# Compulsory exercise for STK4051/9051 - Computational statistics

Spring 2022

Part 1 (of 2)

This is the first part of the compulsory exercise for STK4051/9051, spring semester 2022. The second part of the compulsory exercise will be made available in the end of Mars. (so it is a grate tip to complete the first part by the end of March) The deadline for the complete compulsory exercise (including part 2) is May 5th the complete exercise has to be delivered within the Canvas system. You can get feedback on the first part by sending your solution in mail to the lecturer. Reports may be written in Norwegian or English, and should preferably be text processed (e.g. LaTeX, Word, RMarkdown). Write concisely. Relevant figures need to be included in the report. Copies of relevant parts of machine programs used (in R, or matlab, or similar) are also to be included, perhaps as an appendix to the report.

This first part contains five exercises and comprises five pages (including this front page). Some R-code is available from the course web-page. You are free to use other software, but would then need to translate or write your own code for that part included in the R-script.

Data sets to be used are available on the course webpage, in a standard R save file. Read the corresponding .txt file to understand the structure of data.

| | |
|---|---|
| sparseDataWithErrors.dat | Exercise 1,2,3 |
| blurredSparseDataWithErrors.dat | Exercise 1g (STK 9051 only) |
| optimalTransport.dat | Exercise 4 |
| functionEstimationNN.dat | Exercise 5 |

There will be a Q and A, with respect to the compulsory exercise on the course webpage. The page is updated when questions arise.

**Exercise 1** (Lp-regularization)We will in this exercise consider the problem where the number of parameters we will estimate are of the same size as the number of data. In this case it is common to use regularization to impose additional constraints on the model. In a simplified regression model we have data on the form:

$$y_i = \beta_i + \varepsilon_i, i = 1, \dots, n \tag{1}$$

Where $y_i$ is the data, $\beta_i$ is the parameter and $\varepsilon_i$ is an error term, we will assume that the error term has variations according to a normal distribution with mean zero and unit variance, i.e. $\varepsilon_i \sim N(0,1^2)$.

a) Derive the maximum likelihood estimator for $\beta_i$, $i = 1, \dots, n$.

b) An alternative estimator is derived using a penalized least squares, by solving the optimization problem:

$$\min_{\beta}\{-\ell(\boldsymbol{\beta}|\boldsymbol{y}) + \tfrac{\gamma}{p}\,||\boldsymbol{\beta}||_p^p\} \tag{2}$$

where $\ell(\boldsymbol{\beta}|\boldsymbol{y})$ is the likelihood,

$$||\boldsymbol{\beta}||_p^p = \sum_{i=1}^{n} |\beta_i|^p,$$

and $\gamma$ is a regularization parameter, giving a tradeoff between data fit and the penalty term. The problem (2) have a solution on the form:

$$f_{p,\gamma}(\beta_i) = y_i, i = 1, \dots n \tag{3}$$

This means that the estimator for $\beta_i$ only depend on the corresponding data point $y_i$. The expression for the function $f_{p,\gamma}(\beta_i)$ only depend on $p, \gamma$. Show that

$$f_{p,\gamma}(\beta_i) = \beta_i + \gamma \cdot \mathrm{sign}(\beta_i)|\beta_i|^{p-1}$$

c) For $\gamma$=1 and $\gamma$=0.2, plot the function $f_{p,\gamma}(\beta)$ for $p = 1.1,\ 2,\ 5$ and 100 on the interval $[-5,\ 5]$, plot also the inverse function by flipping the order of the arguments in the plotting function. Give an interpretation of the results.

d) Implement a function which finds the root of expression (3). Use the method of bisection (book page 23). The root of (3) will be the estimator $\hat{\beta}_{\gamma,p}(y)$. What are good starting values for upper and lower bounds? Test the algorithm for $\gamma = 1$, and $p = 1.1$, $p = 2$ and $p = 100$. Evaluate the results for $y$ in the interval $[-5, 5]$, and plot the results in the same form as in c).

e) Use the data set `sparseDataWithErrors.dat` and perform estimation of parameters $\beta_i, i = 1, \dots, n$, from the data $y_i, i = 1, \dots, n$. Compare the result to the ground truth in terms of residual sum of squares. Do the estimation for penalized regression, i.e. compute $\hat{\beta}_{\gamma,p}$ with $\gamma = 1$, $p = 1.1$, $p = 2$ and $p = 100$, compare also to the MLE estimator. For $p = 100$, try also to estimate the parameters $\beta_i$ by using the residuals, i.e. $\hat{\beta}^{\text{Alt}}_{1,100} = y - \hat{\beta}_{1,100}$. Comment on the results. How does MLE measure up?

f) In the linear regression problem for $p = n$, we have data on the form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Where $\boldsymbol{y}$ is the $(n \times 1)$ data vector, $\boldsymbol{\beta}$ is the $(n \times 1)$ parameter vector, $\boldsymbol{X}$ is a $(n \times n)$ design matrix, and $\boldsymbol{\varepsilon}$ is the $(n \times 1)$ error vector, with $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. How can you use the ADMM algorithm together with the solution to the problem above to derive a solution to the penalized regression problem?

$$\min_{\beta}\{-\ell(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) + \tfrac{\gamma}{p}\,||\boldsymbol{\beta}||_p^p\}$$

g) STK 9051 only. Implement the ADMM algorithm for the Lp-regularization. You can combine your result from 1.d with the ADMM algorithm and code presented for $p = 1$ during lecture. Solve the problem using the data set `blurredSparseDataWithErrors.dat` given on the course page. Use $\gamma = 0.1$, and give the solution for $p = 1.1$ and $p = 2$.

**Exercise 2** (EM-algorithm) We will now assume that the data $Y_i, i = 1, \dots n$, are independent and identically distributed according to the mixture distribution:

$$f(y_i) = p \cdot \phi(y_i; 0,1^2) + (1-p) \cdot \phi(y_i; 0, \tau^2 + 1^2)$$

where $\phi(y_i; \mu, \sigma^2)$ is the normal density with mean $\mu$ and variance $\sigma^2$. We will now consider estimation of the parameters $\theta = (p, \tau^2)$.

a) Give an expression for the likelihood of $\theta$.

b) Introduce the variable $C_i$ which identify which of the two modes $y_i$ belongs to. Give an expression for the complete log-likelihood using the pairs $(C_i, y_i), i = 1, \dots n$

c) Give an expression for $Q(\theta|\theta^{(t)})$, what is the interpretation of $Q(\theta|\theta^{(t)})$. Derive the estimates for $\theta = (p, \tau^2)$, using $Q(\theta|\theta^{(t)})$.

d) Implement the solution you derived in c) as a function, and apply it to the data `sparseDataWithErrors.dat`. What is a good initialization?

e) Compute a bootstrap estimate of the uncertainty of the two parameters, using the function from d. Sample B=1000 times and display scatter plot of the values.

f) Compute the observed information matrix. How can you use the observed information matrix to give an uncertainty estimate for $\theta$? Compare the result to e.

g) Compute the likelihood from 2a, in a dense grid $(p, \tau^2) \in [0.8, 1] \times [50, 130]$. Normalize it by dividing by the maximum value and plot it in a contour plot. Use contours lines [0.01 0.1 0.5 0.95], mark the ML estimator in the plot. Compare the results to e and f.

**Exercise 3**

Assuming that the data $Y_i$ are on the same form as in Exercise 1, i.e.

$$y_i = \beta_i + \varepsilon_i, i = 1, \dots, n,$$

where the parameters are as described for (1). Assume also that $y_i$, follows the mixture distribution as in Exercise 2,

$$f(y_i) = p \cdot \phi(y_i; 0, 1^2) + (1 - p) \cdot \phi(y_i; 0, \tau^2 + 1^2)$$

a) Use a Bayesian interpretation of $\beta_i$,(e.g. assume $\beta_i$ is random) and argue that $P(\beta_i = 0|C_i = 0) = 1$, and that $f(\beta_i|C_i = 1) = \phi(\beta_i; 0, \tau^2)$. Give an expression for $P(\beta_i = 0|y_i = y)$.

b) Argue that the estimator for $\beta_i$(i.e. the conditional expectation of the parameter given the data), is $P(C_i = 1|y_i = y)E(\beta_i\ |y_i = y, C_i = 1)$. Plot this expression as a function of y in the interval $[-5, 5]$, and compare the result with 1.d. Use values: $p = 0.9$, and $\tau^2 = 80$ . Hint: $E(\beta_i\ |y_i = y, C_i = 1) = \frac{\tau^2}{\tau^2+1} y$.

c) For the dataset in `sparseDataWithErrors.dat`, evaluate the estimator using the values $p = 0.9$, and $\tau^2 = 80$ . Compare the result with 1.e, also in terms of residual sum of squares.

**Exercise 4:** (Combinatorial optimization) In this problem we will consider a logistics problem. Assume that you work in a company which want to distribute their product to 20 cities, starting from its home city. In the file `optimalTransport.dat` you will find the locations of all 21 cities the distances between the cities is the time it takes to travel between them. The first city in the file is the home city. We want to find the loop which reduces the traveling time going from the home city, cycling through all cities and returning to the home city. You can modify the scripts found on the home page of the course to solve the problem.

a) Write an optimization using simulated annealing to find the optimal solution. Define your neighborhood using mathematical notation. Define your cooling schedule. Argue that your proposed algorithm can reach all possible states.

b) Implement also a TABU search for the optimal path. In what way does both the simulated annealing and the TABU algorithm differ from a steepest decent algorithm?

c) If your manager ask you whether the solution you propose is the optimal one, how should you reply? How can you improve confidence in your result?

The company considers buying an additional lorry and want to find out how much this will reduce the distribution time. If the two lorries start at the same time the distribution time is defined as the time it takes until both lorries have returned to the home city.

d) Suggest a modification to your simulated annealing algorithm such that you can have two cycles which both must start and end in the home city. Define a neighborhood for this setup, and argue that the algorithm is able to reach to reach all possible states using your proposed neighborhood. You do *not* need to implement the solution.

**Exercise 5:** (Stochastic gradient decent, SGD) In this exercise we will try out the minibatch approach for optimizing neural nets. We will consider a simple 1D situation. The input $x$ and output $y$ are both one dimensional. We will use a model architecture with one hidden layer, and a width of 50. As the activation function, $\sigma(x)$, you should use the ReLu, which is defined as:

$$\sigma(x) = \begin{cases} x \text{ for } x > 0 \\ 0 \text{ for } x \leq 0 \end{cases}$$

Define also the derivative of the activation function to be:

$$\sigma'(x) = \begin{cases} 1 \text{ for } x > 0 \\ 0 \text{ for } x \leq 0 \end{cases}$$

The estimator have the form (also called the architecture):

$$f(x) = \sum_{i=1}^{50} \beta_i \sigma(\alpha_i x + \alpha_{0,i}) + \beta_0$$

We will minimize the squared sum of errors:

$$\text{sse} = \sum_{i=1}^{N} \left(y_i - \hat{f}(x_i)\right)^2$$

a) Adapt the formulas from the lecture slides or the SGD note of Geir Storvik such that these can be used for the specific problem formulation defined above. Present the formulas that are relevant for implementing SGD on the current problem. How many parameters are there in the architecture?

b) Implement the SGD for the architecture above and test it on the data found in `functionEstimationNN.dat`. Test different alternatives for the learning rate and record the test error for each epoch. Use a batch size of 50.

c) Discuss the choices made and comment on the results from b.