官网：https://www.kddcup2015.com/information-introduction.html

帐号：[jiaming.xu@ia.ac.cn](mailto:jiaming.xu@ia.ac.cn) 密码：

【任务】：根据学生之前的行为，预测学生的一个注册ID在未来十天是否会翘课？

【评估】：预测学生翘课的概率，在[0,1]区间内。1 是翘课，0 是未翘课。评估指标：ROC 曲线（AUC）。

【时间】：2015/05/01 - 2015/05/15: Review Problem and Download Data

2015/06/23: Team Merger Deadline

2015/06/29: First Submission Deadline

2015/07/05: End Date

2015/07/10: Winner Announcement

【数据分析】：训练120542 / 测试80362 = 3：2

1. 需要预测的注册ID是由学生ID和课程ID一起生成的，因而每个注册ID只会对应一个学生参加的一门课程。
2. 数据集结构说明

**\*\* train.7z \*\* 训练语料：**

* **enrollment\_train.csv**：每行包括一个注册ID，学生ID和课程ID
  + enrollment\_id – Enrollment ID，注册ID
  + username - Student ID，学生ID
  + course\_id - Course ID，课程ID
* **log\_train.csv**：每行记录一个行为事件，具体内容如下：
  + enrollment\_id - Enrollment ID，注册ID
  + time - Time of the event，事件时间戳
  + source - Event source (server or browser)，时间客户端类型，Server
  + event - In terms of event type, we defined 7 different event types，7种类型，具体如下：
    - 1. problem - Working on course assignments，做作业
    - 2. video - Watching course videos，是否在看视频
    - 3. access - Accessing other course objects except videos and assignments，登录（不包括作业及视频）
    - 4. wiki - Accessing the course wiki，访问WIKI
    - 5. discussion - Accessing the course forum，访问课程论坛
    - 6. navigate - Navigating to other part of the course，浏览课程的其他部分
    - 7. page\_close – Closing the web page，关闭页面
  + **object - The object the student access or navigate to.(For navigate and access event only). 访问的实体ID**
* **true\_trian.csv** – each line contains information about the ground truth of enrollments in the training set.
  + 1st column - Enrollment ID，注册ID
  + 2nd column - Ground truth of dropout (1 for a dropout event and 0 for continuing study)，是否翘课

**\*\* test.7z \*\* 测试语料：**

* test.7z contains testing data. For each file in this archive, the format is the same as the counterpart in the train.7z.

要注意，只是根据注册ID区分的训练和测试，因而有可能同一个学生ID选的不同的课程ID，而被分别分到训练和测试中去

**\*\* object.csv.7z \*\*** 对应event事件中访问的一些课程模块ID及属性，如模块类型、子章节、发布时间、等。

* course\_id - The course to which the module belongs. 模块所属的课程ID
* module\_id - The ID of a course module. 模块ID，每个课程下的模块组件是树结构的，如（课程->章节->视频）
* category - The category of the course module. 模块所述类别
* children - The children modules of the course module. 模块下的子模块
* start - The time that the module was released to students. 模块发布时间

**\*\* sampleSubmission.csv.7z** **\*\*** 一个提交示例

* 1st column- Enrollment ID. 注册ID
* 2nd column - A real-valued probability of dropout. 一个实值的逃课概率

目录

[**零：编程环境** 3](#_Toc421103851)

[**一：特征抽取（有些特征需要测试样本和训练样本 一同处理）** 3](#_Toc421103852)

[**二：模型选择** 4](#_Toc421103853)

[a) 随机森林 4](#_Toc421103854)

[b) 逻辑回归 4](#_Toc421103855)

[a) 【第一梯队特征】 4](#_Toc421103856)

[[MultiNomial LR] Accuracy is 0 4](#_Toc421103857)

[[MultiNomial LR] The area under the ROC curve (AUC) is **0.85489** 4](#_Toc421103858)

[b) 【第二梯队特征】后的交叉验证结果： 4](#_Toc421103859)

[[MultiNomial LR] Accuracy is 0 4](#_Toc421103860)

[[MultiNomial LR] The area under the ROC curve (AUC) is **0.85852** 4](#_Toc421103861)

[c) 支持向量机SVM 4](#_Toc421103862)

[**三：结果及后处理：** 4](#_Toc421103863)

[a) 特征选择，逐个剔除特征，查看性能是否提升； 4](#_Toc421103864)

[b) 设置阈值 [0.001:0.001:0.050]； 4](#_Toc421103865)

[c) 考虑多模型融合是否会提升性能； 4](#_Toc421103866)

[**九八：备注材料**： 5](#_Toc421103867)

[**九九：问题** 5](#_Toc421103868)

**零零：编程环境**

编程语言：Java，Matlab

编译环境：JDK 1.7，Matlab 2011b

**零：数据分析**

1. 总共/训练集合多少学生，总共/训练集合多少门课，训练/测试集合Log的活动时间范围，所有课程Object的发布时间范围；
2. 每门课的退学率、Log活动时间范围、Object的发布时间范围；

course\_id, log\_start, log\_end, object\_start, object\_end;

1 1 1

1 0 0

0 1 1

Object

总共多少人，多少门课，所有的时间范围，每门课的时间分布

**一：特征抽取（有些特征需要测试样本和训练样本 一同处理）**

【第一梯队特征】以下特征来自log\_train.csv和log\_test.csv文件：

* 1. ~~注册ID的活跃时间起始（单位：秒）、终止（单位：秒）及跨度（单位：秒）、（单位：天）~~

其中，时间起始和终止时刻不参与特征决策。

* 1. 注册ID的活跃次数（单位：次）
  2. 注册ID的活跃天数（单位：天）及平均每天的活跃次数（单位：次/天）及平均活跃天占比（%百分比）
  3. 注册ID访问Problem（问题）的频次（单位：次）及比重（%百分比）
  4. 注册ID访问Video（问题）的频次（单位：次）及比重（%百分比）
  5. 注册ID访问Access（问题）的频次（单位：次）及比重（%百分比）
  6. 注册ID访问Wiki（问题）的频次（单位：次）及比重（%百分比）
  7. 注册ID访问Discussion（问题）的频次（单位：次）及比重（%百分比）
  8. 注册ID访问Navigate（问题）的频次（单位：次）及比重（%百分比）
  9. 注册ID访问Page\_close（问题）的频次（单位：次）及比重（%百分比）

以上特征最好结果为0.843，而且由于每个注册ID的特征抽取相互独立，因而验证结果与测试结果完全一致！

【第二梯队特征】以下特征同时来自enrollment\_train.csv、truth\_train.csv和enrollment\_test.csv文件：

* 1. ~~注册ID所属用户的总选课数（单位：门数）~~
  2. ~~注册ID所述课程的选课总人数（单位：人）~~
  3. ~~注册ID所属用户的逃课次数（单位：次）及翘课率（%百分比）~~
  4. ~~注册ID所选课程的逃课人数（单位：人）及翘课率（%百分比）~~

以上特征最好结果为0.598，而且验证结果出现了过拟合至0.971！因而尝试剔除一些特征，仅保留13). 还是不行。因而需要修改策略。修改方案如下：

1. 既然【第二梯队特征】同时涉及到了训练和测试信息，因而我们抽选验证集时也要进行拆分，信息剔除。生成两份特征，一份是专门做验证的，一份是用于测试的，以使得验证集合的测试节诶过能够复现最终测试结果！复现结果显式：前俩特征略微有效，因而保留。
2. 所有课程或用户的逃课率初始值为0.5或一个均值，然后根据逃课记录对基准分进行调整，避免了验证/测试中出现的新课程或用户出现逃课率为0的情况。需要先抽取一个基准分。
3. 注册ID所属用户的当前总选课数（单位：门数）
4. 注册ID所述课程的当前选课总人数（单位：人）
5. ~~注册ID所属用户的旷课评估值（%百分比，默认0.7929），根据旷课记录进行调整~~

double baseDropoutScore = 0.792926946624413;

public float getDropoutCourseScore(){

if (hasDropoutRecord) {

return (float) (baseDropoutScore

+ (getDropoutCourseRate()-baseDropoutScore)\*Math.tanh((double)getDropoutCourseNum()/scaleFactor));

}else {

return (float) baseDropoutScore;

}

}

1. 注册ID所选课程的旷课评估值（%百分比，默认0.7929），根据旷课记录进行调整。计算方法同13).

结果显示，用户的旷课记录特征引入会降低得分，因而在分类决策中不予考虑。

【第四梯队特征】以下特征来自object.csv文件

注册ID活跃的实时延迟性（距离发布的延迟时间）

注册ID活跃距离所选课程最后一次Object发布时间

**二：模型选择**

1. 随机森林
   1. 非概率输出交叉验证结果：

【第一梯队特征】

[Random Forest] Accuracy is 86.2494

[Random Forest] The area under the ROC curve (AUC) is **0.73889**

* 1. **概率输出交叉验证结果：**

【第一梯队特征】

[Random Forest\_P] Accuracy is 0.014933

[Random Forest\_P] The area under the ROC curve (AUC) is **0.84219**

提交结果：0.8427510895561884，可以看到测试和最终结果基本一致。

1. 逻辑回归
   1. 【第一梯队特征】

[MultiNomial LR] Accuracy is 0

[MultiNomial LR] The area under the ROC curve (AUC) is **0.85489**

* 1. 【第二梯队特征】后的交叉验证结果：

[MultiNomial LR] Accuracy is 0

[MultiNomial LR] The area under the ROC curve (AUC) is **0.85852**

1. 支持向量机SVM
   1. 线性SVM交叉验证结果：

【第一梯队特征】

[SVM-Linear] Accuracy is 86.1548

[SVM- Linear] The area under the ROC curve (AUC) is **0.72438**

【第二梯队特征】后的交叉验证结果：

[SVM-Linear] Accuracy is 86.2494

[SVM-Linear] The area under the ROC curve (AUC) is 0.72633

**三：结果及后处理：**

1. 特征选择，逐个剔除特征，查看性能是否提升；
2. 设置阈值 [0.001:0.001:0.050]；
3. 考虑多模型融合是否会提升性能；

**九八：备注材料**：

1. NIPS Workshop on Data Driven Education (2013) <http://lytics.stanford.edu/datadriveneducation/>

里面有很多Mooc相关的大数据分析论文可以参考。

1. KDD的一些经验博文：

KDD Cup2012亚军简单回顾<http://blog.csdn.net/cserchen/article/details/7657152>

KDDCup 2011回忆录-来斯惟<http://licstar.net/archives/89>

The Art of Lemon队的KDD CUP 2011 Track 2解决方案-刁瑞<http://diaorui.net/archives/184>

1. KDD历年竞赛官网

**九九：问题**

1. 试试RNN，查看2015/05/20微博：<http://t.cn/R2y10ej>
2. 特征输入每个注册ID对发布内容的响应时间
3. 特征13). 虽然在特征分类中失效，但是考虑后处理，利用评估得分进行后处理，如，大于0.9则认为逃课，小于0.1则认为不会逃课。