Exercise 7

Fundamentals of Machine Learning 8 ECTS
Heidelberg University WiSe 20/21

Catherine Knobloch, Elias Olofsson, Julia Siegl

February 9, 2021

2 Proof - Ridge Regression - Primal vs. Dual (10 pts)

In the primal formulation, the ridge regression problem takes the following form

$$\widehat{\beta} = \text{argmin}_\beta \, \|\mathbf{y} - X\beta\|_2^2 + \tau\|\beta\|_2^2, \tag{1}$$

where X is an $N \times D$ matrix, β is a D-dimensional vector and $\mathbf{y}$ is an N-dimensional vector. As you saw in the lecture, the optimal $\widehat{\beta}$ is given by

$$\widehat{\beta} = \left(X^T X + \tau\mathbb{1}_D\right)^{-1} X^T\mathbf{y}. \tag{2}$$

Here $\mathbb{1}_D$ is the D -dimensional unit matrix. You also know that the dual formulation of the problem is given by

$$\widehat{\alpha} = \text{argmax}_\alpha \, -\alpha^T \left(XX^T + \tau\mathbb{1}_N\right)\alpha + 2\alpha^T\mathbf{y}, \tag{3}$$

with the solution for $\widehat{\alpha}$

$$\widehat{\alpha} = \left(XX^T + \tau\mathbb{1}_N\right)^{-1}\mathbf{y}. \tag{4}$$

For each feasible α, a corresponding feasible β is given by

$$\beta = X^T\alpha. \tag{5}$$

Prove that the optimal $\widehat{\beta}$ corresponds to the optimal $\widehat{\alpha}$

$$\widehat{\beta} = X^T\widehat{\alpha}. \tag{6}$$

Hint: Prove the following lemma (e.g. using the SVD of X), which will be useful in your derivation

$$\left(X^T X + \tau\mathbb{1}_D\right)^{-1} X^T = X^T\left(XX^T + \tau\mathbb{1}_N\right)^{-1}. \tag{7}$$

To prove the relation in Eq.(6), we use the individually derived expressions for the optimal $\widehat{\beta}$ and $\widehat{\alpha}$ in Eq.(2) and Eq.(4) and substitute these into Eq.(6), obtaining

$$\left(X^T X + \tau\mathbb{1}_D\right)^{-1} X^T y = X^T\left(XX^T + \tau\mathbb{1}_N\right)^{-1} y, \tag{8}$$

where we can simplify the equation by removing the response vector y from each side, yielding

$$\left(X^T X + \tau\mathbb{1}_D\right)^{-1} X^T = X^T\left(XX^T + \tau\mathbb{1}_N\right)^{-1}, \tag{9}$$

which is the same equation as the one given in the hint, Eq.(7). To prove that this equality holds, we left-multiply Eq.(9) with $\left(X^T X + \tau\mathbb{1}_D\right)$ and right-multiply with $\left(XX^T + \tau\mathbb{1}_N\right)$, which then gives

$$X^T\left(XX^T + \tau\mathbb{1}_N\right) = \left(X^T X + \tau\mathbb{1}_D\right)X^T. \tag{10}$$

By finally multiplying the X^T:s into each parenthesis, we can trivially see that

$$\text{LHS} = X^T XX^T + \tau X^T = X^T XX^T + \tau X^T = \text{RHS}, \tag{11}$$

for any training set X and any regularization parameter τ. We have thus shown that the relation in Eq.(6) holds true, given the expressions Eq.(2) and Eq.(4) for the optimal $\widehat{\beta}$ and $\widehat{\alpha}$ of the ridge regression problem. □