Machine Learning (SS 23)
Assignment 04: Linear Regression

Team Members:

- Likhit Jain, 3678905, M.Sc. Computer Science
- Tareq Abu El Komboz, 3405686, M.Sc. Informatik
- Serge Kotchourko, 3309449, M.Sc. Informatik

1. Linear Regression

Question (a) Linear regression can include nonlinear features. Why is it still called linear regression? In what sense is it linear?

Answer Because a linear function is still sought after, best describing the linear relation between the (possible transformed, non-linear) input/predictor variable(s) and its response variable(s), i.e. still searching for a linear relationship between these parameters.

Question (b) For calculating optimal parameters $\hat{\beta}$ the inverse of $X^T X$ has to be calculated. When would this matrix be singular?

Answer A matrix is singular, iff its determinant is 0, i.e. if $\det(X^T X) = 0$ as $X^T X \in \mathbb{R}^{n \times n}$. Note, if $X \in \mathbb{R}^{n \times n}$, then:

$$\begin{aligned}
\det X^T X &= \det X^T \det X \\
&= \det X \det X \\
&= \det^2 X \\
\Rightarrow \det^2 X &\stackrel{!}{=} 0 \text{ iff } \det X = 0
\end{aligned}$$

By the Fundamental Theorem of Algebra, it follows:

$$\det X = \prod \lambda_i, \text{ where } \lambda_i \text{ is an eigenvalue}$$

Thus, $X^T X$ (and if X a square matrix) is non-singular, iff $\forall i : \lambda_i \neq 0$ or $\ker(X^T X) = \emptyset$, i.e. $X^T X \vec{x} = 0$ has no solution, which can be checked by Gaussian Elimination.

Question (c) Suppose that attempting to optimize the weights $\hat{\beta}$ is unsuccessful because the matrix $X^T X$ is singular. Describe how you would alter the matrix X to prevent this issue.

Answer The easiest way is to add some noise/small error to some predictor variable, i.e. use a transform $\varphi(x) = x + \epsilon$ for some small error $0 \lesssim \epsilon$. The probability that $\hat{X}^T \hat{X}$ is singular will be small, as singular matrices are a small subset of square matrices, but non-zero, hence a check is needed. Also note, that the resulting matrix $\hat{X}$ should be checked such that it is not ill-behaved in this setting and it should be kept in mind, that this error term might also be a factor in the performance of the linear regression model and thus also needs to be "adjusted". To check the for ill-behaviour between the in and output a check can be made, like for example condition number.

2. Regression for Time-Series Prediction

Question (a) Formally define a linear regression model to estimate the function f.

Answer Because we are still searching for a linear dependency between the predictor variables y_t, x_t and response variable y_{t+1}, we can define y_{t+1} as follows:

$$\begin{aligned} y_{t+1} &= f(x_t, y_t) \\ &= X_t \beta_t + Y_t \alpha_t \end{aligned}$$

where x_t and y_t are the t-th row of X and Y respectively and

$$\begin{aligned} X_t &= \begin{pmatrix} x_t \\ x_t \end{pmatrix} \in \mathbb{R}^{2\times 5},\ \beta_t = \begin{pmatrix} \beta_{1,t}, \beta_{2,t} \end{pmatrix} \in \mathbb{R}^{5\times 2} \text{ (i.e. } \beta_{1,t}, \beta_{2,t} \in \mathbb{R}^5) \\ Y_t &= \begin{pmatrix} y_t \\ y_t \end{pmatrix} \in \mathbb{R}^{2\times 2},\ \alpha_t = \begin{pmatrix} \alpha_{1,t}, \alpha_{2,t} \end{pmatrix} \in \mathbb{R}^{2\times 2} \text{ (i.e. } \alpha_{1,t}, \alpha_{2,t} \in \mathbb{R}^2) \end{aligned}$$

Therefore, we can also define f as follows:

$$\begin{aligned} y_{t+1} &= f(x_t, y_t) \\ &= \begin{pmatrix} x_{t,1}, x_{t,2}, \dots, x_{t,5}, y_{t,1}, y_{t,2} \\ x_{t,1}, x_{t,2}, \dots, x_{t,5}, y_{t,1}, y_{t,2} \end{pmatrix} \begin{pmatrix} \beta_{1,t,1}, \beta_{2,t,1} \\ \dots \\ \beta_{1,t,5}, \beta_{2,t,5} \\ \alpha_{1,t,1}, \alpha_{2,t,1} \\ \alpha_{1,t,2}, \alpha_{2,t,2} \end{pmatrix} \end{aligned}$$

Question (b) Update the function signature of f to match the described prediction task. Then, formally define the linear regression model including the lagged variables.

Answer The same idea as above holds here, but the time step is reduced, i.e. $t \in \{1, \dots, T-1\}$, as x_{-1} is not defined (might be extended to use some kind method, like backfill). Hence, the following function f is only defined over these time steps. Now, we extend the above function f with the predictor variable x_{t-1} as lag, which is basically "just adding another feature":

$$\begin{aligned} y_{t+1} &= f(x_t, x_{t-1}, y_t) \\ &= \begin{pmatrix} x_{t,1}, \dots, x_{t,5}, x_{t-1,1}, \dots, x_{t-1,5}, y_{t,1}, y_{t,2} \\ x_{t,1}, \dots, x_{t,5}, x_{t-1,1}, \dots, x_{t-1,5}, y_{t,1}, y_{t,2} \end{pmatrix} \begin{pmatrix} \beta_{1,t,1}, \beta_{2,t,1} \\ \dots \\ \beta_{1,t,5}, \beta_{2,t,5} \\ \gamma_{1,t,1}, \gamma_{2,t,1} \\ \dots \\ \gamma_{1,t,5}, \gamma_{2,t,5} \\ \alpha_{1,t,1}, \alpha_{2,t,1} \\ \alpha_{1,t,2}, \alpha_{2,t,2} \end{pmatrix} \end{aligned}$$

Question (c) Please proceed with this task by using the provided Jupyter notebook.

Answer See uploaded Notebook