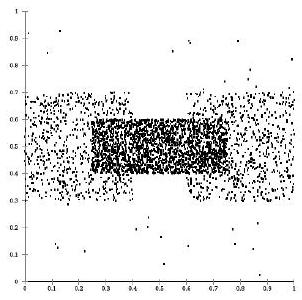
**3. Exercise 9**

**4. Exercise 9-1 OPTICS Plot**

1. For the data below we got computed the reachability diagram to the right.



With a nälve understanding of hierachical clustering, wouldn't we have expected three valleys in the plot?

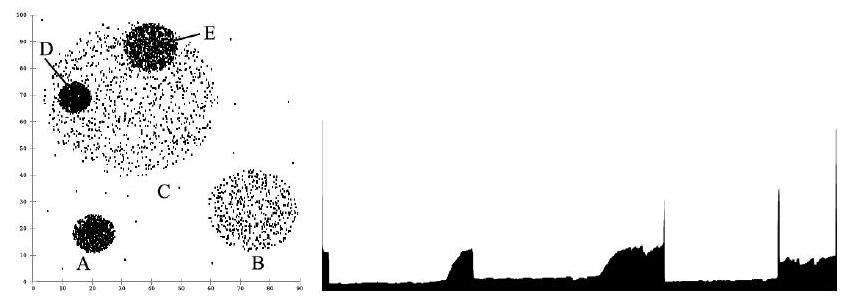
Explain, why this is not the case and why the plot, instead, looks as it does and accurately describes the density structure of the data.

**5. Suggested solution:**

The clusters to the left and to the right are density-connected by the more dense, central cluster. OPTICS aims at finding clusters starting from the more dense parts. Once the seed list contains objects from the densest part, it can extend the structure towards both sides simultaneously.

This is an extreme case of the "density-linkage" effects (for single-link: the single-linkeffect, in OPTICS, choosing a larger minPts is a way to reduce this effect but only excluding densities below the threshold).

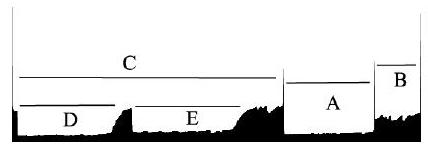
Understanding the density-based cluster-model, we indeed have here only one cluster of lower density that contains a cluster of higher density. 2. For this dataset (left) we have the reachability plot (right).



Mark in the reachability plot which areas relate to the clusters , and .

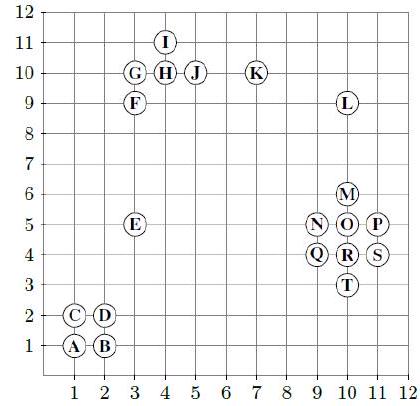
**6. Suggested solution:**

Solution:



**7. Exercise 9-2 Outlier Scores**

Given the following 2 dimensional data set:



As distance function, use Manhattan distance .

Compute the following (without including the query point when determining the ):

LOF using for the points and

LOF using for the points and

distance using for all points.

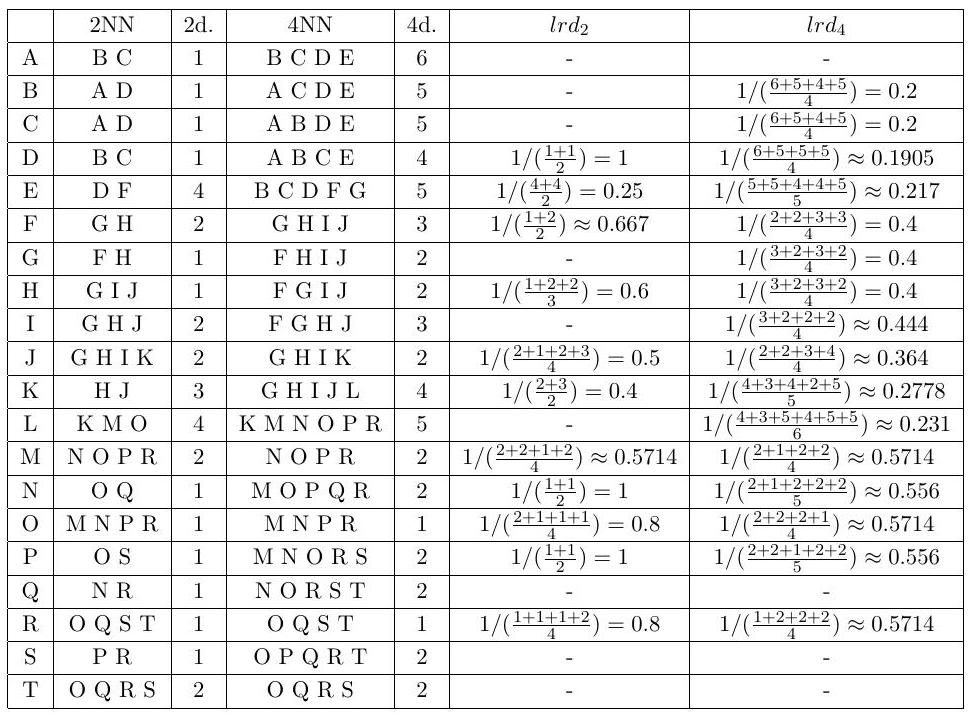
distance using for all points.

aggregated distances for and for all points

(aggregated distance averaged sum of the distances to all the )

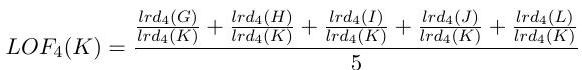
**8. Suggested solution:**

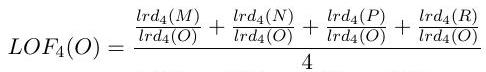
We find and distance for all points using and respectively. We then calculate the local reachability density .



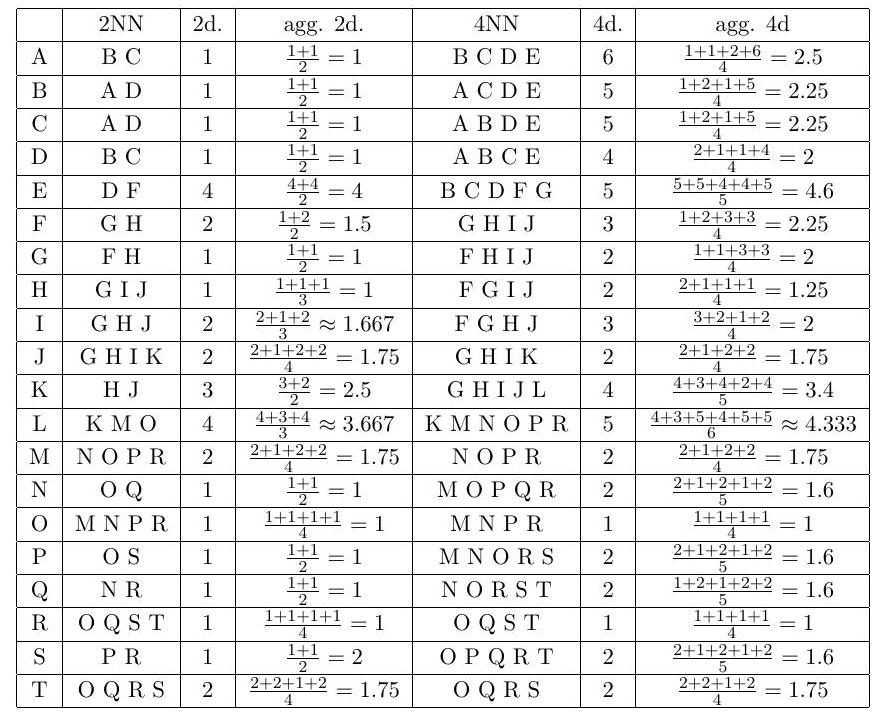
Recall the formula for the local outlier factor :

scores for





Aggregated distances for and



**9. Exercise 9-3 Evaluation of Outlier Scores**

A data set with known outliers + was evaluated using two outlier detection methods and . The results of the methods are given in the table below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Object |  |  |  |  |  |  |  |  |  |  |
| Label |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Evaluate both outlier detection methods and using the following metrics:

Precision, Recall and F-Measure, assuming that the top ranked outliers were classified as outliers.

Average Precision for , assuming that the top ranked outliers were classified as outliers.

Draw the ROC curve, and compute the area under curve (AUC) measure.

**10. Suggested solution:**

Sorted w.r.t. :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Label |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

Sorted w.r.t. :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Label |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

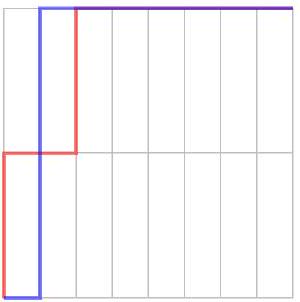
ROC curves:

Recall for ROC curves:

each TP in the ranking: one step up

* each FP in the ranking: one step to the right

comparison of two rankings: area under the curve (ROC



Area in both cases:

**11. Exercise 9-4 Decision Trees**

Predict the risk class of a car driver based on the following attributes:

. Time since getting the driving license years, years, years

Gender (male, female)

Residential area (urban, rural)

For your analysis you have the following manually classified training examples:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Person | Time since license | Gender | Area | Risk class |
| 1 |  |  | urban | low |
| 2 |  |  | rural | high |
| 3 |  |  | rural | low |
| 4 |  |  | rural | high |
| 5 |  |  | rural | high |
| 6 |  |  | rural | high |
| 7 |  |  | urban | low |
| 8 |  |  | urban | low |

1. Construct a decision tree based on this training dataset. Use information gain for selecting the split attributes. Build a separate branch for each attribute. The decision tree shall stop when all instances in the branch have the same class, you do not need to apply a pruning algorithm.

**12. Suggested solution:**

Remember: split of by selection of attribute in partitions :

where is the probability of randomly selecting an example in class , and is the number of classes.

We can see that are low and are high. Calculating the entropy, we get

Now we want to calculate the information gain for each attribute.

* IG 'Time since license'
* 1-2 years: persons

years: persons

years: persons 3,5

Thus, we can calculate the information-gain for 'Time since license'.

**13. IG 'Gender'**

persons

persons

Thus, we can calculate the information-gain for 'Gender'.

* IG 'Area'
* urban: persons

rural: persons

Thus, we can calculate the information-gain for 'Area'.

Area has the largest information-gain.

Split 2, right branch:

entropy

* IG 'Time since license'
* 1-2 years: persons 4,6

years: person 2

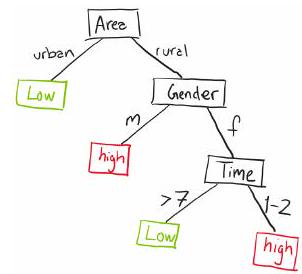
years: persons 3,5

IG 'Gender'

persons

persons 3,4

You can choose one of the two attributes arbitrarily, and sketch the resulting tree.



1. Apply the decision tree to the following drivers:

Person A: , f, rural

Person B: , urban

Person C:

Suggested solution:

Person A: , f, rural high

Person B: , urban low

Person C: , urban low