CARLETON UNIVERSITY
Department of Systems and Computer Engineering

SYSC/BIOM 5405 – Pattern Classification and Experiment Design
Fall Semester 2018

Working in pairs, each team will develop a pattern classification system for the same pattern classification challenge using the same training data. This task will encompass elements from the entire course, ranging from experiment design, to feature extraction, to classification techniques, to reporting classification accuracy. This project may require teams to learn concepts outside of the scope of the lectures and each team will deepen their expertise about one or more methods.

### PROJECT DESCRIPTION

What is unique about this year's challenge is that you will be predicting a healthy or diseased state from relatively simple time series data collected on patients walking a short distance. There are samples from subjects with one of three neuro-degenerative diseases contained in the sample data: *Amyotrophic lateral sclerosis* (ALS), *Parkinson's disease*, and *Huntington's disease*. In patients suffering from these degenerative diseases, motor control progressively fails over the course of the disease. There are some promising treatments under investigation, but no cure is widely available. Treatment focuses on improving symptoms.

In this study, force-sensitive resistors were installed in footwear with a recording device attached to the ankle. Output is roughly proportional to force. Subjects were either "healthy" or suffered from from one of the three diseases. For each subject, there are two time-series, one for each foot, measured at 300 Hz for 5 minutes as the subject walked along a course. There is some additional information available for each subject, though some of these additional data points were missing for some subjects. The course was not the same for all subjects.

Your goals, based on these time-series data, are two-fold:

1. to predict "healthy" `H` or "diseased" `D` state, and
2. to predict the specific disease (`ALS`, Huntington's `HD`, or Parkinson's `PD`) or declare a subject healthy `H`.

The first problem should be relatively easier, as a two-class problem with distinct changes in motor activation under disease conditions.

There are a relatively small number of samples to work with.

## 1 Evaluation

Teams will be evaluated on

- the quality of all *deliverables* below,
- the *accuracy* obtained on the final unlabeled test dataset, and
- the *correctness* of your accuracy prediction over the test dataset.

## 2 Deliverables

1. **Project Proposal** detailing the pattern classification approach that you plan to use, including a source for an implementation of your chosen method. Your presentation must include a description of 1 feature that you intend to use in your approach.

   This will be a 4 minute presentation with approximately 6 slides.

   You will be evaluated on the quality of your presentation and your progress to date. (In other words, demonstrate that you've started working, have a software framework in place, and understand the problem.)

2. **The Pitch** describing your approach, your predicted accuracy, and how you computed it. Pitch your method as the best approach.

   This will be a 4 minute presentation with approximately 6 slides.

   Slides should cover

   - a quick review of your method,
   - any pre-processing and feature extraction or class imbalance handling,
   - describe your approach to dealing with time-series data,
   - describe your training/testing protocol, including your meta-learning strategy,
   - provide your estimated accuracy *including the standard deviation of your estimate* and describe your methodology for estimating your "true"/expected accuracy.

   You *must* include two confusion matrices in your presentation: one for the two-class `H/D` classification problem, and one for the four-class `H/ALS/HD/PD` performance.

   You will be evaluated on the quality of your presentation and content.

   At the conclusion of this class, all groups will be provided with the blind test data set.

3. **Final Report** describing the method that you have chosen to use, the source of the implementation of your method, details on training techniques and parameters used, any preprocessing of the data and feature extraction, a discussion of your approach for the two-class and four-class problems and your testing procedures, an estimate of prediction accuracy with and without meta-learning, and a discussion of the actual accuracy achieved over the blind test dataset. This report should be approximately 10 pages, double-spaced including figures/tables.

## 3 Schedule

**Project Launch**: November 13, 2018 (competition announced)

**Project Proposal**: November 20, 2018 (presentations submitted via cuLearn)

**Project Pitch**: November 27, 2018 (presentations submitted via cuLearn, blind test data released)

**Submit Predictions**: November 30, 2018 **3pm** (classification of blind test data)

**Results Announced**: December 4, 2018 (winners glorified; prizes distributed)

**Final Report**: December 21, 2018 **6pm** (submitted via cuLearn)

## 4 The Dataset

The data sets will consist of a training set `train.zip` and a blind test set `blind.zip`.

The training set consists of tab separated files with two columns, the left column for left foot force and the right column for right foot force during 5 minutes of walking. The files are named as `<type><`$N$`>.tsv`. There is also a `train.txt` file which contains additional information for each subject including age, height, weight, gender, and gait speed. There are 43 subjects. Each subject is identified by a unique number $N$ and their disease/health: `control1.tsv` (healthy subject), `als2.tsv` (ALS), `hunt3.tsv` (Huntington's disease), `park4.tsv` (Parkinson's disease).

The blind test set is structured in the same way but all 20 subjects are identified as `blind`$N$`.tsv`. The additional information can be found in the file `blind.txt`.

For some subjects, information in the additional data is missing. These fields are denoted with a "`?`".

## 5 Detailed Instructions

- Phase 1: Determine Approach
  - All teams will choose a *unique* pattern classification approach.
  - First-come, first-served…ideas include
    1. Bayesian belief networks
    2. feed-forward neural networks
    3. recurrent neural networks
    4. linear discriminants
    5. support vector machines
    6. k-nearest-neighbour
    7. decision trees
    8. radial basis function networks
    9. probabilistic neural networks
    10. genetic algorithms
    11. k-means clustering
    12. hidden Markov models
    13. association mining
    14. logistic regression
    15. your own idea!
  - Find an implementation in any language you like
  - Prepare and deliver project proposal detailing your proposed pattern classification approach and chosen implementation framework. Demonstrate that you understand the problem and have a clear plan on how to solve it.
  - Discuss 1 feature that you intend to use in your method. Please discuss how to extract the scalar feature from the time-series data.

- Phase 2: Develop the pattern classification system
    - Structure your investigation using the following steps:
        1. Data pre-processing: normalization, outlier detection, censoring of bad data, handling of missing data, records of varying length, etc.
        2. Feature extraction/selection: you may wish to generate many features from the time-series data provided to you and ultimately only use a subset in your classifier.
        3. Partition data and establish experiment design: train/validation/test sets, balancing classes (optional).
        4. Train classifier: approach used, parameters required, how they were tuned
        5. Testing and expected accuracy: what is your predict accuracy for each of the two problems, how was it computed, and provide a standard error / standard deviation on your estimate. For example, "My accuracy will be $0.73 \pm 0.04$."
        6. Meta-learning approaches: Implement at least one meta-learning strategy and investigate its effect on accuracy. For example: CME-voting, bagging, boosting.
- Phase 3: Pitch method to class
    - Present to class
    - Predict accuracy you will get on the blind test dataset
        - discuss expected performance both with and without meta-learning, but ultimately choose 1 approach and 1 estimate
        - Include 2 Confusion Tables in your presentation
    - The blind test data is released. Keep in mind the size of the dataset. We have held back approximately 30% of the total data... Beware of runtime issues (you have < 48 hours to process all data!
- Phase 4: Competition
    - Provide single best set of predictions for blind test data to the course instructor.
    - Course instructor will evaluate each submission.
        - Score1 will be overall accuracy $A_{cc} = A_{cc,h/d} \cdot A_{cc,h/als/hd/pd}$
        - Score2 will be probability of observing this accuracy given your estimated accuracy and standard deviation, assuming a normal distribution.
    - Results announced: Laugh, cry, acceptance speeches...; Points for how well you do (score1), points for how close your prediction is to your actual performance on the test data (score2).
- Phase 5: Final report
    - Prepare a final report (10 pages double-spaced including figures) describing entire effort and results. Discuss how you would change your approach now that you have seen the other approaches and now that you know how well you did.

**End of Project Description**