Gustav Sternelöv

732A31 Data Mining - Lab 2

In the lab is three different experiments conducted. In the first experiment is the respective features discretized into three separate bins and for the simpleKMeans clustering is k=3. In the second experiment is *k* changed to 4 and the number of bins is held fixed. For the last experiment is it instead the number of bins that is 4 and the number of clusters is held fixed at k=3.

The number of different states the features are divided into decides which data that is used in the SimpleKMeans clustering. The *k* decides the number of clusters that are created. These clusters are then further analyzed with an association analyses where the best rules with the respective clusters as consequents are the rules of interest.

So, the workflow then is: 1. Construct the bins. 2. Look at the SimpleKMeans clustering briefly. 3. Do the association analysis and look at the best rules for the respective clusters.

3 clusters and 3 bins

The bins

The features in the data set with numerical data is discretized into the following states.

Sepal length

No.	Label	Count
1	'(-inf-5.5]'	59
2	'(5.5-6.7]'	71
3	'(6.7-inf)'	20

Sepal width

No.	Label	Count
1	'(-inf-2.8]'	47
2	'(2.8-3.6]'	88
3	'(3.6-inf)'	15

Petal length

No.	Label	Count
1	'(-inf-2.966667]'	50
2	'(2.966667-4.9...	54
3	'(4.933333-inf)'	46

Petal width

No.	Label	Count
1	'(-inf-0.9]'	50
2	'(0.9-1.7]'	54
3	'(1.7-inf)'	46

The clustering

Next is a brief analysis of the clustering for a SimpleKMeans with K=3 and seed=10 conducted. The number of observations assigned into the respective clusters is given by the table below.

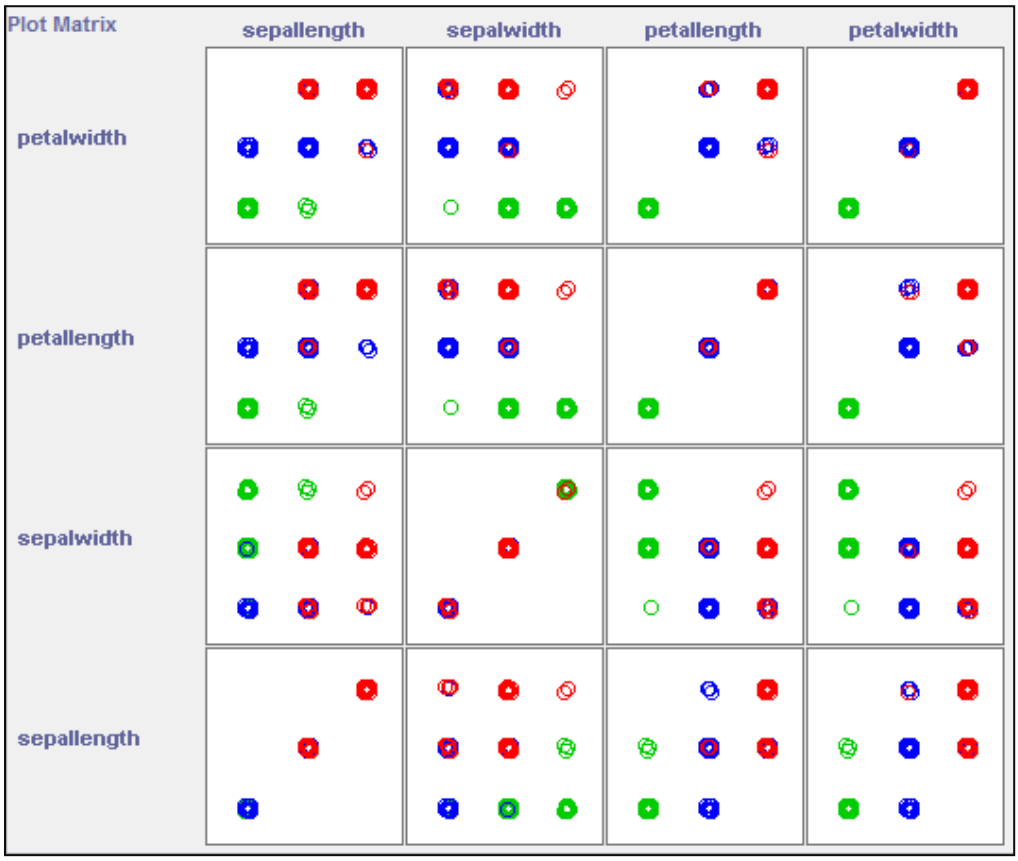
No.	Label	Count
1	cluster1	55
2	cluster2	45
3	cluster3	50

Since the true class is known for every data point the classification performed by the clustering can be compared to the true classes. The table below gives this information and it can be concluded that 9 data points has been clustered incorrectly.

```
  0  1  2  <-- assigned to cluster
  0  0 50 | Iris-setosa
 48  2  0 | Iris-versicolor
  7 43  0 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-setosa
```

A visual analysis of the clustering is performed with help of the following graph were cluster 0 is blue, cluster 1 is red and cluster 2 is green.



A short resume of the characteristics for each cluster:

- Cluster 0 contains the values in the "mid-bin" for petal width and values in the same interval for petal length it seems. Low values for sepal width and mostly low or mid values for sepal length.
- Cluster 1 has low values for petal width and length. High values for sepal width and low values for sepal length.
- Cluster 2 has high values for petal width and petal length. Mid-values for sepal width and mostly high values for sepal length.

The association rules

Minimum support is 10 % and minimum confidence is 90 %. The names of the clusters have changed in the output of the association analysis. Cluster 0 now is cluster 1, cluster 1 is cluster 2 and cluster 2 is now called cluster 3.

Cluster 1 as consequent

The rules in the table below are the three best rules obtained when cluster 1 is the consequent and class attribute not is in the antecedent. A petal length between 2.97 and 4.93, the mid-bin, and a petal width between 0.9-1.7, also the mid-bin, are the attributes in the antecedent in the rule with the highest support and cluster 1 as consequent. The antecedent for the second rule has the mid-bin interval for sepal length together with the objects in the antecedent for the first rule. In the third rule is also the last feature, sepal width, introduced. The first interval for this feature, -inf to 2.8, is combined with the mid-interval for petal width.

To summarize the rules it is concluded that cluster 1 contains the data points with mid-values for all the features except sepal width (lowest interval).

Rule	Support	Confidence
petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' 48 ==> cluster=cluster1	48 %	100 %
sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' ==> cluster=cluster1	33 %	100 %
sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]' ==> cluster=cluster1	31 %	100 %

Cluster 2 as consequent

With cluster 2 as consequent the antecedent for the first rule contains the highest intervals for the features petal length and petal width. The second rule combines the mid-interval for sepal width with petal width and the third combines sepal width with petal length.

Hence, the second cluster consists of data points with high values for petal length and petal width and mid-values for sepal width.

Rule	Support	Confidence
petallength='(4.933333-inf)' petalwidth='(1.7-inf)' ==> cluster=cluster2	40 %	100 %
sepalwidth='(2.8-3.6]' petalwidth='(1.7-inf)' ==> cluster=cluster2	29 %	100 %
sepalwidth='(2.8-3.6]' petallength='(4.933333-inf)' ==> cluster=cluster2	28 %	100 %

Cluster 3 as consequent

Low values of petal length points to cluster 3 and so does also low values of petal width and the rule where this intervals for the two features are combined. All rule has an support of 50 % and a confidence level of 100 %.

Rule	Support	Confidence
petallength='(-inf-2.966667]' ==> cluster=cluster3	50 %	100 %
petalwidth='(-inf-0.9]' ==> cluster=cluster3	50 %	100 %
petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' ==> cluster=cluster3	50 %	100 %

4 clusters and 3 bins

The bins

In this experiment is the states of the bins exactly the same as in the first experiment.

The clustering

The clusters for a SimpleKMeans with K=4 and seed=10.

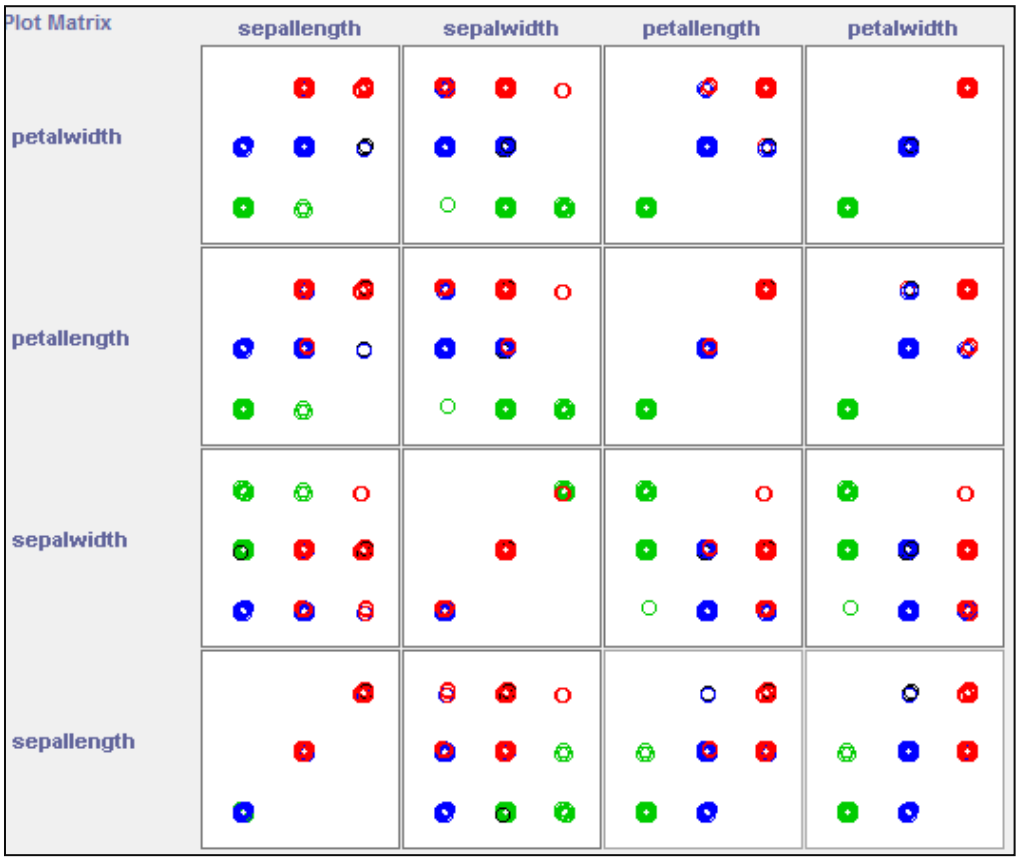
No.	Label	Count
1	cluster1	52
2	cluster2	44
3	cluster3	50
4	cluster4	4

13 of the data points has been clustered incorrectly.

```
   0   1   2   3   <-- assigned to cluster
   0   0  50   0 | Iris-setosa
  45   2   0   3 | Iris-versicolor
   7  42   0   1 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-virginica
Cluster 2 <-- Iris-setosa
Cluster 3 <-- No class
```

Cluster 0 is blue, cluster 1 is red, cluster 2 is green and cluster 3 is black.



A short resume of the characteristics for each cluster:

- Cluster 0 contains the values in the mid-bins for petal width and petal length. For sepal width and sepal length it mostly seem to contain low values.
- Cluster 1 has high values for petal width and petal length. It is slightly harder to analyze the remaining features but it looks like the red cluster in general has mid-values for sepal width and sepal length.
- Cluster 2 has low values for petal width, petal length and sepal length. For sepal width it has mid- or high values.
- The characteristics of cluster 3 are hard to analyze visually since it contains so few data points.

The association rules

Minimum support is 10 % and minimum confidence is 90 %. The names of the clusters have changed in the same way as in the first experiment.

Cluster 1 as consequent

The best rules with cluster 1 as consequent has an support around 30 % and a confidence at 100 %. In the antecedent for the first rule are the mid-bins for sepal length, petal length and petal width. Both the second and the third rule contains the lowest interval for sepal width values. In one case combined with petal width and in the other with petal length.

Thus, the rules for cluster 1 in the consequent has mid-values for all features except sepal width that has the bin with low values associated to the cluster.

Rule	Support	Confidence
sepallength='(5.5-6.7]' petallength='(2.966667-4.933333]' petalwidth='(0.9-1.7]' ==> cluster=cluster1	33 %	100 %
sepalwidth='(-inf-2.8]' petalwidth='(0.9-1.7]' ==> cluster=cluster1	31 %	100 %
sepalwidth='(-inf-2.8]' petallength='(2.966667-4.933333]' ==> cluster=cluster1	30 %	100 %

Cluster 2 as consequent

The first rule has high values of petal length and petal width in the antecedent, a support of 40 % and a confidence of 100 %. In the next rule is the mid-values for sepal width combined with petal width. The third rule combines the first rule with sepal width. For the second and third rule is the support just below 30 % and the confidence is 100 %.

By looking on the best rules the cluster seem to include the data points with high values for petal length and petal width and the mid-values for sepal width.

Rule	Support	Confidence
petallength='(4.933333-inf)' petalwidth='(1.7-inf)' ==> cluster=cluster2	40 %	100 %
sepalwidth='(2.8-3.6]' petalwidth='(1.7-inf)' ==> cluster=cluster2	29 %	100 %
sepalwidth='(2.8-3.6]' petallength='(4.933333-inf)' petalwidth='(1.7-inf)' ==> cluster=cluster2	26 %	100 %

Cluster 3 as consequent

The support is 50 % and confidence 100 % for the three best rules. In the first rule is low values for petal length antecedent and in the second rule low values of petal width. The third rule is a combination of the two first.

Cluster 3 is concluded to be a cluster with data points that has low values for petal length and petal width.

Rule	Support	Confidence
petallength='(-inf-2.966667]' ==> cluster=cluster3	50 %	100 %
petalwidth='(-inf-0.9]' ==> cluster=cluster3	50 %	100 %
petallength='(-inf-2.966667]' petalwidth='(-inf-0.9]' ==> cluster=cluster3	50 %	100 %

Cluster 4 as consequent

To obtain rules for this cluster the support had to been lowered substantially since the cluster only consists of four data points. The earlier minimum support was 10 % and is now 1 %. The minimum confidence on the other hand remains unchanged.

The rules obtained under the mentioned circumstances is shown in the table below. The first has a support of 3 % and says that data points with high values of sepal length and mid-values of sepal width and petal width with a confidence of 100 % belongs to cluster 4. The second rule in the table has a support of 2 % and the only difference compared to the first rule is that petal width is replaced by the mid-bin for petal length.

Rule	Support	Confidence
sepallength='(6.7-inf)' sepalwidth='(2.8-3.6]' petalwidth='(0.9-1.7]' ==> cluster=cluster4	3 %	100 %
sepallength='(6.7-inf)' sepalwidth='(2.8-3.6]' petallength='(2.966667-4.933333]' ==> cluster=cluster4	2 %	100 %

3 clusters and 4 bins

In the next experiment is the performed clustering a SimpleKMeans with K=3 and seed=10 and the number of bins is exceeded from 3 to 4.

The bins

The states for the bins now are the following. Regarding for example sepal length is the main difference that the middle bin is divided into two separate parts and that the first and last bin are slightly smaller than before.

Sepal length

No.	Label	Count
1	'(-inf-5.2]'	45
2	'(5.2-6.1]'	50
3	'(6.1-7]'	43
4	'(7-inf)'	12

Sepal width

No.	Label	Count
1	'(-inf-2.6]'	24
2	'(2.6-3.2]'	84
3	'(3.2-3.8]'	36
4	'(3.8-inf)'	6

Petal length

Petal width

No.	Label	Count
1	'(-inf-2.475]'	50
2	'(2.475-3.95]'	11
3	'(3.95-5.425]'	61
4	'(5.425-inf)'	28

No.	Label	Count
1	'(-inf-0.7]'	50
2	'(0.7-1.3]'	28
3	'(1.3-1.9]'	43
4	'(1.9-inf)'	29

The clustering

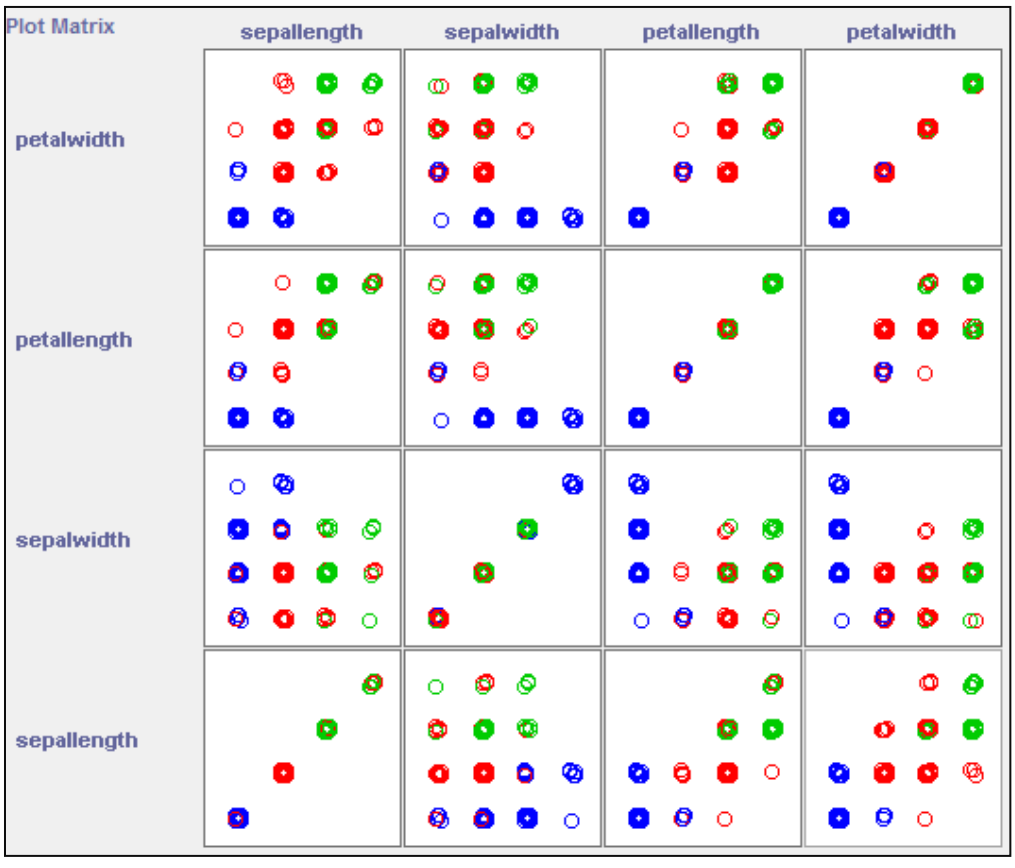
No.	Label	Count
1	cluster1	54
2	cluster2	66
3	cluster3	30

24 of the data points has been clustered incorrectly.

```
  0  1  2  <-- assigned to cluster
 50  0  0 | Iris-setosa
  4 46  0 | Iris-versicolor
  0 20 30 | Iris-virginica

Cluster 0 <-- Iris-setosa
Cluster 1 <-- Iris-versicolor
Cluster 2 <-- Iris-virginica
```

Cluster 0 is blue, cluster 1 is red and cluster 2 is green.



A short resume of the characteristics for each cluster:

- Cluster 0 contains the low values for petal width, petal length and sepal length. For sepal width it contains the high or moderately high values.
- Cluster 1 has either moderately low or moderately high values for petal width. Moderately high values for petal length and moderately low values for sepal width and sepal length.
- Cluster 2 has high values for petal width and high or moderately high for petal length and sepal length. For sepal width the values seem to be either moderately low or moderately high.

The association rules

Cluster 1 as consequent

Support of 50 % and confidence of 100 % for all of the three best rules. The first rule has the lowest values of petal length in the antecedent and the second rule the lowest values of petal width. The last rule has an antecedent that combines the two earlier rules.

Associated to the cluster is low values of petal length and petal width.

Rule	Support	Confidence
petallength='(-inf-2.475]' ==> cluster=cluster1	50 %	100 %
petalwidth='(-inf-0.7]' ==> cluster=cluster1	50 %	100 %
petallength='(-inf-2.475]' petalwidth='(-inf-0.7]' ==> cluster=cluster1	50 %	100 %

Cluster 2 as consequent

The support for the rules is around 30 % and the confidence is 100 % for all rules. In the antecedent for the first rule are moderately high values of both petal length and petal width The second rule includes moderately low values of sepal length together with petal length rule. In the last rule is the same interval of values or sepal length combined with moderately low values of sepal width.

So, the values in cluster 2 is associated with moderately high values for petal length and petal width and moderately low values for sepal length and sepal width.

Rule	Support	Confidence
petallength='(3.95-5.425]' petalwidth='(1.3-1.9]' ==> cluster=cluster2	33 %	100 %
sepallength='(5.2-6.1]' petallength='(3.95-5.425]' ==> cluster=cluster2	32 %	100 %
sepallength='(5.2-6.1]' sepalwidth='(2.6-3.2]' ==> cluster=cluster2	26 %	100 %

Cluster 3 as consequent

The support is just under 20 % for all rules and the confidence is 100 %. High values of petal length and petal width is in the antecedent for the first rule. Petal width with moderately high values of sepal length in the antecedent for the second rule and sepal length together with petal length is the antecedent of the third best rule.

Cluster 3 is associated to high values of petal length and petal width and moderately high values of sepal length.

Rule	Support	Confidence
petallength='(5.425-inf)' petalwidth='(1.9-inf)' ==> cluster=cluster3	19 %	100 %
sepallength='(6.1-7]' petalwidth='(1.9-inf)' ==> cluster=cluster3	18 %	100 %

sepallength='(6.1-7]' petallength='(5.425-inf)' ==> cluster=cluster3	15 %	100 %

Summary

The clustering with four clusters gave the worst clustering results. A logical result since it is known that there are three classes.

Regarding the number of bins so were better results obtained with three bins than with four bins. **Why?**

More alternatives in the last experiment, with four bins, makes it harder to construct as general groups as with three bins.

Clearly harder to separate two of the classes in the last experiment.

The second experiment. Created like a "outlier group"?

Do a cluster analysis, analyze it briefly.

Use association analysis to assist the cluster analysis.

Want to describe the instances grouped in each cluster.

- In each of the 3 (or 4) experiments you should identify at least one (possibly the best) association rule for each of the clusters.
- In addition, try to explain the differences between results in different experiments as well as reasons for why these differences occured.