Active Learning for Computer Vision Curriculum

Project 1: Brief Overview of Machine Learning

Total Points Possible: 50

Datasets: The datasets required for this project are included in the folder. For the feature matrix in each dataset, each row denotes a sample and each column denotes a feature

Problem 1 (25 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 = 5 on this dataset. Use the simple Euclidean distance measure to compute the distance between two samples.

Train an SVM classifier with a polynomial kernel with parameter 2 on the training set and test on the test set. You need to train one SVM for each class; for predicting a test sample, use the maximum of the values returned by all the SVMs to decide the final class.

Report the percentage accuracy on the test set using each method.

Problem 2 (25 points)

The *seeds* dataset contains measurements of geometrical properties of kernels belonging to different varieties of wheat. It has 210 data samples and each sample is described by 7 attributes (features).

Implement the k-means clustering algorithm on this dataset. Use the simple Euclidean distance to compute the distance between any two samples. Start with a random initialization of the centroids and iterate until convergence. The algorithm is assumed to have converged if the number of iterations exceeds 100 **OR** the change in the sum of squared errors (SSE) between two successive iterations is less than 0.001. Run the algorithm for k = 3, 5 and 7. For each value of k, run the algorithm with 10 random initializations of the centroids. Report the average SSE value (averaged over the 10 initializations) for each value of k.

Note: You are **NOT** allowed to use any built-in function for k-means in your implementation. You are allowed to use the *pdist* function to compute the Euclidean distance.