Computer Assignment 1

Ali Bemani **Ali.Bemani@hig.se**
Oscar Bautista Gonzalez **Oscar.Bautista.Gonzalez@hig.se**
Martin Hellkvist **Martin.Hellkvist@angstrom.uu.se**

Solutions for **Computer Assignment 1** by Group 2 are in this document provided in the form of photocopies of handwritten notes.

Problem 1

We consider the following minimization problem of finding the optimal parameter vector $\boldsymbol{w}$ for the regularized least-squares (LS) problem:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i \in [N]} (y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w})^2 + \lambda \|\boldsymbol{w}\|_2^2, \tag{1}$$

where we have scalar desired responses $y_i \in \mathbb{R}$, and vector valued regressors $\boldsymbol{x}_i \in \mathbb{R}^{d \times 1}$ and parameter vector $\boldsymbol{w} \in \mathbb{R}^{d \times 1}$.

Problem 1. a)

Q: Find the closed form solution of the problem.

Solution: The optimization problem is strictly convex, because we have that

$$\frac{1}{N} \sum_{i \in [N]} (y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w})^2$$

is convex, and $\lambda \|\boldsymbol{w}\|_2^2$ is strictly convex. Note that therefore, the solution $\boldsymbol{w}^*$ is unique. We determine $\boldsymbol{w}^*$ by setting the gradient of the objective function in (1) to zero, and solving for $\boldsymbol{w}$. For ease of exposition, we write the problem in matrix form with the following definitions:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{bmatrix}. \tag{2}$$

And the corresponding reformulation of (1):

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|_2^2. \tag{3}$$

The gradient of the righthand side with respect to $\boldsymbol{w}$ is

$$\frac{2}{N} \left(-\boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} + \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{w} + \lambda N \boldsymbol{I}_d \boldsymbol{w} \right). \tag{4}$$

Setting the gradient to zero yields the equation

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda N \boldsymbol{I}_d)\boldsymbol{w}^* = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}, \tag{5}$$

which can be solved using matrix inversion:

$$\boldsymbol{w}^* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda N \boldsymbol{I}_d)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{6}$$

Problem 1. b)

Q: Consider "Individual household electric power consumption" dataset ($N = 2075259$, $d = 9$) and find the optimal linear regressor from the closed-form expression.

Solution: We load the data using the `pandas` python library.

The dataset consists of 9 columns:

- `Date`
- `Time`
- `Global_active_power`
- `Global_reactive_power`
- `Voltage`
- `Global_intensity`
- `Sub_metering_1`
- `Sub_metering_2`
- `Sub_metering_3`

The raw dataset is missing some values (25979 to be exact) in the column `Sub_metering_3`. We pre-process the data by removing the data points containing these missing values using the method `pandas.DataFrame.dropna()`.

We will here present the result of finding the optimal parameter vector $\boldsymbol{w}^*$ of estimating the values `Global_active_power` using the other columns, excluding `Date` and `Time`, i.e., the last six columns to estimate the third one. We exclude the Date and Time because they are not decimal numbers. There could of course be relevant information there, to estimate the other data, as what season in the year, or time of day, since we can expect power consumption to vary with that.

We use the regularization parameter $\lambda = 1$

We measure the performance of the solution sing the normalized mean square error (NMSE), which we define by

$$\epsilon = \frac{1}{N} \sum_{i \in [N]} \frac{(y_i - \hat{y}_i)^2}{\|\boldsymbol{y}\|_2^2},$$

where y_i is the i^{th} value of the third column `Global_active_power`, and $\hat{\boldsymbol{y}}_i$ is the estimate of it using the solution to (1), i.e., $\hat{y}_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}^*$.

Using the six columns to estimate the third, we obtain an NMSE of $\epsilon = 3.8 \times 10^{-3}$. In contrast, using only the fourth column to estimate the third, we obtain much worse performance with NMSE $\epsilon = 0.98$, which is more than 200 times larger than the first experiment.

Problem 1. c)

Q: Repeat for "Greenhouse gas observing network" dataset and observe the scalability issue of the closed-form expression.

Solution: Here we have data structured in 16 columns from $N_s = 2921$ measurement sites. We use the first $d = 15$ columns to estimate the 16^{th}. Each column has $N_m = 327$ measurements. We want to find the optimal parameter vector, which fits for all the measurement sites, so we structure the regressor matrix block wise, so the first 327 rows are from the first site, the next block of 327 rows from the second site, etc. Hence, the regressor matrix $\boldsymbol{X}$ now has dimensions $\mathbb{R}^{N_m N_s \times d}$, where $N = N_m N_s = 955167$.

We find the optimal parameters $\boldsymbol{w}^*$ in the same way as in **Problem 1. b)**. We again use the regularization parameter $\lambda = 1$.

The NMSE obtained is here $\epsilon = 4.6 \times 10^{-2}$.

Varying the regularization does not change the NMSE significantly.

Problem 1. d)

Q: How would you address even bigger datasets?

Solution: We did not notice any issues regarding computational complexity in the two exercises. As the number of rows N in $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ grows, it is the products $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ that becomes more expensive, but the complexity grows linearly with N so we don't notice it until N is extremely large.

The heaviest duty of **Problem 1. c)** was to read the Greenhouse data, because each site's measurement was in a separate file.