Homework 6:

Practice of Association Rules

Henry Qui

Goktug Cengiz

Mehmet Fatih Cagil

May 21, 2019

1 Read the file: tic_tt. Check the "class" of every variable of tic_tt.

Before reading the file, we had to perform some manual modifications on the header of the dataset. Concretely, there were some missing semicolons between the variables `Catala`, `Castella` and `Angles`. After adding this, we can read the dataset properly setting the first column as the identification of the observations.
The dataset contains 1666 observations and 33 variables. The main type of the variables are logical such as `Tenen.tarifa.plana.`, `TV.de.pagament.`, our response variable `Pagament.a.traves.d.Internet.`, and so on, and categorical variables, such as `Quants.televisors.tenen.`, `Edat, Nivell.ingressos.mensuals.`, and so on, that are representations of ranges of numeric variables already split.

2 Find the profile of the people who do payments by Internet

The variable corresponding to payments by Internet is clearly the logical variable `Pagament.a.traves.d.Internet.` Where value False means the person does not do payments by Internet, and value True means it does.
To analyse the profile of the people who do payments by Internet, we use the function catdes, transforming previously the variables of our dataset into factors. Figure 1 shows the obtained results.

In the figure 1 we can see in decreasing order the variables most linked to the response variable, which is the variables that determines most the value of the response variable. Among the most related variables we can observe that Ha.comprat.per.Internet. is the most related, which completely makes sense because people who has bought on internet are probably used to pay on internet. Some other variables also show interesting features, for example if people has an e-mail address, or use very frequently the computer also indicates they buy on internet. It is also interesting that variables Edat and Sexe are not very linked to our response variable, which also make sense according to our expectations.

Link between the cluster variable and the categorical variables (chi-square test)

	p.value	df
Ha.comprat.per.Internet.	1.921362e-166	1
Ha.comprat.aliments.	2.658975e-24	1
Utilitza.serveis.banca.electronica.	1.436704e-19	1
Amb.quina.frequencia.usa.inet.	5.101735e-19	3
Fa.servir.correu.electronic.	1.392287e-10	1
Amb.quina.frequencia.usa.ordinador.	1.085994e-09	4
Visita.webs.de.l.Adm	3.044356e-09	1
Angles	4.578359e-09	1
Connexio.rapida	1.063271e-08	1
Tenen.tarifa.plana.	1.201904e-07	1
Disposa.de.connexio.a.Internet.a.la.llar.	1.699390e-06	1
Internet.per.a.l.oci.	1.799611e-06	1
Nivell.d.estudis	7.059566e-06	3
Tramits.per.Internet.amb.l.Adm	4.261981e-05	1
Informatica.avancada	1.708200e-04	1
Aplicacions.especifiques	6.033325e-04	1
Activitat.professional	7.058293e-04	7
Estudia.per.Internet.	1.489634e-03	1
Utilitza.programes.ofimatics	2.406055e-03	1
Missatges.mobil	2.924516e-03	5
Fa.teletreball.	3.960133e-03	1
Edat	6.821073e-03	5
Sexe	1.456634e-02	1

Figure 1: chi-square test for cluster and categorical variable

Description of each cluster by the categories

$`FALSE`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Ha.comprat.per.Internet.=FALSE	100.00000	93.873313	81.003584	9.707169e-140	25.164905
Utilitza.serveis.banca.electronica.=FALSE	92.49677	74.247144	69.265233	5.128565e-18	8.650475
Ha.comprat.aliments.=FALSE	88.36127	98.546210	96.236559	7.456444e-16	8.062803
Fa.servir.correu.electronic.=FALSE	96.82540	31.671859	28.225806	1.034788e-12	7.125799
Visita.webs.de.l.Adm=FALSE	94.90617	36.760125	33.422939	2.522673e-10	6.325590
Angles=FALSE	90.46358	70.924195	67.652330	1.423765e-08	5.670506
Connexio.rapida=FALSE	89.05908	84.527518	81.899642	1.078407e-07	5.312990
Tenen.tarifa.plana.=FALSE	90.62958	64.278297	61.200717	1.951001e-07	5.203947

$`TRUE`

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Ha.comprat.per.Internet.=TRUE	72.169811	100.000000	18.996416	9.707169e-140	25.164905
Amb.quina.frequencia.usa.inet.=Diàriament	25.688073	73.202614	39.068100	3.295180e-20	9.208889
Utilitza.serveis.banca.electronica.=TRUE	27.696793	62.091503	30.734767	5.128565e-18	8.650475
Ha.comprat.aliments.=TRUE	66.666667	18.300654	3.763441	7.456444e-16	8.062803
Amb.quina.frequencia.usa.ordinador.=Diàriament	19.219653	86.928105	62.007168	3.643686e-13	7.268178
Fa.servir.correu.electronic.=TRUE	17.852684	93.464052	71.774194	1.034788e-12	7.125799
Visita.webs.de.l.Adm=TRUE	18.034993	87.581699	66.577061	2.522673e-10	6.325590
Angles=TRUE	22.437673	52.941176	32.347670	1.423765e-08	5.670506
Connexio.rapida=TRUE	26.237624	34.640523	18.100358	1.078407e-07	5.312990
Tenen.tarifa.plana.=TRUE	20.554273	58.169935	38.799283	1.951001e-07	5.203947

Figure 2: first 8 lines of the results

In the figure 2, second part of the results,we can see some statistics explanation for relation between the variables and the response variable with their concrete values. We can observe some information like the all the people that pays on internet has bought on internet, and among the people that bought on internet, a 72% has paid on internet, which is very high and log to interpret.

3 Convert the tic_tt file to a transactions file.

Converting the dataset into a transaction files means to find rules associated to each level of the response, which in our case is a logical variable. It consists in converting all categorical variables into binary indicators, then every row will be considered a transaction, with this we can extract rules from these transactions. Concretely, for each id of an individual, each logical variable that has value true or each value of its categorical variable will be assigned as an item of its transaction.

4 Define the parameters and run the apriori.

The values we have set for the minimum support is 0.01, we took a very small value because we want to consider mostly all the transactions, even their frequency is very low, otherwise, taking into account that the dataset is not too big, if we set a high value for the minimum support, some interesting rules that are not frequent by chance can be filtered out.
For the minimum confidence, we thought that 0.4 is a proper value, we cannot set a very low value because if we obtain rules that has low confidence means that they are not too reliable, which are useless for us. By the other hand, if we set a value quite big, we can filter out a lot of interesting rules, because we have to take into account that rules with confidence greater than 0.5 means it is a highly reliable rule, however big part of transactions has a confidence lower than 0.5.
Lastly, we have define the maximum size of itemsets as 5, we are not interesting in rules with too many items, too many items means the rule will be too concrete, but the idea is to obtain minimally general rules.

5 List the 10 most frequent itemsets

We can see the 10 most frequent itemsets in the figure 3. It completely make sense that languages variables such as `Castella` and `Catala` are part of the most frequent itemsets, we can interpret that the individuals are mostly spanish speakers. It is also very obvious that most of the people nowadays need to use office programs, have an email, have internet at home, and use the computer in a daily basis.

```
                                                                items   support
13                                                          {Castella} 0.7921147
14                                        {Utilitza.programes.ofimatics} 0.7894265
12                                        {Fa.servir.correu.electronic.} 0.7177419
11                          {Disposa.de.connexio.a.Internet.a.la.llar.} 0.6899642
10                                              {Internet.per.a.l.oci.} 0.6827957
9                                                {Visita.webs.de.l.Adm} 0.6657706
1036                          {Utilitza.programes.ofimatics,Castella} 0.6272401
8                      {Amb.quina.frequencia.usa.ordinador.=Diàriament} 0.6200717
1034 {Utilitza.programes.ofimatics,Fa.servir.correu.electronic.} 0.5949821
7                                                           {Catala} 0.5806452
```

Figure 3: most frequent itemsets

6 List the first 10 rules sorted by the lift

These are the rules with the highest lift we have obtained. The lift of a rule indicates that the rule behaves better than an independence when it is higher than 1. This means that these are the most "interesting" rules we have obtained, because the left hand side and right hand side itemsets are not very independent, it exists a relation which can be useful for us to obtain conclusions. We can observe that all these rules has very small support value The results are compound mainly by people of 65 and 99 years and profession of `Altres inactius`. We can take interpretations like: man who has `Altres inactius` as profession tends to be between 65 and 99 years.

```
     lhs                                                    rhs                                                    support    confidence lift     count
[1]  {Sexe=Home,Activitat.professional=Altres inactius}     => {Edat=Edat(65,99]}                                  0.01344086 0.6250000  26.82692 15
[2]  {Per.oci,Edat=Edat(65,99],Internet.per.a.l.oci.}       => {Activitat.professional=Altres inactius} 0.01075269 0.9230769  25.12570 12
[3]  {Edat=Edat(65,99],Internet.per.a.l.oci.}               => {Activitat.professional=Altres inactius} 0.01344086 0.8823529  24.01722 15
[4]  {Per.oci,Edat=Edat(65,99]}                             => {Activitat.professional=Altres inactius} 0.01254480 0.8750000  23.81707 14
[5]  {Castella,Activitat.professional=Altres inactius}      => {Edat=Edat(65,99]}                                  0.01433692 0.5161290  22.15385 16
[6]  {Activitat.professional=Altres inactius}               => {Edat=Edat(65,99]}                                  0.01881720 0.5121951  21.98499 21
[7]  {Edat=Edat(65,99]}                                     => {Activitat.professional=Altres inactius} 0.01881720 0.8076923  21.98499 21
[8]  {Utilitza.programes.ofimatics,Edat=Edat(65,99]}        => {Activitat.professional=Altres inactius} 0.01075269 0.8000000  21.77561 12
[9]  {Edat=Edat(65,99],Sexe=Home}                           => {Activitat.professional=Altres inactius} 0.01344086 0.7894737  21.48909 15
[10] {Activitat.professional=Altres inactius,Internet.per.a.l.oci.} => {Edat=Edat(65,99]}                           0.01344086 0.5000000  21.46154 15
```

Figure 4: first 10 rules sorted by the lift

7 List the 10 rules according the lift, where the Consequent is "Pagament.a.través.d.Internet."

We are looking for the kind of features a person person who pay on internet have. In our case we can see that people who speaks english, studies by internet, buy on internet and use internet for entertainment has high probability to pay on internet, also the people who have specific applications, works as entrepreneur, has bought on internet and use bank electronic services, and so on. With this results we have proved again that people who has bought on internet is really a good indicator that they are also used to buy on internet because the item `Ha.comprat.per.Internet.` is appearing in all the rules.

```
     lhs                                                rhs                                support confidence     lift count
[1]  {Angles,
      Estudia.per.Internet.,
      Ha.comprat.per.Internet.,
      Internet.per.a.l.oci.}                         => {Pagament.a.traves.d.Internet.} 0.01344086          1 7.294118    15
[2]  {Aplicacions.especifiques,
      Activitat.professional=Empresaris,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01254480          1 7.294118    14
[3]  {Nivell.d.estudis=Superiors,
      Activitat.professional=Empresaris,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01164875          1 7.294118    13
[4]  {Tenen.tarifa.plana.,
      Activitat.professional=Empresaris,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01254480          1 7.294118    14
[5]  {Catala,
      Activitat.professional=Empresaris,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01254480          1 7.294118    14
[6]  {Tenen.tarifa.plana.,
      Per.oci,
      Activitat.professional=Empresaris,
      Ha.comprat.per.Internet.}                      => {Pagament.a.traves.d.Internet.} 0.01164875          1 7.294118    13
[7]  {Informatica.avancada,
      Nivell.d.estudis=Superiors,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01344086          1 7.294118    15
[8]  {Tenen.tarifa.plana.,
      Informatica.avancada,
      Nivell.d.estudis=Superiors,
      Ha.comprat.per.Internet.}                      => {Pagament.a.traves.d.Internet.} 0.01075269          1 7.294118    12
[9]  {Missatges.mobil=Menys de 5,
      Sexe=Home,
      Ha.comprat.per.Internet.,
      Tramits.per.Internet.amb.l.Adm}                => {Pagament.a.traves.d.Internet.} 0.01254480          1 7.294118    14
[10] {Connexio.rapida,
      Nivell.ingressos.mensuals=NS/NC,
      Ha.comprat.per.Internet.,
      Utilitza.serveis.banca.electronica.}           => {Pagament.a.traves.d.Internet.} 0.01254480          1 7.294118    14
```

Figure 5: first 10 rules sorted by the lift