# DASC 5420: Assignment 2

**Due Date: April 5, 2024**

**Total Mark: 50**

---

Make an **R Markdown** file and produce a **pdf** file by solving the following problems. In the title of your file make sure you include your name and student ID number.

1. (10 marks)

   This question involves the use of cross-validation in classification on the *German Credit Risk*. Use **R** to complete these tasks.

   *German Credit Risk* data:

   - Description: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
   - Description: https://www.kaggle.com/datasets/uciml/german-credit

   The data is uploaded to Moodle (https://moodle.tru.ca/mod/folder/view.php?id=2461851). You can download the data from Moodle page. The purpose of this analysis is to implement a machine learning algorithm to predict the credit risk (good or bad) of a consumer in the German market. In the data, *Creditability* variable indicates the credit risk and has two levels 0 (bad) and 1 (good).

   (a) Write the goal of this data analysis. List the inputs and output. Do some exploratory data analysis (EDA) first. Process any necessary data transformation. Explain why you are using that transformation. This could include:

   - Feature scaling such as standardizing or normalizing the data.
   - Selecting or removing certain values (such as outliers or missing values).

   (b) Build a classifier to predict the creditability of a consumer using an appropriate machine learning algorithm.

   (c) Print out your algorithm performance. Choose the right metric(s) for judging the effectiveness of your prediction. Interpret the results in your own words.

   (d) Iterate and improve your algorithm performance. To do this:

   i) Write down the function to implement the k-fold cross-validation (k-fold CV) by yourself for your selected algorithm. Use k=5 or 10. Interpret the results in your own words.

   ii) Write down the function to implement the leave-one-out cross-validation (LOOCV) by yourself for your selected algorithm. Interpret the results in your own words.

   (e) Compare the results in (d) with the results obtained using the cross-validation function in R in **caret** or **boot** package.

2. (5 marks)

This question involves the use of K-Nearest Neighbour (KNN) on the *red wine quality* data set from the UCI repository. Use **R** to complete these tasks. Make sure you included all the **R** codes.

*Red Wine Quality* data:

- Description: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

The data is uploaded to Moodle (https://moodle.tru.ca/mod/folder/view.php?id=2461851). You can download the data from the above data file link or directly from Moodle page.

(a) Write the goal of this data analysis. List the inputs and output. Do some exploratory data analysis (EDA) first. Process any necessary data transformation. Explain why you are using that transformation. This could include:

- Feature scaling such as standardizing or normalizing the data.
- Selecting or removing certain values (such as outliers or missing values).

(b) Build a KNN (K-Nearest Neighbour) classifier to predict wine quality using red wine quality data set. To get a better result, you may need to think to reduce the categories of the outcome.

(c) Apply cross-validation. Which kind of cross-validation do you think is appropriate? Find the optimal value of K? You can use the *train* function under **caret** package in R for this.

(d) Print out your algorithm performance. Choose the right metric(s) for judging the effectiveness of your prediction. You should evaluate the model performance using the Confusion Matrix.

(e) Interpret the results in your own words. What do you learn from the analysis of the wine quality data set?

3. (5 marks)

Consider a classification problem with a large number of inputs, as may arise, for example, in genomic or proteomic applications. For example, consider a simple classifier applied to some two-class data such as a scenario with $N = 50$ samples in two equal-sized classes, and $p = 3000$ quantitative inputs (standard Normal) that are independent of the class labels. The true (test) error rate of any classifier is 48.9%. Now, we have selected 100 inputs from 3000 inputs having the largest correlation with the class labels over all 50 samples and then used a logistics regression classifier, based on just these 100 inputs. Finally, we use 5-fold cross-validation to estimate the unknown tuning parameters and to estimate the prediction error of the final model. And then over 50 simulations, we found the average cross-validation error rate was 2.9% which is far lower than the true error rate of 48.9%.

Is this a correct application of cross-validation? If not, then what has happened? How do you correctly carry out cross-validation in this example to estimate the test set performance of this classifier? Can you justify these scenarios via a small simulated data experiment?

4. (5 marks)

This question involves the use of Bootstrap on simulated data. This is based on the example given in the Bootstrap section of Unit 6 notes.

Suppose that we wish to invest a fixed sum of money in two financial assets (say, Apple, IBM) that yield returns of X and Y, respectively, where X and Y are random quantities. We will invest a fraction $\alpha$ of our money in X, and will invest the remaining $(1-\alpha)$ in Y. We wish to choose $\alpha$ to minimize the total risk, or variance, of our investment. In other words, we want to minimize

$$Var(\alpha X + (1-\alpha)Y).$$

It can easily be shown that the value that minimum value of $\alpha$ is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}, \tag{1}$$

where $\sigma_X^2 = Var(X), \sigma_Y^2 = Var(Y)$, and $\sigma_{XY} = \text{Cov}(X, Y)$.

Perform bootstrap on this example to see the variability of the sample estimator $\hat{\alpha}$ over 1000 simulations (data sets) from the true population and to estimate the standard deviation of $\hat{\alpha}$. Also calculate bootstrap bias estimate and a basic bootstrap confidence interval for $\alpha$. Please ensure that the results are reproducible (i.e, setting a seed in R).

Hints:

- To perform a bootstrap analysis for this example you first write a function to generate an $(X, Y)$ pair as the underlying population. You can generate $(X, Y)$ pair, say $n = 100$ simulated returns for investments X and Y, as follows:
  i. Generate $Z_1$ and $Z_2$ from standard normal distribution,
  ii. Set $X = Z_1$ and then generate $Y = \rho Z_1 + (1-\rho^2)Z_2$, where $\rho = 0.4$ is the correlation coefficient of X and Y.
- You need to write a function to estimate $\alpha$ using the above equation (1). Calculate the true estimate of $\alpha$ using the simulated returns from the previous step.
- Then draws 1000 bootstrap samples from the true population with replacement and calculates an estimate of $\alpha$ from these bootstrap samples.

5. (10 marks)

This question involves the use of Regularized linear regression on the *prostate cancer* data set. In this *prostate cancer* study 9 variables - including age, log weight, log cancer volume, etc. - were measured for 97 patients. We will now construct a model to predict the 9th variable a linear combination of the other 8. A description of this dataset can be found in Hastie, Tibshirani & Freedman's Elements of Statistical Learning book. Use **R** to complete these tasks. Make sure you included all the **R** codes.

*"The data for this example come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate-specific antigen (PSA) and a number of clinical measures,*

*in 97 men who were about to receive radical prostatectomy. The goal is to predict the log of PSA (lpsa) from a number of measurements including log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).”*

The data is uploaded to Moodle (https://moodle.tru.ca/mod/folder/view.php?id=2461851). You can download the data from the above data file link or directly from Moodle page.

(a) **Visualizing the data:** Download the prostate cancer dataset from Moodle and then create a “scatterplot matrix”, i.e. a set of subplots which plots each variable against every other variables, as is done in the lecture slides (or see Hastie et al., page 3, The elements of statistical learning, https://hastie.su.domains/ElemStatLearn/).

(b) **Ridge regression:**

(i) First, split the data into an outcome vector ($\mathbf{y}$) and a matrix of predictor variables ($\mathbf{X}$) respectively:

```
# load data first
y <- prostate[ , 9]
X <- prostate[ , - 9]
```

and then set both variables to have zero mean and standardize the predictor variables to have unit variance.

(ii) Choose the first 65 patients as the training data. The remaining patients will be the test data.

(iii) Write your own code for ridge regression starting from the following skeleton:

```
ridge <- function(X, y, lambda) {
???
return(theta)
}
```

Compute the ridge regression solutions for a range of regularizers ($\lambda$). Plot the values of each in the y-axis against ($\lambda$) in the x-axis. This set of plotted values is known as a regularization path. Your plot should look like Figure 1.

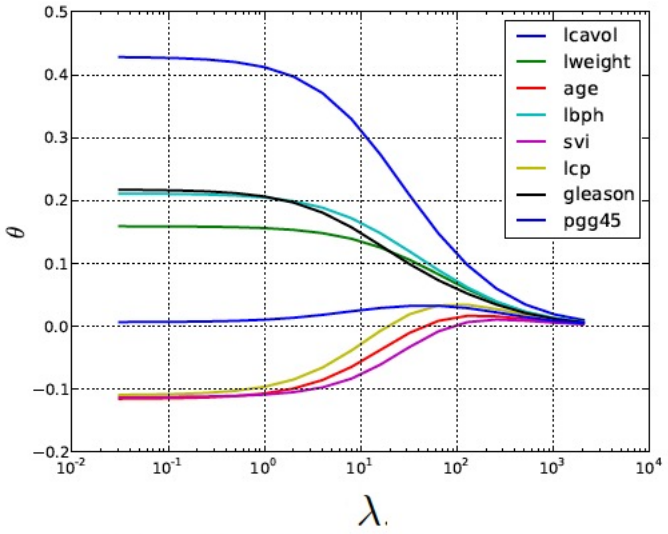


Figure 1: Regularization path for ridge regression.

(iv) For each computed value of $\theta$, compute the train and test error. Remember, you will have to standardize your test data with the same means and standard deviations before you can make a prediction and compute your test error since *ridge regression assumes the predictors are standardized and the response is centred!*
Choose a value of $\lambda$ using cross-validation. What is this value? Show all your intermediate cross- validation steps and the criterion you used to choose $\lambda$. Plot the train and test errors as a function of $\lambda$. Your plot should look like Figure 2.

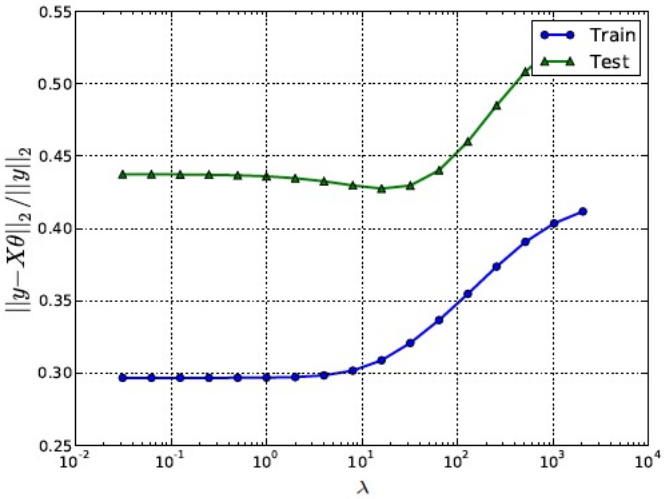


Figure 2: Relative error of the ridge estimator against regularization parameter $\lambda$

(v) For the best $\theta$, plot separately (using subplots) the train and test error as a function of the patient number. That is, for each patient show the actual response and the prediction.

(c) **Lasso regression:** We will now implement the Lasso and try this code out on the *prostate cancer* data.

We know that the most popular approach for fitting lasso and other penalized regression models is to employ coordinate descent algorithms (aka "shooting" method), a less beautiful but simpler and more flexible alternative. The idea behind coordinate descent is, simply, to optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached.

(i) Implement the coordinate descent for solving Lasso. The coordinate descent algorithm is implemented in the R package **glmnet**. You can use **glmnet** or **caret** package in R to solve this part. You should look at "Unit 7: Regularization" lecture slides (data application part) for a better understanding.

(i) Find the solutions and generate the plots from (iii – v) of the previous question, but now using this new Lasso estimate.

(d) Compare the results obtained from Ridge and Lasso regression. What do you learn from the analysis of the *prostate cancer* data?

6. (10 marks)

Consider the multiple linear regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon,$$

where the input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the output vector $\mathbf{y} \in \mathbb{R}^n$. We know the least squares estimate of parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is

$$\hat{\boldsymbol{\theta}}_{LS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

To get the solution of $\hat{\boldsymbol{\theta}}_{LS}$, we need the inversion of $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$. This can lead to problems if the system of equations is poorly conditioned. A solution to this problem is to use the ridge regression estimate which can be obtained by solving the following objective function

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2 + \lambda ||\boldsymbol{\theta}||_2^2,$$

where $\lambda > 0$ is a scalar.

(a) Show that the ridge regression estimate of $\boldsymbol{\theta}$ can be written as $\hat{\boldsymbol{\theta}}_R = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d\right)^{-1} \mathbf{X}^\top \mathbf{y}$, where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

(b) Show that $\hat{\boldsymbol{\theta}}_R$ is not an unbiased estimate of $\boldsymbol{\theta}$.

(c) Find the Bias, Variance and MSE of ridge estimate, $\hat{\boldsymbol{\theta}}_R$.

(d) **Prediction:** Given $x_{new}$, a new vector of input variables from the test data, derive $\hat{Y}_{new}$ and $\text{Var}(\hat{Y}_{new})$ under ridge regression.

7. (5 marks)

(i) Let $y_1, y_2, \ldots, y_n$ is a random sample from a Poisson distribution with mean $\theta$. Let the prior distribution of $\theta$ be Gamma with shape parameter $a > 0$ and rate parameter $b > 0$.

(a) Derive the posterior distribution of $\theta$.

(b) Determine the predictive distribution for a new data point $y_{new}$.

(ii) Let $y_1, y_2, \ldots, y_n$ be a random sample from a Bernoulli distribution with probability of success $\theta$. Let the prior distribution of $\theta$ be beta distribution with shape parameters $a, b > 0$.

(a) Derive the posterior distribution of $\theta$.

(b) Determine the predictive distribution for a new data point $y_{new}$.

Good Luck!