exponential random variates of average value equal to 2 actually generates exponential variates.

1. Extract N_{tr} exponential random variates with the same average value equal to 2 using through the CDF inversion method.
2. Draw a QQ-plot to compare your draws against the quantiles of the exponential distribution with the same average value.
3. What happens if, instead, you draw your QQ-plot against the quantiles of an exponential distribution with a different average value? And what if you increase (e.g., $2\times$, $4\times$, ...) the number of exponential draws? Discuss.