Assignment 4 PLS, ANNs

To be completed in groups of UP TO THREE (3)
Due: March 27, 2023 @11:59pm.
Grading: 5% of course grade (45 points available)

Submission Instructions

Please submit this assignment *electronically* before the due date. Late submissions will **not** be accepted. Submit via the A2L dropbox for the appropriate assignment. Be sure that you have the names and student numbers of all students on the front page of your submission. Submit your answers as a **single .pdf** file including all relevant figures, tables, and math. You may include relevant code embedded in the report, but you **must submit a .zip along with your report** that includes all your code for the assignment.

Please upload your files with the **naming convention** `A0X_macID1_madID2_macID3`, where the first McMaster ID is for the person uploading the submission.

Up to 10 points may be deducted from your submission for sloppy or otherwise unprofessional work. This is rare, but possible. The definition of unprofessional work may include:

- Low-resolution screenshots of figures and tables.
- Giving no context to an answer relating to the task (*i.e.* "See code" or "113.289" with no units, context, or discussion whatsoever).
- Clear changes in author denoted by format changes, blatant writing style changes, or other factors that may deduct from the cohesiveness of the report.
- Failing to provide references for work that is obviously not yours (particularly bad cases will be considered as academic dishonesty).

It's a shorter assignment because Kipling. Don't @ me :P

Problem 1 [30]

In this problem, we would like to get exposed to the automated ANN training tool available in MATLAB.

Previously, Mahir Jalanko made a summary presentation on how to use the ANN toolbox in MATLAB. Review Mahir's Deep Learning Toolbox MATLAB PowerPoint provided in this assignment's companion. We would like to use this toolbox to create a predictive ANN for a coffee roasting process that estimates product quality as a function of roasting temperatures at different roasting stages.

Consider the data set coffee_production.csv in the assignment 4 companion. This data set contains measured process temperatures from an industrial roasting process for over 29,000 batches of coffee. The first 15 columns of data are the roast temperatures (in °F) at different times during a three stage roasting process. The last two columns contain the elevation that the coffee was grown at and a humidity measurement of the typical growing conditions for that batch. The data set quality.csv is the measured quality of each batch of coffee as determined via a variety of metrics such as aroma, colour, and taste. We seek a model that predicts quality as a function of growing conditions and roasting conditions.

YOUR TASK for this problem is to fit an ANN to this data set that accurately predicts coffee quality given a set of process measurements. Explore the ANN tool built into MATLAB and use it to experiment with the number of nodes in your hidden layer, how much data you use for testing and validation, and other design choices. **Keep the number of hidden layers to one**. Your grade will be based on how you describe your thinking and what you tried, if it worked, and why. **Create ANNs with 3, 5, 10, and 20 hidden layer nodes**. Keep the activation functions as their default and use the default solver (Levenberg-Marquardt). To help you along, here are some questions you might want to answer:

- The ANN toolbox reserves some data for "testing" and some for "validation." What is the difference?
- How does increasing the number of hidden nodes in your network affect model performance?
- The toolbox generates error histograms for each model. How do those histograms change as the number of hidden nodes increases?
- One of the default plots available in the ANN tool is called a "Regression" plot (function plotregression). This displays the model output ($\hat{y}$) on the y-axis and the true ("target") value y on the x-axis for the training, testing, and validations sets. What are you looking for on this plot to denote a good model fit? Do you notice anything about this plot as you increase the number of nodes in the hidden layer?
- Are there any strange behaviours in the predicted vs. expected regression plots? Does this correspond to outlier data? You might want to consider filtering the data to see if there are any outliers and try again.

The purpose of this question is for you to explore different ANNs and see how to build them using built-in MATLAB tools. Have fun!

Problem 2 [15]

Consider the dataset `wastewater.xlsx` from the A4 data companion for this problem. This dataset is from real laboratory measurements of 26 different industrial wastewater samples. The dataset contains 26 observations (rows), seven variables and one outcome variable. The 'SNR outcome' is measured for each wastewater sample using a time consuming 'Specific Nitrification Rate' (SNR) test which is directly related to the toxicity of this wastewater sample towards biological organisms. A quick way to determine toxicity of wastewater samples was developed by preserving seven different strains of bacteria and then testing the response of these seven bacteria strains on each wastewater sample, which are then used to predict the 'SNR outcome'. **NOTE: NO KNOWLEDGE OF WASTEWATER TREATMENT IS NEEDED TO SOLVE THIS QUESTION**. The responses are tabulated in columns 2-8 in `wastewater.xlsx`. More information about the dataset is given below:

- Column 1 ('WW-ID') - Wastewater sample ID
- Column 2 to 8 ('A' to 'G') - Response of 7 bacteria strains named 'A' to 'G'
- Column 9 ('SNR outcome') - Measured SNR outcome

1. Divide the dataset into training and testing sets as follows: [**4**]

    - **Training set** – *Observations 1 to 19*
    - **Testing set** – *Observations 20 to 26*

    Develop a PLS model using the training set ONLY with three components and report the values of **w***, **p**, **c** and $\mathbf{R^2}$ for each component.

2. Plot a loadings scatter plot between components 1 and 2 for **w*** and **c** on the same axes. Include appropriate data labels. Make sure to label the axes appropriately. [**3**]

3. Based on the loadings plot, discuss how the scores $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ for the first two components are affected by response from strains A, D & F. Also discuss the relationship between the SNR outcome and the values of $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$. [**4**]

4. Now use the model built above with the training set to **predict** the SNR outcome for the testing set. Show all the calculation steps for wastewater ID '21' to obtain the predicted outcome. Plot the observed values of SNR outcome vs predicted values of SNR outcome for the prediction set. What type of shape of this plot would you expect for a good model? [**5**]

5. Do you think that the model has good predictive ability? Explain using quantitative metrics. Is there a better way to determine the predictive performance of a PLS model on this dataset? Discuss briefly how you would go about testing the predictive performance in a better way and what metrics would you use to determine the predictive ability of the model. [**4**]