TU Dortmund 26.5.2022
Artificial Intelligence / Data Science and Data Engineering **Machine Learning for Complex Data**
Prof. Dr. Erich Schubert
Prof. Dr. Emmanuel Müller **Summer 2022**
Daniel Wilmes

Assignment 6: Gaussian Mixture Modeling, CLIQUE and DBSCAN

Due: Thursday, 2.6.2022

Problem 6-1 Expectation in Gaussian Mixture Models — 7

Given a data set with 100 points consisting of three Gaussian clusters A, B and C and the point p.

Cluster A

- contains 30% of all objects
- has the mean $\mu_A = (2,2)$
- has the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$

Cluster B

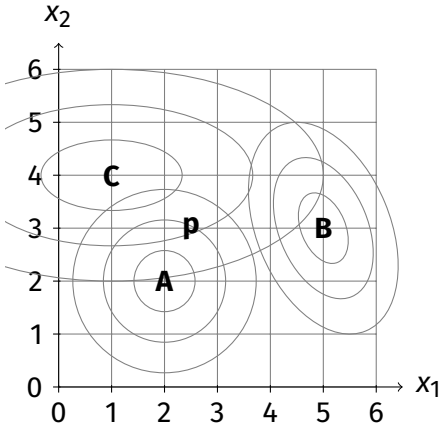
- contains 20% of all objects
- has the mean $\mu_B = (5,3)$
- has the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$

Cluster C

- contains 50% of all objects
- has the mean $\mu_C = (1,4)$
- has the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$

The point p is given by the coordinates $(2.5,3.0)$.

The following sketch is *not exact*, and only gives a rough idea of the cluster locations:



Perform the Expectation-Step (E of EM) of Gaussian Mixture Modeling (GMM).
Show how to compute the following:

a) The squared Mahalanobis distances of p to the mean of each Gaussian

b) The probability density of p with respect to each Gaussian: $P(p|A)$, $P(p|B)$, $P(p|C)$

c) The conditional probabilities $P(A|p)$, $P(B|p)$, $P(C|p)$

d) Which cluster does point p most likely belong to?
How clear is this decision? Justify your answer.

Hint: In the multivariate Gaussian, $|\Sigma|$ denotes the matrix determinant.

Problem 6-2 Grid-based Subspace Clustering with CLIQUE **7**

Consider the four-dimensional dataset:

	d1	d2	d3	d4
A	15	12	16	9
B	14	13	18	3
C	12	14	14	15
D	16	13	19	19
E	5	6	9	4
F	4	11	10	18
G	6	17	8	13
H	6	9	14	16
I	14	19	13	15
J	19	3	15	14

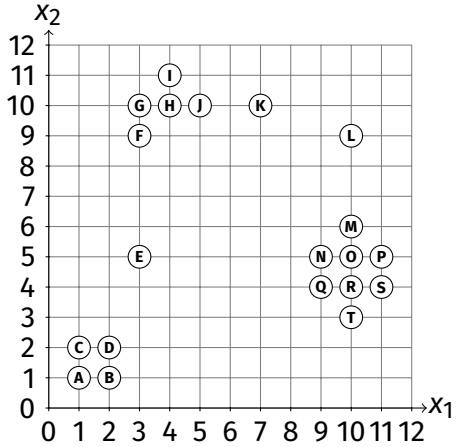
Cluster the dataset using CLIQUE. Use the following parameters:

a) Density threshold = 3 points and number of intervals = 2

b) Density threshold = 5 points and number of intervals = 2

Problem 6-3 DBSCAN **9**

Given the following data set from the lecture:



As distance function use the Manhatten Distance:

$$L_1(x,y) = |x_1 - y_1| + |x_2 - y_2|$$

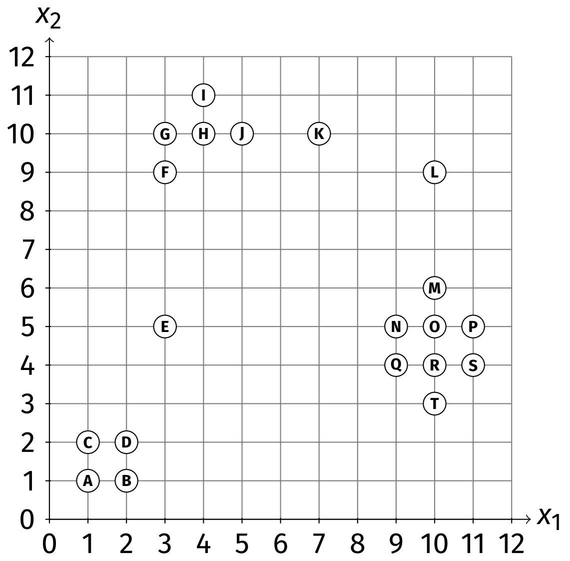
Compute DBSCAN for each of the following parameter settings, and indicate which points are core points, border points and noise points:

- $\varepsilon = 1.1$ and $\text{minPts} = 3$
- $\varepsilon = 1.1$ and $\text{minPts} = 4$
- $\varepsilon = 4.1$ and $\text{minPts} = 5$

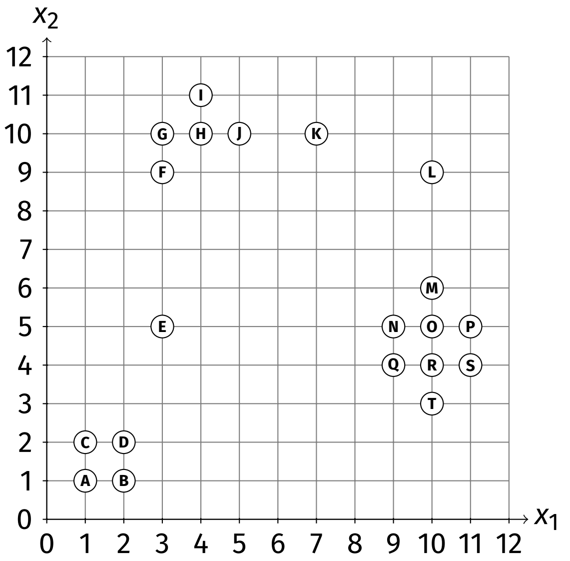
The order in which you chose points, when there are multiple to chose from, is up to you.

Hint: It is recommended to do this on paper with multiple colors.

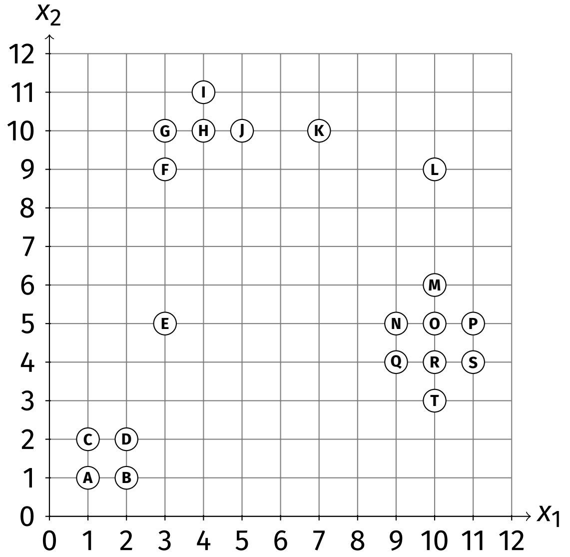
Templates:



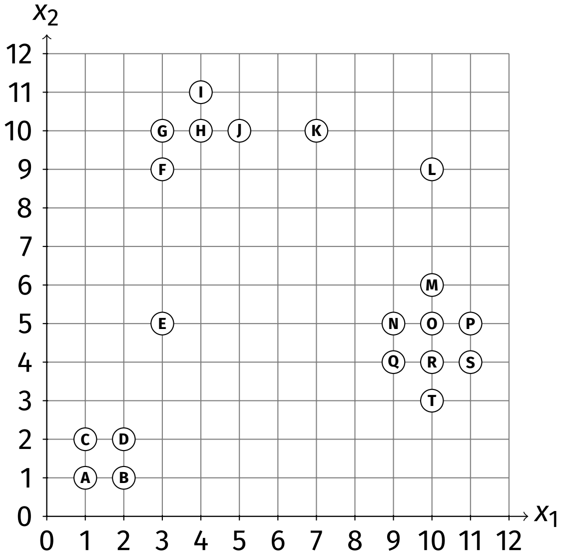
x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T



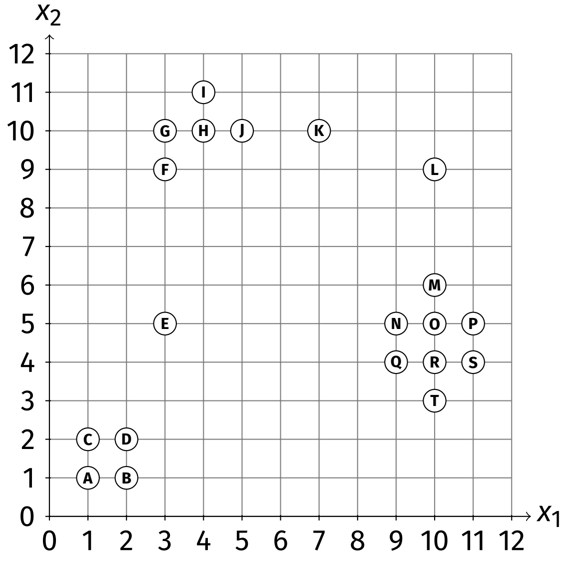
x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T



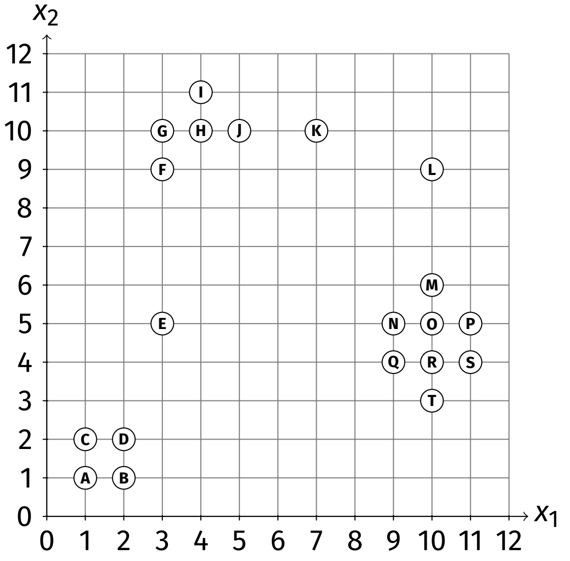
x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T



x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T



x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T



x_2
x_1
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T