MXN442, Modern Computing Techniques

Assignment 1-Part 1

Please answer the following questions in a Markdown file, and submit both the file and the knitted PDF document. Your R code must knit, and it is crucial that you submit the knitted PDF file together with the R Markdown file.

Question 1

This question involves the credit cardit data and assess you for Linear Regression, Model Selection, and Regularization, as well as resampling.(14 marks)

```
# Load required libraries
library(tidyverse)
library(glmnet)
library(caret)
library(boot)
library(broom)
library(tidyverse)
```

a) Fit a multiple linear regression model using all variables to predict the credit card balance. Interpret the coefficients of the significant predictors (2 marks).

b) Perform stepwise model selection using AIC. Compare the selected variables with those from the full model and point out the differences.(2 marks)

c) Implement k-fold cross-validation (k=5) for Ridge and Lasso regression. Plot the cross-validation error as a function of for both methods. Explain the differences between Ridge and Lasso in terms of variable selection.(2 marks)

d) Compare the performance of the four models (full linear regression, stepwise selection, Ridge, and Lasso) using Mean Squared Error (MSE) on the test set.(2 marks)

e) Investigate potential interaction effects in the linear model. Add interaction terms for Income:Limit and Age:Education to the full linear model. Interpret the results and discuss whether these interactions improve the model's performance.(2 marks)

f) Use bootstrapping to estimate the uncertainty in the coefficients of the multiple linear regression model from part a. Calculate and report the 95% confidence intervals for each coefficient. Compare these intervals with the standard errors reported by the summary() function. Explain any differences you observe.(4 marks)

Question 2

This exercise involves the same data set and assess you for Dimension Reduction and Model Comparison.(16 marks)

```
library(pls)
```

a) Perform Principal Component Analysis (PCA) on the numerical variables (Income, Limit, Rating, Cards, Age, Education). How many principal components are needed to explain 80% of the variance in the data? Interpret the first two principal components.(3 marks)

b) Implement Principal Component Regression (PCR) using cross-validation to determine the optimal number of components. Plot the cross-validation error as a function of the number of components.(3 marks)

c) Implement Partial Least Squares (PLS) Regression using cross-validation to determine the optimal number of components. Plot the cross-validation error as a function of the number of components. Compare the results with PCR.(2 marks)

d) Compare the performance of PCR and PLS models using Mean Squared Error (MSE) on the test set. Discuss which model performs best and why.(2 marks)

e) For the best performing model among Ridge and Lasso (if applicable), identify the top 5 most important features based on their coefficient magnitudes. Discuss how they are different by visualising their coefficent values, and report the non-zero coefficients. (6 marks)