Elements of Machine Learning
Exercise Sheet 3
Winter Term 2023/2024

William LaCroix - wila00001@stud.uni-saarland.de - 7038732
Philipp Hawlitschek - phha00002@stud.unisaarland.de - 7043167

Problem 2 (T, 10 Points) Model selection in Linear Regression

Given is a training set consisting of samples $\mathbf{X} = (x_1, x_2, ..., x_N)^T$ with respective regression targets $y = (y_1, y_2, ..., y_N)^T$ where $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. Alice fits a linear regression model $f(\mathbf{x}_i) = w^T x_i$ to the dataset using the closed-form solution for linear regression (normal equations).

Bob has heard that by transforming the inputs x_i with a vector-valued function Φ, he can fit an alternative function $g(\mathbf{x}_i) = \mathbf{v}^T \Phi(\mathbf{x}_i)$, using the same procedure (solving the normal equations). He decides to use a linear transformation $\Phi(\mathbf{x}_i) = \mathbf{A}^T \mathbf{x}_i$, where $\mathbf{A} \in \mathbb{R}^{D \times D}$ has full rank.

1. [4pts] Show that Bob's procedure will fit the same function as Alice's original procedure, that is $f(x) = g(x)$ for all $x \in \mathbb{R}^D$ (given that w and v minimize the training set error).

Alice fits model $f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i$
Bobbo fits model $g(\mathbf{x}_i) = \mathbf{v}^T \Phi(\mathbf{x}_i) = \mathbf{v}^T \mathbf{A}^T \mathbf{x}_i$
Since $\mathbf{A}$ is full rank, then $\mathbf{A}^{T,-1} \mathbf{A}^T = \mathbf{I}$. So if we let $\mathbf{v}^T = \mathbf{w}^T \mathbf{A}^{T,-1}$, then $g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{A}^{T,-1} \mathbf{A}^T \mathbf{x}_i = \mathbf{w}^T \mathbf{x}_i = f(\mathbf{x}_i)$

2. [3pts] Can Bob's procedure lead to a lower training set error than Alice's if the matrix $\mathbf{A}$ is not of full rank i.e. non-invertible? Explain your answer.

No, Bob's non-invertible procedure cannot lead to lower training error. If Bob's transformation is not full rank, then the transformed features will be of reduced dimensionality, leading to a loss of information/expressiveness. Similar to PCA feature selection, a reduced-rank transformation is a projection onto a lower dimensional space, which necessarily captures less of the variance from higher dimensional data.

3. [3pts] Explain if Alice and Bob will fit the same function or not in the case they perform ridge regression with the same regularization strength λ.

Since ridge regression punishes the model weights, if (as before) we define $\mathbf{v}^T = \mathbf{w}^T \mathbf{A}^{T,-1}$, and both Bob and Alice use the same regularizer λ, then they will end up with the same function, and the choice of linear transformation does not matter.