Quora Answer Classifier

Quora uses a combination of machine learning (ML) algorithms and moderation to ensure high-quality content on the site. High answer quality has helped Quora distinguish itself from other Q&A sites on the web.

Your task will be to devise a classifier that is able to tell good answers from bad answers, as well as humans can. A good answer is denoted by a +1 in our system, and a bad answer is denoted by a -1.

Input Format:

The first line contains N, M. N = Number of training data records, M = number of parameters. Followed by N lines containing records of training data. Then one integer q, q = number of records to be classified, followed by q lines of query data
Training data corresponds to the following format:
(:)*

Query data corresponds to the following format:
(:)*

The answer identifier is an alphanumeric string of no more than 10 chars. Each identifier is guaranteed unique. All feature values are doubles.

The data is extracted from real production data, and thus will not include the raw form of the answers. We, however, have extracted as many features as we think are useful, and you can decide which features make sense to be included in your final algorithm. The actual labelling of a good answer and bad answer is done organically on our site, through both human moderators as well as our own classifier.

Output Format:

For each query, you should output q lines to stdout, representing the decision made by your classifier, whether each answer is good or not:

Constraints:

0 0 0
Sample Input:

5 23
2LuzC +1 1:2101216030446 2:1.807711 3:1 4:4.262680 5:4.488636 6:87.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.891820 12:0 13:1 14:0 15:0 16:0 17:1 18:1 19:0 20:2 21:2.197225 22:0.000000 23:0.000000
LmnUc +1 1:99548723068 2:3.032810 3:1 4:2.772589 5:2.708050 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:4.727388 12:5 13:1 14:0 15:0 16:1 17:1 18:0 19:0 20:9 21:2.833213 22:0.000000 23:0.000000
ZINTz -1 1:3030695193589 2:1.741764 3:1 4:2.708050 5:4.248495 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.091042 12:1 13:1 14:0 15:0 16:0 17:1 18:1 19:0 20:5 21:2.564949 22:0.000000 23:0.000000
gX60q +1 1:2086220371355 2:1.774193 3:1 4:3.258097 5:3.784190 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.258097 12:0 13:1 14:0 15:0 16:0 17:1 18:0 19:0 20:5 21:2.995732 22:0.000000 23:0.000000
5HG4U -1 1:352013287143 2:1.689824 3:1 4:0.000000 5:0.693147 6:0.000000 7:0.000000 8:0.000000 9:0 10:1 11:1.791759 12:0 13:1 14:1 15:0 16:1 17:0 18:0 19:0 20:4 21:2.197225 22:0.000000 23:0.000000

2
PdxMK 1:340674897225 2:1.744152 3:1 4:5.023881 5:7.042286 6:0.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.367296 12:0 13:1 14:0 15:0 16:0 17:0 18:0 19:0 20:12 21:4.499810 22:0.000000 23:0.000000
ehZ0a 1:2090062840058 2:1.939101 3:1 4:3.258097 5:2.995732 6:75.000000 7:0.000000 8:0.000000 9:0 10:0 11:3.433987 12:0 13:1 14:0 15:0 16:1 17:0 18:0 19:0 20:3 21:2.639057 22:0.000000 23:0.000000

Sample Output:

PdxMK +1
ehZ0a -1

You will be given a relative large input dataset with its corresponding output to finetune your program with your ML libraries.

Scoring
Only one very large dataset will be given for this problem as input to your program for scoring. This input data set will not be revealed to you.

Output for every classification is awarded points separately. The score for this problem will be the sum of points for each correct classification. To prevent naive solution credit (outputting all +1s, for example), points are awarded only after X correct classifications, where X is number of +1 answers or -1 answers (whichever is greater).