Assignment 2

Total Marks: 50 **Due Date: 30/04/2023**

The problem for this assignment is adapted from an earlier real-life problem on predicting performance of new employees based on demographic information and test scores. The variables are self-explanatory, and the last column (performance) is the one that we want to predict.

1.	Which independent variables can be used as inputs for a Neural Network or an SVM and why?	**2**
2.	Convert the inputs identified in Q1 into standard form (zero mean and unit variance). Can all data points be used? If not, then devise a strategy to deal with missing information for the rest of the assignment.	**3**
3.	Perform Kolmogorov-Smirnov test (http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test) to rank the features according to their discrimination ability for classes BP and LP.	**5**
4.	Pick the top two features and plot them in a 2-d space with separate markers for different classes. Are the classes linearly separable using these two features?	**2**
5.	Use Fisher Discriminant Analysis on the classes BP and LP to project the variables identified in Q1 into a 1-d space. Plot the posterior $p(\omega_i\|x)$ against x for the two classes. Plot the boundary that will minimize the error for this two-class problem. Comment on the LDA projection in light of the results of Q3.	**3**
6.	Build a Neural Network to distinguish between the three classes (ignoring MD which stands for Missing Data). Start with 1 hidden layer and 3 hidden neurons and sigmoid activation function. Train it using a training algorithm of your choice on randomly selected 80% of the points. Validate the results on the 20% of the points not used in training. Plot the error on the training set and validation set against # epochs. Comment on these results. You may use the in-built commands from MATLAB or any other machine learning package for this.	**10**
7.	Repeat using different number of hidden nodes and plot minimum error on validation set vs. # hidden nodes. Comment on these results.	**5**
8.	For the optimal number of hidden nodes, interpret the input-to-hidden node weight matrix and its relation to results of Q3 and Q5.	**10**
9.	Repeat step 6 using SVM. Select a Gaussian Kernel, and use a grid search (http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf). Comment on the results, and compare them to Q3, Q5 and Q8.	**10**

Instructions:

Include code:

a. At the end of the report, include an Appendix with clearly marked sections for code for different parts of the assignment.
b. Code can be in MATLAB or Python.
c. Code should be properly indented (http://net.tutsplus.com/tutorials/html-css-techniques/top-15-best-practices-for-writing-super-readable-code/).
d. Each line of the code should have a corresponding comment in your own words to explain why the line is included, and what it does.

Notes:
Answers are to be typed. Any handwritten responses or snapshots or pics won't be considered at all.

All code and data sets (training and validation) must be included in appendices.