Classification of Take-out Orderers by Expectation Maximization Algorithm.

- Topic: Research on UCAS students' online take-out ordering
- Group members: 王华强、刘蕴哲、杨钊、高云聪

Introduction

Recent developments in the mobile Internet have heightened the need for take-out. The fast-growing Big Data technology and the Mobile Internet technology have seen the rapid development of take-out market(iiMedia Research, 2016). According to the report of Institute of Frontier Industry Research in 2016, the whole market of online take-out selling will reach at 118 billion yuan per year at the end of 2017. It is easy to predict that the take-out market will grow even faster in the following few years (iiMedia Research, 2017). In the meantime, college students make a cosiderable contribution to the hypergrowth of take-out. TrustData found a strong relationship between the number of take-out orders and the number of universities, which also suggests that students consumed a large number of online take-out(TrustData, 2017). However, among all the exist studies, no math model of university students' take-out ordering pattern has been developed. Despite many reports focusing on the whole market of take-out, none of them is detailed enough to show the buying pattern on campus. There have also been reports made by other universities' students which are based on rather casual questionnaires and guesses (Li, 2015), which is not enough to reveal the whole pattern of campus online take-out buying.

This paper seeks to find out the pattern of online take-out buying and the relationships between all these factors by analyisng the data obtained from UCAS's Yuquan campus. After the analysis of the results of 107 questionnaires, the result clearly showed two different activity patterns of take-out orderers. Based on the above analysis results, we provide a method to conjecture one's take-out ordering behavior according to his daily routine, with which useful suggestions can be offered for the university's logistics department and the take-out providers.

Methods

Large amount of data is required to support a new behavior pattern. Consequently, an efficient way of data collecting is required. Finally, we adopted the method of online questionnaire. An online questionnaire enables us to directly obtain digital data through the backstage, which is convenient to manipulate especially when the sample scale is large.
In order to obtain an overall conclusion, we tried to cover a wide range of questions in the questionnaire. However, a redundant questionnaire is time-consuming which will lead to a significantly low recovering rate. Consequently, the quantity of questions is supposed to be limited. The contents of the questionnaire are listed as follows:

1. Online ordering frequency
2. Online ordering time preference
3. Online ordering platform
4. Online ordering price range and the proportion to life expenses
5. Main factors considered in online ordering
6. Main reasons for choosing takeout
7. Degree of concern about takeout hygiene
8. Parents' attitude towards takeout
9. Students' general evaluations of takeout.
10. Conditions in which the students will give up ordering online

The options were carefully set that we took many factors into consideration. Take the question about online ordering price range as an example. The lowest discount price of most stores is set in the range of 20-30 yuan in eleme app. A part of the students possibly order what they want only. However, most of the students, in order to use the red envelopes, have to purchase food valued over 35 yuan. The final cost will consequently exceed 25 yuan. As a result, we set the price range as below: <15 rmb/share", "15-25 rmb/share", "25-50rmb/share", ">50 rmb/share". These details enable us to obtain well-directed data.

The object of the survey is the undergraduates in the Yuquanlu campus of the University of Chinese Academy of Sciences (hereinafter referred to as UCAS). In order to include a large number of participants, we posted our online questionnaire on major social platforms including QQ and WeChat. We finally received in total 105 sets of questionnaire answers, of which 103 are valid. 2 responses were eliminated because though the participants confirmed that they ordered take-out, they called for deliveries at a frequency of null. Though

the rest of answers are valid, drawbacks still exist. Due to the feature of online questionnaire, we have no access to the number of people who have viewed the questionnaire. As a result, the recovery rate of the questionnaire remains unknown. In addition, the title of the questionnaire is Take-out in UCAS. We suspect that undergraduates who never conducted online ordering probably overlooked the questionnaire. In conclusion, it is not a perfect sampling of undergraduates in UCAS. Consequently, we are not able to conclude the ratio of take-out users according to the data as intended. In addition, we have to admit that the meaning of options in a few questions were not explicit enough for participants. They can be variously interpreted, which brings us some problems. We also found that some of the questions were poorly related to each other. As a result, it was difficult for us to conclude rules according to the data. We then adopted a powerful tool named EM algorithm.

The Expectation–Maximization (EM) algorithm is able to rule out redundant information. We expected to explore possible relationships among various factors. EM algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables (Wikipedia). It conducts an expectation (E) step and a maximization (M) step in turn repeatedly. To make it simple, we can describe the expectation (E) step as a step to guess the result, while maximization (M) step is a step for the algorithm to check and correct the guess with the data. Through the process, it is able to find a latent pattern and determine the distribution of variables in the pattern.

Result

1. Line of thinking

The analyzation for the gained data can be sorted in to four stages. At the beginning, the data were manually analyzed to see if some basic rules can be find. Secondly, different algorithms were applied in order to find the suitest one for the following research. In this stage, the EM algorithm was proved to the the best one. After that, the EM algorithm was applied directly, and the result was checked manually to see if some improvements can be taken. Finally, the EM algorithm's parameter and the source data were changed according to the result of the last stage, and the result of second trail was taken as the final result.

2. Basic analyzing of the data

We first finished basic analyzing of the data. Having collected enough information, we exported spreadsheet data. Then we made pie charts and bar charts based on the data. Before setting out to find intern relationships among data, we ruled out some invalid information. Some of the questions were not set properly. Consequently, their results have no reference value.

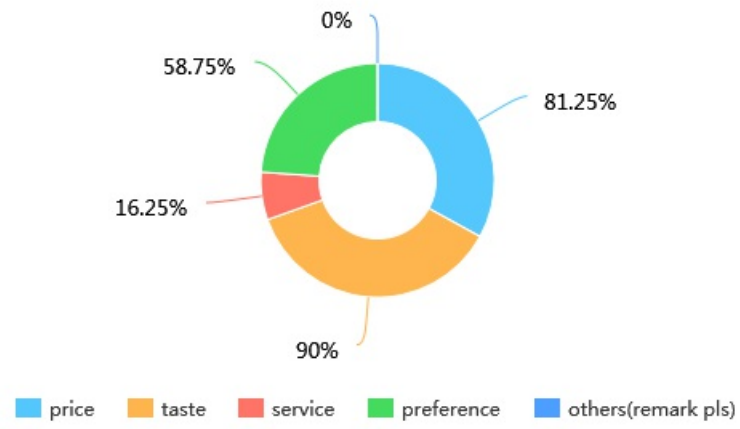

Figure 2.2. What factors will you consider when choosing take-out?

The options of the question were not properly set because the meaning of the word preference was not explicit enough for participants. As a result, the fourth option has various interpretations. In conclusion, the result of the question is invalid strictly.

To cover all possible factors which affect students' take-out ordering behavior, we set a large number of questions. In addition, some of the questions have weak relation between each other. Yet we cannot just judge them as redundant. Therefore, it is quite difficult to conclude rules manually according to the simply-processed data.

3. Preliminary analysis

As it is impossible to give out rules from the simply-processed data, our group chose to use machine learning algorithm to analyze these

data. Before formally start mining the data, we did some preliminary studies to find a correct direction. For there are too many algorithms to choose, we chose to test every valid algorithm in Weka (Waikato Environment for Knowledge Analysis) using the initial parameter provided by the software, and compare each one for the best result.

Algorithm	Bayes-Net	Naive-Bayes-Net	Logistic	SMO	DecisionTable	J48	EM
Kappa	0.0382	0.0243	-0.0724	0.1181	0.0003	0.0709	x

result

The result above revealed that all the classify algorithm returned result with kappa less than 0.2, some of the even had negative kappa value which means the results were even worse that the result of random classify. The same result also happened in the associate algorithms. Surprisingly, the EM algorithm, although it could not offer a clear classification for specially appointed attribute (hence its kappa can not be calculated), it did provide some meaningful findings with the disordered data. As a result, the EM algorithm was used for the further mining job.

4. Applying the Expectation Maximization Algorithm

To rule out redundant information, we are going to analyze the influence factors with the EM algorithm. Expectation–Maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables (岳佳, 2007). We expect to explore possible relationships among various factors with the help of EM.

The EM Algorithm ,as the name suggests, is an algorithm used for getting the classification of a known data set which have the max expectation (Do, C. B., & Batzoglou, S. (2008)). In this experiment, we used the EM algorithm included in Weka (Waikato Environment for Knowledge Analysis) by the University of Waikato.

At the very beginning, we apply the EM algorithm directly to the pre-processed data, the first test used the following parameters:

```
weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-
slots 1 -S 100
```

The head of the results are listed as follow.

```
EM
Number of clusters selected by cross validation: 2
Number of iterations performed: 14
                               Cluster
Attribute                            0       1
                                (0.56)  (0.44)
============================================
WHERE-ELM
  yes                           29.1075 35.8925
  no                             17.871   1.129
  [total]                       46.9785 37.0215
WHERE-BD
  yes                            9.6428 16.3572
  no                            37.3356 20.6644
  [total]                       46.9785 37.0215
  ......
  (23 attributes unlisted)
```

Table 4.1 Head of the first test's result

The result includes a large sum of data, which is quite hard to find some significant results in it. However, the EM algorithm did succeed in part the raw data into two different sets with similar size. For most of the attributes in the result list, the difference is not obvious. However, for some of the binary attributes, the classification matrix showed that the 2 sets generated by this algorithm have huge difference in those properties. We picked out these attributes as following:

```
Attributes picked out in the result of EM

WHY-TOGETHER
  no                    28.7305 32.2695
  yes                   17.9241  5.0759
  [total]               46.6546 37.3454
WHY-OUTOFLUNCHTIME
  yes                   29.7774 12.2226
  no                    16.8772 25.1228
  [total]               46.6546 37.3454
AT_TIME
  c                     19.2888 27.7112
  a                     25.4579  9.5421
  b                      2.9078  1.0922
  [total]               47.6546 38.3454
WHY-AWFULCANTEEN
  no                    25.9073  5.0927
  yes                   20.7473 32.2527
  [total]               46.6546 37.3454
WHY-CROWDEDCANTEEN
  no                    35.1047 10.8953
  yes                   11.5498 26.4502
  [total]               46.6546 37.3454

......
```

Table 4.2 Attributes picked out

All the attributes listed above had obvious difference in the two groups, Since this was only the basic result generated by the EM algorithm with the initial parameters, we could claim that there were still vast improvement space for it.

5. Improve the Algorithm: Further Data Process

5-1. General View

In the first trail, we take 25 factors into consideration. However, according to the result, only some of them shows great differences in the two sets, which showed other attributes had no significant contribution for the classification. For instance:

```
WHY-WANNAEATBETTER (The participant choose taken-out for he wants to eat better)
  no                    31.5739 25.4261
  yes                   16.4403 10.5597
  [total]               48.0142 35.9858

WHY-AWFULCANTEEN (The participant choose taken-out for he can not bear the canteen)
  no                    25.9073  5.0927
  yes                   20.7473 32.2527
  [total]               46.6546 37.3454
```

From the matrix above, it can be concluded that the two group of participants share similar distribution of the percentage of the people claims that they choose taken-out for he wants to eat better. However, the two groups differ greatly in the attitude towards the school canteen. In the left group, only half of its members claims that the canteen is awful, while in the other group, more than 80% of participants hold the idea that the canteen is unbearable. In fact, there are 15 attributes differ greatly between the two groups.

Yet it was only the first trail with the raw data and initial parameters, doing further process of these attributes would undoubtly improve the outcome. We took two measures then: deleting useless data and merging similar options.

5-2. Delete Useless Attributes

For some of the attributes, the choices gathered together in single choices.

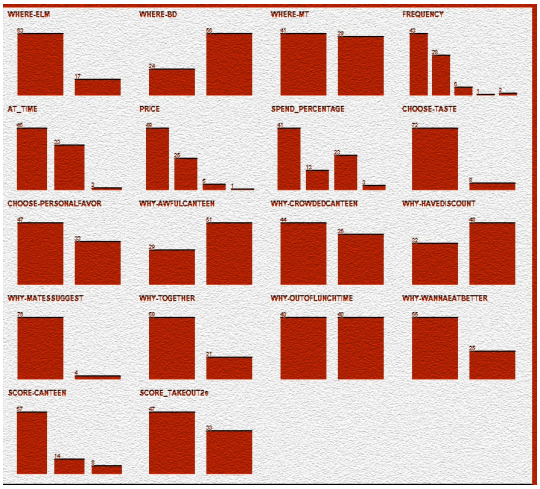

Figure 5.1. Pre-processed data

For instance, in attributes like WHY-MATESUGGEST (which asked if the responder choose to order take-out because of other peoples' suggest), is greatly imbalanced. As the graph suggests, few participants order take-out because others' suggest. In this situation, this attribute can be deleted in the next EM test, for it can hardly provide any useful information for classification, so according to the principle of EM. , deleting them will not harm the general result. Also, if left untouched, these imbalance attribute will introduce more Randomness into the result of EM. Attributes with similar conditions are:

```
WHY-IAMRICH
CHOOSE-ELSE
WHERE-ELSE
WHY-MATESSUGGEST
WHY-TASTE
CHOOSE-SERVICE
WHY-NOOUTDOOR
```

Table 5.2. Attributes chosen

These attributes were removed before the next turn of EM began.

5-3. Merge Similar options

Not only the common pattern will make blur of the results, but also one question with many different options can also add up to the difficulty of data analyze. To avoid this, we chose to merge the options of such questions.

Take the score for canteen attribute for instance. In the questionnaire, the question asking participants to make a score foe the canteen was designed as a Likek scale (Likert summated rating scale) to make out the difference between the slight difference in attitudes towards the canteen. Unfortunately, this design also made the result too complicated so that the algorithm could not use this key to do classify works correctly. To be more specific, for instance, the algorithm does not know the relationship between "dislike" and "hate", hence it algorithm considered these 2 options as 2 completely different emotions, while in fact they both represent negative emotions, only to be different in levels. Therefore, merge such options together as negative attitude can greatly improve the outcome of the algorithm (林东方, 2012).

Attributes with merged options are listed below. These attributes used to have no less than 5 options, while they have no more than 3 options after merging.

```
SCORE-CANTEEN
  neutral              33.9477 26.0523
  positive              10.721   5.279
  negative              2.9859  7.0141
  [total]              47.6546 38.3454
```

```
SCORE_TAKEOUT2s
  neutral               29.6341 19.3659
  positive              17.0205 17.9795
  [total]               46.6546 37.3454
```

Table 5.3. Attributes with merged options

The same method was also applied to "PRICE", "AT-TIME", "FREQUENCY" and "SPEND-PERCENTAGE".

6. Final Result by Second Trail

After processing the data, the EM algorithm was applied once again to the adjusted data, which lead to the final result as following:

```
*Clustering model (full training set)

EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 20

                      Cluster
Attribute                  0       1
                      (0.58)  (0.42)
=====================================

WHY-CROWDEDCANTEEN
  no                  36.4262  9.5738
  yes                  11.588  26.412
  [total]             48.0142 35.9858
AT_TIME
  c                   20.3408 26.6592
  a                   25.7288  9.2712
  b                    2.9446  1.0554
  [total]             49.0142 36.9858

......

SCORE-CANTEEN
  neutral             35.1152 24.8848
  positive            10.8694  5.1306
  negative             3.0295  6.9705
  [total]             49.0142 36.9858
SCORE_TAKEOUT2s
  neutral             30.8225 18.1775
  positive            17.1917 17.8083
  [total]             48.0142 35.9858

Time taken to build model (full training data) : 0.48 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      45 ( 56%)
1      35 ( 44%)

Log likelihood: -11.44594
```

Table 5.4. Clustering model

**The whole Clustering model is too long to list in paper's main body. Full version of the clustering model can be accessed in Appendix.*

For some of the questions, the raw result was in format like "a, b, c, d, e", their meanings are as following:

```
FREQUENCY
A-（never）
B-（no more than three times a month）
C-（several times a week）
D-（almost once a day）
E-（two or three times a day）

AT-TIME
A-（mostly on weekends）
B-（mostly on weekdays）
C-（whenever I want）

PRICE
A- <15 rmb/share
B- 15-25 rmb/share
C- 25-50 rmb/share
D- >50 rmb/share

SPEND-PERCENTAGE
A- <15%
C- 30%-50%
B- 15%-30%
D- >50%
```

Table 5.5. Check list for several questions

Above is the full result generated by EM algorithm. Compared with the first result, this result divided the participants into two groups differ more significantly.

This result can be repeated using the data set with the following commands on Weka:

```
weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-
slots 1 -S 100
```

7. Analysis for the Final Result

Combining the result of the EM algorithm and the result of the basic analyzing, we can roughly divide the take-out orderers in UCAS into two groups. Generally, the first group can be described as people who have negative attitudes towards the canteen, most of them highly evaluated take-out, and enjoyed ordering take-out together with their roommates. Meanwhile, most of them also believes that the canteen is too crowded some times.

The other group of people share different characteristic with the first group of people. While half of them claim that they choose take-out because the canteen is too awful (which is far less than the first group, in which almost all the people believes that the awful canteen is the reason for them to order take-out), they order take-out mainly because of they missed the time for lunch or dinner. These people order take-out mainly at the weekends, and they are less likely to share take-out with their friends.

To sum up, the first group of people can be defined as those who do not like the school canteen so that they choose to take take-out, while another group, which is similar in size with the first group, order take-out mainly because they want food out of the meal time and dinner time.

Discussion

By using the method of supervised machine learning, our study summarized the behavior pattern of the UCAS take-out orderers for the first time.

In the model build by the EM algorithm, we can see two kinds of groups with obvious differences in online-ordering behavior, one of which is people who are not satisfied with the university canteen and the other is people who order take-out only for its efficiency. Compared with other researches about the university take-out ordering , this research provides a quantified model for take-out ordering, which is far more convincing and more repeatable. By using this model, we can speculate one person's taken-out ordering pattern from limited information.

Some suggestions can be given from the result of our study:

1. Recommendations for the stores:

Most students are more willing to accept 15-25 yuan / share takeout, stores can try to keep the profit while controlling the price of 15-25 yuan / single, which might increase the sales. At the same time, the hygiene safety of food is one of the most important areas which students are concerned about. Over 90% of the students worry about the hygiene safety of takeout to varying degrees. Stores should also pay attention to food hygiene and packaging forms, providing customers with a more safe and comfortable dining experience.

2. Recommendations for canteens:

Most of the students' evaluation of college canteen is not well, mainly because of poor food taste and crowded eating environment. The canteen should periodically add some new dishes that students like to keep students from boredom and can also increase the meal time to meet the needs of more students. In fact, most of the students' expectation of college canteen is very high. Over 40% of the students said they could give up takeout as long as the canteen improves their food and service. Therefore, the change of college canteen may even affect the takeout pattern of many students.

3. Recommendations for students:

It turns out that more than half of students almost formed a habit of ordering online, and the most reason for them to choose takeout is laziness: they just don't want to go out. we should not ignore our own health because of the convenience that technology development brings to us. A healthy body requires more exercise instead of just lying on the bed.

However, limitation for this research do exists. The greatest challenge is from the data we can use. We suspected that the activity of take-out ordering is also related to the academic achievement and the physical fitness of the orderer. However, for two limitation factors listed as follow, the doubt cannot be discussed in this paper.

The first limitation factor is the size of the questionnaire. Experience shows that too many questions will make the participants feel board so that the quantity for the questionnaire retrieved will drop sharply. In our questionnaire there were already 14 questions so it is unwise to add more questions to it.

The second limitation factor is about the privacy of the participants. For some of the participants may not want to offer their GPA, it is hard to establish the link between take-out ordering and grades. Meanwhile, because the questionnaire is anonymous, it is also impossible to establish the link between the participants with the publicized information about grades.

As a result, the research is only about the take-out ordering itself. Yet we believe further research about the relationship between take-out ordering and academical and physical performance will lead to more exciting findings.

References

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm?. Nature biotechnology, 26(8), 897.

iiMedia Research. (2016). Research report for Chinese take-out industry(2016).

iiMedia Research. (2017). Research report for Chinese take-out industry(2017).

Wikipedia. (2017) Expectation–maximization algorithm. [Online] Available from: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm [Accessed: 10th January 2018]

林东方. (2012). 基于EM算法的不完全测量数据的处理方法研究. (Doctoral dissertation, 中南大学).

李鲁静. (2015). 大学生网络外卖消费现状及发展研究. 商场现代化(2), 25-25.

岳佳. (2007). 基于EM算法的模型聚类的研究及应用. (Doctoral dissertation, 江南大学).

赵耀. (2016). 大学生外卖消费现况及其影响因素分析——以安徽财经大学为例. 江苏商论(20), 164-165.

Appendix

Appendix-1 *Full Clustering model**

```
=== Clustering model (full training set) ===

EM
==

Number of clusters selected by cross validation: 2
Number of iterations performed: 20

                        Cluster
Attribute                    0       1
                        (0.58)  (0.42)
=======================================
WHERE-ELM
  yes                  30.3819 34.6181
  no                   17.6323  1.3677
  [total]              48.0142 35.9858
WHERE-BD
  yes                   9.4627 16.5373
  no                   38.5514 19.4486
  [total]              48.0142 35.9858
WHERE-MT
  no                   17.1953 25.8047
  yes                  30.8189 10.1811
  [total]              48.0142 35.9858
FREQUENCY
  c                    17.1175 27.8825
  b                    28.8838  1.1162
  d                     1.0127  6.9873
  e                     1.0013  1.9987
  a                     2.9989  1.0011
  [total]              51.0142 38.9858
AT_TIME
  c                    20.3408 26.6592
  a                    25.7288  9.2712
  b                     2.9446  1.0554
  [total]              49.0142 36.9858
PRICE
  b                     27.734  23.266
  c                    16.2951 10.7049
  a                     3.9873  3.0127
  d                     1.9979  1.0021
  [total]              50.0142 37.9858
SPEND_PERCENTAGE
  a                    40.0709  2.9291
  c                     1.7132 13.2868
  b                     6.2307 18.7693
  d                     1.9994  3.0006
  [total]              50.0142 37.9858
CHOOSE-PRICE
  yes                  35.7326 31.2674
  no                   12.2816  4.7184
  [total]              48.0142 35.9858
CHOOSE-PERSONALFAVOR
  yes                  27.6935 21.3065
  no                   20.3207 14.6793
  [total]              48.0142 35.9858
WHY-AWFULCANTEEN
  no                    26.908   4.092
  yes                  21.1061 31.8939
```

```
    [total]                48.0142 35.9858
WHY-CROWDEDCANTEEN
    no                     36.4262  9.5738
    yes                     11.588  26.412
    [total]                48.0142 35.9858
WHY-HAVEDISCOUNT
    yes                     18.976  15.024
    no                     29.0382 20.9618
    [total]                48.0142 35.9858
WHY-TOGETHER
    no                     29.8723 31.1277
    yes                    18.1419  4.8581
    [total]                48.0142 35.9858
WHY-OUTOFLUNCHTIME
    yes                    30.1958 11.8042
    no                     17.8183 24.1817
    [total]                48.0142 35.9858
WHY-WANNAEATBETTER
    no                     31.5739 25.4261
    yes                    16.4403 10.5597
    [total]                48.0142 35.9858
SCORE-CANTEEN
    neutral                35.1152 24.8848
    positive               10.8694  5.1306
    negative                3.0295  6.9705
    [total]                49.0142 36.9858
SCORE_TAKEOUT2s
    neutral                30.8225 18.1775
    positive               17.1917 17.8083
    [total]                48.0142 35.9858

Time taken to build model (full training data) : 0.48 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      45 ( 56%)
1      35 ( 44%)

Log likelihood: -11.44594
```

Table A-1. Full Clustering model