2
Trackers
Google Analytics
Google Tag Manager

note **25** views

HW2: Drug Activity Prediction

Available on miner.vsnet.gmu.edu via VPN or on Campus

HW2: Drug Activity Prediction

Published Date:
Sept. 18, 2016, 8:49 p.m.
Deadline Date:
Oct. 3, 2016, 4:30 p.m.
Description:
Description:
****

This is a team assignment with maximum size of 2. Deadline is 10/03/2016 4:29 PM EST.
****

Overview and Assignment Goals:

The objective of this assignment are the following:

- Use/implement a feature selection/reduction technique.
- Experiment with various classification models.
- Think about dealing with imbalanced data.
- 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 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_drugs.data: Training set (a sparse binary matrix, patterns in lines, features in columns: the number 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 number of non-zero features are provided).
example_entry.csv: A sample submission with 350 entries randomly chosen to be 0 or 1.

Rules:

- This is an team assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes 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. However, you should be able to explain these methods and their choice in sufficient detail.
- You are allowed 5 submissions in a 24 hour cycle.

Deliverables:

- Valid Submissions to the Miner.vsnet.gmu.edu website
- **Blackboard Submission of Source Code and Report:**
-
  - Create a folder called HW2_LastName1_LastName2
  - 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. Be sure to include the following in the report:

1.
   1. Team Name(s) 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.
   - Archive your parent folder (.zip or .tar.gz) and submit via Blackboard for HW1.

Grading:

Grading for the Assignment will be split on your implementation (50%), report (20%) and ranking results (30%).

hw2

Updated 5 days ago by Huzefa Rangwala

followup discussions *for lingering questions and comments*

● Resolved ○ Unresolved

Shane Armstrong 5 days ago
Hey Everyone,

Since this is a possible team assignment, I was wondering if anyone is looking for a partner. I would really enjoy working with someone else to gain insight into someone else's perspective on approaching the problem. I would like to work with someone that could meet up face to face, preferably on the weekends or after class. Let me know if you're interested in working with someone else as well. I plan on starting on this later on this week.

- Shane

Darron Fuller 4 days ago Shane,

I am planning on coding my own implementation but I am interested in discussing approach. I have an event on campus from 8am to 3pm and could possibly meet afterward.

Darron

● Resolved ○ Unresolved

Nathan Obert 5 days ago
Good morning, I think I remember miner needed a code update. Just curious if it's updated and happy. Thx

nathan

Huzefa Rangwala 5 days ago Yes. it's all set

● Resolved ○ Unresolved

Nathan Obert 4 days ago
Hi, I had a few questions on F1 Score and Binary Classification. I totally understand we are looking specifically for the positive entries however...

Questions:

#1 Are we penalized for false positives? (declaring a negative is positive?) -- couldn't we just make everything positive?

#2 Are we penalized for false negatives? (declaring a positive as negative?) -- I'm assuming this is super important

I thought I remember hearing something about not carrying about the negative ones, but it seems like you have to care about them if you we are to avoid false negatives and/or false positives.

I'm cross validating my data, and depending on how i tune things i see different values for either positives identified %, or overall accuracy % or even false positives or false negatives and i'm just trying to figure out how i'm trying to fit everything.

Thanks!

-Nathan

Nathan Obert 4 days ago https://youtu.be/fcO9820wCXE So is this what we are doing? I think video is helpful

Huzefa Rangwala 4 days ago does that answer your good questions?

Huzefa Rangwala 4 days ago >I'm cross validating my data, and depending on how i tune things i see different values for either positives identified %, or overall accuracy % or even false positives or false negatives and i'm just trying to figure out how i'm trying to fit everything.

You should use F1 score while cross validating.

● Resolved ○ Unresolved

Kyle Jackson 3 days ago
Out of curiosity, is it correct to say that the specific receptor has been identified and we are now trying to predict which compounds bind to the receptor based on known compounds and their activity?

Huzefa Rangwala 3 days ago Yes that's correct