Phi Coefficients For Benoulli Distributed Data

Andriy Fedorenko

Friday, February 06, 2015

Introduction

Given a dataset of transactions (100k rows) from a large retailer, we should find groups of items that are frequently purchased together. Each row is a single transaction with a '1' denoting that item was in the transaction. The dataset can be found at http://bit.ly/1sCk2vd:

Solution:

We need to find groups of items that are frequently purchased together.It means that data variables, represented by the items, have to be high correlated.

```
##   id item_0 item_1 item_2 item_3 item_4 item_5 item_6 item_7 item_8
## 1  0      0      1      0      0      0      0      0      0      0
## 2  1      0      0      0      1      0      1      0      0      0
```

```
dim(ini.data)
```

```
## [1] 100000     51
```

Correlation between variables have meaning of similarity for Bernoulli distributed data and can be calculated as Phi coefficients based on Pearson formula

$$S_{Phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

The object function performs following procedures:

- sample n=500 cases from the data and remove "id" variable
- remove all "only-zero" conataining cases

```
 dim(sampled.data)
```

```
## [1] 456  50
```

- calculate the Phi coefficients using hetcor() function

```
recived.het<-hetcor(sampled.data,digits=3)
```

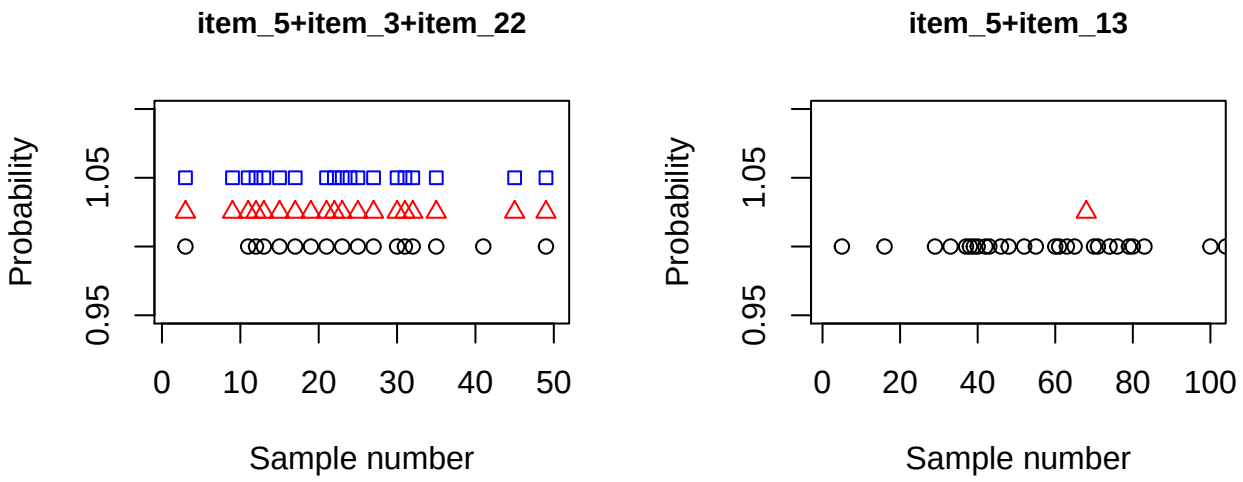
Results:

The table of coefficients shows that among 50 items 16 are correlated.

```
recived.variables
```

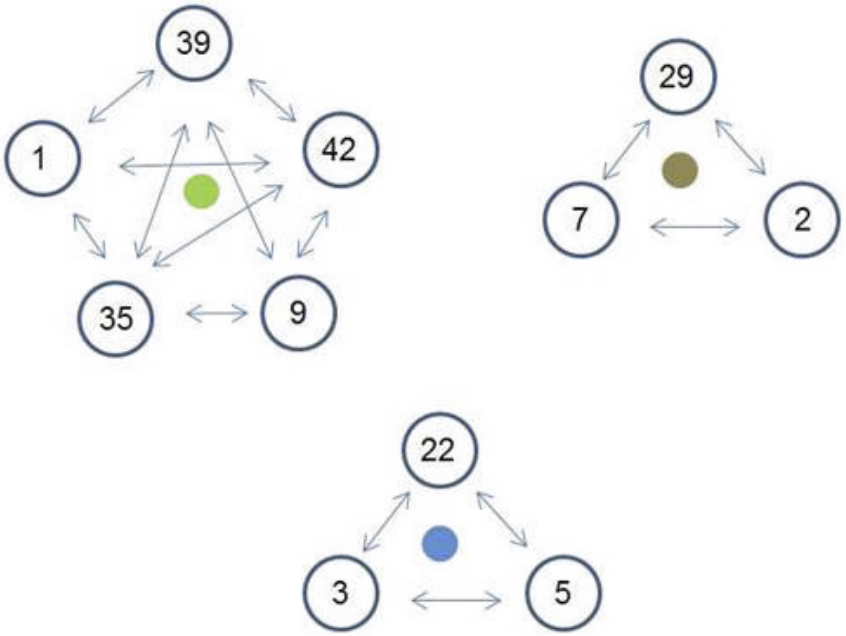
```
##         row col
## item_3    3   5
## item_2    2   7
## item_1    1   9
## item_3    3  22
## item_5    5  22
## item_2    2  29
## item_7    7  29
## item_1    1  35
## item_9    9  35
## item_1    1  39
## item_9    9  39
## item_35  35  39
## item_1    1  42
## item_9    9  42
## item_35  35  42
## item_39  39  42
```

For example, if we plot item_5, item_3 and item_22 we can find that in majority of cases they all show up approximately in the same samples and with the same frequency (to make data points clearer, I assigned value 1.025 to item_3 and 1.05 to item_22, left panel).



In contrast, item_13 appears rarely compare to item_5 and in different samples, illustrating that these variables are not correlated (right panel).

item_number	Item_numer
35	1
35	9
39	1
9	1
42	1
39	9
42	9
42	39
39	35
42	35
7	2
29	2
29	7
22	3
5	3
22	5



Rearranged correlated variables can be split into three separate groups. Each variable is high correlated with another one in the same group. The "green" group consists of 5 variables 1,39,42,35,9; the "brown" group consists of 3 variables 7,29,2; the "blue" group consists of 3 variables 22,3,5. It suggests, that if customer wants to buy item 2, the chance of buying the item 29 together with 7th and 2nd is very high (in fishing store a person will buy a fishing rod together with a reel and line).

This report was generated with R (3.1.1) and pander (0.5.1) on x86_64-w64-mingw32 platform.