Exercises week 36

Contents

September 4-8, 2023

Date: **Deadline is Sunday September 10 at midnight**

Overarching aims of the exercises this week

This set of exercises form an important part of the first project. The analytical exercises deal with the material covered last week on the mathematical interpretations of ordinary least squares and of Ridge regression. The numerical exercises can be seen as a continuation of exercise 3 from week 35, with the inclusion of Ridge regression. This material enters also the discussions of the first project.

Exercise 1: Analytical exercises

The aim here is to derive the expression for the optimal parameters using Ridge regression. Furthermore, using the singular value decomposition, we will analyze the difference between the ordinary least squares approach and Ridge regression.

The expression for the standard Mean Squared Error (MSE) which we used to define our cost function and the equations for the ordinary least squares (OLS) method, was given by the optimization problem

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}\left\{\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right\}.$$

which we can also write as

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}\sum_{i=0}^{n-1}\left(y_i-\tilde{y}_i\right)^2=\frac{1}{n}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2,$$

where we have used the definition of a norm-2 vector, that is

$$||\boldsymbol{x}||_2=\sqrt{\sum_i x_i^2}.$$

By minimizing the above equation with respect to the parameters $\boldsymbol{\beta}$ we could then obtain an analytical expression for the parameters $\boldsymbol{\beta}$.

We can add a regularization parameter λ by defining a new cost function to be optimized, that is

a) $C(\beta) = \frac{1}{m}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Proof. $C(\beta) = \frac{1}{m}\sum_{i=0}^{m-1}\left(y_i - \sum_{j=0}^{n-1} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=0}^{n-1}\beta_j^2$

$$\frac{\partial C(\beta)}{\partial \beta_k} = \frac{2}{m}\sum_{i=0}^{m-1}\left(y_i - \sum_{j=0}^{n-1} x_{ij}\beta_j\right)\left(-x_{ik}\right) + 2\lambda\beta_k$$

$$\frac{\partial C(\beta)}{\partial \beta} = -\frac{2}{m}X^T(y - X\beta) + 2\lambda I\beta = 0$$

$$-\frac{2}{m}X^T y + \frac{2}{m}X^T X\beta + 2\lambda I\beta = 0$$

$$\frac{2}{m}X^T X\beta + 2\lambda I\beta = \frac{2}{m}X^T y$$

$$\beta\left(\frac{\cancel{2}}{m}X^T X + \cancel{2}\lambda I\right) = \frac{\cancel{2}}{m}X^T y$$

$$\hat{\beta} = (X^T X + \underbrace{m\lambda}_{\lambda} I)^{-1} X^T y$$

$$\min_{\beta\in\mathbb{R}^p}\frac{1}{n}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2+\lambda||\boldsymbol{\beta}||_2^2$$

which leads to the Ridge regression minimization problem. One can require as part of the optimization problem that $||\boldsymbol{\beta}||_2^2\le t$, where t is a finite number larger than zero. We will not implement that here.

a) Expression for Ridge regression

Show that the optimal parameters

$$\hat{\boldsymbol{\beta}}_{\mathrm{Ridge}}=\left(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

with $\boldsymbol{I}$ being a $p\times p$ identity matrix with the constraint that

$$\sum_{i=0}^{p-1}\beta_i^2\le t,$$

with t a finite positive number. In the optimization, we will not require that the latter is satisfied.

The ordinary least squares result is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

b) The singular value decomposition

Here we will use the singular value decomposition of an $n\times p$ matrix $\boldsymbol{X}$ (our design matrix)

$$\boldsymbol{X}=\boldsymbol{U\Sigma V}^T,$$

to study properties of Ridge regression and ordinary least squares regression. Here $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of dimensions $n\times n$ and $p\times p$, respectively, and $\boldsymbol{\Sigma}$ is an $n\times p$ matrix which contains the singular values only. This material was discussed during the lectures of week 35.

Show that you can write the OLS solutions in terms of the eigenvectors (the columns) of the orthogonal matrix $\boldsymbol{U}$ as

$$\tilde{\boldsymbol{y}}_{\mathrm{OLS}}=\boldsymbol{X}\boldsymbol{\beta}=\sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{y}.$$

For Ridge regression, show that the corresponding equation is

$$\tilde{\boldsymbol{y}}_{\mathrm{Ridge}}=\boldsymbol{X}\boldsymbol{\beta}_{\mathrm{Ridge}}=\boldsymbol{U\Sigma V^T}\left(\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T+\lambda\boldsymbol{I}\right)^{-1}(\boldsymbol{U\Sigma V^T})^T\boldsymbol{y}=\sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\frac{\sigma_j^2}{\sigma_j^2+\lambda}\boldsymbol{y},$$

with the vectors $\boldsymbol{u}_j$ being the columns of $\boldsymbol{U}$ from the SVD of the matrix $\boldsymbol{X}$.

Give an interpretation of the results. Section 3.4 of Hastie et al's textbook gives a good discussion of the above results.

Exercise 2: Adding Ridge Regression

b) $\tilde{y}_{ols} = X\beta = X(X^TX)^{-1}X^Ty \overset{X = U\Sigma V^T}{=}$

$$= U\Sigma V^T\left[(U\Sigma V^T)^T(U\Sigma V^T)\right]^{-1}(U\Sigma V^T)^T y =$$

$$= U\Sigma V^T\left[V\Sigma^T \overset{=1}{\boxed{U^T U}}\Sigma V^T\right]^{-1} V\Sigma^T U^T y = U\Sigma V^T (V\Sigma^T\Sigma V^T)^{-1} V\Sigma^T U^T y =$$

$$= U\Sigma V^T \cdot \frac{1}{V\Sigma^2 V^T} \cdot V\Sigma^T U^T y \qquad \text{where } \Sigma^2 = \Sigma^T\Sigma \in \mathbb{R}^{p\times p} : \Sigma^2 = \begin{bmatrix} \sigma_0^2 & & & \\ & \sigma_1^2 & & \\ & & \ddots & \\ & & & \sigma_{p-1}^2 \end{bmatrix}$$

$$= U\Sigma V^T \cdot \frac{1}{\underset{1 =}{\boxed{VV^T}}\Sigma^2} \cdot V\Sigma^T U^T y = U\Sigma V^T \cdot \frac{1}{\Sigma^2} \cdot V\Sigma^T U^T y =$$

$$= U\Sigma \frac{V^TV}{\Sigma^2} \Sigma^T U^T y = UU^T y = \sum_{j=0}^{p-1} u_j u_j^T y$$

$\tilde{y}_{ridge} = X\beta_{ridge} = X(X^TX + \lambda I)^{-1}X^Ty =$

$$= U\Sigma V^T (V\Sigma^T \overset{=1}{\boxed{U^T \cdot U}}\Sigma V^T + \lambda I)^{-1} \cdot V\Sigma^T U^T \cdot y =$$

$$= U\Sigma V^T (V\Sigma^2 V^T + \lambda I)^{-1} \cdot V\Sigma^T U^T y =$$

$$= U\Sigma V^T (VD^2V^T + \lambda VV^T)^{-1} \cdot V\Sigma^T U^T y = U\Sigma V^T \overbrace{\left[V(\Sigma^2 + \lambda I)V^T\right]^{-1}}^{= V^{-T}(\Sigma^2+\lambda I)^{-1}V^{-1} = (V^T)^T(\Sigma^2+\lambda I)^{-1}\cdot V^T} \cdot V\Sigma^T U^T y =$$

$$= U\Sigma \cdot (\Sigma^2 + \lambda I)^{-1} \cdot \Sigma^T U^T y = \frac{U\Sigma^2 U^T}{\Sigma^2 + \lambda I} \cdot y = \sum_{j=0}^{p-1} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^T y$$

This exercise is a continuation of exercise 3 from week 35, see https://compphysics.github.io/MachineLearning/doc/LectureNotes/_build/html/exercisesweek35.html. We will use the same function to generate our data set, still staying with a simple function $y(x)$ which we want to fit using linear regression, but now extending the analysis to include the Ridge regression method.

In this exercise you need to include the same elements from last week, that is

1. scale your data by subtracting the mean value from each column in the design matrix.
2. perform a split of the data in a training set and a test set.

The addition to the analysis this time is the introduction of the hyperparameter λ when introducing Ridge regression.

Extend the code from exercise 3 from week 35 to include Ridge regression with the hyperparameter λ. The optimal parameters $\hat{\beta}$ for Ridge regression can be obtained by matrix inversion in a similar way as done for ordinary least squares. You need to add to your code the following equations

$$\hat{\boldsymbol{\beta}}_{\mathrm{Ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

The ordinary least squares result you encoded last week is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

Use these results to compute the mean squared error for ordinary least squares and Ridge regression first for a polynomial of degree five with $n = 100$ data points and five selected values of $\lambda = [0.0001, 0.001, 0.01, 0.1, 1.0]$. Compute thereafter the mean squared error for the same values of λ for polynomials of degree ten and 15. Discuss your results for the training MSE and test MSE with Ridge regression and ordinary least squares.

By Morten Hjorth-Jensen