## Project

**Due: May. 28, 2016.**

Choose one topic from the following topics.

Up to 15 groups per topic are allowed, 4 students per group. First come first serve. **Please send the members of your group and the topic you choose to** [**datamining2016@163.com**](mailto:datamining2016@163.com) **as soon as possible.**

**Submission:**

* Source code of your algorithm. **Implement it in C/C++ (recommended)**.
* Project report (including the interpretation of the algorithm, your implementation details, evaluation strategy and performance analysis results). Algorithm optimization or improvement is a plus.
* Please upload your source code and project report to our class website.

Your work will be evaluated by:

1. Accuracy
2. Scalability
3. Efficiency
4. Collaboration
5. Novelty

#### Classification for Physical Particles

The goal in this task is to propose a classification method that differentiates between two types of particles generated in high energy collider experiments. It is a binary classification problem with 78 attributes. The training set has 33333 examples, and the test set is of size 16667. Your task is to provide a prediction for the test set that optimizes the accuracy.

\*\*\* Accuracy:

We use the usual definition of accuracy -- the number of cases predicted correctly, divided by the total number of cases. An accuracy of 1.00 is perfect prediction. Accuracy near 0.00 is poor.

**Dataset**

* phy\_train.dat: Training data for the quantum physics task (33,333 train cases)
* phy\_test.dat: Test data without class labels for the quantum physics task (16, 667 test cases)

Each line in the training and the test file describes one example. The structure of each line is as follows:

* The first element of each line is an EXAMPLE ID that uniquely describes the example. You will need this EXAMPLE ID when you submit results.
* The second element is the class of the example. Positive examples are denoted by 1, negative examples by 0. Test examples have a "?" in this position. This is a balanced problem so the target values are roughly half 0's and 1's.
* All following elements are feature values. There are 78 feature values in each line.
* Missing values: columns 22, 23, 24 and 46, 47, 48 use a value of "999" to denote "not available", and columns 31 and 57 use "9999" to denote "not available". These are the column numbers in the data tables starting with 1 for the first column (the case ID numbers). If you remove the first two columns (the case ID numbers and the targets), and start numbering the columns at the first attribute, these are attributes 20, 21, 22, and 44, 45, 46, and 29 and 55, respectively. You may treat missing values any way you want, including coding them as a unique value, imputing missing values, using learning methods that can handle missing values, ignoring these attributes, etc.

The elements in each line are separated by whitespace.

1. **Coauthor Relationship**

The DBLP data set (www.informatik.uni-trier.de/~ley/db) consists of over one million entries of research papers published in computer science conferences and journals. Among these entries, there are a good number of authors that have coauthor relationships. Please propose a method to efficiently mine a set of coauthor relationships that are closely correlated (e.g. often coauthoring papers together). In addition, please propose a measure which can convincingly uncover close collaboration patterns based on the mining results and the pattern evaluation measures, such as support, confidence, lift, etc.

1. **学习意图感知及推荐系统**

##### 意义和作用

信息的时代使我们有机会享受丰富的电子教育资源。但是，人们又正在面临着另一个严重的问题——信息过载。每个人可以接触到的课程资源、学术论文、学位论文、电子图书、期刊杂志太多太多，远远超过了个人可以分析并从中获取知识的能力。由于研究生处在科学研究的起步阶段，“被淹没在知识的海洋里”的感觉尤为强烈，往往不知道读什么、怎么读。结果，浪费了大量的时间和精力，走了很多的弯路。

如何能够有目标地、有选择地、正确地利用这些资源是一个亟待解决的问题。目前研究生主要采用在网络上搜索相关论文的方法进行学习，随意性高，质量难以预测，而且费时费力。而目前流行的推荐系统，大多依靠用户的输入或者长期的兴趣，不能及时准确地感知学生的学习意图。因此，学习意图感知和推荐系统是提高研究生培养水平的一个重要手段。

##### 系统目标

此系统通过信息检索和数据挖掘等技术将学术资源进行有效地管理，利用数据挖掘技术捕获或预测每位学习者的学习意图和兴趣，主动地向学习者推荐与其兴趣相关的科教资源。这样，可以提高信息的到达率，提高扩展阅读的效果，提高学习者在网络环境下学习的质量。

##### 系统功能

初次登录的新用户：使用本系统的浏览器，填写研究兴趣,系统主动从学术资源库中推荐相关的学术信息(包括学术论文、课件、图书、毕业论文、科研报告等)。

老用户：通过对用户历史行为的分析，预测用户当前的兴趣偏好，主动从资源库中推荐相关的学术信息(包括学术论文、课件、图书、毕业论文、科研报告等)。

4. **Social Media Mining**

The goal is to apply the state-of-the-art network analysis tools and algorithms to an application. Some interesting topics:

Social network analysis

Information diffusion in social media

Recommendation system 

Community detection

A scalable algorithm needs to be implemented for processing massive graphs. Experimental evaluation will be performed on an interesting dataset. A novel algorithm or model proposed will be considered as a plus.

This topic will be evaluated based on: 

1. The technical quality of the work: Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insight about the problem and/or algorithms? 
2. Significance: Did the authors choose an interesting or a "real" problem to work on, or only a small "toy" problem? Is this work likely to be useful and/or have impact? 
3. The novelty of the work, and the clarity of the write-up.