EEL6935 – SPRING 2019 – DEEP LEARNING
Homework Assignment 2

In this assignment, you will learn how to train a neural network for binary classification problem on a moderately sized dataset.

<u>Preparation:</u> First you will need to download the "spambase_mod.csv" file included with this homework assignment. The dataset is provided in CSV format. This is a 4601 x 58 dataset where there are 4601 samples (observations) of 57 features (x1, x2, …, x57) and 1 binary class label (y) at the last column. The last column denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

<u>Objective:</u> We want to predict label "y" using features 1 through 57. In order to do that, we will first split dataset into test and train sets. A neural network model will be built by using training data and training labels. After the training, features in the test data will be used to predict the labels of test data.

<u>Cross-validation:</u> **We will not use the same kind of cross-validation we did in the 1st homework.** Instead of folding the dataset, we will simply spare the first 4000 instances as the training dataset and the last 601 instances as the testing dataset. Since the dataset is already shuffled, you don't need to re-shuffle it again. However, normalization is required among the features **(i.e., you need to make sure each column on your dataset is zero mean – unity variance within itself).** The weights of the network should be initialized as **small non-zero random numbers**. There are four hyper parameters of the model such as 1) L2 regularization strength of parameters, 2) learning rate, 3) number of hidden layers and 4) number of corresponding hidden neurons. We will fix the regularization strength at 0.1 and learning rate also at 0.1. The experiments will be repeated only among varying number of hidden layers and neurons with the following parameters:

No of Neurons	5 – 25
No of Layers	1 – 2

That means 4 experiments in total, [5] [5 5] [25] [25 25]**,** will be performed and **accuracy** will be reported on the test set. **The accuracy is defined as the number of correctly classified <u>test instances</u> divided by the total number of test instances.** For example, if you classify 300 samples correctly, your accuracy is 300/601 on this test set. In order to obtain statistically meaningful results, each one of the 4 experiments should be repeated 10 times (remember that you are randomly initializing the network) and the average score should be reported for each repetition.

<u>Goal:</u> You will display a **single figure** that shows the average classification accuracies for each of the 4 models. See Fig. 1 for an example of this. (Note: This is an example, your results may vary!)

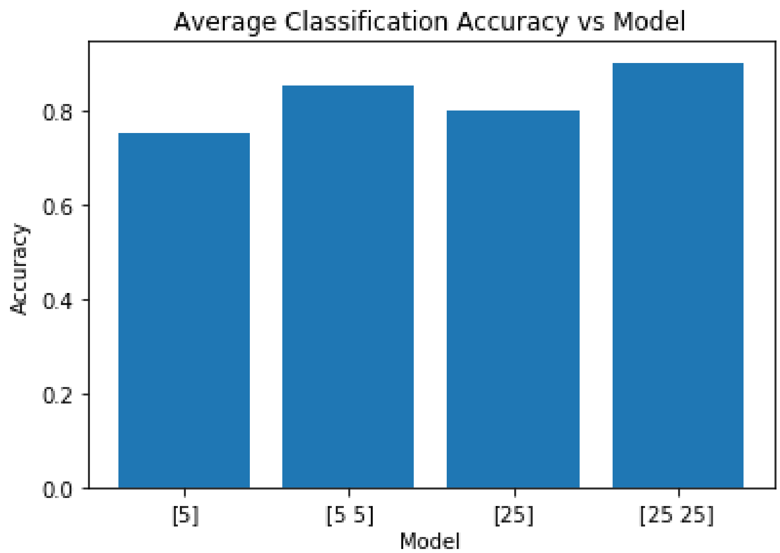

Figure 1: An example showing the formatting of the required figure. Your results will vary.

<u>Deliverables:</u> This assignment has two deliverables:

1) Your single Jupyter notebook, prepared as explained in the explanation video uploaded earlier this semester. Please read and follow all instructions, and ask questions if you are not certain about anything.

<u>IMPORTANT INFORMATION:</u>

Remember what we discussed in class about the HW assignments:

- ⇨ All code should be contained within a **single** Jupyter Notebook, do not submit multiple files.
- ⇨ Your code must run **without modification** and produce all, **and only**, the required results and figures asked for in the **Goal** section.
- ⇨ Dataset files should not be modified. The name of the files should not be changed.
- ⇨ Your code should read dataset files without an absolute path, and your notebook **MUST** be able to run if they are in the same folder.
- ⇨ Your Jupyter notebook should be saved as follows: last_name_first_initial_hwX.ipynb
    - o **For example, for me it would be: uysal_i_hw2.ipybn**