STAT 6650 Homework 1

María Victoria Liendro
Master's in Data Engineering
mzl0201@auburn.edu

Spring 2025

1) I have never worked much with Machine Learning, so I am excited to understand the core concepts in it and to apply it in a project. I feel comfortable withe Bias-Variance concepts and metrics. I feel like I may need more practice with the Statistical Learning Theory, also because I am not sure if we should be able to do all the math or just understanding the concept.

2) Augmenting the matrixes, we get this:

$\mathbf{X_2}=\begin{vmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I_{p\times p}} \end{vmatrix}$ $\mathbf{y_2} = \begin{vmatrix} \mathbf{y} \\ \mathbf{0_p} \end{vmatrix}$

Now, replacing in the Least Square Regression expression for estimates:

$$\hat{\beta} = (\mathbf{X_2}^T\mathbf{X_2})^{-1}\mathbf{X_2}^T\mathbf{y_2}$$

$$\hat{\beta} = ([\mathbf{X}^T, \sqrt{\lambda}\mathbf{I}] \ \ [\mathbf{X}, \sqrt{\lambda}\mathbf{I}])^{-1}[\mathbf{X}^T, \sqrt{\lambda}\mathbf{I}] \ [\mathbf{y}, 0_p]$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Which is the expression for Ridge Regularization estimates.

Coding Section

2) a) Pre-processing the data helps us to prepare our data to analyze. For example, we could avoid data with missing values. Two techniques used for feature scaling are Min-Max Normalization and Standardization. The first, will scale the data for that feature within a range, normally between 0 and 1. The latter, will center the data to zero mean and scale by unit variance, in this exercise this will be used.

2) f) The best value obtained for the validation set was using Lasso the best alpha = 3.856, it has the highest $\mathbf{R^2}$ (0.0737) and lowest RMSE (93.999). Ridge best performance was with alpha = 2212.216, its $\mathbf{R^2}$ (0.069) and RMSE (94.192).

2) g) In Figure 1 we have the coefficients for Ridge and in Figure 2 the coefficients for Lasso, both with a vertical line on the best performance of alpha, explain on 2f.

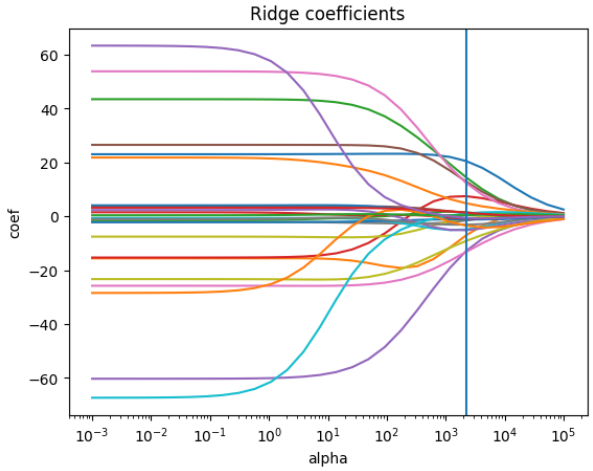

Figure 1

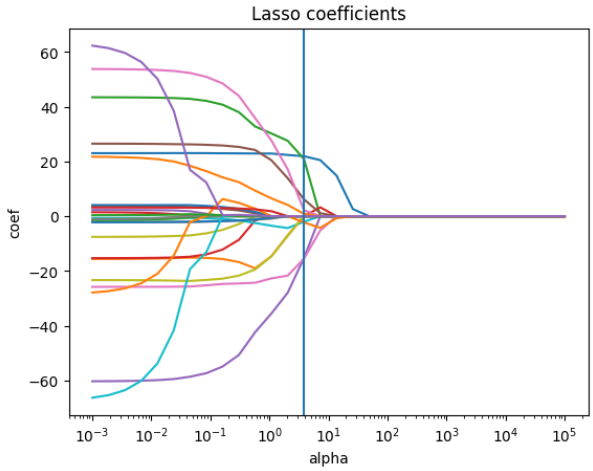

Figure 2

2) h) We can observe that Lasso's coefficients converge to 0 much faster than Ridge coefficients. We can also note that the best R2 needs much smaller values of alpha for Lasso. Finally, small values of alpha make both Ridge and Lasso perform similarly.

3) b) Using the best value from the previous point, 2212, and setting MAX_EPOCHS to 1000 and the BATCH_SIZE to 2048, we got Figure 3, where we can see the different learning rates, when it was bigger tan 0.0001 it would oscillate too much, making it hard to plot. In Figure 4 we set the LEARNING_RATE to 0.000001 and try different batches.

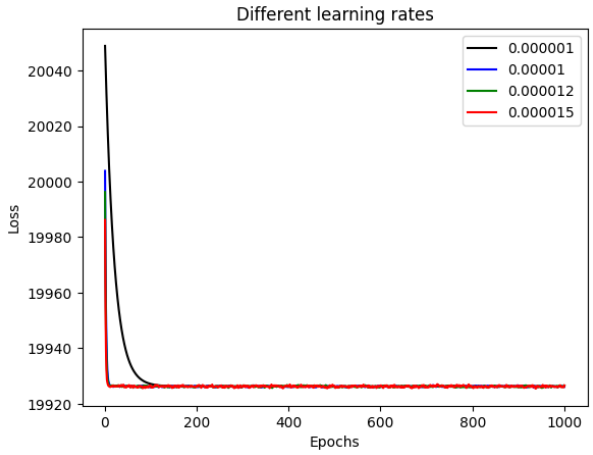

Figure 3

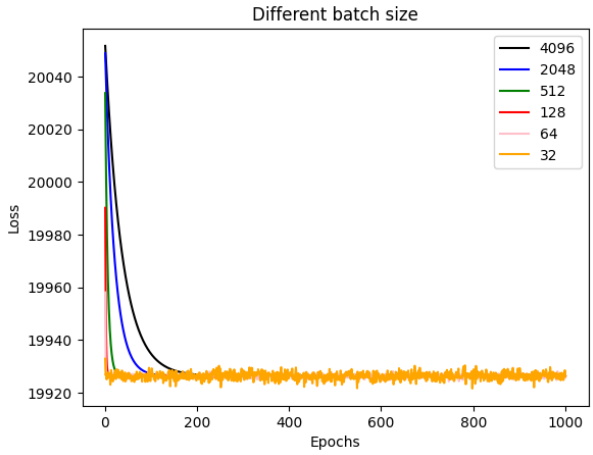

Figure 4