Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

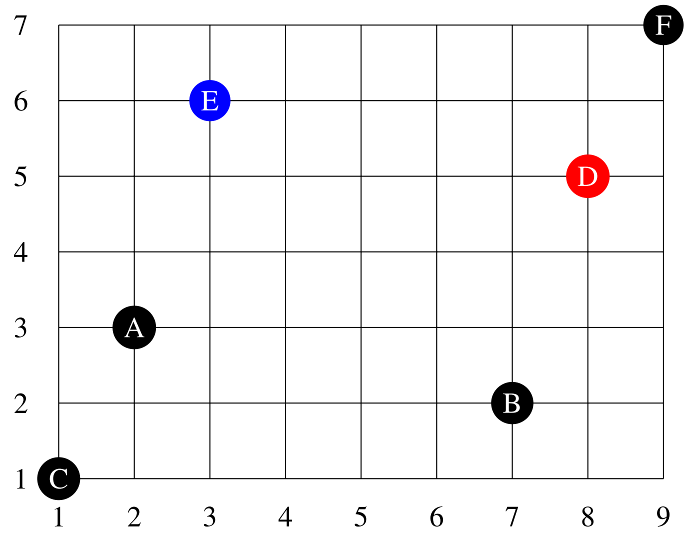
Knowledge Discovery and Data Mining I
WS 2018/19

Exercise 6: k-Medoid, EM, DBSCAN

Exercise 6-1 K-Medoid (PAM)

Consider the following 2-dimensional data set:

	A	B	C	D	E	F
x_1	2	7	1	8	3	9
x_2	3	2	1	5	6	7



(a) Perform the first loop of the PAM algorithm ($k = 2$) using the Manhattan distance. Select D and E (highlighted in the plot) as initial medoids and compute the resulting medoids and clusters.
Hint: When $C(m)$ denotes the cluster of medoid m, and M denotes the set of medoids, then the total distance TD may be computed as

$$TD = \sum_{m \in M} \sum_{o \in C(m)} d(m, o)$$

(b) How can the clustering result $C_1 = \{A, B, C\}, C_2 = \{D, E, F\}$ be obtained with the PAM algorithm ($k = 2$) using the weighted Manhattan distance

$$d(x, y) = w_1 \cdot |x_1 - y_1| + w_2 \cdot |x_2 - y_2|?$$

Assume that B and E are the initial medoids and give values for the weights w_1 and w_2 for the first and second dimension respectively.

Exercise 6-2 Convergence of PAM

Show that the algorithm PAM converges.

Exercise 6-3 Assignments in EM-Algorithm

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

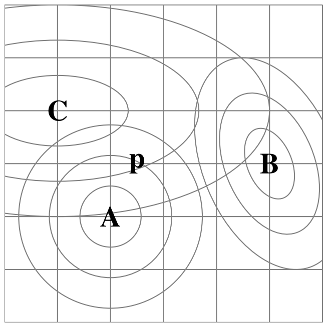
The cluster A contains 30% of all objects and is represented using the mean of all his points $\mu_A = (2,2)$ and the covariance matrix $\Sigma_A = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$.

The cluster B contains 20% of all objects and is represented using the mean of all his points $\mu_B = (5,3)$ and the covariance matrix $\Sigma_B = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$.

The cluster C contains 50% of all objects and is represented using the mean of all his points $\mu_C = (1,4)$ and the covariance matrix $\Sigma_C = \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix}$.

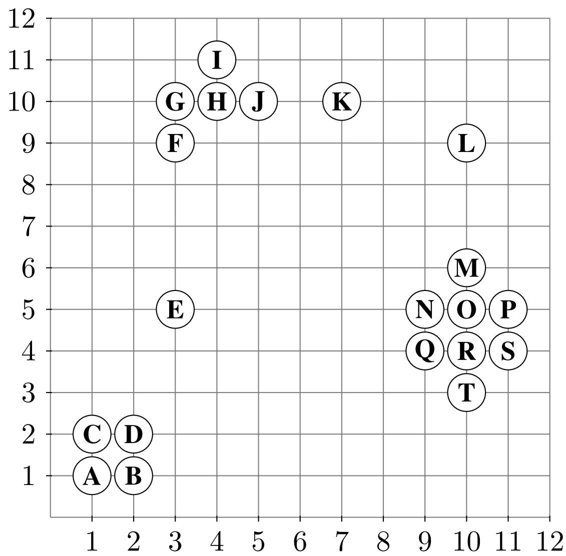
The point p is given by the coordinates $(2.5, 3.0)$.
Compute the three probabilities of p belonging to the clusters A, B and C.

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



Exercise 6-4 DBSCAN

Given the following data set:



As distance function, use Manhattan Distance:

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

Compute DBSCAN and indicate which points are core points, border points and noise points.
Use the following parameter settings:

- Radius $\varepsilon = 1.1$ and *minPts* $= 2$
- Radius $\varepsilon = 1.1$ and *minPts* $= 3$
- Radius $\varepsilon = 1.1$ and *minPts* $= 4$
- Radius $\varepsilon = 2.1$ and *minPts* $= 4$
- Radius $\varepsilon = 4.1$ and *minPts* $= 5$
- Radius $\varepsilon = 4.1$ and *minPts* $= 4$

Exercise 6-5 Properties of DBSCAN

Discuss the following questions/propositions about DBSCAN:

- Using *minPts* $= 2$, what happens to the border points?
- The result of DBSCAN is deterministic w.r.t. the core and noise points but not w.r.t. the border points.
- A cluster found by DBSCAN cannot consist of less than *minPts* points.
- If the dataset consists of n objects, DBSCAN will evaluate exactly n ϵ-range queries.
- On uniformly distributed data, DBSCAN will usually either assign all points to a single cluster or classify every point as noise. k-means on the other hand will partition the data into approximately equally sized partitions.