FYS-STK4155 Week 36

Edvard B. Rørnes, Isak O. Rukan, Anton A. Brekke

September 5, 2024

Exercise 1

a) Show that the optimal parameters for Ridge regression are given by

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1}$$

Solution: The cost function for ridge regression is given by

$$C_{\text{Ridge}} = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^2 \tag{2}$$

Taking the derivative of the cost function w.r.t. $\boldsymbol{\beta}$ and setting it to zero whilst using $\frac{\partial \boldsymbol{a}^2}{\partial \boldsymbol{a}} = 2\boldsymbol{a}^T$ and the result from the previous exercise set:

$$\frac{\partial(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^T(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})}{\partial \boldsymbol{s}} = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^T\boldsymbol{A}$$

we have

$$\begin{aligned}
0 = \frac{\partial C_{\text{Ridge}}}{\partial \boldsymbol{\beta}} &= -\frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T \\
&= \frac{2}{n}(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{y}^T\boldsymbol{X}) + 2\lambda\boldsymbol{\beta}^T \\
0 &= \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I}) - \boldsymbol{y}^T\boldsymbol{X} \\
\boldsymbol{\beta}^T &= \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \tilde{\lambda}\boldsymbol{I})^{-1} \\
\boldsymbol{\beta} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}$$

where we defined $\tilde{\lambda} \equiv n\lambda$, renamed $\tilde{\lambda} \to \lambda$ and used that the matrix in the parenthesis is a symmetric matrix and thus its inverse must also be symmetric. The constraint requirement $\boldsymbol{\beta}^2 \leq t$ for some $t < \infty$ is just a requirement so that we can choose our arbitrary parameter $\lambda \geq 0$ to be sufficiently small s.t. the cost function (2) does not diverge.

b) Show that for OLS the solution in terms of the eigenvectors of the orthogonal matrix U is given by

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

and that the corresponding equation for Ridge is given by

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

where $\boldsymbol{u}_i$ are the columns of $\boldsymbol{U}$ from the SVD of the matrix $\boldsymbol{X}$. Give an interpretation of the results.

Solution:

Using the orthogonality of $\boldsymbol{U}$ and $\boldsymbol{V}$, $(\boldsymbol{A}\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A}^T$ and $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ we have

$$\begin{aligned}
\tilde{\boldsymbol{y}}_{\text{OLS}} &= \boldsymbol{X}\boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}^T)^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma}^T)^{-1}\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{y} = \sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\boldsymbol{y}
\end{aligned}$$

where the last equality holds due to $\boldsymbol{U}$ being orthogonal. The next case is similar but now we need to use that $\boldsymbol{\Sigma}$ is diagonal and that the inverse of a diagonal matrix contains the inverse element of on the diagonal.

$$\begin{aligned}
\tilde{\boldsymbol{y}}_{\text{Ridge}} &= \boldsymbol{X}\boldsymbol{\beta}_{\text{Ridge}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T((\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T+\lambda\boldsymbol{I})^{-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^T+\lambda\boldsymbol{I})^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{V}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}+\lambda\boldsymbol{I})\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y}\\
&= \boldsymbol{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{y}\\
&= \sum_{j=0}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j^T\frac{\sigma_j^2}{\sigma_j^2+\lambda}\boldsymbol{y}
\end{aligned}$$

where once again the last step is valid due to the orthogonality of $\boldsymbol{U}$ and σ_j are the elements on the diagonal of $\boldsymbol{\Sigma}$. Since $\lambda \geq 0$ then this added factor compared to OLS is ≤ 1. The larger λ is the smaller this factor becomes and is the so-called a "shrinkage" factor.

Exercise 2

For low values of λ we see that both Ridge and OLS are practically the same no matter the polynomial degree. For deg = 5 we see that as λ increases both the train and test data MSE for Ridge increase. Here we are underfitting the data and thus when λ increases we are effectively underfitting the data even more. This can be seen to generally be true for deg = 10 as well but here there is more variance when trying out different seeds. For deg = 15 however we see that increasing λ actually decreases the test MSE whilst still slightly increasing the training MSE. In this case it is because we are overfitting with the data. Since large λ corresponds to penalizing large coefficients, this effectively works to reduce the overfitting. For the largest value of λ this eventually overshoots and we essentially go back to underfitting the data as with the lower polynomial degrees.