CS-470 Machine Learning (OHT1 Spring-2017)

Physical Activity Classification

Problem Statement & Background: For your semester project you are required to classify 5 different physical activities using the artifacts introduced into time-series waveforms of respiratory sensors. The problem background is given in the the paper draft uploaded to LMS (titled: `ProjectPaper`). Please read the project paper and then come back to this document.

WELCOME BACK..!! After reading the paper you should have a fairly good idea of the problem background. I have completed the difficult part of the problem for you i.e., I have extracted wavelet based signatures from the sensor outputs. Download, the `ProjectData.mat` file from LMS. This file contains 129 activity sessions; the wavelet score curves for accelerated ($p'[n]$) and normal ($\hat{p}[n]$) breathing for each activity session are supplied in vectors `HiCurve` (accelerated breathing) and `LoCurve` (normal breathing). In addition to score curves you are also given the class labels of the activity sessions (in vector ClassLabels, numbered 1 to 5). Each session contains only one type of physical activity; an example of the score curves obtained during a coughing session are shown in Fig. 1.

Objective: Use any classifier of your choice (I recommend SVMs) to classify the 5 activity patterns.

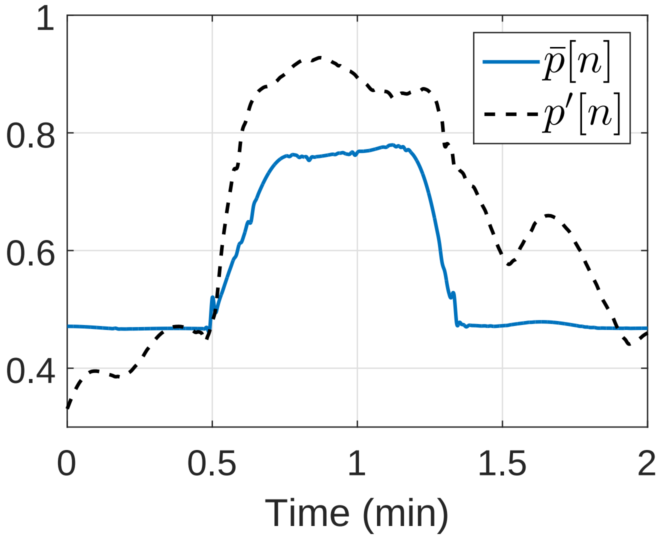

Fig. 1: Wavelet score curves for the artifact induced during a coughing session.

Preprocessing: You need to convert the scores into valid features or a format that acceptable to a classifier. This can be accomplished by the following steps:

- Concatenate the vectors `HiCurve` and `LoCurve` to form one feature vector.
- Make all feature vector equal length by discarding a few samples at the start and end.

Evaluation Protocol and Metrics: The number of activity sessions is limited therefore; you should use 4-fold cross-validation. More specifically:

- Split your data, *randomly*, in to 4 non-overlapping folds (or partitions).
- Train on 3 folds and test on the 4^{th} fold.
- Repeat this process until all four folds have been in the test set one-by-one.
- Average your accuracy over the 4-folds and store it.
- Repeat 4-fold cross-validation (with randomly created partitions) at least 10 times and present the overall average accuracy.

The activity sessions are somewhat unbalanced therefore, in addition to accuracy, you should also use an evaluation metric called the Balanced-Error-Rate (BER), which is more suitable for unbalanced data. The BER can be calculated as below:

$$BER = \frac{1}{M}\sum_{i=1}^{N} \frac{\left(\sum_{j=1}^{M} A_{ij}\right) - A_{ii}}{\sum_{j=1}^{N} A_{ij}}$$

Here, A_{ij} denotes the element of the confusion matrix A at row-i and column-j. It indicates the number of test vector of class-i that are predicted to belong to class-j. N and M represent the total number of test vectors and classes respectively. *NOTE:* BER also needs to be averaged over atleast 10 runs of 4-fold cross-validation.

Deliverables:

- Project report in IEEE-Conference format summarizing the problem and the approach you employed. Also present your average accuracy and BER. **Maximum length:** 04 pages.
- Your code attached as an appendix.

Rules:

- Due Date: XX June 2017 (after End Semester Exam).
- Maximum Members per group: 02.
- Plagiarism of any form shall be dealt with strictly.