Due Date: See Webcampus
How to submit: Webcampus

General Guidelines:

- Please prepare a **typed** report that describes what you did. The report should be as concise as possible while providing all necessary information required to replicate your plots.
- For each problem, please provide, at the end of your report, a commented version of your python code files. **Python Notebook files are preferred. You may put the codes for all the problems in a SINGLE ipynb file with necessary texts to separate each problem.**

P2-1. Decision Tree

Use the Iris dataset embedded in scikit-learn:

```
from sklearn import datasets
iris = datasets.load_iris()
```

(a) Develop a decision tree based classifier to classify the 3 different types of Iris (Setosa, Versicolour, and Virginica). To build your classifier, you can use:

sklearn.tree.DecisionTreeClassifier

(See https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)

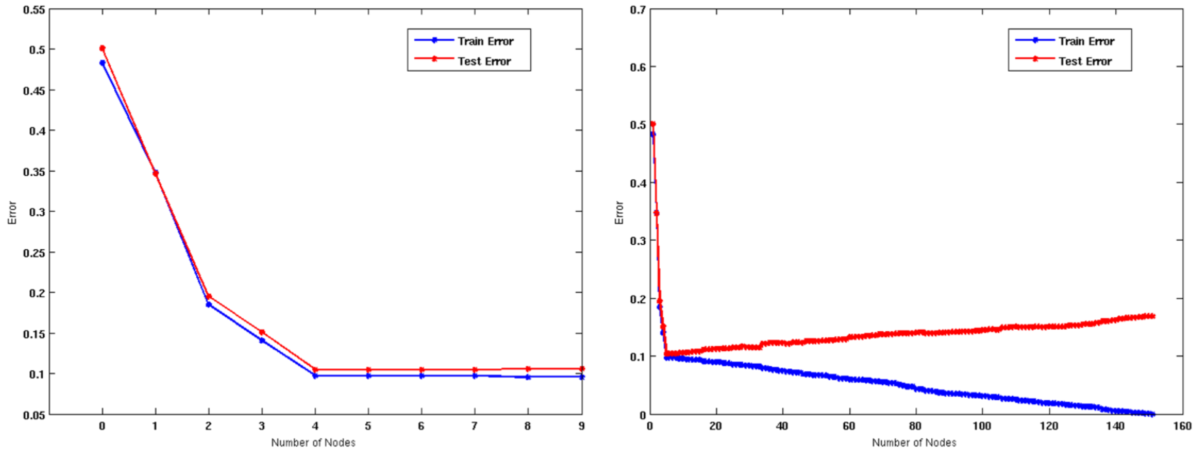
Instructions: When training the decision tree, use 5-fold cross validation. To make your training dataset balanced, pick 10 samples from each Iris type as test set so that the training dataset will contain 40 samples from each Iris type. In total, you will have 120 samples in the training dataset and 30 samples in the test dataset.

Refer to sklearn.model_selection.StratifiedKFold (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html)

(b) Optimize the hyperparameters (e.g., max_depth, min_samples_leaf, etc.) of your decision tree to maximize the classification accuracy. Once the hyperparameters are determined, show the confusion matrix of one decision tree with the accuracy close to the average accuracy in the 5-fold cross validation. Plot this decision tree.

P2-2. Model Overfitting

Reproduce the figures in slide 61 in Chapter 3, i.e,



(a) Generate the dataset as in slide 56 in Chapter 3:

- Class 1: Generate 5000 instances following a Gaussian distribution centered at (10,10) with covariance $\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$. And generate 200 instances following a uniform distribution in a plane defined by the corner points {(0,0), (0,20), (20,0), (20,20)}.
- Class 2: Generate 5200 instances following a uniform distribution in the same plane.

Plot your dataset.

(b) Randomly select 10% of the data as test dataset and the remaining 90% of the data as training dataset. Train decision trees by increasing the number of nodes of the decision trees until the training error becomes 0. Plot the training errors and the testing errors under different numbers of nodes and explain the model underfitting and model overfitting.

P2-3. Text Documents Classification

Use the 20 newsgroups dataset embedded in scikit-learn:

```
from sklearn.datasets import fetch_20newsgroups
```

(See https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html#sklearn.datasets.fetch_20newsgroups)

(a) Load the following 4 categories from the 20 newsgroups dataset: categories = ['rec.autos', 'talk.religion.misc', 'comp.graphics', 'sci.space']. Print the number of documents in the training dataset and the test dataset. Print the number of attributes in the training dataset.

(b) Optimize the hyperparameters (e.g. max_depth, min_samples_leaf, etc.) of your decision tree to maximize the classification accuracy. Show the confusion matrix of your decision tree.