CSE 572: Data Mining

Fall 2016

Assignment 3 / Mini Project 1

Instructions

- **Submission Deadline**: Wednesday, October 19, 2016 (11:59pm). You will submit this assignment through Blackboard
- There are two problems. Total points possible: 20
- You have to use Matlab for this assignment
- You will work in groups of 2 or 3 for this assignment
- Only one submission is required from each group. Mention the names of all group members in the report (please see details below)

In this assignment, you will study the application of the k-nearest neighbor and neural network classifiers on two real-world classification problems. The datasets to be used for this assignment are uploaded under the "Datasets" folder. X_train, y_train, X_test and y_test denote the training features, training labels, testing features and testing labels respectively. In X_train and X_test, each row denotes a data sample and each column denotes a feature.

Problem 1 (10 points)

The Human Activity Recognition dataset was created from experiments carried out on a group of 30 volunteers to recognize human activities using smart phone data. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz were captured. The data was processed using signal processing algorithms to extract feature vectors of dimension 561. The training set contains 7,352 samples and the test set contains 2,947 samples.

Implement the k-nearest neighbor algorithm with k = 7 on this dataset. Use the simple Euclidean distance measure to compute the distance between two samples. Report the percentage accuracy on the test set.

Problem 2 (10 points)

The VidTIMIT dataset consists of video and audio recordings of 43 subjects reciting short sentences. In this assignment, we will use a subset of the dataset with 25 subjects. We will also use only the video modality. The videos were sliced into images and the discrete cosine transform function was used to extract feature vectors of dimension 100 from each image. The training set contains 3,500 samples and the test set contains 1,000 samples. Our objective is to recognize a subject from a given image.

Part 1

Implement the k-nearest neighbor algorithm with k = 7 on this dataset. Use the simple Euclidean distance measure to compute the distance between two samples. Report the percentage accuracy on the test set.

Part 2

Use the training set to train a feedforward neural network with 1 hidden layer containing 25 neurons. Report the percentage accuracy on the test set. You can use Matlab's in-built neural network related functions for this part. Take a look at the "feedforwardnet" function.

You need to submit (only one submission is required from each group)

- The Matlab code files
- A ReadMe file with clear instructions on how to run your code
- A brief report (1-2 pages) summarizing your findings and the accuracies obtained. **Mention the names of all the group members in the report.**

Submit all the documents as a single zip file through Blackboard.