Adventist University of Central Africa
P.O. Box 2461 Kigali, Rwanda | www.auca.ac.rw | info@auca.ac.rw

Faculty of Information Technology

Department of Big Data Analytics

Assignment III

Course: Data Mining and Information Retrieval

Prepared by:
Godfrey Mawulizo (100901)
Shema Hugor (100763)
Shyaka Kevin (100915)
Gumira Theophile (100920)
Niyonsenga Jean Paul (100888)
Nyirmanzi Jean Claude (100882)

Lecturer: Dr. Pacifique Nizeyimana

June 24, 2025

Question 1: Model Selection Methods

We perform best subset, forward stepwise, and backward stepwise selection on a single dataset, obtaining $p + 1$ models with $0, 1, \ldots, p$ predictors for each method. We address the following questions regarding the models with k predictors.

Background

- **Best Subset Selection**: This method exhaustively searches all possible combinations of predictors for each model size k. For a given k, it finds the model with the smallest Residual Sum of Squares (RSS).
- **Forward Stepwise Selection**: This is a greedy approach. It starts with a null model and iteratively adds the predictor that provides the greatest additional improvement to the model (e.g., largest reduction in RSS) until k predictors are reached. Once a predictor is added, it cannot be removed.
- **Backward Stepwise Selection**: This is also a greedy approach. It starts with the full model (all p predictors) and iteratively removes the predictor that results in the smallest decrease in model fit (e.g., smallest increase in RSS) until k predictors remain. Once a predictor is removed, it cannot be re-added.

Part (a): Smallest Training RSS for k Predictors

Which of the three models with k predictors has the smallest training RSS?

Answer: Best Subset Selection

Explanation: Best subset selection, by definition, examines all possible k-predictor models and chooses the one with the smallest training RSS. Forward and backward stepwise are greedy algorithms, meaning they make locally optimal decisions at each step. This does not guarantee that they will find the globally optimal k-predictor model (the one with the absolute smallest training RSS), which best subset selection does guarantee.

Part (b): Smallest Test RSS for k Predictors

Which of the three models with k predictors has the smallest test RSS?

Answer: It depends, but generally, there is no guaranteed winner.

Explanation: While best subset selection yields the lowest training RSS for a given k, this does not automatically translate to the smallest test RSS. The model with the smallest training RSS might be overfitting the training data, leading to poorer performance on unseen test data. Best subset selection is more prone to overfitting than stepwise methods because it has more flexibility in choosing predictors. More flexible models (like those from best subset) tend to have lower bias but higher variance, while less flexible models (from stepwise) might have higher bias but lower variance. The optimal balance for test RSS depends on the specific dataset and the underlying true relationship between predictors and response. Stepwise methods can sometimes yield better test RSS if they introduce a beneficial amount of regularization by not exploring all possible models and thus preventing severe overfitting. Therefore, you cannot definitively say which method will have the smallest test RSS without actually evaluating them on a test set.

Part (c): True or False Statements

Evaluate the following statements about the predictors in the k-variable and $(k+1)$-variable models:

i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

Answer: True

Explanation: Forward stepwise selection is an additive process. To get the $(k+1)$-variable model, it takes the k-variable model and simply adds one more predictor. It never removes previously selected predictors. So, all predictors from the k-variable model will inherently be present in the $(k+1)$-variable model.

ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

Answer: True

Explanation: Backward stepwise selection is a subtractive process. To go from the full model down to the k-variable model, it removes predictors one by one. To get the $(k+1)$-variable model, it will have removed one fewer predictor than to get the k-variable model. This means the predictors in the k-variable model are precisely the $(k+1)$-variable model minus the one predictor that was removed to reach k. Therefore, the k-variable set is a subset of the $(k+1)$-variable set.

iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

Answer: False

Explanation: There is no guaranteed relationship here. Forward and backward stepwise selection are fundamentally different greedy algorithms. They build their sets of predictors based on different criteria (adding the best vs. removing the worst). A k-variable model from backward stepwise might contain a completely different set of predictors than a $(k+1)$-variable model from forward stepwise.

iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

Answer: False

Explanation: Similar to the previous point, there is no guaranteed subset relationship between the predictor sets chosen by forward and backward stepwise selection for different model sizes. Their greedy approaches can lead to entirely different sets of chosen predictors.

v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

Answer: False

Explanation: Best subset selection does not guarantee this. While it finds the optimal model for each k, the optimal k-predictor model might not be a subset of the optimal $(k+1)$-predictor model. The algorithm exhaustively searches all combinations. It is entirely possible that adding a new predictor to the optimal k-predictor model might not yield the overall

best $(k+1)$-predictor model. The best $(k+1)$-predictor model might require a completely different set of $k+1$ predictors, which could mean dropping one or more predictors from the best k-predictor model to include others.

Question 6: Exploring Ridge and Lasso Objectives

We explore the Ridge (6.12) and Lasso (6.13) objective functions for $p = 1$, plotting them as functions of β_1 and confirming their solutions (6.14 and 6.15).

Part (a): Ridge Regression

Consider the Ridge objective function:

$$(y_1 - \beta_1)^2 + \lambda\beta_1^2$$

with solution:

$$\hat{\beta}_1 = \frac{y_1}{1+\lambda}.$$

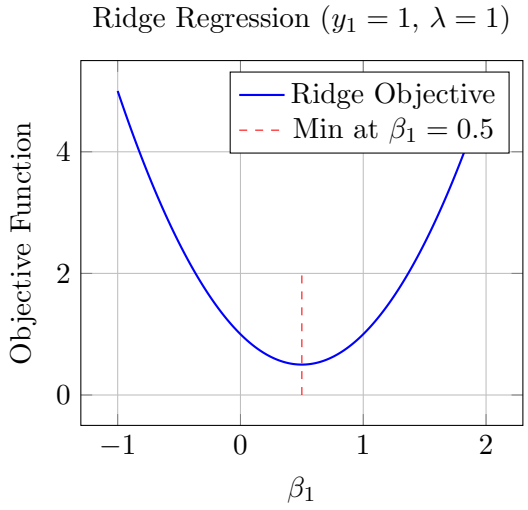
For $y_1 = 1$, $\lambda = 1$, the objective becomes:

$$f(\beta_1) = (1-\beta_1)^2 + \beta_1^2 = 2\beta_1^2 - 2\beta_1 + 1.$$

The analytical minimum is:

$$\hat{\beta}_1 = \frac{1}{1+1} = 0.5.$$

The plot of $f(\beta_1)$ over $\beta_1 \in [-1, 2]$ is shown below, with a vertical line at the minimum $\beta_1 = 0.5$.



The plot is a smooth quadratic function, with the minimum at $\beta_1 = 0.5$, confirming the solution in (6.14).

Part (b): Lasso

Consider the Lasso objective function:

$$(y_1 - \beta_1)^2 + \lambda|\beta_1|$$

with solution:

$$\hat{\beta}_1 = \begin{cases} y_1 - \frac{\lambda}{2} & \text{if } y_1 > \frac{\lambda}{2}, \\ 0 & \text{if } |y_1| \leq \frac{\lambda}{2}, \\ y_1 + \frac{\lambda}{2} & \text{if } y_1 < -\frac{\lambda}{2}. \end{cases}$$

For $\lambda = 1$, we consider two cases:

- **Case 1**: $y_1 = 1$, where $y_1 > \frac{\lambda}{2} = 0.5$. The objective is:

  $$f(\beta_1) = (1 - \beta_1)^2 + |\beta_1|.$$

  The minimum is:

  $$\hat{\beta}_1 = 1 - \frac{1}{2} = 0.5.$$

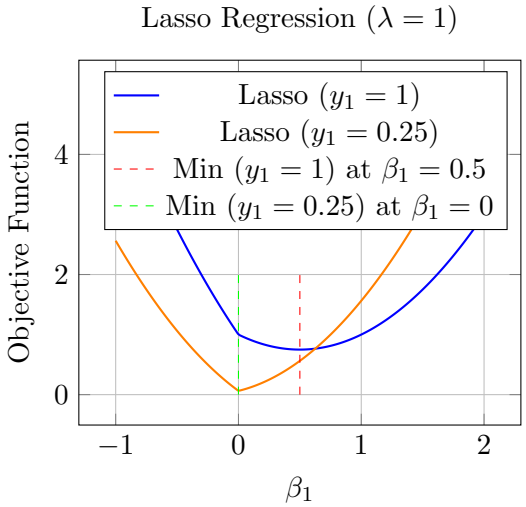
- **Case 2**: $y_1 = 0.25$, where $|y_1| \leq 0.5$. The objective is:

  $$f(\beta_1) = (0.25 - \beta_1)^2 + |\beta_1|.$$

  The minimum is:

  $$\hat{\beta}_1 = 0.$$

The plots of both cases over $\beta_1 \in [-1, 2]$ are shown below, with vertical lines at the minima ($\beta_1 = 0.5$ for $y_1 = 1$, $\beta_1 = 0$ for $y_1 = 0.25$).



The plot shows two functions with kinks at $\beta_1 = 0$ due to the absolute value term. For $y_1 = 1$, the minimum is at $\beta_1 = 0.5$; for $y_1 = 0.25$, the minimum is at $\beta_1 = 0$. These confirm the solution in (6.15).