Computer Lab 6
Computational Statistics

Linköpings Universitet, IDA, Statistik

November 29, 2021

Course code and name:	732A90 Computational Statistics
Lab session:	6.12, 8-10
Submission deadline:	8.12, 23:59
Resubmission deadlines:	resubmission 1: 22.12, 23:59; resubmission 2 for labs 5-6: 21.1
Seminar:	Seminar 3 (first part) on 15.12
Teachers:	Maryna Prus, Filip Ekström, Joel Oskarsson, Martynas Lukosevicius, Shashi Nagarajan, Yifan Ding
Instructions:	This computer laboratory is a part of the examination Create a group report (in English) on the solutions to the lab as a .PDF file. All R codes should be included as an appendix into your report. In the report reference all consulted sources and disclose all collaborations. The report should be handed in via LISAM (or alternatively in case of problems e–mailed to your teacher - see file "lab groups" on lisam).

Exercises originally developed by Krzysztof Bartoszek

Question 1: Genetic algorithm

In this assignment, you will try to perform one-dimensional maximization with the help of a genetic algorithm.

1. Define the function

$$f(x) := \frac{x^2}{e^x} - 2\exp(-(9\sin x)/(x^2 + x + 1))$$

2. Define the function `crossover()`: for two scalars x and y it returns their "kid" as $(x+y)/2$.

3. Define the function `mutate()` that for a scalar x returns the result of the integer division $x^2 \mod 30$. (Operation $\mod$ is denoted in `R` as `%%`).

4. Write a function that depends on the parameters `maxiter` and `mutprob` and:

   (a) Plots function f in the range from 0 to 30. Do you see any maximum value?

   (b) Defines an initial population for the genetic algorithm as $X = (0, 5, 10, 15, \dots, 30)$.

   (c) Computes vector `Values` that contains the function values for each population point.

   (d) Performs `maxiter` iterations where at each iteration

      i. Two indexes are randomly sampled from the current population, they are further used as parents (use `sample()`).

      ii. One index with the smallest objective function is selected from the current population, the point is referred to as victim (use `order()`).

      iii. Parents are used to produce a new kid by crossover. Mutate this kid with probability `mutprob` (use `crossover()`, `mutate()`).

      iv. The victim is replaced by the kid in the population and the vector `Values` is updated.

      v. The current maximal value of the objective function is saved.

   (e) Add the final observations to the current plot in another colour.

5. Run your code with different combinations of `maxiter`$= 10, 100$ and `mutprob`$= 0.1, 0.5, 0.9$. Observe the initial population and final population. Conclusions?

Question 2: EM algorithm

The data file `physical.csv` describes a behavior of two related physical processes $Y = Y(X)$ and $Z = Z(X)$.

1. Make a time series plot describing dependence of Z and Y versus X. Does it seem that two processes are related to each other? What can you say about the variation of the response values with respect to X?

2. Note that there are some missing values of Z in the data which implies problems in estimating models by maximum likelihood. Given is the following model:

$$Y_i \sim \exp\left(X_i/\lambda\right), \qquad Z_i \sim \exp\left(X_i/(2\lambda)\right),$$

where λ is some unknown parameter.
For this model it can be proved analytically that λ^{k+1} ($\lambda^{k+1} = \operatorname{argmax}_\lambda \; Q(\lambda, \lambda^k)$) for EM algorithm for estimation of λ is given by

$$\lambda^{k+1} = \frac{1}{2n}\left(\sum_{i=1,\ldots,n} X_i Y_i + 0.5 \sum_{i=1,\ldots,n,\, i\notin u} X_i Z_i + |u|\lambda^k\right),$$

where u is the set of indexes for missing data for Z, i.e. observation Z_i is missing for all $i \in u$, $|u|$ denotes number of elements in u (number of missing values) and n is the sample size.
Using this information implement this algorithm in `R`. Use $\lambda_0 = 100$ and convergence criterion "stop if the change in λ is less than 0.001". What is the optimal λ and how many iterations were required to compute it?