UNIVERSIDAD
D
CÓRDOBA

INTRODUCCIÓN A LA MINERÍA DE DATOS

Práctica 3: Reglas de asociación

4º Curso - Grado en Ingeniería Informática

UCO - Escuela Politécnica Superior de Córdoba

Manuel Casas Castro - 31875931R

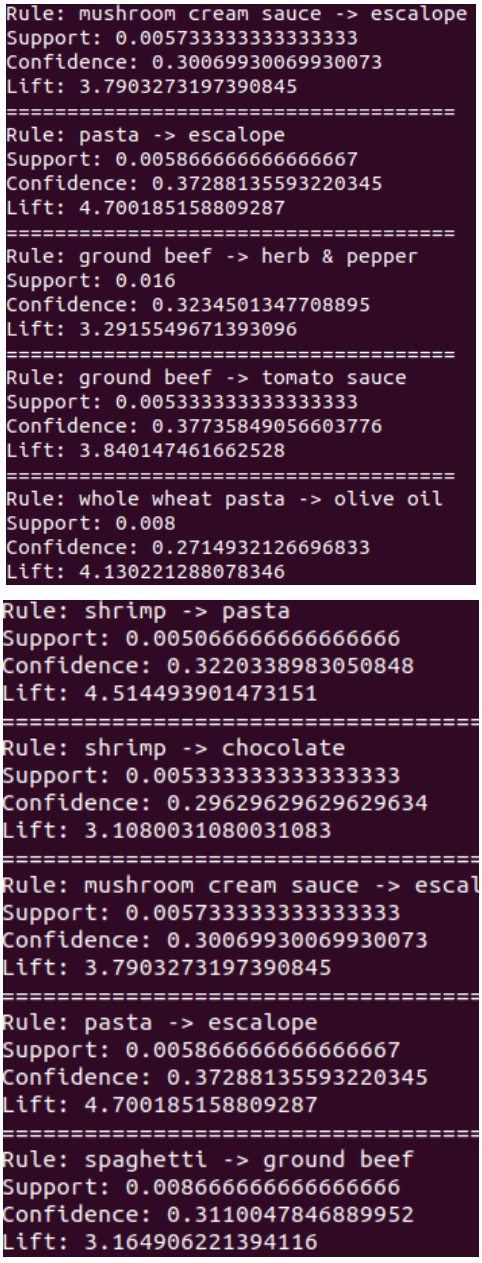
i72cascm@uco.es

i72cascm@uco.es **1**
Ejercicio 1. 3
Ejercicio 2. 7
Ejercicio 3. 9
Ejercicio 4. 10

1.

Utilizando el conjunto de datos “store_data.csv” ejecute el programa “assoc.py” para comprobar su funcionamiento. Intente interpretar las reglas obtenidas e indicar cuáles de ellas son importantes.

Los resultados de la ejecución es la siguiente:

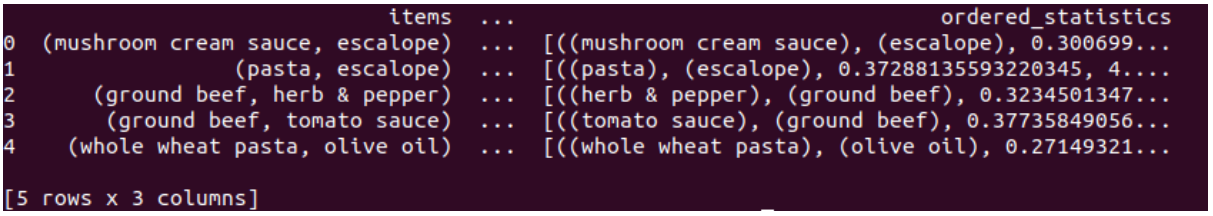


```
Rule: shrimp -> mineral water
Support: 0.0072
Confidence: 0.3068181818181818
Lift: 3.2183725365543547
====================================
Rule: tomatoes -> spaghetti
Support: 0.006666666666666667
Confidence: 0.3184713375796178
Lift: 3.340608436149837
====================================
Rule: ground beef -> spaghetti
Support: 0.005333333333333333
Confidence: 0.3225806451612903
Lift: 3.282706701098612
====================================
Rule: ground beef -> herb & pepper
Support: 0.006666666666666667
Confidence: 0.390625
Lift: 3.975152645861601
====================================
Rule: ground beef -> herb & pepper
Support: 0.016
Confidence: 0.3234501347708895
Lift: 3.2915549671393096
```

```
Rule: spaghetti -> ground beef
Support: 0.0064
Confidence: 0.3934426229508197
Lift: 4.003825878061259
====================================
Rule: ground beef -> tomato sauce
Support: 0.005333333333333333
Confidence: 0.33735849056603776
Lift: 3.840147461662528
====================================
Rule: shrimp -> spaghetti
Support: 0.006
Confidence: 0.5232558139534884
Lift: 3.004914704939635
====================================
Rule: whole wheat pasta -> olive oil
Support: 0.008
Confidence: 0.2714932126696833
Lift: 4.130221288078346
====================================
Rule: shrimp -> pasta
Support: 0.005066666666666666
Confidence: 0.3220338983050848
Lift: 4.514493901473151
```

```
Rule: shrimp -> chocolate
Support: 0.005333333333333333
Confidence: 0.29629629629629634
Lift: 3.1080031080031083
==================================
Rule: spaghetti -> ground beef
Support: 0.008666666666666666
Confidence: 0.3110047846889952
Lift: 3.164906221394116
==================================
Rule: shrimp -> mineral water
Support: 0.0072
Confidence: 0.3068181818181818
Lift: 3.2183725365543547
==================================
Rule: tomatoes -> spaghetti
Support: 0.006666666666666667
Confidence: 0.3184713375796178
Lift: 3.340608436149837
==================================
Rule: ground beef -> spaghetti
Support: 0.005333333333333333
Confidence: 0.3225806451612903
Lift: 3.282706701098612
```

```
Rule: nan -> ground beef
Support: 0.006666666666666667
Confidence: 0.390625
Lift: 3.975152645861601
==================================
Rule: spaghetti -> ground beef
Support: 0.0064
Confidence: 0.3934426229508197
Lift: 4.003825878061259
==================================
Rule: shrimp -> spaghetti
Support: 0.006
Confidence: 0.5232558139534884
Lift: 3.004914704939635
```



```
                            items  ...                                ordered_statistics
0  (mushroom cream sauce, escalope)  ...  [((mushroom cream sauce), (escalope), 0.300699...
1                (pasta, escalope)  ...  [((pasta), (escalope), 0.37288135593220345, 4....
2      (ground beef, herb & pepper)  ...  [((herb & pepper), (ground beef), 0.3234501347...
3       (ground beef, tomato sauce)  ...  [((tomato sauce), (ground beef), 0.37735849056...
4    (whole wheat pasta, olive oil)  ...  [((whole wheat pasta), (olive oil), 0.27149321...

[5 rows x 3 columns]
```

Tras la ejecución del script aportado, se han generado un total de 28 reglas, de estas reglas, podemos destacar aquellas reglas cuyo lift es mayor al resto. Las reglas con mayor lift:

-pasta -> escalope

-shrimp -> pasta

-whole wheal pasta -> olive oil

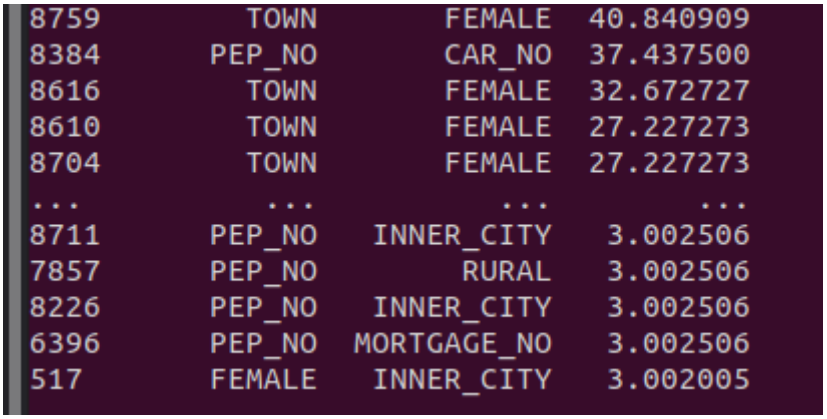
2.

Usando los conjuntos “titanic.csv” y “bank-data-final.arff” ejecute de nuevo el programa de generación de reglas, ordenando los resultados según su valor de lift. Interprete las reglas que se obtienen indicando su evaluación objetiva de interés.

Mostramos los conjuntos de reglas ordenados por su lift
-titanic.csv

	Antecedente	Consecuente	Lift
23	Q	3	6.068182
25	female	Q	5.779221
17	1	female	5.594778
46	female	Q	5.591931
31	female	Q	5.504020
34	Dead	Old	5.393939
9	1	Dead	5.191667
32	female	Adult	5.128546
39	female	Old	5.056818
6	female	Q	4.701739
12	1	Old	4.377049
43	female	Unknown	4.223729
0	1	Old	4.120370
35	female	Old	3.912088
24	female	Q	3.852814
4	Q	3	3.852814
5	Q	Survived	3.852814
16	1	Dead	3.792614
11	1	female	3.787234
26	male	Q	3.768631
13	1	female	3.765109
38	Dead	Old	3.755274
40	female	Adult	3.682759
28	female	Unknown	3.676115
1	1	Dead	3.650391
41	male	2	3.612013
36	female	Unknown	3.585818
21	Unknown	male	3.572248
10	1	Old	3.490196
37	female	C	3.476562
15	1	Dead	3.461111
18	1	Old	3.433642
2	1	Old	3.371212
42	female	2	3.332373
29	female	Old	3.325031
44	female	C	3.316770
30	Q	Dead	3.250238
3	1	female	3.150871
33	male	Adult	3.097724
27	Young	Q	3.090278
8	1	female	3.079979
7	1	Survived	3.074947
20	female	S	3.074005
19	female	Survived	3.062840
45	Dead	male	3.055680
14	1	Young	3.047945
22	Q	Dead	3.000749

-bank.csv



8759	TOWN	FEMALE	40.840909
8384	PEP_NO	CAR_NO	37.437500
8616	TOWN	FEMALE	32.672727
8610	TOWN	FEMALE	27.227273
8704	TOWN	FEMALE	27.227273
...	...	...	...
8711	PEP_NO	INNER_CITY	3.002506
7857	PEP_NO	RURAL	3.002506
8226	PEP_NO	INNER_CITY	3.002506
6396	PEP_NO	MORTGAGE_NO	3.002506
517	FEMALE	INNER_CITY	3.002005

Podemos observar las reglas ordenadas en orden de mayor a menor lift.
En el dataset titanic.csv podemos destacar reglas como que si es un pasajero es de tipo Q, es tipo 3 (regla con el mayor lift) o que si el pasajero es mujer es de tipo Q.

En cuanto al método bank.csv, podemos observar reglas como que si una persona es de ciudad, entonces será mujer ó que si el atributo “pep” es igual a NO, entonces la persona no tendrá coche (reglas con mayor lift).

3.

Seleccione al menos un nuevo conjunto de datos de los suministrados en los repositorios habituales. Para que se pueda usar el método a priori es necesario que los conjuntos no contengan variables numéricas.

Se ha modificado el conjunto weather.csv para que solo posea variables categóricas para este experimento:

```
sunny,hot,low,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,cold,high,FALSE,yes
rainy,cold,high,FALSE,yes
rainy,cold,low,TRUE,no
overcast,cold,low,TRUE,yes
sunny,hot,high,FALSE,no
sunny,cold,low,FALSE,yes
rainy,cold,high,FALSE,yes
sunny,cold,low,TRUE,yes
overcast,cold,high,TRUE,yes
overcast,hot,low,FALSE,yes
rainy,hot,high,TRUE,no
```

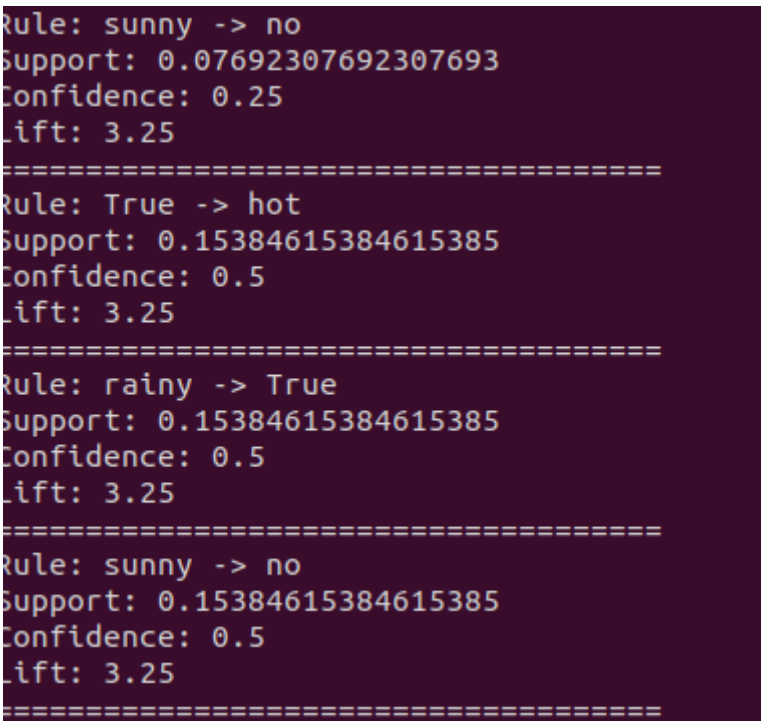
A continuación se muestran reglas ordenadas por lift al ejecutar el script en este nuevo dataset:

	Antecedente	Consecuente	Lift
25	high	True	4.333333
23	yes	cold	4.333333
47	cold	overcast	4.333333
37	high	sunny	3.250000
28	sunny	high	3.250000
29	overcast	high	3.250000
30	rainy	True	3.250000
31	sunny	True	3.250000
32	low	True	3.250000
33	yes	cold	3.250000
34	cold	low	3.250000
35	sunny	cold	3.250000
36	rainy	high	3.250000
38	overcast	high	3.250000
26	sunny	high	3.250000
39	overcast	low	3.250000
40	cold	False	3.250000
41	cold	sunny	3.250000
42	sunny	False	3.250000
43	False	overcast	3.250000
44	False	overcast	3.250000
45	cold	yes	3.250000
46	cold	True	3.250000
48	cold	sunny	3.250000
49	True	no	3.250000
27	high	True	3.250000

Donde podemos observar reglas como que si la humedad es alta el atributo play es igual a verdadero o que si el atributo play es verdadero indica que hará frío (reglas con valor más alto de lift).

4.

Seleccione las 5 mejores reglas usando las medida de confianza y lift, para obtener 10 reglas. Compare las reglas obtenidas y comente qué información puede obtener de algunas de las reglas.



Aplicando el script dado para la práctica, podemos obtener la confianza para cada una de las reglas obtenidas, de las cuales se puede observar que la confianza oscila para las reglas entre 0.25 y 0.5, lo cual puede ser debido a la poca cantidad de patrones que posee este dataset y por tanto concluimos que no tenemos reglas fiables.