ML: Algo & Theory
SS 18

1	2	3	4	$\sum$

Florence Lopez (3878792),
florence.lopez@student.uni-tuebingen.de
Jennifer Them (3837649),
jennifer.them@student.uni-tuebingen.de

Assignment 3

(Abgabe am 08. Mai 2018)

Exercise 1

a.) Compute the derivative $\frac{\delta f}{\delta X}$, where $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$:

$$f(X) = a^T X = \begin{bmatrix} 2 & -1 & 5 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 2x_1 - x_2 + 5x_3.$$

$\frac{\delta f}{\delta X} f(X) = \begin{bmatrix} \frac{\delta f}{\delta x_1} & \frac{\delta f}{\delta x_2} & \frac{\delta f}{\delta x_3} \end{bmatrix}$ with $\frac{\delta f}{\delta x_1} = 2$, $\frac{\delta f}{\delta x_2} = -1$, $\frac{\delta f}{\delta x_3} = 5$.
This leads to $\frac{\delta f}{\delta X} = \begin{bmatrix} 2 & -1 & 5 \end{bmatrix} = a^T$.

b.) Compute the derivative $\frac{\delta f}{\delta X}$, where $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$:

$$f(X) = X^T A X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 5 & 3 \end{pmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{pmatrix} x_1 + 2x_2 \\ 5x_1 + 3x_2 \end{pmatrix} = x_1^2 + 3x_2^2 + 7x_1x_2.$$

$$\frac{\delta f}{\delta X} f(X) = \begin{bmatrix} \frac{\delta f}{\delta x_1} & \frac{\delta f}{\delta x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 + 7x_2 & 7x_1 + 6x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \cdot \begin{pmatrix} 2 & 7 \\ 7 & 6 \end{pmatrix} = x^T \cdot (A + A^T).$$

Exercise 2

a.) We have to show that $X^t X$ is positive semidefinite and that $\operatorname{rank}(X^t X) = \operatorname{rank}(X)$.

1. A matrix X is positive semidefinite, if for all $v \in \mathbb{R}^n$ the following holds: $v^t X v \geq 0$. In this case we will need $v^t X^t X v \geq 0$. Since $v^t X^t = Xv$, we will have to prove that $XvXv \geq 0$. $XvXv \geq 0 \Leftrightarrow \langle Xv, Xv \rangle \geq 0$. The multiplication of $v^t X^t$ gives us an $1xn$ vector and the multiplication of Xv gives us an $nx1$ vector. Therefore we will have a single value at the end of the whole multiplication. The scalar product of $\langle Xv, Xv \rangle$ equals $||Xv||^2$, which will always be ≥ 0, since all values of Xv multiplied with Xv will be positive, even if the single entries of Xv are negative. Therefore $X^t X$ is a positive semidefinite matrix.
2. $\operatorname{rank}(X^t X) = \operatorname{rank}(X)$. This relation is called the rank-nullity theorem.

b.) We have to show that $X(\omega^* - \omega_0) = 0$ with $\omega_0 = (X^tX)^+X^tY$. To show this, we have to show that $\omega^* = \omega_0$, because this would be the solution to $X(\omega^* - \omega_0) = 0$. Since ω^* is the optimal solution for the regression problem, we have to show that $\min_\omega ||Y - X\omega||^2 = (X^tX)^+X^tY$. To show this, we will compute the derivative of $||Y - X\omega||^2$ in the following steps:

1. $||Y - X\omega||^2$ can be represented in matrix notation: $(Y - X\omega)^t \cdot (Y - X\omega) \Leftrightarrow (Y^t - X^t\omega^t) \cdot (Y - X\omega) \Leftrightarrow Y^tY - \omega^tX^tY - Y^tX\omega + \omega^tX^tX\omega \Leftrightarrow Y^tY - 2\omega^tX^tY + \omega^tX^tX\omega$. The last step could take place, because of the following equation: $\omega^tX^tY = (\omega^tX^tY)^t = Y^tX\omega$.
2. To find the optimal solution, we have to derivate the above equation: $\frac{\delta}{\delta\omega}Y^tY - 2\omega^tX^tY + \omega^tX^tX\omega$. We will split up this equation in 2 parts and derive them.
3. First part: $\frac{\delta}{\delta\omega}\omega^tX^tY = \frac{\delta}{\delta\omega}Y^tX\omega = Y^tX = X^tY$.
4. Second part: $\frac{\delta}{\delta\omega}\omega^tX^tX\omega = \frac{\delta}{\delta\omega}\omega^tX^t + \frac{\delta}{\delta\omega}X\omega = X^tX\omega + XX^t\omega^t = 2X^tX\omega$, because $XX^t\omega^t = X^tX\omega$.
5. If we put all parts together, we get the overall derivation and set it to zero: $\frac{\delta}{\delta\omega}Y^tY - 2\omega^tX^tY + \omega^tX^tX\omega = 0 \Leftrightarrow -2X^tY + 2X^tX\omega = 0 \Leftrightarrow 2X^tX\omega = 2X^tY \Leftrightarrow X^tX\omega = X^tY \Leftrightarrow \omega = (X^tX)^+X^tY$, which is the solution for ω_0.

Therefore one can say that $\omega^* = \omega_0 = (X^tX)^+X^tY$, which leads to $X(\omega^* - \omega_0) = 0$.

c.) -

Exercise 3

a.) Inserting all 4 X, Y-pairs we get the following LGS with 4 equations and $\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}$:

$$\begin{array}{rr}
\text{I)} & \omega_1 + \omega_2 = 3 \\
\text{II)} & \omega_1 - \omega_2 + 2\omega_3 = 1 \\
\text{III)} & 2\omega_1 + 3\omega_2 - \omega_3 = 7 \\
\text{IV)} & -\omega_1 + 2\omega_2 - 3\omega_3 = 0 \\
\hline
\text{I)} & 3 - \omega_1 = \omega_2 \\
\omega_2 \text{ in II)} & \omega_1 - (3 - \omega_1) + 2\omega_3 = 1 \\
\Leftrightarrow & 2\omega_1 + 2\omega_3 = 4 \\
\Leftrightarrow & 2 - \omega_1 = \omega_3 \\
\hline
\omega_2, \omega_3 \text{ in III)} & 2\omega_1 + 3(3 - \omega_1) - (2 - \omega_1) = 7 \\
\Leftrightarrow & 2\omega_1 + 9 - 3\omega_1 - 2\omega_1 = 7 \\
\Leftrightarrow & 2\omega_1 - 3\omega_1 + \omega_1 = 0 \\
\Leftrightarrow & 0 = 0
\end{array}$$

This means that ω_1 can be chosen freely. We chose $\omega_1 = 1$, which leads to $\omega_2 = 2$ and $\omega_3 = 1$ and therefore we have $\omega = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$ as first optimal solution. We can obtain other optimal solutions, since we can chose ω_1 freely. We also chose the $\omega_1 = 2$, which led to $\omega_2 = 1$ and $\omega_3 = 0$. The last optimal solution we obtained was with $\omega_1 = 3$, which led to $\omega_2 = 0$ and $\omega_3 = -1$. So we obtain 3 different optimal solutions:

$\omega^1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \omega^2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$ and $\omega^3 = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}$.

b.) Predicting $\hat{Y}$ with $X_5 = (0, 2, -2)$:

- ω^1: $\hat{Y} = X_5 \cdot \omega^1 = (0, 2, -2) \cdot \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = 2.$
- ω^2: $\hat{Y} = X_5 \cdot \omega^2 = (0, 2, -2) \cdot \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = 4.$
- ω^2: $\hat{Y} = X_5 \cdot \omega^3 = (0, 2, -2) \cdot \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} = 2$

Predicting $\hat{Y}$ with $X_6 = (1, 0, 0)$:

- ω^1: $\hat{Y} = X_6 \cdot \omega^1 = (1, 0, 0) \cdot \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = 0.$
- ω^2: $\hat{Y} = X_6 \cdot \omega^2 = (1, 0, 0) \cdot \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} = 2.$
- ω^2: $\hat{Y} = X_6 \cdot \omega^3 = (1, 0, 0) \cdot \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} = 3$

The predictions do not match, because we found the optimal solutions based on the freely chosen ω_1, therefore the solutions are not the same for the 3 different optimal solutions found in (a).

c.) Inserting all 4 X, Y-pairs we get the following LGS with 4 equations and $\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix}$:

$$
\begin{aligned}
\text{I)} \quad & 2\omega_1^2 + 2\omega_2^2 + \omega_3^2 = 3 \\
\text{II)} \quad & 2\omega_1^2 + 3\omega_3^2 = 1 \\
\text{III)} \quad & 3\omega_1^2 + 4\omega_2^2 = 7 \\
\text{IV)} \quad & 3\omega_2^2 - 2\omega_3^2 = 0 \\
\hline
\text{II)} \quad & 2\omega_1^2 + 3\omega_3^2 = 1 \\
\Leftrightarrow \quad & \frac{1 - 2\omega_1^2}{3} = \omega_3^2 \\
\omega_3^2 \text{ in IV)} \quad & 3\omega_2^2 - 2 \cdot (\frac{1 - 2\omega_1^2}{3}) = 0 \\
\Leftrightarrow \quad & \frac{2}{3} + \frac{4}{3} \cdot \omega_1^2 = 3\omega_2^2 \\
\Leftrightarrow \quad & \frac{2}{9} + \frac{4}{9} \cdot \omega_1^2 = \omega_2^2 \\
\hline
\omega_2^2, \omega_3^2 \text{ in I)} \quad & 2\omega_1^2 + 2 \cdot (\frac{2}{9} + \frac{4}{9} \cdot \omega_1^2) + \frac{1 - 2\omega_1^2}{3} = 3 \\
\Leftrightarrow \quad & \frac{20}{9}\omega_1^2 = \frac{20}{9} \\
\Leftrightarrow \quad & \omega_1^2 = 1 \\
\Leftrightarrow \quad & \omega_1 = 1
\end{aligned}
$$

This leads to $\omega_2 = \sqrt{\frac{1}{3}} \approx 0.58$ and $\omega_3 = \sqrt{\frac{-1}{3}} = \frac{i}{\sqrt{3}}$, which leads to $\omega = \begin{bmatrix} 1 \\ 0.58 \\ \frac{i}{\sqrt{3}} \end{bmatrix}$ as an optimal solution for the ridge regression problem.

Exercise 4

a.) see code in jupyter notebook.

b.) see code in jupyter notebook.

c.) see code in jupyter notebook. The MSE is the highest for $\lambda = 10$. The plots can be seen in the jupyter notebook.

d.) see code in jupyter notebook. As one can see in the plots the linear regression with $\lambda = 0$ gives us the same prediction as the ridge regression with $\lambda = 0.001$.

e.) see code in jupyter notebook. As one can see the MSE is approximately the same for all different λ and even for the linear regression. The best λ could be $\lambda = 0.001$, since it produces a marginally smaller error than the other ones, but it seems like they all lead to the same result in the end.

f.) -