Assignment 1

Purpose

The purpose of this assignment is to assist you with mastering the topics in Modules 1, 2 and 3.

Learning Outcomes

This assignment aligns to the following module learning outcomes:

- Module 1
  - Describe what it means by predictive modeling. (CLO 2, 3)
  - Explain the role of loss function in supervised algorithms. (CLO 2, 3)
- Module 2
  - Explain why there is a need for regularization in the context of supervised predictive models. (CLO 1, 2, 3)
  - Explain how regularization can improve models' generalization capabilities and to prevent overfitting. (CLO 1, 2, 3)
- Module 3
  - Explain how regularization can be applied to linear models. (CLO 1, 2, 3)
  - Distinguish between various types regularized linear models (i.e., Lasso, Ridge and Elastic net). (CLO 1, 2, 3)
  - Demonstrate how regularized linear models can be implemented and optimized. (CLO 4)
  - Select and filter attributes by constructing Lasso regression models for variable selection. (CLO 4)

Instructions

Part A

Please read the following questions carefully and answer each question.

QA1. What is the main purpose of regularization when training predictive models? (10% of total points)

QA2. What is the role of a loss function in a predictive model? And name two common loss functions for regression models and two common loss functions for classification models. (10% of total points)

QA3. Consider the following scenario. You are building a classification model with many hyper parameters on a relatively small dataset. You will see that the training error is extremely small. Can you fully trust this model? Discuss the reason. (10% of total points)

QA4. What is the role of the lambda parameter in regularized linear models such as Lasso or Ridge regression models? (10% of total points)

Part B

This part of the assignment involves building generalized linear regression models to answer a number of questions. We will use the Carseats dataset that is part of the ISLR package (you need to install and load the library). We may also need the following packages: caret, dplyr and glmnet

Let's start by loading these libraries:

```
library(ISLR)
## Warning: package 'ISLR' was built under R version 4.0.3
library(dplyr)
## Warning: package 'dplyr' was built under R version 4.0.2
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(glmnet)
## Warning: package 'glmnet' was built under R version 4.0.2
## Loading required package: Matrix
## Loaded glmnet 4.0-2
library(caret)
## Warning: package 'caret' was built under R version 4.0.3
## Loading required package: lattice
## Loading required package: ggplot2
## Warning: package 'ggplot2' was built under R version 4.0.2
```

For this assignment, we only need the following attributes: "Sales", "Price", "Advertising", "Population", "Age", "Income" and "Education". The goal of the assignment is to build models to predict the sales of the carseats ("Sales" attribute) using the other attributes.

We can use the dplyr select function to select these attributes.

```
Carseats_Filtered <- Carseats %>% select("Sales", "Price",
"Advertising","Population","Age","Income","Education")
```

QB1. Build a Lasso regression model to predict Sales based on all other attributes ("Price", "Advertising", "Population", "Age", "Income" and "Education"). What is the best value of lambda for such a lasso model? (Hint1: Do not forget to scale your input attributes – you can use the caret preprocess() function to scale and center the data. Hint 2: glment library expect the input attributes to be in the matrix format. You can use the as.matrix() function for converting) (20 % of total points)

QB2. What is the coefficient for the price (normalized) attribute in the best model (i.e. model with the optimal lambda)? (15% of total points)

QB3. How many attributes remain in the model if lambda is set to 0.01? How that number changes if lambda is increased to 0.1? Do you expect more variables to stay in the model (i.e., to have non-zero coefficients) as we increase lambda? (15% of total points)

QB4. Build an elastic-net model with alpha set to 0.6. What is the best value of lambda for such a model? (10% of total points)

General Submission Instructions:

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

You may submit a Microsoft Word (.doc/.docx) document as an attachment using the Canvas Assignment tool, or you may copy and paste your answer into the provided box within the Assignment tool. If you attach a document for your assignment, be sure to include your name in the text of the document and in the name of the document.

- You can only submit once, so make sure you are completely finished before submitting and that you attach the correct word .doc/.docx file.
- Submissions sent by email will NOT be accepted.

Due dates are listed in the Assignment Schedule document.