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.

What is unique about this year's challenge, is that you will be provided with a small body of labelled training data and a large body of unlabelled training data. We have created an oracle that can provide the true class label for any unlabelled sample. You will be permitted to request the actual class label for **up to 1000** of these unlabelled data. It is up to you to decide which samples to elucidate, using concepts from *active learning*.

Thank you to Mr. Waldo Paz-Rodriguez for preparing the dataset and the oracle for this project!

Evaluation

Teams will be evaluated on:

- The quality of all deliverables (see below)
- The accuracy obtained on the final unlabeled test dataset (measured by *min(Sn,Pr)*)
- 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. 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 ~7 slides describing your approach, your predicted accuracy, 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 approach to active learning. How did you select your 1000 additional labelled training points?
    e. Describe your training/testing protocol, including your meta learning strategy
    f. 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 *min(Sn,Pr)*.
        i. You must include a precision-recall curve in your presentation.

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 active learning approach 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 ~10 pages, double-spaced including figures/tables.

Schedule

Wednesday 15 March:	Competition announced.
Monday 20 March:	Project proposal presentations (submit via CULearn)
Monday 3 April:	Pitch presentations (submit via CULearn). Blind test data released.
3pm Tuesday 4 April:	Classification of blind test data submitted to instructor.
Wednesday 5 Apr:	Results announced. Winners glorified. Prizes distributed.
Monday 17 April:	Final reports submitted electronically via CULearn.

The dataset

- The dataset is a collection of protein windows centered on lysine residues. Positive windows correspond to sites that are known to be ubiquitinated (i.e. an addition of a ubiquitin peptide to the protein at that lysine residue). Negative windows correspond to sites on proteins that we assume are not ubiquitinated (i.e. sites on proteins that have been examined for ubiquitination, but were not observed to be modified, are assumed to never be modified).
- Your task is to identify sites which are ubiquitinated.
- After applying CD-HIT to remove nearly identical sequence windows (sites), we have 16897 positive and 102930 negative sites
- Application of ProtDCal to each site results in 19584 computational descriptors (features) for each site. Filtering by information gain, this number was reduced to 435 features. You have been provided with these 435 features for each site.
- We have withheld 20% of the total data as a blind test set. The labels of these data will never be released.
- Of the remaining 80% of data, you have been provided with 30% labelled training data
- Leaves 70% unlabelled. You will only receive the feature data for these samples.
- The feature data is available in a ZIP file on the CULearn course website.
- We have created an all-knowing oracle at http://bioinf.sce.carleton.ca/5405-ORACLE. You can request to label up to 1000 unlabelled sites. You can make as many requests as you like, so long as the total number of sites requested does not exceed 1000, summed over all requests.

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 active learning strategies for maximizing your classification accuracy by selecting samples to be labelled.
- 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 select a subset of the 435 features provided to you.
    - 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. "the minimum of my sensitivity and precision 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 the blind test dataset
    - discuss expected performance both with & without meta-learning, but ultimately choose 1 approach and 1 estimate
    - Include a **precision-recall curve** in your presentation
  - The blind test data is released. Keep in mind the size of the dataset (we have held back 20% of the total data)… Beware of runtime issues (you have 24 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 accuracy (min(Sn,Pr))
    - 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.