

FYS-STK4155 Week 36

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

September 1, 2024

Exercise 1

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

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

Solution: The cost function for ridge regression is given by

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

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

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

we have

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

where we defined $\tilde{\lambda} \equiv n\lambda$, renamed $\tilde{\lambda} \rightarrow \lambda$ 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 \mathbf{U} is given by

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

and that the corresponding equation for Ridge is given by

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

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

Solution:

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

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

where the last equality holds due to \mathbf{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.

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

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

Exercise 2