Information Retrieval and Data Mining SS 2017
Prof. Dr.-Ing. Sebastian Michel
MSc. Koninika Pal
TU Kaiserslautern, FB Informatik – Lehrgebiet Informationssysteme
http://dbis.informatik.uni-kl.de
Exercise Sheet 6: Hand out 26.06.2017, Due 11.07.2017 - 16:00

Assignment 1: K-Means Clustering (1 P.)

(a) The k-means algorithm uses the cluster centroid as the mean of the points in the cluster,

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x.$$

Show that the centroid minimizes the sum of squared errors for the cluster in the k-means algorithm.

(b) Consider the following 2-dimensional data points.

	x	y
x_1	1	2
x_2	2	2
x_3	4	2
x_4	1	1
x_5	2	1
x_6	4	1

Apply the k-means algorithm with $k = 2$ where the initial seeds (centroids) are x_2 and x_5. Discuss the problem of k-mean algorithm based on your result.

Assignment 2: Hierachical and Density-Based Clustering (1 P.)

(a) Consider the data in Figure 1. Answer to the following questions assuming that we are using the Euclidean distance and that $\varepsilon = 2$ and **minpts** $= 3$.

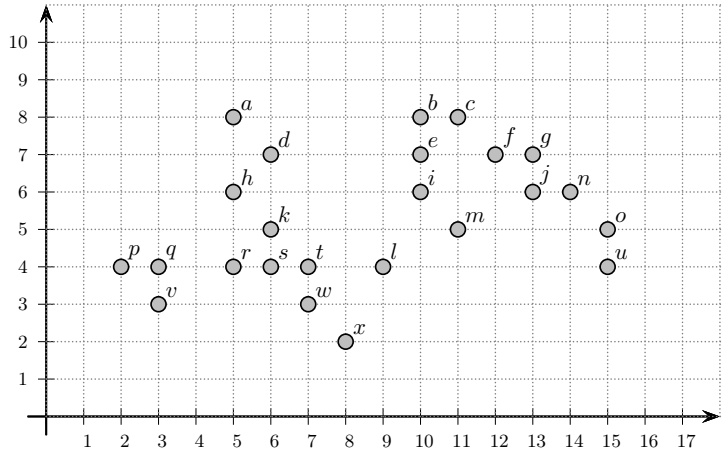

Figure 1: Points in a space

(i) List all the core points.

(ii) Is a directly density-reachable from d?

(iii) Is o density-reachable from i? Show the complete chain or where it breaks.

(iv) Is l density-connected to x? Show the intermediate points that make them density-connected or that break the condition.

(b)

	A	B	C	D	E
A	0	2	4	3	5
B		0	4	3	4
C			0	2	4
D				0	6
E					0

Consider the above distance matrix and compute the hierarchical clustering, breaking ties arbitrarily, using the following inter-cluster similarity:

(i) MIN and (ii) Group Average

Draw the dendrogram diagrams, clearly showing the orders in which the points are merged, and compare the results.

Assignment 3: Classification (1 P.)

Consider the training set of 9 documents, d_1 to d_9 given in following table. Each document is presented as a term vector which presents number of appearances of the terms in the document. Each of these eight terms in term vector are considered here as features $(f_1, f_2, \ldots, f_8)$ for classification task.

Doc.	Term Vector								Category
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	class
d_1	3	2	0	0	0	0	0	1	c_1 (Algebra)
d_2	1	2	3	0	0	0	0	0	c_1 (Algebra)
d_3	0	0	0	3	3	0	0	0	c_2 (Calculus)
d_4	0	0	1	2	2	0	1	0	c_2 (Calculus)
d_5	0	0	0	1	1	2	2	0	c_3 (Stochastics)
d_6	1	0	1	0	0	0	2	2	c_3 (Stochastics)
d_7	0	1	1	0	0	0	0	0	c_1 (Algebra)
d_8	0	0	1	0	0	1	1	0	c_3 (Stochastics)
d_9	0	0	0	1	1	2	2	0	c_3 (Stochastics)

(a) Construct a decision tree for the binary classification of the category c_3 ("*Stochastics*"), i.e., the tree decides whether a new document belongs to this category or not. Use binary splits. Determine the *information gain* for each inner node of this decision tree, and always select the split based on the maximum information gain.

Classify document $d_{10} = (1, 0, 1, 1, 1, 2, 2, 1)$ using your decision tree.

(b) Classify the document $d_{11} = (0, 0, 1, 1, 1, 1, 3, 1)$ on the basis of the training set given above using the Bayesian classifier. Assume that the term frequencies in the documents presented in term vector follow a multinomial distribution and also that terms are used as features in the classification. Hence, for calculating the posterior probabilities, use the first equation presented in lecture 3 on slide 16 and to find the likelihood use the maximum likelihood estimate.

Hint: The equations presented in the Lecture 3 slide 16 and 17 are to be adapted to the classification task. For example, $P(t_i|d)$ in slide 17 is needed to be adapted to find likelihood s.t. the probability that i^{th} term in term vector, i.e., f_i appears in a document where document is classified as c_k, denoted as $P(f_i$ occurs in $d|d \in c_k)$.