Problem representation:

In my training procedure, I generate seven training attributes from the raw material. In the raw dataset, it contains the following features, event data, event name, event type, time and category. Actually, these data are not directly usable unless we do the data preprocessing. In addition, we can retrieve gender and age directly from the category. Other attributes, such as matchBeforeYear and marathonBeforeYear are derived from event type and time, and each of them represent the number of marathon and match before the prediction year. Also, there is another attribute, which is last year attendance representing exactly whether they attend the last year competition.

In addition, although these data are discrete data, but they haven’t been normalized. As Naïve Bayes needs normalization data, so that I manually normalize the data. For example, I category age from teenager to old, match or marathon before year as 0 or 1 distinguished from whether they attend marathon or match before.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gender | Age | AvgMarTime | MatchBeforeX | MarathonBeforeX | lastYearAttendance |
| 0-1 | 0-4 | 0-3 | 0-1 | 0-1 | 0-1 |

Training Method:

In my part, I use Naïve Bayes algorithm as the classification method, which needs the training data to be categorical.

And I use the Laplace smoothing method to improve the naïve Bayes algorithm in order to deal with the case when we can’t find such X.

The classification attribute is whether they would participate in the next year marathon.

Actually, at the beginning, I use all the attributes that lying above to predict the output. Especially, I categorize matchBeforeX and MarathonBeforeX into 5 types according to the absolute number of matches. It runs an acceptable result on the test set, which given a 72% correct rate on 2015 data. However, when it comes to 2016 data, the outcomes are ridiculous, in which there are only 17 entries having a participate result. Then I try to adjust the attribute better fil the real data. Then it comes out, when I delete the gender feature, the correct rate of the model jump from 72% to 78%. However, the final outcome of 2016 marathon still works terrible.

Then I think about other problems of the features, from 2015 – 2016, most entries would change their value on Match/MarathonBeforeX, however this change may lead to a wrong interpretation. There exists a hidden problem, although the attribute change, however the property of the entry doesn’t change in fact. So I try to make Match/MarathonBeforeX binary attributes, where a change means an actual change.

Result:

The validation error and training error of the algorithm is (7.2698048220436275, 0.7839513259097692).

K-fold validation is used on the validation set, and the number of k is 10.

Discussion:

As Naïve Bayes use categorical data, it has a natural drawback, which cannot trace the increase in matchBeforeYear and marathonBeforeYear accurately as good as continuous data. So that it seems has a worse prediction than logistic regression. The second drawback of this approach is that, most of the output that predicting 1 on 2016 have 0 on 2015. This reason is that people both participate in 2014 and 2015 marathon are quite few, so that when comes to the generalization data, that part of effect seems been offset by other predictions.