BIOM5405/SYSC5405 W16 Term Project

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 with regards to one or more methods.

Evaluation

Teams will be evaluated on:

- The quality of all deliverables (see below)
- The accuracy obtained on the final unlabeled test dataset (measured by *classification accuracy*)
- The correctness of your accuracy prediction over the test dataset

Deliverables

1) A **project proposal presentation** detailing the pattern classification approach that you plan to use, including a source for an implementation of your chosen method. Each team should also describe one feature that they have extracted from the data. This will be a **6 minute presentation** with ~8 slides.

2) **The pitch** consisting of a presentation with ~10 slides describing your approach, your predicted accuracy, and how you computed it. Each group will be given **6 minutes** to pitch their method as being the best approach. At the conclusion of this class, all groups will be provided with the blind test data set. Slides should cover:
    a. Quickly review of your method/implementation
    b. Describe your experiment design
    c. Describe any pre-processing of the data, including feature selection/extraction and class imbalance issues (if relevant)
    d. Describe your training/testing protocol, including your meta learning strategy
    e. Provide your estimated accuracy (including the standard deviation of your estimate) and describe your methodology for estimating your "true" accuracy (i.e. the accuracy you should expect when applied to new test data). Here "accuracy" is measured as the *classification accuracy*. **Please include a confusion matrix between the six classes**.

3) A **final report** detailing the method that you have chosen to use, the source of the implementation of your method, details on training techniques and parameters used, any pre-processing of the data and feature extraction, discussion of 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 ~10 pages, double-spaced including figures/tables.

Schedule

Thursday 10 March:	Competition announced.
Tuesday 22 March:	Project proposal presentations (submit via CULearn)
Tuesday 5 April:	Pitch presentations (submit via CULearn). Blind test data released.
3pm Wednesday 6 April:	Classification of unlabelled data submitted to instructor.
Thursday 7 Apr:	Results announced. Winners glorified. Prizes distributed.

Monday 18 April: Final reports submitted electronically via CULearn.

The dataset

- The dataset is a collection of images of mouths. The images come from 4 subjects. The data from the 5th subject is being withheld as test data. The images represent six states of the mouth: closed, open (tongue inside mouth), and tongue extended UP, DOWN, LEFT, and RIGHT. During data collection, subjects were asked to move their tongue slightly and to tilt their faces left, right, up and down to approximately 20 degrees. Six different lighting conditions were used: light from the front-high, front-low, side-high, side-low, top-high, top-low. You are being provided with approximately 4000 images for each of the six states. You do not have to use all images. Your primary task is to identify the mouth state (O/C/U/D/L/R) in each image of the unlabelled test data.
- You may wish to extract subsamples (e.g. representative subsets of images) from each class.
- You may wish to exclude records that appear to be invalid/noisy/unrepresentative.
- The data is available in a ZIP file on the CULearn course website.

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
  - Learn about feature extraction from images (i.e. how to create meaningful feature vectors from images). MATLAB, OpenCV, or other tools may be useful here.
  - Implement at least one feature extraction technique and apply it to all images. **Show the distribution of this feature for all six classes.**
  - Prepare and deliver project proposal detailing proposed pattern classification and feature extraction approach and implementation.
- **Phase 2: Develop pattern classification system**
  - Structure your investigation using the following steps:
    - Data pre-processing
      - Normalization, outlier detection, censoring of bad data, etc.
      - Handling of missing data, records of varying length, etc
    - Feature extraction

        - You may wish to convert image data into one or more scalar features. This is not a course on image processing, but you will have to learn about feature extraction from images.
    - Partition data & establish experiment design
        - Train/validation/test sets, balancing classes (optional), etc.
    - Train classifier
        - What approach used, what parameters required, how were they tuned, etc
    - Testing & expected accuracy
        - What is predicted accuracy, how was it computed, provide a standard error / standard deviation on your estimate (e.g. "my 6-class accuracy will be 0.73 ± 0.04")
    - Meta-learning approaches
        - Implement at least one meta-learning strategy (e.g. CME-voting, bagging, boosting), and investigate its effect on accuracy
- **Phase 3: Pitch method to class**
    - Present to class
    - Predict accuracy you will get on test dataset
        - discuss expected performance both with & without meta-learning, but ultimately choose 1 approach and 1 estimate
        - Include a 6-class confusion table in your presentation
    - The blind test data is released. There will be approximately 1000 images of each class from a 5^{th} subject. Beware of runtime issues (you have 24 hrs to process all images!)
- **Phase 4: Competition**
    - Provide single best set of predictions for unlabelled data to course instructor.
    - Course instructor will evaluate each submission.
        - Score1 will be 6-class overall accuracy
        - 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.