CS-E4650 Methods of Data MIning

Home assignment 2

Rosales Rodríguez, Pablo - 914769

Contents

Exercise 1 **2**

Exercise 2 **6**

Exercise 3 **7**

Exercise 4 **11**

Exercise 5 **13**

Exercise 1

The provided data set contains two features and the ground-truth label for each of the points. In the figure 1, there is the representation of the points.

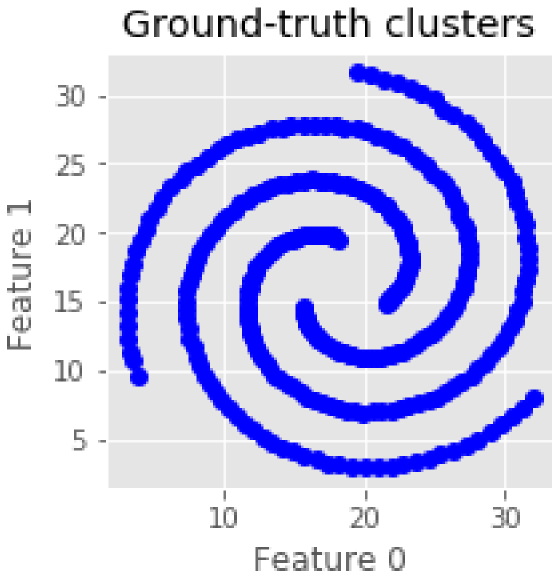

Figure 1: Representation of the data set.

Given the points in figure 1, they clearly compose a spiral, hence it is expected that spectral clustering will perform better than k-means in the given case. Two more representations of the initial data set are introduced in figures 2 and 3. In these cases, including also the ground-truth labels.

k-means clustering

K-means is one of the most popular clustering methods. It is based on solving an optimization problem in which the distance from each of the points to the representative or centroid of its group is the minimum. The k stands for the number of clusters in which the data must be grouped. As stated in the Data Mining Textbook, by Charu C. Aggarwal (Aggarwal, 2015), the optimization problem has the following expression:

$$Dist(\bar{X}_i, \bar{Y_j}) = ||\bar{X}_i, \bar{Y_j}||_2^2$$

The clusterization was performed using the built in functions in the *Sklearn* Python Library, as following:

```
kmeans = KMeans(n_clusters=k, random_state=0).fit(X)
centroids = kmeans.cluster_centers_
labels = kmeans.labels_
```

The graphic result is presented in the figure 4. As it is depicted, the performance of k-means is not good. The optimizaton technique previously described is not accurate with arbitrary-shaped clusters.

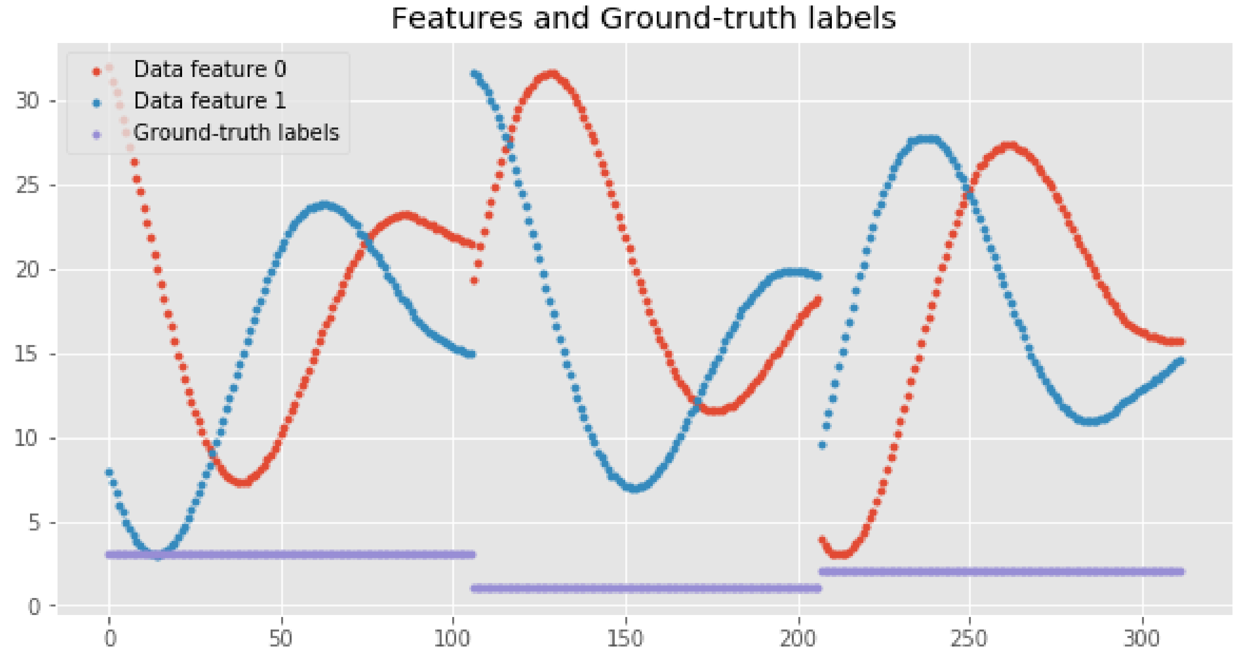

Figure 2: Representation of the data set with ground-truth labels.

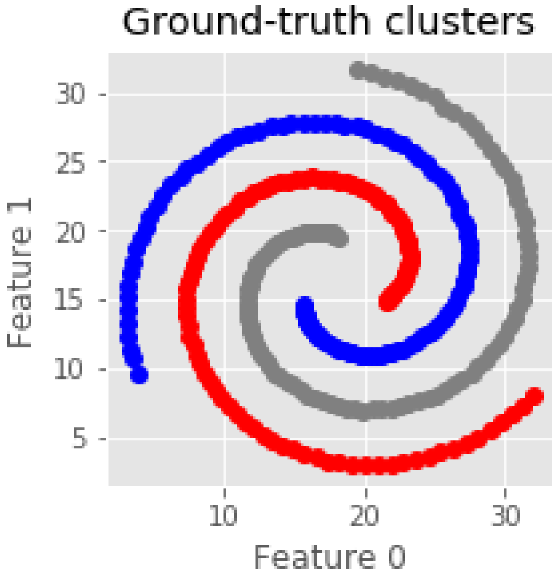

Figure 3: Representation of the data set with the proposed clusters.

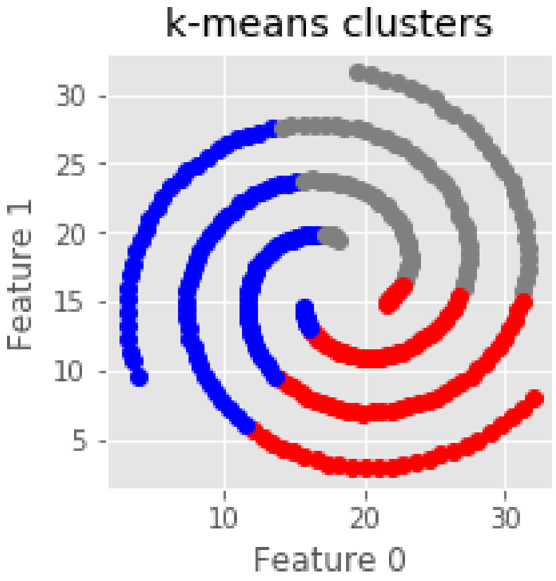

Figure 4: Representation of the data set clustered using k-means.

Spectral clustering

Spectral clustering is a better method for data sets with a specifical geometric shape, as the given one. It is based on mapping the data set into a new dimensional representation in which k-means can be applied (Aggarwal, 2015). The result is showed in the following figure. As expected, the performance is good. Actually, the obtained clusterization is not different from the provided one presented in figure 3. Again, the built in function in *Sklearn* was used, as following:

```
kernel_gamma = 1.0
spectral = SpectralClustering(n_clusters=K, gamma=kernel_gamma).fit(X)
spectral_labels = spectral.labels_
```

In the previous code chunk, the value for gamma was set to 1. In the original expression for the kernel, ther hyperparamenter is sigma, as in the following equation:

$$k(x,y) = exp(\frac{-||x-y||_2^2}{2\sigma^2})$$

However, when using the built in function, the expression is the following:

$$k(x,y) = exp(-\gamma||x-y||_2^2)$$

The Gaussian kernel represents a similarity function. The standard deviation, σ, is a measure of how grouped the data is, so by fixing the point x, increasing the value of the standard deviation, the number of similar y points to x is higher. For very extreme values of sigma, then, the performance is worse. If a higher value of sigma is chosen, such as $\sigma = 2$, the obtained clusterization is similar to k-means: because of the higher standard deviation, more points are considered similar to x and the algorithm finds very difficult to respect the geometric spiral shape of the data set. In the oposite case, some of the clusters found by the method are really small because of the low standard deviation of the gaussian function.

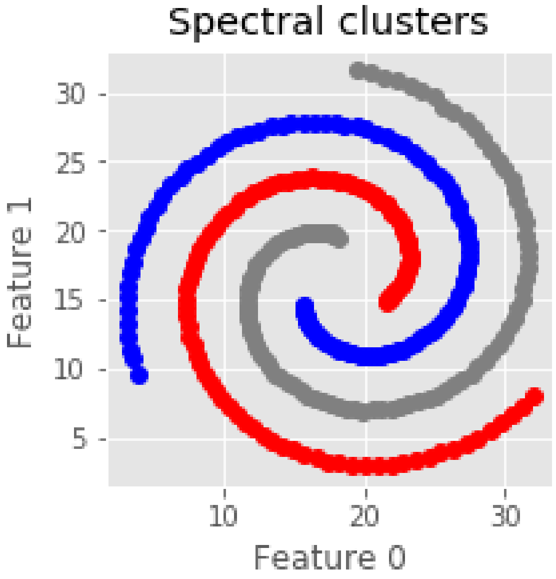

Figure 5: Representation of the data set clustered using spectral clustering.

Performance metrics

Silohuette Coefficient

It represents how high the "intra-cluster" similarity and "inter-cluster" are. It takes values inside [-1, 1]. As stated in (Aggarwal, 2015), the equation for computing the Silhouette coefficient is the following:

$$S_i = \frac{Dmin_i^{out} - Davg_i^{in}}{max(Dmin_i^{out}, Davg_i^{in})}$$

Where, $Davg_i^{in}$ is the average of the distances measured from X_i to all the points within its cluster. Same computation is performed between X_i and the points in different clusters. Then, the minimum of those is selected as $Dmin_i^{out}$ (Aggarwal, 2015). The greater the result of the coefficient, the more robust the clusterization is, meaning that there is high "intra-cluster" similarity and low "inter-cluster" similarity.

- K-means clustering = 0.361456276818138
- Spectral clustering = 0.0013442973442779936

Considering the previous description, according to the Silhouette Coefficient, k-means clustering would be a better option than spectral clustering. However, this is not true, since this coefficient is not an appropriate metric for measuring the performance with the given data set. Silhouette coefficient considers the intra-cluster and inter-cluster distances as main parameters, meaning that it is a good expression for non-shaped data sets, but the proposed one is a spiral, so there will be points within the same cluster whose distance is too high, as well as points in different clusters that are really close. For this reason, the value obtained for Spectral clustering is so low.

Davies-Bouldin

Davies-Bouldin index has a more complex expression than the previous ones. It first computes the similarity as the ratio of the average of the "intra-cluster" distances and the average of the "inter-clusters" distances. Then, the average of the most similar clusters is calculated (Scikit-learn developers, 2007-2020). The lower the score, the better the clusterization is. * K-means clustering = 0.8895007416113797 * Spectral clustering = 5.882022552277642

As it happened with the previous score, Davies-Bouldin one does not represent the performance in the given case.

Normalized Mutual Information

- K-means clustering = 0.0007031008001133277
- Spectral clustering = 1.0

The obtained Normalized Mutual Information for the Spectral Clustering yields a result of 1, meaning that the performed clusterization is perfect and so, it is possible to conclude, by visual comparison between the figures 5 and 3, that this is the appropriate metric for the proposed data set.

Exercise 2

Let K be the kernel matrix:

$$K_{ij} = k(x_i, x_j) = exp(\frac{-||x_i - x_j||^2}{2\sigma^2})$$

Where x_i is the i-th data point. Given the clustering of data:

$$c_{ij} = \begin{cases} 1 \text{ if } x_i \text{ and } x_j \text{ are in the same cluster} \\ 0 \text{ otherwise} \end{cases}$$

The validation index is:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{i \neq j} c_{ij} K_{ij}}{\sum_{j \neq i} K_{ij}}$$

Where n is the number of data points.

The value of the previous metric for the different clusterization methods is:

- K-means clustering = 0.9556022746652073
- Spectral clustering = 0.9999999999999173

The value of the kernel depends on the distance between the two considered points x_i and x_j. The greater the distance, the lower the value of the kernel, hence, the smaller the denominator is and the closer the metric is to one. In the numerator, only those points within the same cluster are considered, so its value is lower than that of the denominator. For this reason, in general, this metric gives values that are close to one, always smaller, but, as previously mentioned, since the distance between the points inside the spectral clusters is usually higher, the value of the score is also higher and closer to the unit. In general, this is a more appropriate metric for the given data set than the Silhouette coefficient or the Davies-Bouldin score.

The main disadvantage when using τ metric is the computation of the kernel matrix:

- From the point of view of the computational cost, because it is necessary to go through all the points and compare one with the others.
- From the point of view of its dependency on the parameter σ whose modification leads to different results.

As an alternative index, the same calculation can be performed, substituting the kernel matrix by the euclidean distance adjacency matrix. In this case, the euclidean distance between all the points is computed. As it happened before, the value of τ will be smaller than one, since only those points within the same cluster are considered in the numerator, while all of them are taken into account in the denominator. Clusters in which the points are more dispersed yield higher values of K_{ij}, providing a higher denominator. However, they also provide a higher numerator, hence, as in the previous metric, the higher the score, the better the performance. Although, as it happened with the Silhouette coefficient or the Davies-Bouldin score, this metric would not be the most appropriate for the given data set, because points have a specific geometric distribution. This is the reason why the obtained values of the score are low and far from one, which would be the optimal result:

- K-means clustering = 0.2104823227051633
- Spectral clustering = 0.32361489326474285

Exercise 3

Leverage and lift values

In the following table there is a summary of the values for each of the rules presented in the exercise. Leverage and lift are two metrics that measure the strength of dependence. They have the following equations:

$$leverage = \delta(A=a, B=b) = P(A=a, B=b) - P(A=a)P(B=b)$$

$$lift = \gamma(A=a, B=b) = \frac{P(A=a, B=b)}{P(A=a)P(B=b)}$$

Num	fr(X)	fr(X,C)	P(X)	P(C)	P(X,C)	Leverage	Lift
1	300	125	0.300	0.3	0.125	0.0350	1.3889
2	500	150	0.500	0.3	0.150	0.0000	1.0000
3	500	400	0.500	0.7	0.400	0.0500	1.1429
4	342	240	0.342	0.7	0.240	0.0006	1.0025
5	2	2	0.002	0.7	0.002	0.0006	1.4289
6	500	352	0.500	0.7	0.352	0.0020	1.0057
7	260	100	0.260	0.3	0.100	0.0220	1.2821
8	120	32	0.120	0.3	0.032	-0.0040	0.8889
9	240	100	0.240	0.3	0.100	0.0280	1.3889
10	80	32	0.080	0.3	0.032	0.0080	1.3333
11	200	100	0.200	0.3	0.100	0.0400	1.6667
12	251	203	0.250	0.7	0.203	0.0273	1.1554

Two rules, A and B, show positive statistical dependence if $P(A,B) > P(A)P(C)$, hence $\delta > 0$ or $\gamma > 1$. Given this consideration, the two rules that must be pruned in this exercise are the second one and the eighth one.

Mutual information

Including the computation of the mutual information for the remaining rules, the following table is obtained.

Num	fr(X)	fr(X,C)	P(X)	P(C)	P(X,C)	Leverage	Lift	n x Mutual Information
1	300	125	0.300	0.3	0.125	0.0350	1.3889	19.43559
3	500	400	0.500	0.7	0.400	0.0500	1.1429	34.85155
4	342	240	0.342	0.7	0.240	0.0006	1.0025	0.00549
5	2	2	0.002	0.7	0.002	0.0006	1.4289	1.03038
6	500	352	0.500	0.7	0.352	0.0020	1.0057	0.05496
7	260	100	0.260	0.3	0.100	0.0220	1.2821	8.39876
9	240	100	0.240	0.3	0.100	0.0280	1.3889	14.2019

Num	fr(X)	fr(X,C)	P(X)	P(C)	P(X,C)	Leverage	Lift	n x Mutual Information
10	80	32	0.080	0.3	0.032	0.0080	1.3333	2.84666
11	200	100	0.200	0.3	0.100	0.0400	1.6667	32.26840
12	251	203	0.250	0.7	0.203	0.0273	1.1554	14.46150

According to the following rule:

$$\text{Prune if: } n.MI < 1.5$$

The rules 4, 5 and 6 must be removed, so the resulting table is:

Num	fr(X)	fr(X,C)	P(X)	P(C)	P(X,C)	Leverage	Lift	n x Mutual Information
1	300	125	0.30	0.3	0.125	0.0350	1.3889	19.43559
3	500	400	0.50	0.7	0.400	0.0500	1.1429	34.85155
7	260	100	0.26	0.3	0.100	0.0220	1.2821	8.39876
9	240	100	0.24	0.3	0.100	0.0280	1.3889	14.2019
10	80	32	0.08	0.3	0.032	0.0080	1.3333	2.84666
11	200	100	0.20	0.3	0.100	0.0400	1.6667	32.26840
12	251	203	0.25	0.7	0.203	0.0273	1.1554	14.46150

Evaluate overfitting among remaining rules

Considering the remaining rules, one by one:

Rule 1:

$$P(X,C) = P(smoking, heartdisease) = 0.125$$
$$P(X) = P(smoking) = 0.3$$
$$P(C) = P(heartdisease) = 0.3$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(heartdisease|smoking) = \frac{P(smoking, heartdisease)}{P(smoking)} = \frac{0.125}{0.3} = 0.4167$$

Rule 3:

$$P(X,C) = P(sports, \neg heartdisease) = 0.4$$
$$P(X) = P(sports) = 0.5$$
$$P(C) = P(\neg heartdisease) = 0.7$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(\neg heartdisease|sports) = \frac{P(sports, \neg heartdisease)}{P(sports)} = \frac{0.4}{0.5} = 0.8$$

Rule 7:

$$P(X,C) = P(female, stress, heartdisease) = 0.100$$
$$P(X) = P(female, stress) = 0.260$$
$$P(C) = P(heartdisease) = 0.3$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(heartdisease|female, stress) = \frac{P(female, stress, heartdisease)}{P(female, stress)} = \frac{0.1}{0.26} = 0.3846$$

Rule 9:

$$P(X,C) = P(smoking, coffee, heartdisease) = 0.100$$
$$P(X) = P(smoking, coffee) = 0.240$$
$$P(C) = P(heartdisease) = 0.3$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(heartdisease|smoking, coffee) = \frac{P(smoking, coffee, heartdisease)}{P(smoking, coffee)} = \frac{0.1}{0.24} = 0.4167$$

Considering the rule 1, in which:

$$P(C|Y) = \frac{P(Y,C)}{P(Y)} \rightarrow P(heartdisease|smoking) = \frac{P(smoking, heartdisease)}{P(smoking)} = \frac{0.125}{0.3} = 0.4167$$

It happens that:

$$P(C|Y) \geq P(C|X)$$

So the rule 9 can be pruned.

Rule 10:

$$P(X,C) = P(smoking, sports, heartdisease) = 0.032$$
$$P(X) = P(smoking, sports) = 0.08$$
$$P(C) = P(heartdisease) = 0.3$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(heartdisease|smoking, sports) = \frac{P(smoking, sports, heartdisease)}{P(smoking, sports)} = \frac{0.032}{0.08} = 0.4$$

As in the previous case, considering rule 1:

$$P(C|Y) = \frac{P(Y,C)}{P(Y)} \rightarrow P(heartdisease|smoking) = \frac{P(smoking, heartdisease)}{P(smoking)} = \frac{0.125}{0.3} = 0.4167$$

It happens that:

$$P(C|Y) \geq P(C|X)$$

So the rule 10 can be pruned.

Rule 11:

$$P(X,C) = P(stress, smoking, heartdisease) = 0.100$$
$$P(X) = P(stress, smoking) = 0.200$$
$$P(C) = P(heartdisease) = 0.3$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(heartdisease|stress, smoking) = \frac{P(smoking, stress, heartdisease)}{P(smoking, stress)} = \frac{0.1}{0.2} = 0.5$$

Considering rule 1:

$$P(C|Y) = \frac{P(Y,C)}{P(Y)} \rightarrow P(heartdisease|smoking) = \frac{P(smoking, heartdisease)}{P(smoking)} = \frac{0.125}{0.3} = 0.4167$$

In this case, $P(C|Y) \ngeq P(C|X)$, so rule 1 cannot be used for pruning rule 11.

Rule 12:

$$P(X,C) = P(female, sports, \neg heartdisease) = 0.203$$
$$P(X) = P(female, sports) = 0.251$$
$$P(C) = P(\neg heartdisease) = 0.7$$

Applying the Bayes' Rule:

$$P(C|X) = \frac{P(X,C)}{P(X)} \rightarrow P(\neg heartdisease|female, sports) = \frac{P(female, sports, heartdisease)}{P(female, sports)} = \frac{0.203}{0.251} = 0.8088$$

Considering rule 3:

$$P(C|Y) = \frac{P(Y,C)}{P(Y)} \rightarrow P(\neg heartdisease|sports) = \frac{P(sports, \neg heartdisease)}{P(sports)} = \frac{0.4}{0.5} = 0.8$$

Since $P(C|Y) \ngeq P(C|X)$, rule 12 cannot be pruned using rule 3.

Remaining rules:

Finally, the remaining rules are:

Num	fr(X)	fr(X,C)	P(X)	P(C)	P(X,C)	Leverage	Lift	n x Mutual Information
1	300	125	0.30	0.3	0.125	0.0350	1.3889	19.43559
3	500	400	0.50	0.7	0.400	0.0500	1.1429	34.85155
7	260	100	0.26	0.3	0.100	0.0220	1.2821	8.39876
11	200	100	0.20	0.3	0.100	0.0400	1.6667	32.26840
12	251	203	0.25	0.7	0.203	0.0273	1.1554	14.46150

Rule tea → heart disease

Considering the following information:

- $$fr(tea) = 390$$

- $$fr(tea, \neg heartdisease) = 283$$

- $$fr(tea, smoking) = 40$$

From Rule 1:

- $$fr(smoking) = 300$$

- $$fr(smoking, heartdisease) = 125$$

Now, the following can be computed:

$$P(heartdisease|smoking) = \frac{P(smoking, heartdisease)}{P(smoking)} = \frac{0.125}{0.3} = 0.4167$$

$$fr(heartdisease, \neg smoking) = fr(heartdisease)P(\neg smoking) - n\delta_{rule1} = 300(1 - 0.3) - 35 = 175$$

$$P(heartdisease, \neg smoking) = 0.175$$

$$P(heartdisease|\neg smoking) = \frac{P(heartdisease, \neg smoking)}{P(\neg smoking)} = \frac{0.175}{1 - 0.3} = 0.25$$

Since tea is independent of heart disease given smoking:

$$P(heardisease|tea, smoking) = P(heartdisease|smoking) = 0.4167$$

Since tea is independent of heart disease given $\neg smoking$

$$P(heartdisease|tea, \neg smoking) = P(heartdisease|\neg smoking) = 0.25$$

Finally, the expectation for $fr(tea, heartdisease)$ can be computed as:

$$fr_{exp}(tea, heartdisease) = fr(tea, smoking, heartdisease) + fr(tea, \neg smoking, heartdisease) =$$
$$= fr(tea, smoking)P(heartdisease|smoking) + fr(tea, \neg smoking)P(heartdisease|\neg smoking) =$$
$$= 40 \cdot 0.4167 + (390 - 40) \cdot 0.25 \approx 104$$

The obtained frequency is much lower than the provided one, hence the rule is not a strong statistical supposition.

Exercise 4

a)

Initially, a python script transforms all the names in the file into integers, by searching first all of them and removing those appearing more than one time. Then, using this vector of features, the script reads the file again and assigns an integer to each of the names.

```python
import pandas as pd
import numpy as np

fp = open('worlddiscr.names')
fo = open('worlddiscr_int.names', 'w')



words= [word.strip() for line in fp.readlines() for word in line.split(',') if word.strip()]
different_words = list(dict.fromkeys(words))

fp.close()
fp = open('worlddiscr.names')
it = 0
for line in fp.readlines():
    for word in line.split(','):
        if word.strip():
            index = different_words.index(word.strip())
            fo.write('%d' %index)
            fo.write(' ')
    if it > 0:
        fo.write("\n")
    it = it+1



fo.close()
fp.close()
```

Using the Linux command shell, the Kingfisher program is executed and the first one hundred best rules are extracted. The command used is:

```
./kingfisher -i worlddiscr_int.names -k227 -M-10 -t3
```

The obtained rules are evaluated using $ln(pf)$, which is the logarithm of the Fisher's exact p-value. It is a significance measure for the rules. The lower its value, the better the rule.

Rules related to ex-colonies:

There are two rules in which ex-colonies are involved:

$$11, 21, 22 \rightarrow 8$$

$$largestInfantMortality, exColony, young \rightarrow shortestLife$$

$$21 \rightarrow \neg 19$$

$$exColony \rightarrow mostLiterate$$

Rules related to corruption

There is no rule related to those countries that are the most corrupted ones. Neither for the opposite case (least corrupted countries).

Rules related to the length of the compulsory education

Related to those countries with a length of the compulsory education lower than 9 years:

$$52 \to 37$$

$$compulsoryEducation < 7y \to compulsoryEducation < 9y$$

So, having a length of the compulsory education smaller than 7 years implies having a length of the compulsory education lower than 9 years.

Rules related to the oil

There are no rules, among the 100 best ones, related to the oil.

Exercise 5

Maximal

When considering the maximal, it is not possible to find all positive statistical associations. The maximal, as stated in the textbook (Aggarwal, 2015) has the following definition: *A frequent itemset is maximal at a given minimum support level minsup, if it is frequent, and no superset of it is frequent.* In the given case, *minsup* can be interpreted as the minimum frequency min_{fr}. Hence, the maximal contains all the subsets that are frequent, meaning that they have a support higher than the minimum. However, the support of these subsets cannot be extracted from the maximal, so there is no enough information to compute the probability rules that lead to determine the statistical associations, since two variables X and C are positively associated if $P(X,C) > P(X)P(C)$ or, in other words, if they have a leverage greater than 0, and the leverage can be understood as: $leverage(X \to C) = support(X \to C) - support(X)support(C)$.

Considering the example of the figure 6, the maximals are (AB) and (BCD), so from them, the subsets can be extracted:

- $(AB) \to (A), (B)$
- $(BCD) \to (BC), (BD), (CD)$
  - $(BC) \to (B), (C)$
  - $(BD) \to (B), (D)$
  - $(CD) \to (C), (D)$

However, even though the frequencies for (AB) and (BCD) are known, there is no way to extract those of the subsets to compute the leverage and define the potential statistical association.

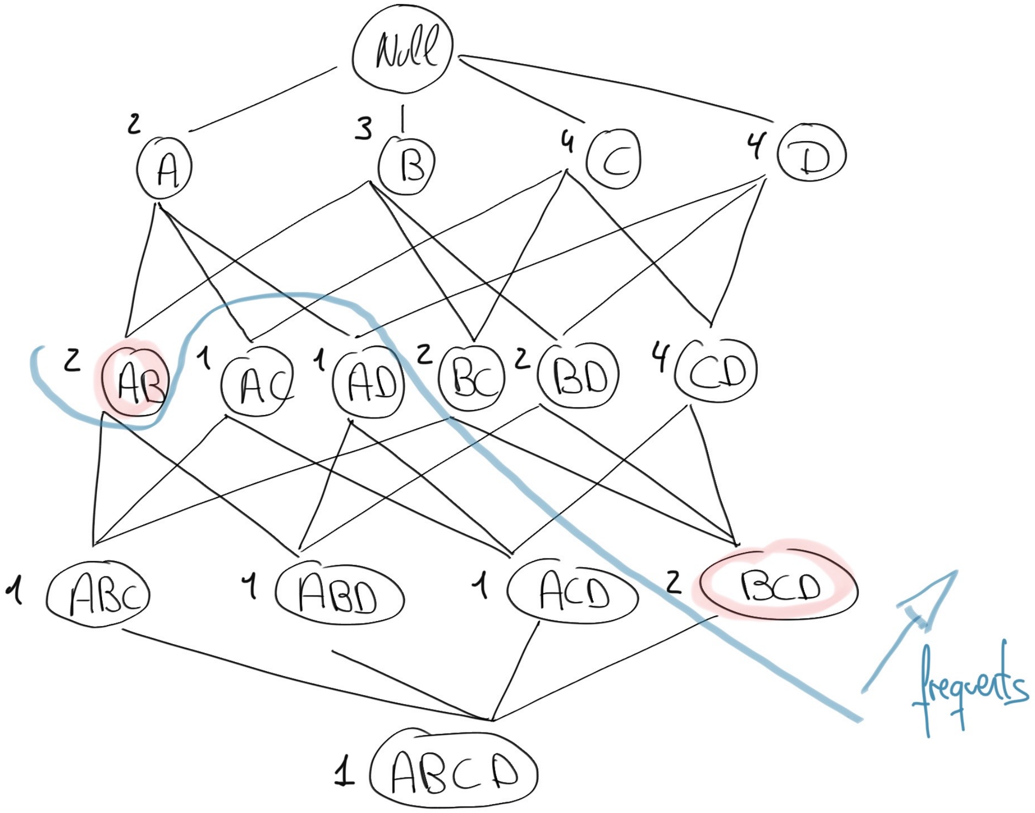

Figure 6: Maximals with $min_{fr} = 2/n$ (red).

Closed

Defining the closed frequent itemsets as *itemsets for which none of their supersets have exactly the same support count as the considered one* (Aggarwal, 2015). In the figure 7, there is the representation of the previous example.

- Knowing that (BCD) and (B) are closed sets with frequencies 2 and 3 respectively, the supersets of (B) which are also subsets of (BCD), (BC) nad (BD) must have lower frequency than (B) (because B is a closed set) and must be frequent without being closed sets, so the only alternative is frequency 2.
- (CD) is another closed set with frequency 4, so (D) must have frequency lower or equal to 4, otherwise it would also be a closed set and it is not. The same situation happens with (C).
- (AB) is another closed set with frequency 3, hence (A) must have frequency lower or equal to 3, otherwise, it would also be a closed set and it is not.

Making use of the previous information, the statistical dependencies can be derived.

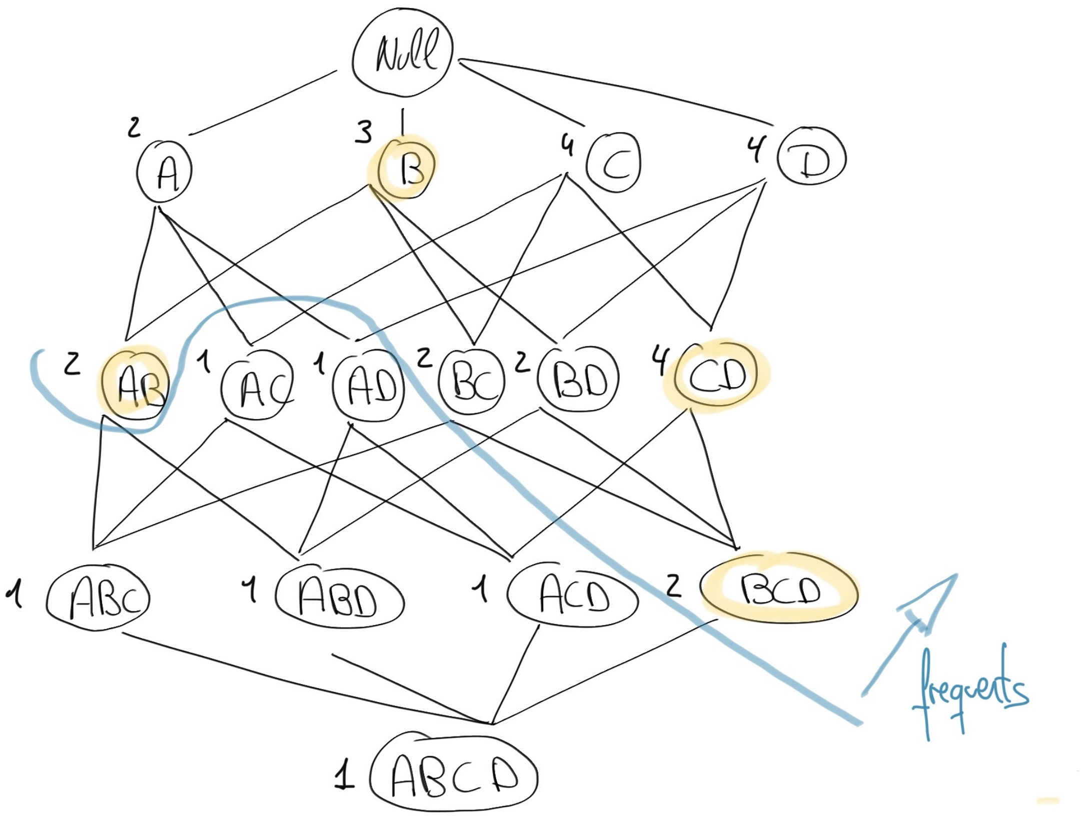

Figure 7: Closed with $min_{fr} = 2/n$ (yellow).

Free

As it happened before with the closed sets, the free sets enable the computation of the positive statistical associations since the first row of variables will always be composed by free sets, hence the frequency of the following ones can be computed.

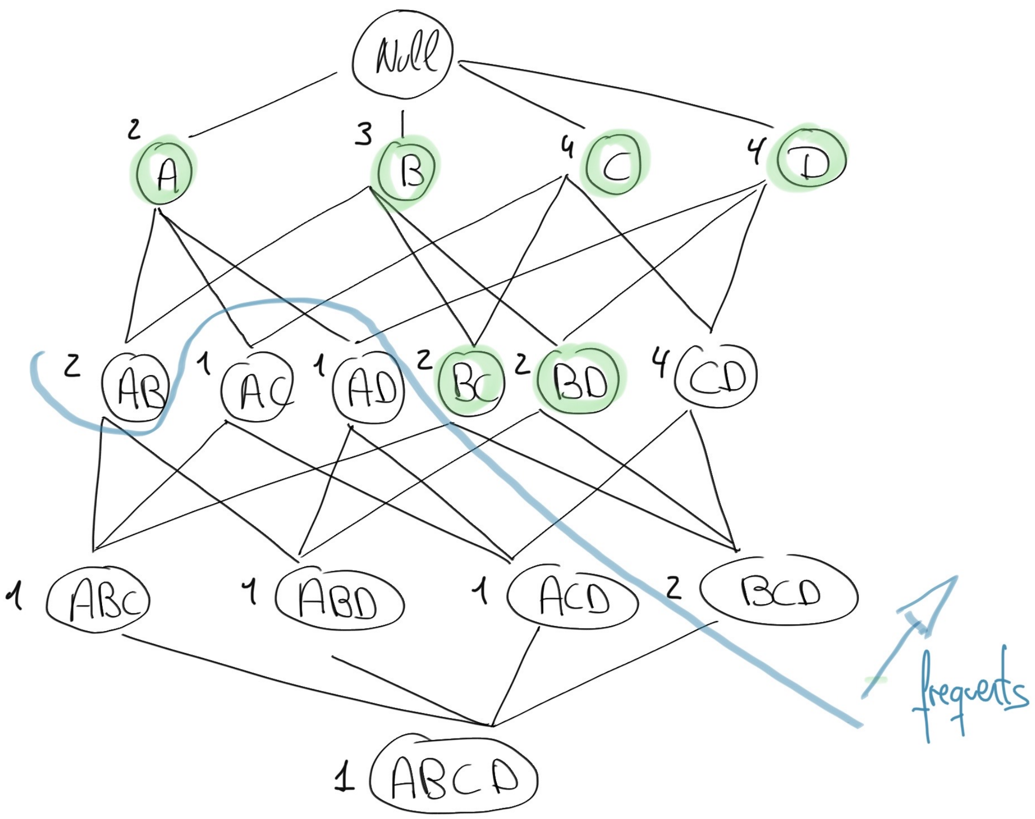

Figure 8: Free with $min_{fr} = 2/n$ (green).

Overfitted rules given maximal, closed or free sets

There is no possibility to detect overfitted rules given only maximal, closed or free sets. This happens because, even though, in many of the mentioned subsets, it is possible to obtain the support for all the frequent items, it is not possible to compute their confidence. Actually, the property of *Confidence Monotonicity* is only valid for itemsets belonging to the same superset, so it can not be applied to some of the cases.