**CS697A – Topic in Computer Science – Machine Learning**

**Summer 2021**

**Assignment 3 (10 points)**

Due date : July 30, 2021 Friday at 11:00pm

**PURPOSE:**

Review: Ch9 (Decision Trees), Ch11 (Multilayer Perceptrons)

**WHERE&HOW TO SUBMIT ASSIGNMENTS:**

Please submit through the class Blackboard site. Put all hw files, including the Jupyter Python notebook you used, the datafiles, a pdf answering the questions for the homework, into a directory named studentID\_HW3 Zip and submit file studentID\_HW3.zip

Include your name, ID, your groupID and your groupmates’ names and IDs in all files. Each group member must submit the same file.

Fill in expected row and submit as a separate excel file, studentID\_HW3\_expected.xlsx in your zipped file.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Question | Q1 | Q2 | Q3 | Q4 | Q5 | TOTAL |
| Max | 1 | 2 | 2 | 2 | 3 | 10 |
| Expected |  |  |  |  |  |  |

**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). You can use any library for this homework.

**DATA:**

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

<https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/>

test set: optdigits.tes

training set: optdigits.tra

**Q1 [1pts]** Create the following **training datasets** from the optdigits.tra set:

X25: Randomly chosen N=25 instances from each class.

X100: Randomly chosen N=100 instances from each class.

**Q2 [2pts]:** Decision Trees, classification:

Use library sklearn.tree.**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. Report the training and test errors for each depth value and the training set. How does the best depth value change as the number of instances change?

**Note:** Check the depth of the your trained decision tree, by e.g. plotting the tree.

**Q2 [2pts]:** Repeat Q1 using XGBoost and LightGBM algorithms. Experiment with at least 5 different parameter settings to see their effect on training and test errors. How do best parameters change as the #training instances change?

**Q3 [2pts]:** Multilayer Perceptrons: classification

Use library sklearn.neural\_network.**MLPClassifier**.

For the neural network (MLP) determine the value of the best **hidden\_layer\_sizes** (experiment with 1 and 2 hidden layers and 2, 5, 10 hidden units in each layer) that results in the best test error for each of the training data sets you created.

**Q4 [3pts]:** **Regression for digit completion**: regression

Use library sklearn.neural\_network.**MLPRegressor**

Using only the data for class 6 and class 9 in X100 for training, use the first 48 features as inputs and predict the next 16 features, i.e. create a neural network with 16 outputs.

Report the test error (use only the instances from classes 6 and 9).

Which pixels are easier to predict?

(Clarification, each of your models will have the same set of features, namely features 1…48.)

