HW2 - Drug Activity Prediction

Published Date:
26 Sep 2024, 11:06 a.m.

Deadline Date:
28 Oct 2024, midnight

Description:

Description:

****

This is an individual assignment with maximum size of 1

****

Overview and Assignment Goals

The objectives of this assignment are the following:

- Use/implement a feature selection/reduction technique.
- Experiment with various classification models.
- Think about dealing with imbalanced data.
- Use F1 Scoring Metric

Detailed Description

Develop predictive models that can determine given a particular compound whether it is active (1) or not (0).

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.).

The goal of this competition is to allow you to develop predictive models that can determine given a particular compound **whether it is active (1) or not (0).** As such, the goal would be develop the best binary classification model.

A molecule can be represented by several thousands of binary features which represent their topological shapes and other characteristics important for binding.

Since the dataset is imbalanced the scoring function will be the F1-score instead of Accuracy.

Caveats:

+ Remember not all features will be good for predicting activity. Think of feature selection, engineering, reduction (anything that works)

+ The dataset has an imbalanced distribution i.e., within the training set there are only 78 actives (+1) and 722 inactives (0). No information is provided for the test set regarding the distribution.

+ Use your data mining knowledge learned till now wisely to optimize your results.

Data Description

The training dataset consists of 800 records and the test dataset consists of 350 records. We provide you with the training class labels and the test labels are held out. The attributes are binary type and as such are presented in a sparse matrix format within train.dat and test.dat

Train data: Training set (a sparse binary matrix, patterns in lines, features in columns: the index of the non-zero features are provided with class label 1 or 0 in the first column).

Test data: Testing set (a sparse binary matrix, patterns in lines, features in columns: the index of non-zero features are provided).

Format example: A sample submission with 350 entries randomly chosen to be 0 or 1.

Rules

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files, source code, or writing will result in honor code violation.
- Feel free to use the programming language of your choice for this assignment.
- While you can use libraries and templates for dealing with this problem, remember implementation is 70% of the grade. There should still be programming needed even if you choose to use existing packages. You should be able to explain the methods and their choice in sufficient detail.
- Implementation will be graded based on the quality of your code, the amount of effort put in for classifier/model selection, scalability, etc. You are required to try **Decision Tree or Naïve Bayes**, whichever gets you a better F1 score. You are **not** allowed to use other classifiers. Justify the choice of your method via experiments and report the results using **tables**. Submit your best predictions. Summarize your findings in the report.
- Your results should be reproducible. If we find that we cannot reproduce your results, or if the description in your report does not match what your code does, you will receive penalty on the assignment, and this may result in honor code violation.
- You are allowed 10 submissions in a 24 hour cycle.

Deliverables

- Valid Submissions to the Miner.vsnet.gmu.edu website
- **GradeScope Submission of Source Code, Canvas Submission of Report**
    - Create a folder called HW2_netid
    - Create a subfolder called src and put all the source code there.
    - Create a subfolder called Report and place a 2-Page, single-spaced report describing details regarding the steps you followed for feature selection and classifier model development. Also report your experimental results from different classifiers/models, including the running times. Be sure to include the following in the report:

    1. Name Registered on miner web-site.
    2. Rank & F1 score for your submission (at the time of writing the report).
    3. Your Approach
    4. Your methodology of choosing the approach and associated parameters.

Files:

- *Train Data:* Download File (/files/uploaded_files/1727363177_5546074_1708449772_9574642_train_data.txt)
- *Test Data:* Download File (/files/uploaded_files/1727363177_5640502_1708449772_970008_test_data.txt)
- *Format File:* Download File (/files/uploaded_files/1727363177_568396_1708449772_9768624_format_example.txt)

rangwala at cs.gmu.edu (mailto:rangwala@cs.gmu.edu)

GEORGE MASON UNIVERSITY