Theoretical Foundations of Machine Learning

Course Project

1- **MNIST**

MNIST digits dataset will be utilized. It is a dataset of handwritten numbers from 0 to 9. MNIST has a training set of 60,000 examples, and a test set of 10,000 examples. It can be downloaded from: http://yann.lecun.com/exdb/mnist/

K Nearest Neighbors (KNN) is a classifier that finds the class of the test sample based on the distance of it from the training samples. It finds the K training samples with smallest distance to the test sample. The dominant class in the K points is then selected as the test point class.

Project steps are as follows:

a. Load **MNIST** dataset. **(Phase 1)**

b. Apply **HOG** features to the images (sklearn) **(Phase 2)**

c. Implement **KNN** with 'K' as a parameter and Euclidian distance. **(Phase 3)**

d. There must be another 2 models to be implemented (of your choice), one is mandatory and the other is bonus.

2- **CIFAR-10**

CIFAR-10 is a dataset of 60000 color images downsized to 32*32 categorized in 10 classes. Each class contains 6000 samples. The dataset is divided into 50000 images for training and 10000 images for testing. It can be downloaded from: https://www.cs.toronto.edu/~kriz/cifar.html

Support Vector Machines (SVM) is a linear classifier that can be applied to nonlinearly separable data through utilizing its kernel functions, Polynomial and RBF.

Project steps are as follows:

a. Load **CIFAR-10** dataset. **(Phase 1)**

b. Prepare the train-validation and test portions. **(Phase 1)**

c. Apply **Central Moments** features 'C' to the images. **(Phase 2)**

- C is the moments order where (p + q <= C). C should be entered as a parameter.
- Ex: c=3 □ M_{00} , M_{10} , M_{01} , M_{11} , M_{20} , M_{02} , M_{12} , M_{21} , M_{30} , M_{03}
- So, feature vector length will be 10
- Moments can be computed through the following equation:
  - $M_{pq} = \sum_x \quad \sum_y \quad (x - \underline{x})^p \, (y - \underline{y})^q \, I(x, y)$
  - $\underline{x}$ *is the mean of* x *dimension*
  - $\underline{y}$ *is the mean of* y *dimension*

- I(x, y) is the pixel value of the image at coordinates (x, y).

d. Implement SVM with different kernel functions. **(Phase 3)**

e. **There must be another 2 models to be implemented (of your choice), one is mandatory and the other is bonus.**

3- <u>Medical Cost Personal</u>

It is a dataset for regression tasks. It consists of 1300+ records containing persons medical data and the target is "charge" column. The goal is make a model that fits these data and predicts the charge for the new persons that the medical insurance should cover. The data is to be divided to 1000 samples for both training and validation and the rest is for testing. The dataset can be downloaded from:
https://www.kaggle.com/mirichoi0218/insurance

<u>Project steps are as follows:</u>

1. Load the dataset. **(Phase 1)**
2. Prepare the train-validation and test portions. **(Phase 1)**
3. Apply any preprocessing or features that you find suitable for the data. **(Phase 2)**
4. Apply 3 different models and compare between them, 2 mandatory and the 3rd is bonus. **(Phase 3)**

NOTE:

1. A comment on the results and on the comparison of the three models applied should be given.
2. It is expected that the selected models should be experimented with different hyper-parameters.
3. At the final results comparisons, proper metrics should be selected such as: precision, recall, F1 measure (F-Score), …
4. Error analysis should be stated such as: correctly/wrongly classified examples. Reasons and suggested improvements should also be included.
5. Teams should select one of these ideas.
6. Besides selecting one idea of the above, students can select different idea.
7. A Proposal with the selected idea(s) and team names should be sent on TAs mails.
8. <u>If the selected idea is different from the above ideas and is irrelevant, the one selected from above ideas will be confirmed.</u>
9. <u>Students are totally responsible of their chosen idea</u>.

Deliverables:

I. Phase 0, proposal delivery. 》 Tuesday Dec. 28 (by email to TAs)
II. Phase 1 》 Tuesday Jan. 4

III. Phase 2 》 Tuesday Jan. 11
IV. Phase 3 》 (Full project Delivery), Tuesday Jan. 18