About

Due

Monday 3/4/19, 11:59 PM CST

Goal

This homework focuses on vector quantization and classification. More specifically, you should do 1) data slicing, 2) vector clustering, 3) making histograms, and 4) building a multi-class classifer. We also encourage you to do in-depth experimentation and analysis.

Code and External Libraries

The assignment can be done using any language.

You may use packages for k-means, for nearest neighbors, and for whichever classification method you choose.

Problems

Total points: 100

Obtain the actitivities of daily life dataset from the UC Irvine machine learning website (https://archive.ics.uci.edu/ml/datasets/Dataset+for+ADL+Recognition+with+Wrist-worn+Accelerometer, data provided by Barbara Bruno, Fulvio Mastrogiovanni and Antonio Sgorbissa). Ignore the directories with MODEL in the name. They are duplicates.

(a) Build a classifier that classifies sequences into one of the 14 activities provided and evaluate its performance using average accuracy over 3 fold cross validation. To do the cross validation, divide the data for each class into 3 folds separately. Then, for a given run you will select 2 folds from each class for training and use the remaining fold from each class for test. To make features, you should vector quantize, then use a histogram of cluster center. This method is described in great detail in the book in section 9.3 which begins on page 166. You will find it helpful to use hierarchical k-means to vector quantize. You may perform the vector quantization for the entire dataset before doing cross validation.

You may use whatever multi-class classifier you wish, though we'd suggest you use a decision forest because it's easy to use and effective.

You should report (i) **the average error rate over 3 fold cross validation** and (ii) **the class confusion matrix of your classifier for the fold with the lowest error (averaged over the class)**, i.e. just one matrix for the 3 folds.

(b) Now see if you can improve your classifier by (i) modifying the number of cluster centers in your hierarchical k-means and (ii) modifying the size of the fixed length samples that you see.

Submission

Submission will be through gradescope:

Your submission for this homework should include:

1. Page 1 (40 pts) Experiment table

Table listing the experiments carried out with the following columns. Size of the fixed length sample Overlap (0-X%) K-value Classifier Accuracy. We expect you to have tried at least 2 values of K and at least 2 different lengths of the windows for quantization. Note: For K-means please also list if you used standard K-means or hierarchical.

2. Page 2 (28 pts) Histograms

Histograms of the mean quantized vector (Histogram of cluster centres like in the book) for each activity with the K value that gives you the highest accuracy. (Please state the K value)

3. Page 3 (22 pts) Confusion matrix

Class confusion matrix from the classifier that you used. Please make sure to label the row/colums of the matrix so that we know which row corresponds to what.

4. Page 4 (10 pts) A screenshot of your code

The page should contain snippets of code demonstrating:

i) Segmentation of the vector

ii) K-means

iii) Generating the histogram

iv) Classification

5. Page 5+ Screenshots of all your souce code.