# MATLS 4ML3 - Assignment 4

## To be completed individually or in a group of two

## Due: Monday December 13, 2021 @ 11:59pm (Submit to dropbox on Avenue)

## Grading: 12.5% of the course (100 points available + 5 BONUS)

**For this Assignment, use the Jupyter notebook that I provided which load the datasets for both questions and do images resize for Problem 2.**

**Problem 1 [60 marks]**

For this question, we will use the young's modulus dataset to predict Young's Modulus for different alloys using its features. In the Jupyter notebook, I have already loaded the data, removed the Alloy column from the dataset (we will not use it as a feature), shuffled the dataset, and split it into X and y for you.

Your tasks are the followings:

1. Split the dataset into training (80%) and testing (20%) datasets using *train_test_split* with *random_state = 1*, then normalize the dataset (both X and y) between 0 and 1. **[4 marks]**
2. Fit the data using Linear Regression with 5-cross validation to predict the Young's Modulus and report the MSE on each fold. Discuss the results and explain why we need to use cross validation on this training dataset to evaluate the performance of the model? **[8 marks]**

   **HINT: If you use cross_vald_score, set the scoring = *'neg_mean_squared_error'*. This will give you the negative of the MSE, so you will have to take the absolute value of the results to compute the MSE. (Use this hint for the next part too)**

   **For the next two question use 5-cross validation**
3. Fit the data using Ridge (L2) Linear Regression with 5-cross validation to predict the Young's Modulus using regularization values equal to [0, 0.001, 0.01, 0.1,1] and report the mean and standard deviation of the MSE on the 5-fold validation sets for the different lambda values. Explain how to interprets the MSE mean and standard deviation, and which regularization value gives the best results. **[12 marks]**

   **HINT: If you use cross_vald_score, set the scoring = *'neg_mean_squared_error'*. This will give you the negative of the MSE, so you will have to take the absolute value when computing the mean and standard deviation of the results.**
4. Fit a neural network with 5-cross validation to predict the Young's Modulus using dropout layer with dropout rate equal to [0, 0.05, 0.1, 0.25, 0.5] and report the mean and standard deviation of the MSE on the 5-fold validation sets for the different dropout. which dropout rate value gives the best results. **[16 marks].**

**The neural network architectures should be 1 hidden layer with 20 neurons using relu activation function. The output layer activation function should be chosen appropriately. The loss function is MSE. Use Adam Optimizer with 0.001 Learning rate and batch_size equal 16. Train the model using 500 epochs (no early stopping). Finally, initialize the weights using glorot_normal and the bias using zeros. Leave everything else to the default value.**

5. Compute the number of the neural network's parameters in part 4. You need to show the steps of how you calculated the number of parameters (You can use *model.summary()* to confirm that you answer is correct). **[6 marks]**
6. Plot the neural network using both *plot_model* and *ann_viz* functions. **[6 marks]**
7. Using the best hyperparameter (regularization value) obtained from parts 3 and 4, train the Ridge Linear Regression and the ANN models again using all the training dataset and compute the MSE of the training and testing dataset and discuss which model is the better. **[8 marks]**

**Problem 2 [40 marks + 5 BONUS]**

For this question we will use the garbage images dataset classify the images into different classes [cardboard, glass, metal, paper, plastic, trash].

I have provided you with the code that will load the data into X and y, you just need to change the directory based on where you save the data.

Note that running the cells will take some time so just be patient. If you find you are running to memory issues, you will need to reset your kernel and re-run all the cells.

1. Split the dataset into training (80%) and testing (20%), normalize the X matrix, and generate a 6 by 6 subplot to see some of the training images with their labels (36 images) **[4 marks]**
2. Redo the same steps as in part 1 but plot the resized images using the block of the code to resize the image. The idea of resizing the image is to reduce the number of pixels, so our CNN is trained much faster. Compare the same images when plotted with the original number of pixels (part 1) versus with the reduced number of pixels. **[6 marks]**
3. Train CNN to classify the resized images dataset using categorical crossentropy as loss function with adam optimizer that uses 0.0001, epochs = 50, and batch size = 32. The CNN should use 20% of the training dataset as validation set to stop training if the validation accuracy does not improve for 5 epochs (so you should track the loss and the accuracy). Also, make the CNN return the best weight by setting restore_best_weight to be true. Report the loss and accuracy of training and testing at the best epoch, and show the model summary with the number of parameters of the model.

   The CNN details are shown below:

   Input layer, CONV layer with 32 3x3 filters, max pooling layer with 2x2 filter, dropout layer with 0.2 dropout rate, CONV layer with 64 3x3 filters, max pooling layer with 2x2 filter, dropout layer with 0.2 dropout rate, CONV layer with 128 3x3 filters, max pooling layer with 2x2 filter, dropout layer with 0.2 dropout rate, CONV layer with 128 3x3 filters, max pooling layer with 2x2 filter, dropout layer with 0.2 dropout rate, flatten layer, dense layer with 512 neurons, dropout layer with 0.2 dropout rate, output layer. The activation function for all the CONV and dense layers are relu, and the activation function of the output layer should be chosen appropriately. **[20 marks]**

4. Plot the training and validation loss function and accuracy as a function of number of epochs and discuss the results. **[5 marks]**
5. Generate 6 by 6 subplot on the testing data and report the predicted and actual class. **[5 marks]**
6. Compute the confusion matrix of the testing dataset and discuss the results **[5 marks BONUS]**