WI4630 STATISTICAL LEARNING – ASSIGNMENT 1

The answers to this assignment should be submitted in a single PDF file by the stated deadline.

Question 1. Suppose that we are interested in an outcome variable $\boldsymbol{y}$ that depends linearly on two sets of features (input variables) represented below by design matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Suppose that our data is generated according to the following model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1^\star + \boldsymbol{X}_2\boldsymbol{\beta}_2^\star + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ is a random disturbance term, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote matrices consisting of features of dimension $n \times k_1$ and $n \times k_2$ respectively, and the unknown model parameters $\boldsymbol{\beta}_1^\star$ and $\boldsymbol{\beta}_2^\star$ are elements of $\mathbb{R}^{k_1}$ and $\mathbb{R}^{k_2}$ respectively. Moreover, we may assume that all features are deterministic and linearly independent. However, only the features present in $\boldsymbol{X}_1$ are available to us, that is, $\boldsymbol{X}_2$ is not observed. We estimate the model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ by means of ordinary least squares (OLS), resulting in the estimated parameter

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}.$$

(a) Derive the bias and the variance of $\hat{\boldsymbol{\beta}}_1$.

Now consider instead the situation where the outcome variable $\boldsymbol{y}$ only depends on the features $\boldsymbol{X}_1$, i.e., the data is generated according to

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1^\star + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. We now assume that both sets of features $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are available to us and consider the OLS estimation of the following model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

The resulting OLS estimator for the parameters $\boldsymbol{\beta}_1$ is now given by (you do *not* have to show this)

$$\tilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^T\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{M}_2\boldsymbol{y},$$

where $\boldsymbol{M}_2$ is the matrix that describes the projection onto the space orthogonal to the column space of $\boldsymbol{X}_2$ and is given by

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T.$$

(b) Derive the bias and the variance of $\tilde{\boldsymbol{\beta}}_1$. *Hint:* What properties do you know of projection matrices?

(c) Show that the variance (i.e., the covariance matrix) of $\tilde{\boldsymbol{\beta}}_1$ is larger than the variance of the OLS estimator $\hat{\boldsymbol{\beta}}_1$ in the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$. For two positive semi-definite matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, we say that $\boldsymbol{\Sigma}_1$ is greater than $\boldsymbol{\Sigma}_2$ if their difference $\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ is positive semi-definite; this is often denoted as $\boldsymbol{\Sigma}_1 \succ \boldsymbol{\Sigma}_2$.

Hint: You may use that if $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are two positive semi-definite matrices then we have that $\boldsymbol{\Sigma}_1 \succ \boldsymbol{\Sigma}_2$ if and only if $\boldsymbol{\Sigma}_1^{-1} \prec \boldsymbol{\Sigma}_2^{-1}$.

Question 2. In this problem we consider the high-dimensional linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, $\boldsymbol{X}$ denotes a given deterministic matrix of dimension $n \times p$, and $\boldsymbol{\beta} \in \mathbb{R}^p$ with $p = n$. Assume that in our data pre-processing step we orthonormalize the features, such that $\boldsymbol{X}$ is an orthogonal matrix, i.e., $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$. Consider the penalised least squares estimator $\hat{\boldsymbol{\beta}}(\lambda)$ which minimises

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda J(\boldsymbol{\beta}), \tag{1}$$

where $J(\boldsymbol{\beta})$ denotes some penalty term and $\lambda > 0$ is a given tuning parameter.

(a) Let the penalty term be given by $J(\boldsymbol{\beta}) = \sum_{j=1}^n \mathbb{1}_{\beta_j \neq 0}$. Show that the components of the solution $\hat{\boldsymbol{\beta}}^{(0)}(\lambda)$ are given by $\hat{\beta}_j^{(0)}(\lambda) = \bar{\beta}_j \mathbb{1}_{\{|\bar{\boldsymbol{\beta}}_j| > \sqrt{\lambda}\}}$, where $\bar{\boldsymbol{\beta}}$ denotes the OLS estimator.

(b) Let the penalty term be given by $J(\boldsymbol{\beta}) = \sum_{j=1}^n |\beta_j|$. Show that now the solution of the penalised least squares problem is given by

$$\hat{\beta}_j^{(1)}(\lambda) = \begin{cases} \bar{\beta}_j + \lambda/2 & \text{if } \bar{\beta}_j < -\lambda/2 \\ 0 & \text{if } -\lambda/2 \leq \bar{\beta}_j \leq \lambda/2 \\ \bar{\beta}_j - \lambda/2 & \text{if } \bar{\beta}_j > \lambda/2 \end{cases}, \tag{2}$$

where again $\bar{\boldsymbol{\beta}}$ denotes the OLS estimator.

For the remainder of this exercise we do *not* assume that $\boldsymbol{X}$ is orthogonal.

(c) Let the penalty term be given by $J(\boldsymbol{\beta}) = \sum_{j=1}^n \beta_j^2$. Show that $\hat{\boldsymbol{\beta}}_\lambda^{(2)} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}$.

(d) Prove that the variance of $\hat{\boldsymbol{\beta}}_\lambda^{(2)}$ is smaller than that of the classical OLS estimator. Does this contradict the Gauss-Markov Theorem? Explain your reasoning.

(e) Consider the general high-dimensional setting with $p \geq n$. Why is it sensible to impose a penalty term on the least-squares criterion? Also comment on the effect of the penalty term on the predictive performance of the estimator.

Question 3. In this exercise we consider a high-dimensional logistic regression model. Suppose that we are given a training set consisting of independent observations $(\boldsymbol{x}_i, y_i)_{i=1}^n$, with $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$. The logistic regression model postulates

$$\mathbb{P}(y_i = 1|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^T\boldsymbol{x}_i}}, \quad \text{for } i = 1, \cdots, n.$$

We consider the high-dimensional situation, i.e., $p > n$. In the theoretical part of this question we will study two problems that can arise in these high-dimensional settings. In the computational part of this question we will estimated a logistic regression model with real and simulated data and see how these problems are circumvented in practice.

(a) Derive the log-likelihood function $\ell(\boldsymbol{\beta})$ and its gradient $\nabla_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})$.

Let X denote the $n \times p$ matrix consisting of all features, i.e.,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

Note that $\boldsymbol{X}^T$ then denotes the matrix where every every column j denotes the features of observation j, i.e., $\boldsymbol{X}^T = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\right]$.

(b) Prove that if $\operatorname{rank}(\boldsymbol{X}^T) = n$, the maximum likelihood estimator does not exist. *Hint:* Examine the first-order conditions that the maximum likelihood estimator must satisfy.

Another complication that can arise in high-dimensional settings is the so-called separation problem. Suppose that the training set is linearly separated by some hyperplane through the origin, i.e., there exists some vector $\boldsymbol{w}$ such that

$$y_i = \begin{cases} 0 & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i < 0 \\ 1 & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i \geq 0. \end{cases} \tag{3}$$

(c) Prove that the maximum likelihood estimator diverges to infinity. *Hint:* Consider maximum likelihood estimators of the form $\hat{\boldsymbol{\beta}} = c\boldsymbol{w}$.