National University
of Computer & Emerging Sciences, Chiniot-Faisalabad Campus

Dr. M Gufran Khan
Dr. Hashim Yasin

CS4104 – Applied Machine Learning
Assignment No. 2

Instructions:

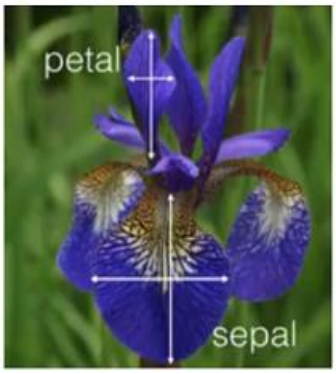
1. Submit your assignment in **soft** as well as in **hard form (Report)** within due date and time. Soft form does **not** mean the photos of the hardcopy (You need to submit source files). Late submission will result in deduction of marks.
2. Report should necessarily contain the discussion, comments and conclusion about the solution. Without report, you will not get full marks.
3. Mention your name and roll number clearly on your document.
4. Name your zip or other folder/file that you want to submit according to the following format: **AML_A2_RollNo_FirstName**
5. Try to solve the each task of the assignment by your own.
6. No excuse or resubmission is permissible.
7. There is no any restriction of the language for the programming tasks.

Question No. 1: Decision Tree & Random Forest

---

a) What is motivation behind ensemble methods? (3p)
b) How Random Forest training and inference works? Give pseudo code. (7p)

c) Iris dataset has 50 samples for each of three different species of Iris flower (total number of samples is 150). For each data sample, you have sepal length, sepal width, petal length and petal width and a species name (class/label). Figure below shows Iris flower and features in dataset:




```
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7.2.8.4.5.1.3.Iris-versicolor
```

Iris flower: sepal length, sepal width, petal length and width

Perform the following tasks:

- Load the given Iris dataset
- Split it into a training set and a test set
- Preprocess the data
- Build a Decision Tree classifier
- Then train a Random Forest classifier
- How much better does it perform compared to the Decision Tree classifier? (10p)

Question No. 2: Support Vector Machine (SVM)

a) What is a support vector? Derive the objective function of support vector machines (SVM) for linearly separable data. (6p)
b) Differentiate between soft margin and hard margin classifier. (4p)

c) In this question we will be using the popular MNIST dataset, which is a set of 70,000 small images of digits handwritten digits (https://www.kaggle.com/datasets/hojjatk/mnist-dataset). Each image is labeled with the digit it represents. There are 70,000 images, and each image has 784 features. This is because each image is 28 × 28 pixels, and each feature simply represents one pixel's intensity, from 0 (white) to 255 (black). Figure below shows some digits from the MNIST dataset:

Perform the following tasks:

- Download MNIST dataset.
- Train a linear SVM classifier on the MNIST. Since SVM classifiers are binary classifiers (you will need to use one-versus-all to classify all 10 digits), and also report its training accuracy.
- Scale (standardize) the data first, retrain an linear SVM classifier and also report its training accuracy.
- Now retrain the classifier with a non-linear SVM using Radial Basis Function (RBF) Kernel (aka Gaussian Kernel), and also report its training accuracy.
- Now make predictions on the above models for each class on test data and report your results. (10p)