BIOM5405/SYSC5405 F19 Term Project

Working in teams of three, 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.

This year's project is about protein methylation. (https://en.wikipedia.org/wiki/Protein_methylation) This is a type of post-translational modification that can alter a protein's structure, stability, or function. You are provided with two files of *descriptors* (aka features), extracted using ProtDCal (doi:10.1002/pro.3673), and you are tasked with distinguishing between those protein windows that are ('P' in last column) or are not ('N' in last column) methylated. This is a difficult problem, and you will have to work hard to do better than random. Part of this difficulty comes from the class imbalance inherent in the dataset.

Thank you to Dr. Yasser Ruiz Blanco for preparing the dataset!

Evaluation

Teams will be evaluated on:

- The quality of all deliverables (see below)
- The accuracy obtained on the final unlabeled test dataset
  - Measured by *maximum achievable recall at a precision of at least 50% (Re@Pr50)*
- The correctness of your *Re@Pr50* 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. This will be a 5 **minute presentation** with ~6 slides. You will be evaluated on the quality of your presentation and your progress to date (i.e. demonstrate that you've started working, have a software framework in place, understand the problem, etc.)

2) **The pitch** consisting of a presentation with ~6 slides describing your approach, your predicted Re@Pr50, and how you computed it. Each group will be given **5 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 Re@Pr50 (including the standard deviation of your estimate) and describe your methodology for arriving at this estimate (i.e. the Re@Pr50 you expect to achieve when your method is applied to new test data).
      i. You must include full P-R curves for your method with/without meta-learning.

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, a discussion of your approach and your testing procedures, an estimate of prediction Re@Pr50 with and without meta-learning, and a discussion of the actual Re@Pr50 achieved over the blind test dataset. This report should be ~10 pages, double-spaced including figures/tables.

Schedule

Tuesday 19 Nov:	Competition announced.
Tuesday 26 Nov:	Project proposal presentations (submit via CULearn)
Tuesday 3 Dec:	Pitch presentations (submit via CULearn). Blind test data released.
3pm Wed 4 Dec:	Classification of blind test data submitted to instructor.
Thursday 5 Dec:	Results announced. Winners glorified. Prizes distributed.
Monday 16 Dec:	Final reports submitted electronically via CULearn.

The dataset

- The dataset is a collection of ProtDCal descriptors derived from sequence windows centred on lysines known to be methylated (‘P’) or assumed to never be methylated (‘N’).
  - Each line of the file gives the 29 pre-computed descriptors for one sequence window
  - You are provided with two files:
    - csv_result-Descriptors_Calibration.csv, containing 4996 rows
    - csv_result-Descriptors_Training.csv, containing 19988 rows
- We have withheld 5150 rows of the total data as a blind test set. You can assume that the class imbalance is roughly equal on all three files. The labels of the blind test data will never be released.

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. convolutional neural networks
    5. linear discriminants
    6. support vector machines
    7. k-nearest-neighbour
    8. decision trees
    9. gradient-boosted decision trees
    10. decision forests
    11. radial basis function networks
    12. probabilistic neural networks
    13. genetic algorithms
    14. k-means clustering
    15. hidden Markov models
    16. association mining
    17. logistic regression
    18. 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.
- **Phase 2: Develop the 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/selection
            - You may wish to generate new features from the data provided to you or to select only a subset in your classifier.
        - Partition data & establish experiment design
            - Train/validation/test sets, balancing classes (optional), etc.
        - Train classifier
            - What approach used, what parameters required, how they were tuned, etc.
        - Testing & expected accuracy
            - What is predicted Re@Pr50, how was it computed, provide a standard error / standard deviation on your estimate (e.g. "the maximum achievable recall at a precision of at least 50% will be 0.43 ± 0.04")
        - Meta-learning approaches
            - Implement at least one meta-learning strategy (e.g. CME-voting, bagging, boosting), and investigate its effect on performance.
- **Phase 3: Pitch method to class**
    - Present to class
    - Predict accuracy you will get on the blind test dataset
        - Discuss expected performance both with & without meta-learning, but ultimately choose 1 approach and 1 estimate
    - The blind test data is released. Keep in mind the size of the dataset (~5K rows)… Beware of runtime issues (you have ~28 hrs 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 Re@Pr50
        - *Score2* will be probability of observing this Re@Pr50, 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.