COMP3032 – Machne Learning
Assignment One (20 marks)
Due date: 5:00 pm Wednesday, 27 September 2023

Main objective:

- This assignment is to apply supervised and unsupervised machine learning techniques.
- You will have the opportunity to employ various regression models for predictions and classifications. Additionally, you will be able to utilize the cross-validation approach for model selection and perform PCA to reduce dimensionality.

Task1 (12 marks):

1. Blood pressure dataset *pressure.csv* contains examples of systolic pressures and other features from different persons.
2. Create polynomial regression models, to predict systolic pressure using the SERUMCHOL feature, for degrees vary from 1 to 14. Perform 10-fold cross validation. Calculate its square roots of the mean square errors (RMSE), and the mean RMSE. Display the mean RMSEs for the 14 different degrees. Produce a cross validation error plot using the mean RMSE with 1 to 14 different degrees.
3. Select the best degree, and explains why briefly. Print its intercept and coefficients.
4. Create a multiple linear regression model to predict systolic pressure using all the relevant features. Print its coefficients. Perform 10-fold cross validation. Calculate its square roots of the mean square errors (RMSE), and the mean RMSE, and display the mean RMSE.
5. Build a ridge regression model of the above (i.e. item 4) using $\alpha = 0.1$. Print its coefficients. Perform 10-fold cross validation. Calculate its square roots of the mean square errors (RMSE), and the mean RMSE, and display the mean RMSE.
6. Select the best model of the three, and explains why briefly.

Task2 (8 marks):

1. This task involves MNIST Digit Classification using PCA and Logistic Regression. Load the renowned MNIST ('mnist_784') dataset, which consists of a large collection of handwritten digit images. Your task is to reduce the number of features first, and then build a binary classification model to distinguish between the digit "6" and all other digits (not "6").
2. Prform Principal Component Analysis (PCA) on the feature data to reduce its dimensionality while retaining 88% of the overall explained variance ratio.
3. Split the data into training and testing sets. A common split ratio is 80% training and 20% testing.

4. Create a Logistic Regression model using the reduced feature dataset.
5. Use this model to predict the language for the training dataset and the testing dataset.
6. Print the number of principal components preserved. Print the prediction accuracy (proportion of correct predictions) of your model on the training set. Print the prediction accuracy, the confusion matrix, and the misclassified digits (i.e. wrong predicitons) of your model on the testing set.
7. What do you think of the model generated (good, underfit, overfit)? Briefly explains why

Documentation:

1. You should write a readme file which contains:
   (a) your name and student ID
   (b) How to run your code
   (c) A simple description of your solution logic of the program
   (d) test runs and outputs (You can include the screen shots)
   (e) your answer to the questions (Task 1 Q3 and Q6, Task 2 Q7)
   (f) the limitations if your program does not output the expected result
2. Your code should contain necessary comments to explain what the code is accomplishing and how.

Submission:

ALL related files (such as the readme, python program and data) should be zipped into a single file StudentID.zip and submitted via vUWS. You are required to demonstrate your program during the next scheduled lab session after the deadline. Please note

1. It is students' responsibility to ensure that they can upload successfully their submissions before the deadline.
2. students' responsibility to ensure that their programs are runnable on the schools lab machines.
3. It is students' responsibility to ensure that they keep a copy of their submission.
4. No email submissions will be accepted