COMP9417 - Machine Learning
Homework 1: Linear Regression & Friends

Introduction In this homework, we will explore Linear Regression and its regularized counterparts, LASSO and Ridge Regression, in more depth.

Points Allocation There are a total of 25 marks. The available marks are:

- Question 1 a): 4 marks
- Question 1 b): 2 marks
- Question 2 a): 1 mark
- Question 2 b): 1 mark
- Question 2 c): 3 marks
- Question 2 d): 3 marks
- Question 2 e): 2 marks
- Question 2 f): 2 marks
- Question 2 g): 1 mark
- Question 3 a): 2 marks
- Question 3 b): 2 marks
- Question 3 c): 2 marks

What to Submit

- A single PDF file which contains solutions to each question. For each question, provide your solution in the form of text and requested plots. For some questions you will be requested to provide screen shots of code used to generate your answer — only include these when they are explicitly asked for.
- **.py file(s) containing all code you used for the project, which should be provided in a separate .zip file.** This code must match the code provided in the report.
- You may be deducted points for not following these instructions.
- You may be deducted points for poorly presented/formatted work. Please be neat and make your solutions clear. Start each question on a new page if necessary.

- You **cannot** submit a Jupyter notebook; this will receive a mark of zero. This does not stop you from developing your code in a notebook and then copying it into a .py file though, or using a tool such as **nbconvert** or similar.
- We will set up a Moodle forum for questions on this homework. Please read the existing questions before posting new questions. Please do some basic research online before posting questions. Please only post clarification questions. Any questions deemed to be *fishing* for answers will be ignored and/or deleted.
- Please check the Moodle forum for updates to this spec. It is your responsibility to check for announcements about the spec.
- Please complete your homework on your own, do not discuss your solution with other people in the course. General discussion of the problems is fine, but you must write out your own solution and acknowledge if you discussed any of the problems in your submission (including their name and zID).
- As usual, we monitor all online forums such as Chegg, StackExchange, etc. Posting homework questions on these site is equivalent to plagiarism and will result in a case of academic misconduct.

When and Where to Submit

- Due date: Week 3, Sunday **June 20th**, 2021 by **11:55pm**.
- Late submissions will incur a penalty of 20% per day (from the ceiling, i.e., total marks available for the homework) for the first 5 days. For example, if you submit 2 days late, the maximum possible mark is 60% of the available 25 marks.
- Submission must be done through Moodle, no exceptions.

Question 1. Simple Linear Regression

(a) Consider a data set consisting of X values (features) $X_1, \ldots, X_n$ and Y values (responses) $Y_1, \ldots, Y_n$. Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$ be the output of running ordinary least squares (OLS) regression on the data. Now define the transformation:

$$\widetilde{X}_i = c(X_i + d),$$

for each $i = 1, \ldots, n$, where $c \neq 0$ and d are arbitrary real constants. Let $\widetilde{\beta}_0, \widetilde{\beta}_1, \widetilde{\sigma}$ be the output of OLS on the data $\widetilde{X}_1, \ldots, \widetilde{X}_n$ and $Y_1, \ldots, Y_n$. Write equations for $\widetilde{\beta}_0, \widetilde{\beta}_1, \widetilde{\sigma}$ in terms of $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}$ (and in terms of c, d), and be sure to justify your answers. Note that the estimate of error in OLS is taken to be:

$$\hat{\sigma} = \sqrt{\frac{\hat{e}^T \hat{e}}{n - p}},$$

where $\hat{e}$ is the vector of residuals, i.e. with i-the element $\hat{e}_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the i-th prediction made by the model, and p is the number of features (so in this case $p = 2$).

(b) Suppose you have a dataset where X takes only two values while Y can take arbitrary real values. To consider a concrete example, consider a clinical trial where $X_i = 1$ indicates that the i-th patient receives a dose of a particular drug (the treatment), and $X_i = 0$ indicates that they did not, and Y_i is the real-valued outcome for the i-th patient, e.g. blood pressure. Let $\overline{Y}_T$ and $\overline{Y}_P$ indicate the sample mean outcomes for the treatment group and non-treatment (placebo) group, respectively. What will be the value of the OLS coefficients $\hat{\beta}_0, \hat{\beta}_1$ in terms of the group means?

What to submit: For both parts of the question, present your solution neatly - photos of handwritten work or using a tablet to write the answers is fine. Please include all working and circle your final answers.

Question 2. LASSO vs. Ridge Regression

In this problem we will consider the dataset provided in `data.csv`, with response variable Y, and features $X_1, \ldots, X_8$.

(a) Use a pairs plot to study the correlations between the features. In 3-4 sentences, describe what you see and how this might affect a linear regression model. *What to submit: a single plot, some commentary.*

(b) In order for LASSO and Ridge to be run properly, we often rescale the features in the dataset. First, rescale each feature so that it has zero mean, and then rescale it so that $\sum_{i=1}^{n} X_{ij}^2 = n$ where n denotes the total number of observations. *What to submit: print out the sum of squared observations of each of the 8 (transformed) features, i.e. $\sum_i X_{ij}^2$ for $j = 1, \ldots, 8$*

(c) Now we will apply ridge regression to this dataset, recall that ridge regression is defined as the solution to the optimisation:

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}.$$

Run ridge regression with $\lambda = \{0.01, 0.1, 0.5, 1, 1.5, 2, 5, 10, 20, 30, 50, 100, 200, 300\}$. Create a plot with x-axis representing $\log(\lambda)$, and y-axis representing the value of the coefficient for each feature in each of the fitted ridge models. In other words, the plot should describe what happens to each of the coefficients in your model for the different choices of λ. For this problem you are permitted

to use the `sklearn` implementation of Ridge regressiom to run the models and extract the coefficients, and base matplotlib/numpy to create the plots but no other packages are to be used to generate this plot. In a few lines, comment on what you see, in particular what do you observe for features $3, 4, 5$?

What to submit: a single plot, some commentary, a screen shot of the code used for this section. Your plot must have a legend, and you must use the following colors: ['red', 'brown', 'green', 'blue', 'orange', 'pink', 'purple', 'grey'] for the features X1,..,X8 in the plot.

(d) In this part, we will use Leave-One-Out Cross Validation (LOOCV) to find a good value of λ for the ridge problem. Create a fine grid of λ values running from 0 to 50 in increments of 0.1, so the grid would be: $0, 0.1, 0.2, \ldots, 50$. For each data point $i = 1, \ldots, n$, run ridge with each λ value on the dataset with point i removed, find $\hat{\beta}$, then get the leave-one-out error for predicting Y_i. Average the squared error over all n choics of i. Plot the leave-one-out error against λ and find the best λ value. Compare your results to standard Ordinary Least Squares (OLS), does the ridge seem to give better prediction error based on your analysis? Note that for this question you are **not** permitted to use any existing packages that implement cross validation, you must write the code yourself from scratch. You must create the plot yourself from scratch using basic matplotlib functionality.

What to submit: a single plot, some commentary, a screen shot of any code used for this section.

(e) Recall the LASSO problem:

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}.$$

Repeat part (c) for the LASSO. *What to submit: a single plot, some commentary, a screen shot of the code used for this section. You must use the same color scheme as in part (c).*

(f) Repeat the leave-one-out analysis of part (d) for the LASSO and for a grid of λ values $0, 0.1, \ldots, 20$. Note that `sklearn` will throw some warnings for the $\lambda = 0$ case which can be safely ignored for our purposes. *What to submit: a single plot, some commentary, a screen shot of the code used for this section.*

(g) Briefly comment on the differences you observed between the LASSO and Ridge. Which model do you prefer and why? Provide reasonable justification here for full marks. *What to submit: some commentary and potentially plots if your discussion requires it.*

Question 3. Sparse Solutions with LASSO

In this question, we will try to understand why LASSO regression yields sparse solutions. Sparse means that the solution of the LASSO optimisation problem:

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

has most of its entries $\hat{\beta}_j = 0$, which you may have observed empirically in the previous question. To study this from a theoretical perspective, we will consider a somewhat extreme case in which we take the penalty term λ to be very large, and show that the optimal LASSO solution is $\hat{\beta} = 0_p$, the vector of all zeroes. Assume that $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$ and the optimisation is over $\beta \in \mathbb{R}^p$.

(a) Consider the quantity $|\langle Y, X\beta \rangle|$. Show that $|\langle Y, X\beta \rangle| \leq \max_j |X_j^T Y| \sum_j |\beta_j|$, where X_j denotes the j-th column of X.

(b) We will now assume that λ is very large, such that it satisfies:

$$\lambda \geq \max_j |X^T Y|.$$

Using the result of part (a), and the assumption on λ, prove that $\hat{\beta} = 0_p$ is a solution of the LASSO problem.

(c) In the previous part, we showed that $\hat{\beta} = 0_p$ is a minimizer of $\ell(\beta)$. Prove that $\hat{\beta}$ is the unique minimizer of $\ell(\beta)$, i.e. if $\beta \neq 0_p$, then $\ell(\beta) > \ell(0_p)$. **hint: consider the two cases:** $\|X\beta\|_2 = 0$ **and** $\|X\beta\|_2 > 0$

What to submit: For all parts of the question, present your solution neatly - photos of handwritten work or using a tablet to write the answers is fine. Please include all working and circle your final answers. Note that if you cannot do part (a), you are still free to use the result of part (a) to complete parts (b) and (c).