时间序列分析数据集

资料来源 ：网络资料

The Time Series Analysis Task

A company is interested in detecting when an account goes ' bad ' . If they do not catch bad behavior early, the company will be liable for the additional damage. However, if the company unjustly cuts off a customer account, they stand to lose revenue.

Some accounts start out as bad, some never go bad, and some go bad over time.

You are given the time series of records for accounts, and must determine when an account goes ' bad ' .
That is, each record has a binary label - 0 for ' good ' , 1 for ' bad ' .
Once there is a record with a ' bad ' label, the remaining records in time series are also bad.

Your tasks reduces to finding the first record with a ' bad ' label.

The following sets of data are available for the Classification Task:
They are .zip files, so use Winzip on Windows, or unzip (not gunzip) on Linux.

Data Set (.zip files)	File Size (uncompressed)	# Accounts	Data Features Contained in Each Record			
			AccountID	RecordID	Features 3-41	Record Label
Training Set	101 MB (678 MB)	2,528	X	X	X	X
Quiz Set	50 MB (330 MB)	1,265	X	X	X	
Test Set	49 MB (324 MB)	1,265	X	X	X	
Time Series Quiz Set Labels	5.0 KB	1,265				X
Time Series Test Set Labels	5.0 KB	1,265				X

Note that these data sets are the same for both the classification task and the time series task.

Training Set

You are given account information (i.e. one time series of records) for each of 2,528 accounts.

The training set has 42 tab-delimited columns.
The first two columns are the account ID and record ID.
The next 39 columns are features that take on real, integer, and Boolean values.
The last column is the binary record label: 0 for good, 1 for bad.

The data looks something like this:

```
1     1    -0.5670.412  ... 37 additional features ...0
1     2    0.735  1.041  ... 37 additional features ...1
...
2     1    0.277  -0.987... 37 additional features ...1
2     2    0.871  1.923  ... 37 additional features ...1
...
25286760.021  -2.123... 37 additional features ...0
```

Test Set

You are given account information (i.e. one time series of records) for each of 1,265 test accounts.

The test set has 41 tab-delimited columns.
The first two columns are the account ID and record ID.
The next 39 columns are features that take on real, integer, and Boolean values.

However, the data is missing the record label columns.

The Final Test Set looks something like this:

```
1     1     -0.1471.37    ... 37 additional features ...
1     2     0.045  1.001  ... 37 additional features ...
...
2     1     0.923  0.841  ... 37 additional features ...
2     2     0.817  -0.221... 37 additional features ...
...
126510121.451  -0.950... 37 additional features ...
```

Quiz Set

The quiz set is just for fun.
It is only here to give you feedback on your system. It does, however,
have the same format and number of examples as the real final set.

You can submit answers to the quiz set on the submission page
and see your scores immediately on the leaderboard page.

Submitting Answers

Your job is to create a text file containing exactly 1,265 lines where each line contains one integer between 1 and 10,050. This integer represents your prediction for the first record id where the account goes ' bad ' .

Your solution file should look some thing like:
103
756
1
...
982

This file should then be submitted here.

Scoring Answers

Evaluation Metric: Average Squared Error.

For each test account, we will calculate the average squared distance between your estimated time (i.e. record id) and the true time at which the bad behavior first occurs.
The equation is given by:

Error(guess , correct) = mean_i [(guess_i - correct_i) ^ 2]

where *guess* is a vector of your predictions and *correct* is a vector of the true record ids of the first ' bad ' labels.

If the account is good (i.e. the time series has no records with a ' bad ' label), then the true value will be set to the last record id plus 1. For example, if a good account has 676 records, then the true value will be 677. Also, it the account starts out as ' bad ' (i.e. the first record is ' bad '), the true value will be 1.

Scoring Example

For example suppose you submit the following answer file for 10 test examples:

504
677
25
214
499

8024
1012
600
1
871

The total number of records for each account and the record id where the first bad label occurs are shown in the second and third columns of the following table.

Your Prediction	# Records	First Bad Record	Score
504	860	861	127,449
677	676	677	0
25	43	1	576
214	318	241	729
499	500	501	4
8024	8050	8051	729
1012	1011	1012	0
600	921	522	6,084
1	25	1	0
871	871	872	1
Average Squared Distance:			13,557

In our example, the accounts 1,2,5,6,7, and 10 are all ' good ' accounts and thus have a true value of one plus the last record id.
Also note that accounts 3 and 9 start out as bad.