Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II

SS 2017

Exercise 8: Recap

Exercise 8-1 General Questions

(a) Name the two phenomena introduced as the high-dimensional challenge in our lecture. (2-Punkte)

(b) Statistical measures are used in many areas, for example, high-dimensional feature selection and stream classification. Name three quality measures introduced in our lecture.

(c) Name the two major methods for trajectory mining introduced in our lecture. Briefly describe them using one sentence each.

(d) Describe two window models in data streams. (2 Punkte)

(e) Hoeffding trees do not forget. What are the implications caused by this? (2 Punkte)

(f) Both feature selection methods and feature reduction methods result in a reduced feature space F' over the original feature space F. How are the features in F' related to the original feature space F? (2 Punkte)

Exercise 8-2 Feature Selection

Determine the most informative subspace using Branch-and-Bound in combination with the inconsistency criterium.

ID	attribute X	attribute Y	attribute Z	class
A	2	red	yes	1
B	3	red	yes	1
C	3	green	yes	1
D	4	green	yes	2
E	1	red	yes	2
F	1	green	yes	2

Exercise 8-3 PCA

(a) Find the PCA decomposition of the matrix

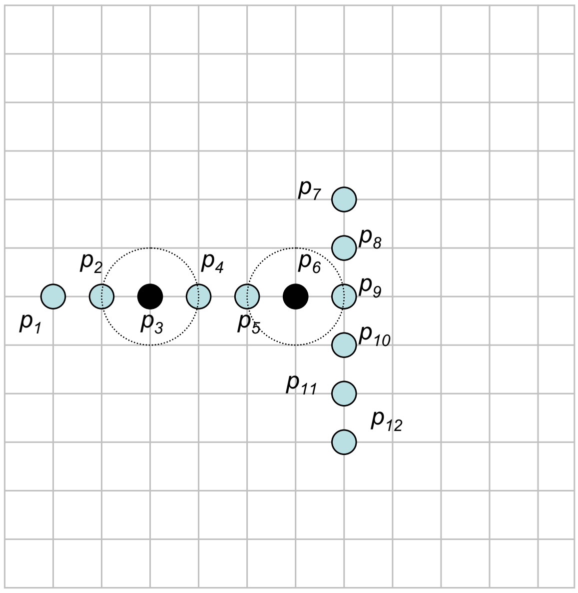
$$A = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

(b) Project the matrix A to 1-D space with the largest eigenvalue.

Exercise 8-4 Density-based Projected-Clustering (PreDeCon)

The algorithm PreDeCon is closely related to 4C. Instead of the expensive PCA, it uses variance analysis and a weighted Euclidean distance function: For the points in a candidate's ϵ-neighborhood, each dimension whose variance is below δ is weighted more heavily (κ).

Consider the 2D data set shown below. Assume the width of the grid to be 1 unit, use the Euclidean distance function to determine a point's ϵ-neighborhood.



Calculate, if p_3 and p_6 are core points. Assume the following parameter values: $minPts = 3, \epsilon = 1, \delta = 0.25, \lambda = 1, \kappa = 100$

Exercise 8-5 Dynamic Time Warping

Given the following two time series: $X = (3, 5, 9, 2, 3, 6, 3)$ and $Y = (3, 4, 6, 10, 1, 3, 2, 7, 4)$, compute the following distances:

(a) Uniform Time Warping Distance D^2_{UTW}

(b) Dynamic Time Warping DTW^2

(c) k-Dynamic Time Warping Distance D^2_{k-DTW} where $k = 3$ (Optional)

Visualize the optimal alignment between the time series.

Sequence after scaling:

index i	1	...	8	9	10	...	15	...	19	...	22	...	28	29	...	36	37	...	43	...	46	...	50	...	55	...	57	...	63
$x_{\lceil i/m \rceil}$																													
$y_{\lceil i/n \rceil}$																													
$x_{\lceil i/m \rceil} - y_{\lceil i/n \rceil}$																													

		3	**5**	**9**	**2**	**3**	**6**	**3**
	0	inf	inf	inf	inf	inf	inf	inf
3	inf	0	4	40	41	41	50	50
4	inf	1	1	26	30	31	35	36
6	inf	10	2	10	26	35	31	40
10	inf	59	27	3	67	75	47	80
1	inf	63	43	67	4	8	33	37
3	inf	63	47	79	5	4	13	13
2	inf	64	56	96	5	5	20	14
7	inf	80	60	60	30	21	6	22
4	inf	81	61	85	34	22	10	7

		3	**5**	**9**	**2**	**3**	**6**	**3**
	0	inf	inf	inf	inf	inf	inf	inf
3	inf	0	4	40	41	inf	inf	inf
4	inf	1	1	26	30	31	inf	inf
6	inf	10	2	10	26	35	31	inf
10	inf	59	27	3	67	75	47	80
1	inf	inf	43	67	4	8	33	37
3	inf	inf	inf	79	5	4	13	13
2	inf	inf	inf	inf	5	5	20	14
7	inf	inf	inf	inf	inf	21	6	22
4	inf	inf	inf	inf	inf	inf	10	7

Exercise 8-6 Discrete Wavelet Transformation

(a) Given the following time series $1, 2, 3, 2, 4, 6, 5, 0$, compute its discrete wavelet transformation.

1	2	3	2	4	6	5	0

(b) Remove the last four digits and compute the inverse discrete wavelet transformation.

Exercise 8-7 Cluster Features

Consider the following data set with features X and Y:

ObjID	Cluster	X	Y
1	A	1	2
2	A	2	2
3	A	2	3
4	A	0	3
5	B	7	-1

ObjID	Cluster	X	Y
6	B	5	0
7	B	6	1
8	C	0	-2
9	C	1	-3
10	C	1	-2

(a) Compute the BIRCH Cluster Features for the following clusters A, B and C.

(b) We now insert a new feature vector $a = (1, -1)$. A cluster absorbs a new instance if the radius of the cluster does not exceed $T = \sqrt{2} \approx 1.41421$ after adding the instance. Compute whether a can be absorbed by any of the existing clusters. If a can be absorbed, add a to the cluster. If a cannot be absorbed, build a new cluster.

Exercise 8-8 DenStream

The fading function used in DenStream algorithm is $f(t) = 2^{-\lambda \cdot t}$, where $\lambda > 0$.

(a) Given the speed of arriving point (the number of point arrived per unit time) is v, compute the overall weight of all points in DenStream algorithm when time $t \to \infty$.

(b) Denstream algorithm stores a set of potential core micro cluster (p-mc). The period to check the weight of those micro-clusters and perform puring is the minimal time span for a p-mc fading into an outlier. Given the threshold of p-mc is μ, determine the micro-cluster maintenance period of DenStream algorithm.

Exercise 8-9 Hoeffding trees

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

- Time since getting the driving license ($1-2$ years, $2-7$ years, >7 years)
- Gender (male, female)
- Residential area (urban, rural)

These are the first 8 examples.

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

- Incrementally construct a Hoeffding tree for this example.
  Use information gain and $\delta = 0.2$ and $N_{\text{min}} = 2$.
- Compute the value of δ at which the tree would still consist of the leaf only.

Exercise 8-10 Cohen's Kappa

Given the following convergence matrix at time points $t = 1, 2, 3$:
Gegeben seien die folgenden Konfusionsmatrizen zu den Zeitpunkten $t = 1, 2, 3$:

$t = 1$

	positiv	negativ
positiv	37	14
negativ	17	32

$t = 2$

	positiv	negativ
positiv	65	8
negativ	7	20

$t = 3$

	positiv	negativ
positiv	90	4
negativ	5	1

Calculate the Accuracy and the Cohen's Kappa value.
Berechnen Sie Accuracy und Cohen's Kappa.