CS697A – Topic in Computer Science – Machine Learning
Summer 2020

Assignment 3 (12.5 points)
Due date : July 19, 2020 Sunday at 11:00pm

PURPOSE:

Review: Ch9 (Decision Trees), Ch8 (Nonparametric Methods), Ch17 (Combining Learners).

WHERE TO SUBMIT ASSIGNMENTS:

Please submit through the class Blackboard site. Please zip and upload all your files using filename studentID_HW3.zip. Submit a zip file of the Jupyter Python notebook you used, the datafiles and also a pdf answering the questions for the homework.

POLICY:

Collaboration in the form of discussions is acceptable, but you should write your own answer/code by yourself. Cheating is highly discouraged for it could mean a zero or negative grade from the homework. If a question is not clear, please let me know (via email, during office hour or in class). Do not use a library unless it is a very basic one or it is indicated otherwise.

DATA:
Read:
https://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits

Download the dataset and read the description carefully:
https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/
test set: optdigits.tes
training set: optdigits.tra

Create the following **training datasets** from the optdigits.tra set:
X100_C69: Randomly chosen N=100 instances from class 6 and randomly chosen N=100 instances from class 9.
X100_CAll : Randomly chosen N=100 instances from each of the classes.
X500_69: similar to X100_69, but for N=500.
X500_CAll : similar to X500_69, but for N=500.

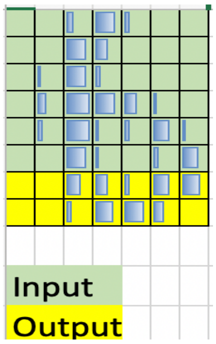
Note that when you use X100_C69 or X500_69 for training, for testing, you should only use the instances that belong to classes 6 and 9 in the test set optdigits.tes.

Questions:

Q1 [4pts]: Decision Trees, classification: Use **sklearn.tree** library's DecisionTreeClassifier algorithm. For the DecisionTreeClassifier determine the value of the **tree depth** parameter (experiment with depth=2, 3, 5, 10) that results in the best test error for each of the 4 training data sets you created. How does the best depth value change as the number of instances and classes change?

Q2 [4pts]: Nonparametric Classification: Use **sklearn.tree** library's KneighborsClassifier algorithm. For the KneighborsClassifier determine the value of the best k parameter (experiment with k=1, 3, 5, 9) that results in the best test error for each of the 4 training data sets you created. How does the best k value change as the number of instances and classes change?

Q3 [4.5pts]: Decision Trees, regression for digit completion: Using only the data in X500_69 for training, use the first 42 features as inputs and predict the next 16 features, i.e. create 16 decision tree regression models, using the sklearn library. Report the test error (use only the instances from classes 6 and 9) for each of the 16 regression models. Which pixels are easier to predict?
(Clarification, each of your models will have the same set of features, namely features 1…48.)


Input
Output