Statistical Machine Learning: Coursework 1

General Instructions

- Work in groups of at least one and at most two.
- Include your student number and your name on your assignment.
- In your solutions, you should just present your R output or snippets of R code as you deem appropriate. Make sure to present your results clearly.
- You should submit your work as a single PDF file on Blackboard by Thursday the 30^{th} January at 12 noon.

Additional Information

- There are **40 marks** available in total. Your mark out of 10 will be scaled by 2.5 to give the mark out of 100 that you see on Blackboard.

Problem 1: Theory (20 marks)

(i) In this question, we are interested the below linear model which contains a single predictor.

$$Y = X_0\beta_0 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I).$$

Here Y, X_0 and ϵ are vectors of size $(n \times 1)$ and $\beta_0 \in \mathbb{R}$. X_0 is an intercept, which means that $X_0 = [1, ..., 1]^\top$, and β_0 is assumed to be non-random. It follows that $Y = [y_1, ..., y_n]^\top$ has i.i.d. entries with distribution $y_i \sim N(X_0\beta_0, \sigma^2)$. Let the sample mean of the vector Y be denoted $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.

By minimizing the appropriate Residual Sum of Squares (RSS) or Penalized Residual Sum of Squares (PRSS) objective functions, show the following:

(a) The OLS estimator for this example is given by $\hat{\beta}_0^{\text{OLS}} = \bar{y}$ and the OLS hat matrix $H = \frac{1}{n}\mathbb{1}_{(n\times n)}$ where $\mathbb{1}_{(n\times n)}$ is the $(n \times n)$ matrix containing 1 for each of it's elements.

Note: For this question you may not use the identity $\hat{\beta} = (X^\top X)^{-1}X^\top Y$. You must instead minimize the RSS. **(1 mark)**

(b) The ridge estimator is given by $\hat{\beta}^r_{\lambda,0} = \frac{n}{n+\lambda}\bar{y}$. **(1 mark)**

(c) The LASSO estimator is given by $\hat{\beta}^1_{\lambda,0} = \text{sign}(\bar{y})(|\bar{y}| - \frac{\lambda}{2n})^+$, where:

$$\text{sign}(x) := \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases} \quad \text{and} \quad (x)^+ := \max(x, 0).$$

For this question you may assume, without proof, that the LASSO PRSS has a unique minimizer. Hints for (c):

1. *Note that $||\beta||_1 = sign(\beta)\beta$ is differentiable for $\beta \neq 0$.*
2. *When solving for $\hat{\beta}^1_{\lambda,0}$, consider the two cases $\bar{y} > 0$ and $\bar{y} < 0$ separately; what can we say about the sign of $\hat{\beta}^1_{\lambda,0}$ in each case?*

(7 marks)

(d) Given your answers to parts $(a) - (c)$ show that, for $\delta \geq 0$:

$$\mathbb{P}\big[|\hat{\beta}^{OLS}_0| \leq \delta\big] = \mathbb{P}\bigg[\frac{-\delta - \beta_0}{\sigma/\sqrt{n}} \leq Z \leq \frac{\delta - \beta_0}{\sigma/\sqrt{n}}\bigg],$$

$$\mathbb{P}\big[|\hat{\beta}^r_{\lambda,0}| \leq \delta\big] = \mathbb{P}\bigg[\frac{-\frac{n+\lambda}{n}\delta - \beta_0}{\sigma/\sqrt{n}} \leq Z \leq \frac{\frac{n+\lambda}{n}\delta - \beta_0}{\sigma/\sqrt{n}}\bigg],$$

$$\mathbb{P}\big[|\hat{\beta}^1_{\lambda,0}| \leq \delta\big] = \mathbb{P}\bigg[\frac{-\delta - \beta_0}{\sigma/\sqrt{n}} - \frac{\lambda}{2\sigma\sqrt{n}} \leq Z \leq \frac{\delta - \beta_0}{\sigma/\sqrt{n}} + \frac{\lambda}{2\sigma\sqrt{n}}\bigg],$$

where $Z \sim N(0, 1)$ is a standard Gaussian random variable. **(3 marks)**

(e) Part (d) relates the OLS, ridge and LASSO estimators to probability statements of the form $\mathbb{P}[Z \in I]$ for some interval I. By interpreting and comparing the intervals in (d) to one another, describe the different behavior exhibited by the OLS, Ridge and LASSO estimators in this example. How does this behavior relate to what we have seen in class? *Hint: What happens when $\delta = 0$? Which interval is shortest? Your answer should be no more than three sentences.* **(2 marks)**

(ii) In this question, we are interested in a linear model of the form:

$$Y = X\beta + \epsilon,$$

where Y and ϵ are vectors of size $(n \times 1)$, X is a matrix of size $(n \times p)$ and β is a vector of size $(p \times 1)$. We assume that X and β are not random, that $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$ for $i = 1, ..., n$, and that $X^\top X = D$ where D is an invertible diagonal matrix with non-zero diagonal elements $d_1, ..., d_p \in \mathbb{R}$. We shall investigate properties of the below matrix:

$$\Sigma := \text{Var}(\hat{\beta}^\lambda_1) - \text{Var}(\hat{\beta}^{\text{OLS}})$$

(a) Prove that Σ is a diagonal matrix with diagonal entries dependent only upon $d_1, ..., d_p, \sigma^2$ and λ. *Note: For this question you may use the fact that the OLS and ridge estimators are given by* $\hat{\beta}^{OLS} = (X^\top X)^{-1} X^\top Y$ *and* $\hat{\beta}_1^\lambda = (X^\top X + \lambda I)^{-1} X^\top Y$*, respectively.* **(2 marks)**

(b) Given that $\lambda > 0$, show that Σ is negative definite. **(2 marks)**

(c) Does your answer to part (b) contradict the Gauss-Markov theorem? Explain your answer. *Your answer should be no more than two sentences.* **(2 marks)**

Problem 2: Practical (20 marks)

In this question, we shall investigate a number of regression techniques in `R`. To do so, we shall use a synthetic dataset named `pets.csv` which you can find on the Blackboard *"Assessment, submission and feedback"* page.

This dataset studies the impact of pet ownership on work-from-home productivity and consists of the following variables:

- `Pet Care Time (PCT)`: Average number of hours the employee spent per day with the pet (i.e. walking them, feeding them, etc).
- `Employee Wellness (EW)`: A wellness score measuring the employee's overall health and lifestyle.
- `Distraction Level (DL)`: Employee's self-reported level of distraction while working from home, rescaled and demeaned.
- `Pet Type (PT)`: A categorical factor with three levels: `Dog`, `Cat` and `Ranitomeya Amazonica`.
- `Productivity (Prod)`: A work productivity score lying between 0 (low productivity) and 6 (high productivity).
- `Number of Teams Timeouts (NTT)`: Number of times the employees' microsoft teams account timed out due to inactivity over a 3-day period.

We wish to model the impact of having a pet at home on employee productivity.

(i) To begin, we are going to investigate the linear model.

(a) By conducting your own research, describe the predictors included in each of the following `lm` formulas. Do any of the below models give

the same parameter estimates? If so, which?

(1) `Prod ~ DL + PCT`
(2) `Prod ~ 0 + DL + PCT`
(3) `Prod ~ 1 + DL + PCT`
(4) `Prod ~ DL + PCT + DL^2`
(5) `Prod ~ DL + PCT + I(DL^2)`
(6) `Prod ~ PCT + poly(DL,2)`

Note: For this question, you do not need to include any code in your answer. Your answer should consist of at most 5 sentences.
(2 marks)

(b) Using only base `R` commands, write a function `my_lm`. Your function must:

- take as input two data frames X and Y, representing a design matrix and response vector respectively,
- return as output the $\hat{\beta}^{\text{OLS}}$ estimator as a `matrix`,
- check if X includes an intercept column. If X does not include an intercept column, your code must add one to the model.

You may assume for this part of the exercise that X does not contain categorical data. **(4 marks)**

(c) Using the continuous variables in the `pets` dataset, write code demonstrating that your function `my_lm` gives the same output as `lm`. **(1 mark)**

(d) The term "one-hot encoding" refers to the act of converting a categorical predictor Z to a binary representation, as shown below:

$$Z = \begin{bmatrix} A \\ C \\ B \\ A \\ B \\ B \\ C \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In practice, whenever you enter a factor into `lm`, the factor is being one-hot-encoded in this way behind the scenes.

Write your own function which takes in a dataframe X and checks for non-numeric columns of X. If your code encounters a column with non-numeric data it must do the following: (i) remove the column from X, (ii) one-hot encode the column, (iii) add the one-hot encoded data back into X.

Note: In writing this function, please do not rely on existing R functions that automatically perform one-hot encoding. **(3 marks)**

(ii) For each of the following, write `R` code which runs the specified model and plots the model's predicted values. Each model must use the knots $1.5, 3.5$ and 6:

(a) Using `lm`, fit a piecewise linear model with `PCT` as a predictor and `Prod` as a response. **(1 mark)**

(b) Using `lm` and functions from the `splines` package, fit a linear spline model with `PCT` as a predictor and `Prod` as a response. **(1 mark)**

(c) Using `lm` and functions from the `splines` package, fit a cubic spline model with `PCT` as a predictor and `Prod` as a response. **(1 mark)**

(d) Using `lm` and functions from the `splines` package, fit a natural cubic spline model with `PCT` as a predictor and `Prod` as a response. **(1 mark)**

(iii) In part (ii), we used splines to model the nonlinear relationship between `Prod` and `PCT`. In this question, we shall continue to explore this relationship, but using different methods.

(a) Based on what we have seen in class, suggest a different approach which could be used for modeling the relationship between `Prod` and `PCT`. Use an appropriate `R` package to fit your proposed model and make a plot of the fitted values. **(2 marks)**

(b) Modify your solution to (a) to incorporate the `Pet Type` variable in your model. Then, create separate plots of the fitted values for each type of pet. **(2 marks)**

(c) Suppose that instead of `Prod`, we wished to model `NTT` as the response variable. In one sentence, describe how you would modify your solution to part (b) to reflect this change. Provide the updated code (including plots) reflecting this substitution. **(2 marks)**