**Notes**

Mark the top sheet with your name and the names of any students with whom you collaborated. If you didn't collaborate with anyone, mark your collaborators as "None." Begin each problem on a new sheet of paper.

Remember, your goal is to communicate. Full credit will be given only to correct solutions which are described clearly. Convoluted and obtuse descriptions will receive low marks.

To complete your homework, you may ONLY consult the following material:

1. lecture slides posted on the class webpage
2. course notes you or others took during lecture
3. the required text (CLRS)
4. websites that may clarify the concepts covered in the material

Please stop by my office for any question.

**Name:**

**ID:**

**Problem Set 1 (25 pts)**

Write a python code (20 pts) such that:

* It loads the data in iris1.csv (in exp.zip).
* It performs a 5-fold cross validation (make sure the target label are equally distributed in all the train sets and in the test sets).
* It scales the independent features with mean zero and unit standard deviation.
* It normalizes the scaled result with the norm “l1”.
* It uses the normalized output to train and test a classification tree (the algorithm uses athe entropy function to decide the best split)
* The metric that should be used is the accuracy and the final value that should be printed is the average accuracy among all the folds.

Change your python code (5 pts) in the way that for each folds the train set is further divided in a second train and test used for understanding (according to the accuracy) what is the best criteria for the classification tree (Gini or entropy). Once the best criterion is chosen train the tree with it and test the accuracy result in each fold.