Assignment #3

Elements of Machine Learning

Saarland University – Winter Semester 2024/25

Rabin Adhikari
7072310
raad00002@stud.uni-saarland.de

Dhimitrios Duka
7059153
dhdu00001@stud.uni-saarland.de

2 Problem 2 (Regularization)

2.1 Lasso and Ridge Regression Equations

The Lasso and the Ridge regressions are used to predict a target Y from X as shown in Equations (1) and (2), respectively. To understand which of the two models is better suited for a task, the mathematical equations for these are written as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid \tag{1}$$

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p} \beta_j^2 \tag{2}$$

2.1.1 Behavior of Coefficients with λ

Question: Discuss how the model coefficients (β_j) change as $\lambda \to 0$ and as $\lambda \to \infty$ in both Equations (1) and (2).

Answer: When $\lambda = 0$, both the equations reduce to RSS, which is the training objective of least squares. So, all the parameters of lasso and ridge regression would be the same as those obtained from least squares when there are no constraints in terms of the magnitude of the parameter ($\lambda = 0$). So when $\lambda \to 0$, the constraints decreases and it would be closer to the least squares solution.

When $\lambda \to \infty$, the second part of the loss dominates, which would be minimum when all parameters (except the intercept) of both the regression is zero ($\beta_{j>0} \to 0$). However, for a large value of λ, some parameters of lasso regression are likely to be exactly zero. While ridge would only have zero for a parameter when $\lambda \to \infty$, that doesn't happen in practice, so, for a large value of λ, the L_2 norm of the parameters (except β_0) is nearly zero, but not exactly zero.

2.1.2 Feature Selection and Regularization Method

Question: If we have significantly more independent features than observations and want to perform feature selection, which type of regularization method should we use? (Hint: L_1 or L_2?) What value of λ should be considered, i.e., small or large?

Answer: If we have significantly more independent features than observations, we would typically want to use L_1 regularization because we would like to get rid of some parameters completely. We can achieve that using a large value of λ for L_1 regularization; this would get rid of some of the irrelevant independent features and perform automatic subset selection depending upon the value of λ provided. However, this is not the case for L_2 regularization, the norm of the parameters corresponding to all the features would have non-zero parameters, however large the value of λ (within infinity).

2.2 Likelihood and Posterior in Lasso Regression

Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a $\mathcal{N}\left(0, \sigma^2\right)$ distribution.

2.2.1 Likelihood for the Data

Question: Write out the likelihood for the data.

Answer: Here, let us assume $f\left(x_i\right) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, which is a constant function and this constant shifts the mean of ϵ_i without changing in variance. Since $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$, this transformation would result $y_i \sim \mathcal{N}\left(f\left(x_i\right), \sigma^2\right)$.

So, the likelihood of data can be written as a conditional probability distribution of y_i given x_i as follows.

$$\begin{aligned} p\left(y_i \mid \beta\right) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - f\left(x_i\right)}{\sigma}\right)^2\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right)^2\right) \end{aligned} \tag{3}$$

2.2.2 Posterior with Double-Exponential Prior

Question: Assume the prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b, written as:

$$p\left(\beta\right) = \frac{1}{2b} \exp\left(-\frac{\mid \beta \mid}{b}\right)$$

Write out the posterior for β in this setting.

Answer: The posterior of β can be written as follows.

$$\begin{aligned} p\left(\beta \mid y\right) &= \frac{p\left(y \mid \beta\right) p\left(\beta\right)}{p\left(y\right)} \\ &\propto p\left(y \mid \beta\right) p\left(\beta\right) \\ &\propto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right)^2\right) \cdot \frac{1}{2b} \exp\left(-\frac{\mid \beta \mid}{b}\right) \\ &\propto \frac{1}{2b \cdot \sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right)^2 - \frac{\mid \beta \mid}{b}\right) \end{aligned} \tag{4}$$

2.2.3 Lasso as the Mode of the Posterior

Question: Show that the lasso estimate is the mode for β under this posterior distribution.

Answer: The mode of a distribution is the value of β, corresponding value of which is the maximum of the posterior. Since the log is a monotonically increasing function, the beta corresponding to the maxima in the posterior is the same as that for the logarithm of the posterior. So, we can write the log posterior as follows.

$$\log p\left(\beta \mid y\right) \propto -\frac{1}{2}\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right)^2 - \frac{\mid \beta \mid}{b} - \log\left(2b \cdot \sigma\sqrt{2\pi}\right) \tag{5}$$

Since the last term is constant, maximizing the above value corresponds to minimizing the following expression.

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} \log p\left(\beta \mid y\right) \\
&= \arg\min_{\beta} \left[\frac{1}{2}\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j}{\sigma}\right)^2 + \frac{\mid \beta \mid}{b}\right] \\
&= \arg\min_{\beta} \frac{1}{2\sigma^2}\left[\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \frac{2\sigma^2 \mid \beta \mid}{b}\right]
\end{aligned} \tag{6}$$

Since $\frac{1}{2\sigma^2}$ is a constant, we can write the above expression as follows.

$$\hat{\beta} = \arg\min_{\beta} \left[\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \frac{2\sigma^2 \mid \beta \mid}{b}\right]$$

The term to minimize is the same as that of Equation (1), with $\lambda = \frac{2\sigma^2}{b}$.