**Research Assignment 2: Regression with Ridge and Lasso Regression**

This paper will explore the concept of regression, its use in supervised learning, and the role that two forms of regularization, Lasso and Ridge regression, have in fitting supervised learning models to evaluate data.

According to Jones et al. (2020, Introduction to Clustering), supervised learning is a machine learning technique that detects meaningful information from data already labeled and structured in some manner. One way that data scientists implement supervised learning is with a regression model, which is a method that predicts a value for a dependent variable by modeling the relationship that dependent variable has with one or more independent variables using a set of training data (Tatsat et al., 2020). The regression model is then used on a separate set of testing data, and its accuracy to the testing data is measured to gauge the model’s fit to data it was not trained on (Tatsat et al., 2020). An example of a supervised learning method using a regression model would be in predicting the price of a home given its square footage. In this example, the model’s training data would be historical home sale prices and square footage information. A regression model would find the slope of the line that best fits the training data and would then be used to provide a prediction for a home that has yet to be sold based on its square footage (Ng, 2003).

One popular form of regression modeling is linear regression, which produces a line that models the relationship between the inputs and output. The line is created by minimizing the sum of the squared residuals, which is the function that squares and then sums up all the distances from the data points to the line (Tatsat et al., 2020). Once the line is created from the training data, the model is then compared to a set of new, unseen testing data, and the sum of the squared residuals is measured between the model and the testing data. If the value of the test data’s sum of squared residuals is large, then the model is overfit to the training data, which is a frequent problem with linear regression models that have many independent variables (Tatsat et al., 2020). Overfitting is another way of saying that the data is too tailored to the specific data included in the training set and is not an optimal model to gauge the relationship between the variables. One technique that addresses the problem of overfitting is regularization regression. Regularization introduces a penalty, lambda, to the parameters of a model to introduce higher prediction accuracy and interpretation (Tatsat et al., 2020).

There are two common ways to regularize regression models. The first form of regularization is known as L1 regularization, or Lasso regression. Lasso regression performs the sum of squared residual function like the linear regression model, but it also adds the absolute value of the regression model’s equation times lambda, which serves as the penalty (Tatsat et al., 2020). This addition of the lambda coefficient allows the model to shrink or eliminate the weight of certain independent variables to zero. This is because as lambda increases, the lasso regression model moves the optimal slope of their relationship closer to zero (Starmer, 2018). Past a certain point, the optimal slope of a lasso regression model can drive the weight of some model parameters to zero while keeping other model parameters greater than zero, effectively removing the influence of those features from the regression model that do not sufficiently contribute to the model’s output (Starmer, 2018). This reduces the complexity of the model, addresses improper fit and allows for the selection of inputs that are more important to providing an accurate model of the training data (Tatsat et al., 2020).

The second form of regularization, L2 regularization, is known as Ridge regression. Ridge regression adds an additional factor, lambda times the model’s equation squared to the sum of squared residuals (Tatsat et al., 2020). This additional factor produces a penalty to the regression model that can prevent overfitting but does so without eliminating the influence of any of the independent variables in the way that Lasso regression does. As the lambda penalty increases, the optimal slope of the linear regression model shrinks closer and closer to zero, but never reaches zero (Starmer, 2020). This keeps all the input variables in the model, while still altering their respective effect on the output value.

So, when is it appropriate to use each regularization regression technique? One major difference between the two techniques is their handling of the subset of data features. Lasso regression can eliminate the influence of features of a dataset, while Ridge regression keeps all input variables in the model (Tatsat et al., 2020). Therefore, if the model has unnecessary features that do not appear to be contributing to the model’s shape, Lasso regression may be a more appropriate technique to simplify the regression model and produce a more accurate result. If each of the features of a model appears to be important in its prediction capacity, Ridge regularization is the more appropriate choice.

In conclusion, regression models are important to supervised learning, and are frequently used to build predictive models given a set of historical training data. Sometimes, these models suffer from becoming overfit to their training data, and regularization techniques must be used to make them less complex and more accurate when working with other sets of data. These regularization techniques share the common trait of using the lambda coefficient to penalize data features, and in doing so, facilitate a more accurate model without making changes to the size or scope of the training data.

**References**

Jones, A., Kruger, C., & Johnston, B. (2020). *The Unsupervised Learning Workshop*. Packt Publishing. https://learning.oreilly.com/library/view/the-unsupervised-learning/9781800200708/B15923\_02\_Final\_SMP.xhtml#\_idParaDest-49

Ng, Andrew. (2003). *CS229 Lecture Notes: Machine Learning (Week 1)*. Stanford University. Retrieved from https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf

Starmer, Josh. (2018). *Regularization Part 2: Lasso (L1) Regression*. YouTube. https://www.youtube.com/watch?v=NGf0voTMlcs

Starmer, Josh. (2020). *Ridge vs. Lasso Regression, Visualized!!!*. YouTube. https://www.youtube.com/watch?v=Xm2C\_gTAl8c

Tatsat, H., Puri, S., Lookabaugh, B. (2020). *Machine Learning and Data Science Blueprints for Finance (Chapter 4: Supervised Learning: Models and Concepts)*. Taiwan: O'Reilly Media.