# Machine Learning I: Foundations
# Exercise Sheet 8

Prof. Marius Kloft TA:Billy Joe Franks

30.06.2021
Deadline: 29.06.2021

**1) (MANDATORY) 10 Points**

In the lecture we found a closed form solution for linear ridge regression and we incorporated $b$ afterwards by simply changing the dataset slightly. This however means that $b$ is regularized during optimization. What would happen if we introduce $b$ in a different way? Consider linear ridge regression with offset

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - (X^T\mathbf{w} + \hat{\mathbf{b}})\right\|^2 \tag{1}$$

where $\forall i : \hat{b}_i = b$. $\hat{\mathbf{b}}$ simply copies $b$ into each component. Alternatively the norm could be written as a sum incorporating only $b$, as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \left(y_i - (\mathbf{x}_i^T\mathbf{w} + b)\right)^2 \tag{2}$$

(1) and (2) have the same closed-form solution. Find this solution. Thereby choose the version from the above two that you prefer ((1) or (2)).

So we can rewrite (1) as

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - (X^T\mathbf{w} + b\mathbf{1}_{n\times 1})\right\|^2$$

And then we can rewrite it to a more familiar form as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\mathbf{w}^T I\mathbf{w} + C\left\|\mathbf{y} - \left(\begin{bmatrix} X \\ \mathbf{1}_{1\times n}\end{bmatrix}^T \begin{bmatrix}\mathbf{w} \\ b\end{bmatrix}\right)\right\|^2$$

Now we can replace $\tilde{\mathbf{w}} := \begin{bmatrix}\mathbf{w} \\ b\end{bmatrix}$ and $\tilde{X} := \begin{bmatrix} X \\ \mathbf{1}_{1\times n}\end{bmatrix}$ and set

$$\tilde{I} := \begin{bmatrix} I_d & \mathbf{0}_{d\times 1} \\ \mathbf{0}_{1\times d} & 0\end{bmatrix}.$$

Then the problem becomes

$$\min_{\tilde{\mathbf{w}}\in\mathbb{R}^d} \frac{1}{2}\tilde{\mathbf{w}}^T\tilde{I}\tilde{\mathbf{w}} + C\left\|\mathbf{y} - \left(\tilde{X}^T\tilde{\mathbf{w}}\right)\right\|^2$$

Now this is exactly what we expect the optimization problem to look like, except for $\tilde{I}$. However, following the proof from the lecture, we easily observe the solution to be

$$\mathbf{w}_{RRwo} = \left(\tilde{X}\tilde{X}^T + \frac{1}{2C}\tilde{I}\right)^{-1}\tilde{X}y.$$

Interestingly the only difference from the proposition in the lecture, i.e. just appending 1 to every datapoint, is $\tilde{I}$ instead of $I$, as such we see adding a proper offset is incredibly simple. We just have to change a 1 to a 0.

**2)** Consider the kernel ridge regression optimization problem (Lecture 8.3, Slide 9). Let $\alpha^* \in \mathbb{R}^d$ be the vector that minimizes the loss function. Show that:

$$\alpha^* = \left(K + \frac{1}{2C}\mathrm{I}_{n\times n}\right)^{-1} y.$$

$$\begin{aligned}
\frac{\partial \frac{1}{2}\alpha^\top K\alpha + C||y - K\alpha||^2}{\partial \alpha} &= \frac{\partial \frac{1}{2}\alpha^\top K\alpha}{\partial \alpha} + \frac{\partial C||y - K\alpha||^2}{\partial \alpha} = \mathbf{0} \\
\frac{1}{2}2K\alpha + \frac{\partial C(y - K\alpha)^\top (y - K\alpha)}{\partial \alpha} &= \mathbf{0} \\
K\alpha - 2CK^\top (y - K\alpha) &= \mathbf{0} \\
K\alpha - 2CK(y - K\alpha) &= \mathbf{0} \\
\frac{1}{2C}K\alpha - K(y - K\alpha) &= \mathbf{0} \\
\frac{1}{2C}K\alpha - Ky + KK\alpha &= \mathbf{0} \\
\frac{1}{2C}K\alpha + KK\alpha &= Ky \\
K\left(\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha\right) &= Ky \\
K^{-1}K\left(\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha\right) &= K^{-1}Ky \\
\frac{1}{2C}\mathrm{I}_n\alpha + K\alpha &= y \\
\left(\frac{1}{2C}\mathrm{I}_n + K\right)\alpha &= y \\
\alpha = \left(\frac{1}{2C}\mathrm{I}_n + K\right)^{-1} y &= \alpha^*
\end{aligned}$$

**3)** In the lecture the following solution to ridge regression was stated

$$\mathbf{w}_{RR} = \left(XX^\top + \frac{1}{2C}\mathrm{I}\right)^{-1} Xy.$$

The traditional linear regression has the solution $\mathbf{w}_R = (XX^\top)^{-1}Xy$. The matrix $X \in \mathbb{R}^{n\times d}$ is commonly not invertible. For example, if our problem has more features than entries the traditional linear regression is not defined since $(XX^\top)$ is singular. Ridge regression can solve this problem by adding $\frac{1}{2C}\mathrm{I}$.

a) For which values of $C$ is $\left(XX^\top + \frac{1}{2C}\mathrm{I}\right)$ singular, thus having no solution? (Tip: consider the eingenvalues of $XX^\top$)

Note that $(XX^T)^T = (X^T)^T X^T = XX^T$. Therefore $XX^T$ is symmetric and diagonalizable. Let $\mathbf{v}$ be an eigenvector of $XX^T$ and $\lambda$ be its respective eigenvalue. Then we have:

$$\begin{aligned}
\left(XX^T + \frac{1}{2C}\mathrm{I}\right)\mathbf{v} &= XX^T\mathbf{v} + \frac{1}{2C}\mathrm{I}\mathbf{v} \\
&= \lambda\mathbf{v} + \frac{1}{2C}\mathbf{v} \\
&= \left(\lambda + \frac{1}{2C}\right)\mathbf{v}
\end{aligned}$$

Therefore, $\mathbf{v}$ is also eigenvector of $\left(XX^T + \frac{1}{2C}\mathrm{I}\right)$ with eigenvalue $\left(\lambda + \frac{1}{2C}\right)$. A singular matrix has at least one eigenvalue equal to 0 thus:

$$\begin{aligned}
\lambda + \frac{1}{2C} &= 0 \\
2C\lambda + 1 &= 0 \\
2C\lambda &= -1 \\
C &= -\frac{1}{2\lambda}
\end{aligned}$$

b) Prove that $XX^T$ is positive semi-definite.

Consider for any $\mathbf{v} \in \mathbb{R}^d$

$$\mathbf{v}^T XX^T\mathbf{v} = \left\|X^T\mathbf{v}\right\|^2 \geq 0$$

c) Prove that, for proper choices of $C$, $\left(XX^\top + \frac{1}{2C}\mathrm{I}\right)$ is always invertible.

In part a) we found for $C = -\frac{1}{2\lambda}$ the matrix in question will be singular. However proper choices for $C$ are positive, i.e. $C > 0$. These two facts can only be rectified if $\lambda < 0$, however $XX^T$ is PSD, so $\lambda \geq 0$, thus the matrix in question will be invertible.

**4)** Solve programming task 8.