COSI 126A Homework 2

Due by March 16th

February 27, 2020

Problem 1 (30 points)

Compare your K-means clustering code with the public scikit-learn one in terms of objective function value and execution time on the following data sets.

- MNIST. http://yann.lecun.com/exdb/mnist/
- CIFAR-10. http://www.cs.toronto.edu/ kriz/cifar.html
- LFW. http://vis-www.cs.umass.edu/lfw/

The above three data sets are image data sets. You can use the pixel level features or other well extracted features with the true cluster number. Modify your code and try to beat the scikit-learn one in a fair setting.

Problem 2 (25 points)

Let us re-use the above MNIST data set for classification. Here we focus on the none-preprocessing category, that means the pixel-level feature. Each image is represented by a 1×784 vector. Reimplement linear classifier (1-layer NN) (Test error rate: 12%), K-nearest-neighbors (Test error rate: 5%), SVM + Gaussian Kernel (Test error rate: 1.4%).
You can use scikit-learn codes. Report the parameter settings of the above classifiers and how you get them.

Problem 3 (5 points)

Read this paper titled *Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?*. What are the recommended classifiers for practical use?

Problem 4 (40 points)

Kaggle project (https://www.kaggle.com/chrisfilo/urbansound8k). Provide a detailed report on the audio classification including feature extraction, dataset building, model selection, model update and result analysis.