**Project (Ⅰ): Data Mining Practice on Real World Task**

* Task Description
  + In this project, you will be asked using data mining tools to analyze an Author Citation Networks dataset. The data set can be used for clustering with network and side information, studying influence in the citation network, finding the most influential papers, topic modeling analysis, etc. The lists of the files used in train/test can be downloaded in attached files.
  + Implement an algorithm classify spam. Use your algorithm to learn from the training set and report accuracy on the test set.
* Dataset
  + This dataset is designed for research purpose only. The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. The first version contains 629,814 papers and 632,752 citations. Each paper is associated with abstract, authors, year, venue, and title.
  + Data Description

For each block, each line starting with a specific prefix indicates an attribute of the paper. More specifically,

#\* --- paperTitle  
#@ --- Authors  
#t ---- Year  
#c  --- publication venue  
#index 00---- index id of this paper  
#% ---- the id of references of this paper (there are multiple lines, with each indicating a reference)  
#! --- Abstract

The following is an example:

#\*Information geometry of U-Boost and Bregman divergence  
#@Noboru Murata,Takashi Takenouchi,Takafumi Kanamori,Shinto Eguchi  
#t2004  
#cNeural Computation  
#index436405  
#%94584  
#%282290  
#%605546  
#%620759  
#%564877  
#%564235  
#%594837  
#%479177  
#%586607  
#!We aim at an extension of AdaBoost to U-Boost, in the paradigm to build a stronger classification machine from a set of weak learning machines. A geometric understanding of the Bregman divergence defined by a generic convex function U leads to the U-Boost method in the framework of information geometry extended to the space of the finite measures over a label set. We propose two versions of U-Boost learning algorithms by taking account of whether the domain is restricted to the space of probability functions. In the sequential step, we observe that the two adjacent and the initial classifiers are associated with a right triangle in the scale via the Bregman divergence, called the Pythagorean relation. This leads to a mild convergence property of the U-Boost algorithm as seen in the expectation-maximization algorithm. Statistical discussions for consistency and robustness elucidate the properties of the U-Boost methods based on a stochastic assumption for training data.

* Evaluation Rule

You are supposed to submit a report to describe your method(s), experiment details and the results. Both of the performance of your models and the quality of the report will be considered when grading your project. Of course, good performance might lead to your high score. However, the score of your assignment does not depend only on this. A good report of your work is also important for high score. In a word, **you should put your effort into both the performance of your models and the writing of your report.**

**Project (Ⅱ): Data Mining Practice on Real World Task**

* Task Description
  + In this project, you will be asked using data mining techniques to predict disease classes using genetic microarray data.
* Dataset
  + Gene data is in genes-in-rows format, comma-separated values. Take genetic\_microarray\_data.zip file from the files, and unzip to extract 3 files:
    1. pp5i\_train.gr.csv (training data, 1.7 MB)
    2. pp5i\_train\_class.txt (training data classes)
    3. pp5i\_test.gr.csv (test data, 0.6MB)
  + Instructions

1. Training data: file pp5i\_train.gr.csv, with 7070 genes (no Affy controls) for 69 samples. A separate file pp5i\_train\_class.txt has classes for each sample, in the order corresponding to the order of samples in pp5i\_train.gr.csv. There are 5 classes, labelled EPD, JPA, MED, MGL, RHB.
2. Test data: file pp5i\_test.gr.csv, with 23 unlabelled samples and same genes. You can assume that the class distribution is similar.
   * Your goal is to learn the best model from the training data and use it to predict the label (class) for each sample in test data.

* Important Hints
  + Be sure that you don't use the sample number as one of the predictors. Training data is ordered by class, so sample number will appear to be a good predictor on cross-validation, but it will not work on the test data!
  + One of the MED samples in the training data is very likely misclassified (by a human). So the best result you can expect to get on cross validation is one error (on a MED sample) out of 69. However, this should not affect your accuracy on the test set (all labels there are correct).
  + You can complete the project using only simple steps, but the more advanced steps will give you extra credit and probably higher accuracy.
  + The following steps suggest one way of finding the best model -- you are welcome to make improvements, where you think appropriate.

1. Data Cleaning
2. Selecting top genes by class
3. Find the best classifier/best gene set combination
4. Generate predictions for the test set

* Evaluation Rule

Your trained classifiers will be tested on hold-out test sets. The***average accuracy*** over the task will be used to evaluate your classifiers. You are also supposed to submit a report to describe your method(s), experiment details and the results. Both of the performance of your classifiers and the quality of the report will be considered when grading your project. Of course, good performance might lead to your high score. However, the score of your assignment does not depend only on this. A good report of your work is also important for high score. In a word, **you should put your effort into both the performance of your classifier and the writing of your report.**

**Project (Ⅲ): Data Mining Practice on Real World Task**

* Task Description
  + In this project, you will be asked using data mining approaches to analysis the data which is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The goal is to predict the success of telemarketing calls for selling bank long-term deposits. This dataset was provided by a Portuguese retail bank, which wa s collected from 2008 to 2013, thus including the effects of the recent financial crisis.
  + The classification goal is to predict if the client will subscribe a term deposit (variable y).
  + There are issues you need to take into account:
    - Feature Selection, in the dataset there are features that are useless or have too much noises which can lower the prediction result.
* Dataset
  + The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The description of the dataset is as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set Characteristics: | Multivariate | Number of Instances: | 45211 | Area: | Business |
| Attribute Characteristics: | Real | Number of Attributes: | 17 | Date Donated | 2012-02-14 |
| Associated Tasks: | Classification | Missing Values? | N/A | Number of Web Hits: | 503796 |

* + There are four datasets:   
    1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]  
    2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.  
    3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).   
    4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).   
    The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).
  + Attribute Information:
* Input variables:

# bank client data:  
1 - age (numeric)  
2 - job: type of job (categorical: ‘admin.’, blue-collar’, ‘entrepreneur', ‘housemaid’, ‘management','retired', ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’)  
3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)  
4 - education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')  
5 - default: has credit in default? (categorical: 'no','yes','unknown')  
6 - housing: has housing loan? (categorical: 'no','yes','unknown')  
7 - loan: has personal loan? (categorical: 'no','yes','unknown')  
# related with the last contact of the current campaign:  
8 - contact: contact communication type (categorical: 'cellular', 'telephone')   
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')  
10 - day\_of\_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')  
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.  
# other attributes:  
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)  
14 - previous: number of contacts performed before this campaign and for this client (numeric)  
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')  
# social and economic context attributes  
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17 - cons.price.idx: consumer price index - monthly indicator (numeric)   
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)   
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)  
20 - nr.employed: number of employees - quarterly indicator (numeric)

* Output variable (desired target):  
  21 - y - has the client subscribed a term deposit? (binary: 'yes','no')
* Evaluation Rule

Your trained classifiers will be tested on hold-out test sets. The ***average accuracy*** over the task will be used to evaluate your classifiers. You are also supposed to submit a report to describe your method(s), experiment details and the results. Both of the performance of your classifiers and the quality of the report will be considered when grading your project. Of course, good performance might lead to your high score. However, the score of your assignment does not depend only on this. A good report of your work is also important for high score. In a word, **you should put your effort into both the performance of your classifier and the writing of your report.**

**Project (Ⅳ): Data Mining Practice on Real World Task**

* Task Description
  + In this project, you will be asked using data mining techniques to Predict Diabetes at Early Stage.
* Dataset

|  |  |
| --- | --- |
| * + Abstract: This dataset contains the sign and symptpom data of newly diabetic or would be diabetic patient. |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 520 | **Area:** | Computer |
| **Attribute Characteristics:** | N/A | **Number of Attributes:** | 17 | **Date Donated** | 2020-07-12 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 42564 |

* + Data Set Information:

This has been collected using direct questionnaires from the patients of Sylhet Diabetes   
Hospital in Sylhet, Bangladesh and approved by a doctor.  
**Attribute Information:**

Age 1.20-65   
Sex 1. Male, 2.Female   
Polyuria 1.Yes, 2.No.   
Polydipsia 1.Yes, 2.No.   
sudden weight loss 1.Yes, 2.No.   
weakness 1.Yes, 2.No.   
Polyphagia 1.Yes, 2.No.   
Genital thrush 1.Yes, 2.No.   
visual blurring 1.Yes, 2.No.   
Itching 1.Yes, 2.No.   
Irritability 1.Yes, 2.No.   
delayed healing 1.Yes, 2.No.   
partial paresis 1.Yes, 2.No.   
muscle sti  
ness 1.Yes, 2.No.   
Alopecia 1.Yes, 2.No.   
Obesity 1.Yes, 2.No.   
Class 1.Positive, 2.Negative.

**Project (Ⅴ): Data Mining Practice on Real World Task**

* Task Description
  + In this project, you will be asked using data mining techniques to classify the graph data.
  + Paper: <https://arxiv.org/abs/2003.04819>  
    Github Page: <https://github.com/benedekrozemberczki/karateclub>
* Dataset

The ego-nets of Eastern European users collected from the music streaming service Deezer in February 2020. Nodes are users and edges are mutual follower relationships. The related task is the prediction of gender for the ego node in the graph.

|  |  |  |
| --- | --- | --- |
| Stats | Min | Max |
| Nodes | 11 | 363 |
| Density | 0.015 | 0.909 |
| Diameter | 2 | 2 |

**Below are specific requirements for your submission.**

* Requirement and Submission Guidelines
  + You are supposed to submit an **executable file** along with the **source files** and **a report** (compressed into a .zip or .rar file and named after your student number). If the parameters of your trained classifier are stored in separate files, please also include these files in the compressed file. Note that the training phase is up to you, so please make sure that the submitted classifier is the result of your training. **What is wanted is a trained classifier, not a classifier trainer.**
* How to writing a report
  + Your submission must be prepared in the [Chineses Journal of Software](http://www.jos.org.cn/modelofpaper.htm) paper style (in Chinese), and must be no longer than 6 pages in length, including all figures, tables, references, etc.
  + Your submission should follow the generally accepted style of paper writing: include an introduction section to motivate your problem and algorithm, a section describing your method and how it compares to other work, a section outlining the experiments you ran and the results you obtained (a analysis section is encouraged), and a short conclusions section to sum up what you discovered.
  + Your submission must include at least 5 references to previous published papers or book sections. Please include page numbers for all references to indicate that you actually saw the paper you are referencing even if you didn't read it very carefully.
  + A good report should consider :
    - Clarity of problem statement and description of approach
    - Discussion of relationship to previous work and references
    - Design and execution of experiments
    - Analysis of the results
    - Figures/Tables/Writing: easily readable, properly labeled, informative

**Project (Ⅵ): Data Mining Practice on Real World Task**

* Task Description
  + Given a set of users and their reviews of items, recommendation systems generate ranked lists of items to recommend to individual users. In this project, you will be asked to propose an algorithm for generating product recommendations using Amazon dataset.
* Dataset
  + This dataset was collected by crawling Amazon website and contains product metadata and review information about 548,552 different products (Books, music CDs, DVDs and VHS video tapes).
  + Data Description

For each product the following information is available:

* Title
* Salesrank
* List of similar products (that get co-purchased with the current product)
* Detailed product categorization
* Product reviews: time, customer, rating, number of votes, number of people that found the review helpful

|  |  |
| --- | --- |
| Dataset statistics | |
| Products | 548,552 |
| Product-Project Edges | 1,788,725 |
| Reviews | 7,781,990 |
| Product category memberships | 2,509,699 |
| Products by product group |  |
| Books | 393561 |
| DVDs | 19828 |
| Music CDs | 103144 |
| Videos | 26132 |

Data format:

* **Id:** Product id (number 0, ..., 548551)
* **ASIN**: Amazon Standard Identification Number
* **title:** Name/title of the product
* **group:** Product group (Book, DVD, Video or Music)
* **salesrank:** Amazon Salesrank
* **similar:** ASINs of co-purchased products (people who buy X also buy Y)
* **categories:** Location in product category hierarchy to which the product belongs (separated by |, category id in [])
* **reviews:** Product review information: time, user id, rating, total number of votes on the review, total number of helpfulness votes (how many people found the review to be helpful)
* Evaluation Rule

You are supposed to submit a report to describe your method(s), experiment details and the results. Both of the performance of your models and the quality of the report will be considered when grading your project. Of course, good performance might lead to your high score. However, the score of your assignment does not depend only on this. A good report of your work is also important for high score. In a word, **you should put your effort into both the performance of your models and the writing of your report.**