Big Data Management and Analytics
Assignment 10

Assignment 10-1

Finding similar items

Suppose that the universal set is given by {1,…,10}. Construct minhash signatures for the following sets:

(a) $S_1 = \{3,6,9\}$

(b) $S_2 = \{2,4,6,8\}$

(c) $S_3 = \{2,3,4\}$

1. Construct the signatures for the sets using the following list of permutations:
   - (1,2,3,4,5,6,7,8,9,10)
   - (10,8,6,4,2,9,7,5,3,1)
   - (4,7,2,9,1,5,3,10,6,8)

Assignment 10-1

i. Create first a characteristic matrix for each permutation

(a) Set one column with the shingles

(b) Set columns for each document

(c) Fill the document columns by setting a 1 where there is an occurence for each shingle, and 0 else

Element	S_1	S_2	S_3
1	0	0	0
2	0	1	1
3	1	0	1
4	0	1	1
5	0	0	0
6	1	1	0
7	0	0	0
8	0	1	0
9	1	0	0
10	0	0	0

Assignment 10-1

i. Create first a characteristic matrix for each permutation

Element	S_1	S_2	S_3
1	0	0	0
2	0	1	1
3	1	0	1
4	0	1	1
5	0	0	0
6	1	1	0
7	0	0	0
8	0	1	0
9	1	0	0
10	0	0	0

1st permutation

Element	S_1	S_2	S_3
10	0	0	0
8	0	1	0
6	1	1	0
4	0	1	1
2	0	1	1
9	1	0	0
7	0	0	0
5	0	0	0
3	1	0	1
1	0	0	0

2nd permutation

Element	S_1	S_2	S_3
4	0	1	1
7	0	0	0
2	0	1	1
9	1	0	0
1	0	0	0
5	0	1	0
3	1	0	1
10	0	1	0
6	1	1	0
8	0	1	0

3rd permutation

Assignment 10-1

ii. Compute the minhash for each permutation

Element	S_1	S_2	S_3
1	0	0	0
2	0	1	1
3	1	0	1
4	0	1	1
5	0	0	0
6	1	1	0
7	0	0	0
8	0	1	0
9	1	0	0
10	0	0	0

1st permutation

Select the first occurence of 1 per set and get the elements at which they can be found

Which leads to the following minhash:
$h(S_1) = 3$
$h(S_2) = 2$
$h(S_3) = 2$

Assignment 10-1

ii. Compute the minhash for each permutation

Element	S_1	S_2	S_3
10	0	0	0
8	0	1	0
6	1	1	0
4	0	1	1
2	0	1	1
9	1	0	0
7	0	0	0
5	0	0	0
3	1	0	1
1	0	0	0

2nd permutation

The same procedure for the 2nd permutation…

$h(S_1) = 6$
$h(S_2) = 8$
$h(S_3) = 4$

Assignment 10-1

ii. Compute the minhash for each permutation

Element	S_1	S_2	S_3
4	0	1	1
7	0	0	0
2	0	1	1
9	1	0	0
1	0	0	0
5	0	1	0
3	1	0	1
10	0	1	0
6	1	1	0
8	0	1	0

3rd permutation

…and for the 3rd permutation

Which leads to the following minhash:
$h(S_1) = 9$
$h(S_2) = 4$
$h(S_3) = 4$

This yields the following signatures:

$SIG(S_1) = \{3,6,9\}$
$SIG(S_2) = \{2,8,4\}$
$SIG(S_3) = \{2,4,4\}$

Assignment 10-1

2. Instead of using the previously given permutations use hash functions:

$$h_1(x) = x \; mod \; 10$$
$$h_2(x) = (2x + 1) \; mod \; 10$$
$$h_3(x) = (3x + 2) \; mod \; 10$$

Assignment 10-1

2. Instead of using the previously given permutations use hash functions:

i. Set up a table:

Element	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0			
2	0	1	1			
3	1	0	1			
4	0	1	1			
5	0	0	0			
6	1	1	0			
7	0	0	0			
8	0	1	0			
9	1	0	0			
10	0	0	0			

Assignment 10-1

ii. Compute the hash values (except for zero-rows):

Element	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

	S_1	S_2	S_3
h_1	∞	∞	∞
h_2	∞	∞	∞
h_3	∞	∞	∞

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Update only S_2, S_3!

S_2:
$\min(\infty, 2) = 2$
$\min(\infty, 5) = 5$
$\min(\infty, 8) = 8$

S_3:
$\min(\infty, 2) = 2$
$\min(\infty, 5) = 5$
$\min(\infty, 8) = 8$

Update of 2nd row:

	S_1	S_2	S_3
h_1	∞	2	2
h_2	∞	5	5
h_3	∞	8	8

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9

Update only S_1, S_3!

S_1:
$\min(\infty, 3) = 3$
$\min(\infty, 7) = 7$
$\min(\infty, 1) = 1$

S_3:
$\min(2,3) = 2$
$\min(5,7) = 5$
$\min(8,1) = 1$

Update of 3rd row:

	S_1	S_2	S_3
h_1	3	2	2
h_2	7	5	5
h_3	1	8	1

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Update only S_2, S_3!

S_2:
$\min(2,4) = 2$
$\min(5,9) = 5$
$\min(8,4) = 4$

S_3:
$\min(2,4) = 2$
$\min(5,9) = 5$
$\min(1,4) = 1$

Update of 4th row:

	S_1	S_2	S_3
h_1	3	2	2
h_2	7	5	5
h_3	1	4	1

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Update only S_1, S_2!

S_1:
$\min(3,6) = 3$
$\min(7,3) = 3$
$\min(1,0) = 0$

S_2:
$\min(2,6) = 2$
$\min(5,3) = 3$
$\min(4,0) = 0$

Update of 6th row:

	S_1	S_2	S_3
h_1	3	2	2
h_2	3	3	5
h_3	0	0	1

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Update only S_2!

S_2:
$\min(2,8) = 2$
$\min(3,7) = 3$
$\min(0,6) = 0$

Update of 8th row:

	S_1	S_2	S_3
h_1	3	2	2
h_2	3	3	5
h_3	0	0	1

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

e	S_1	S_2	S_3	$h_1(x)$	$h_2(x)$	$h_3(x)$
1	0	0	0	-	-	-
2	0	1	1	2	5	8
3	1	0	1	3	7	1
4	0	1	1	4	9	4
5	0	0	0	-	-	-
6	1	1	0	6	3	0
7	0	0	0	-	-	-
8	0	1	0	8	7	6
9	1	0	0	9	9	9
10	0	0	0	-	-	-

Update only S_1!

S_1:
$\min(3,9) = 3$
$\min(3,9) = 3$
$\min(0,9) = 0$

Update of 8th row:

	S_1	S_2	S_3
h_1	3	2	2
h_2	3	3	5
h_3	0	0	1

Assignment 10-1

iii. Create a table for all hash functions and sets and initialize them with infinite distance:

	S_1	S_2	S_3
h_1	3	2	2
h_2	3	3	5
h_3	0	0	1

This yields the following signatures:

$$SIG(S_1) = (3,3,0)$$
$$SIG(S_2) = (2,3,0)$$
$$SIG(S_3) = (2,5,1)$$

Assignment 10-1

3. How does the estimated Jaccard similarity from (1.) and (2.) compare with the true Jaccard similarity of the original data? How to reduce deviations in the approximated Jaccard similarities?

RECAP: Jaccard similarity:

$$d_{Jaccard}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

Assignment 10-1

RECAP: Jaccard similarity:

$$d_{Jaccard}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

(a) $S_1 = \{3,6,9\}$
(b) $S_2 = \{2,4,6,8\}$
(c) $S_3 = \{2,3,4\}$

i. Actual Jaccard similarity:

	S_1	S_2	S_3
S_1	-	$^1/_6$	$^1/_5$
S_2	-	-	$^2/_5$
S_3	-	-	-

Assignment 10-1

RECAP: Jaccard similarity:

$$d_{Jaccard}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

(a) $S_1 = \{3,6,9\}$
(b) $S_2 = \{2,8,4\}$
(c) $S_3 = \{2,4,4\}$

ii. Permutation estimated Jaccard similarity:

	S_1	S_2	S_3
S_1	-	$^0/_6$	$^0/_5$
S_2	-	-	$^2/_3$
S_3	-	-	-

Assignment 10-1

RECAP: Jaccard similarity:

$$d_{Jaccard}(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

(a) $S_1 = \{3,3,0\}$
(b) $S_2 = \{2,3,0\}$
(c) $S_3 = \{2,5,1\}$

iii. Hash estimated Jaccard similarity:

	S_1	S_2	S_3
S_1	-	$^2/_3$	$^0/_5$
S_2	-	-	$^1/_5$
S_3	-	-	-

Assignment 10-1

iv. Comparison:

Actual J. similarity

	S_1	S_2	S_3
S_1	-	$1/6$	$1/5$
S_2	-	-	$2/5$
S_3	-	-	-

Perm. est. J similarity

	S_1	S_2	S_3
S_1	-	$0/6$	$0/5$
S_2	-	-	$2/3$
S_3	-	-	-

Hash. est. J similarity

	S_1	S_2	S_3
S_1	-	$2/3$	$0/5$
S_2	-	-	$1/5$
S_3	-	-	-

Estimation of the actual J. similarity is rather poor, why?

→ Too small minhash vectors. Get more permutations of the universal set or more hash functions to extend the minhash vectors!

Assignment 10-2

(a) Describe what a PCA aims for and under what circumstances it is most helpful

From the lecture slides:

- Detect hidden linear correlations
- Remove redundant and noisy features
- Interpretation and visualization
- Easier storage and processing of the data

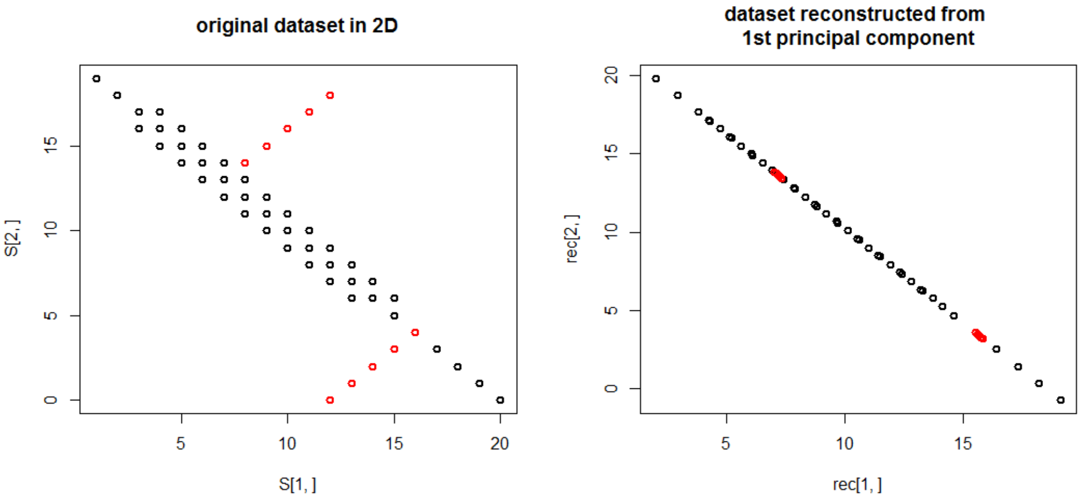
When is PCA most helpful:

- The assumption is that the observed variable can be expressed as a linear combination of the hidden variables $x = \mu + Uw + \epsilon$. If that is not the case, another heuristics should be used (e.g. LDA,RCA etc.)

Assignment 10-2

(b) Which possibly netgative consequences might arise when applying PCA to a dataset of unknown structure?

- Data which is not normed can skew the result. Therefore first norm the data!
- Loss of possibly relevant structures (see red lines within the figures)

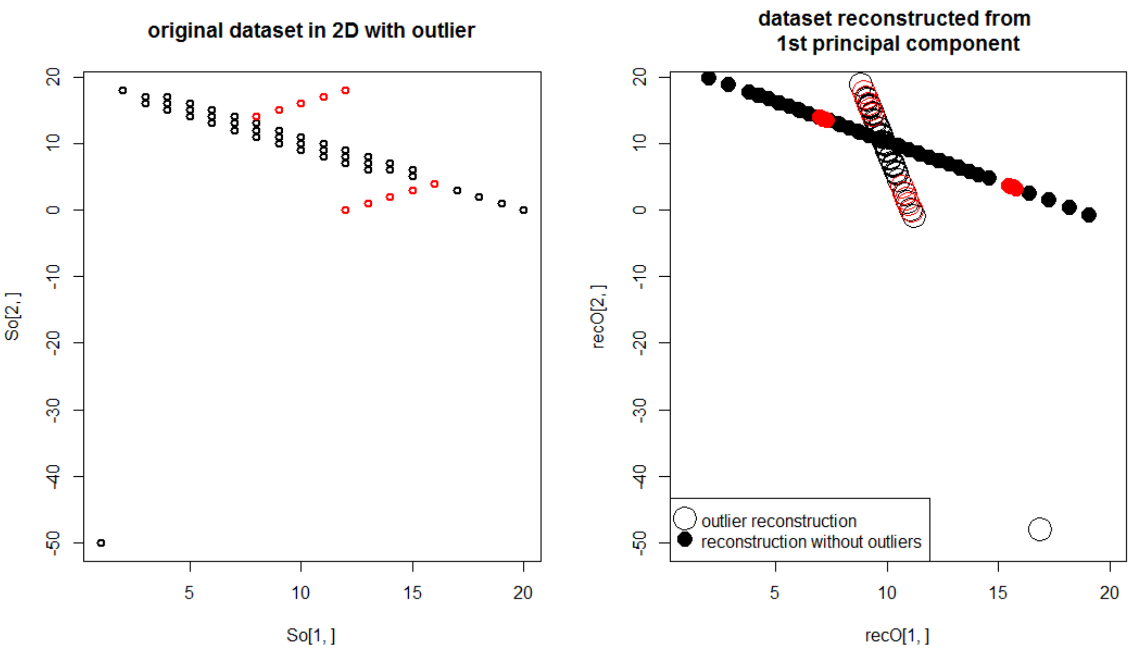


- Solution: subspace clustering / correlation clustering

Assignment 10-2

(b) Which possibly netgative consequences might arise when applying PCA to a dataset of unknown structure?

- Further, problems with outliers may arise, as they may massively skew the PCA transformation:



Assignment 10-3

Consider the $X \in \mathbb{R}^{M\,x\,N}$ matrix containing six data points $X_i \in \mathbb{R}^2$.

$$X = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 5 & 6 \\ 6 & 6 \\ 7 & 6 \end{pmatrix}$$

dim 1   dim 2

Conduct a PCA on the given data, i.e. project the data onto a one-dimensional space. Please state the eigenvectors, eigenvalues, covariance matrix and visualize the data before and after PCA.

Assignment 10-3

i. Center the data by substracting the mean value for each dimension:

$$\hat{\mu} = \frac{1}{N}\sum_i X_i = \binom{4}{3}$$

$$\tilde{X} = \begin{pmatrix} 1-4 & 0-3 \\ 2-4 & 0-3 \\ 3-4 & 0-3 \\ 5-4 & 6-3 \\ 6-4 & 6-3 \\ 7-4 & 6-3 \end{pmatrix} = \begin{pmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{pmatrix}$$

Assignment 10-3

ii. Calculate the covariance matrix $E\left[\left(x - E(X)\right) \cdot \left(X - E(X)\right)^T\right]$:

$$cov(X) \approx \hat{\Sigma} = \frac{1}{N} \tilde{X}^T \tilde{X} = \begin{pmatrix} 4,\overline{7} & 6 \\ 6 & 9 \end{pmatrix} = \begin{pmatrix} {}^{14}/_{3} & 6 \\ 6 & 9 \end{pmatrix}$$

iii. Now compute the eigenpairs (eigenvalues, eigenvectors). Construct the eigendecomposition $\hat{\Sigma} = \hat{\hat{U}}\hat{\hat{S}}\hat{\hat{U}}^T$ with sorted eigenvalues $\hat{\lambda}_j$ in $\hat{\hat{S}}$

Compute the eigenvalues:

$$\det(\hat{\Sigma} - \lambda I) = det\begin{pmatrix} {}^{14}/_{3} - \lambda & 6 \\ 6 & 9 - \lambda \end{pmatrix} = \left({}^{14}/_{3} - \lambda\right) \cdot (9 - \lambda) - 36$$

$$= 14 \cdot 3 - 36 - \tfrac{14+27}{3}\lambda + \lambda^2 =$$

$$\lambda^2 - \frac{41}{3}\lambda + 6 = 0$$

$$\lambda_{1,2} = \frac{{}^{41}/_{3} \mp \sqrt{\left({}^{41}/_{3}\right)^2 - 4 \cdot 6}}{2} = 13.21\ and\ 0.45$$

Assignment 10-3

iii. Now compute the eigenpairs (eigenvalues, eigenvectors). Construct the eigendecomposition $\hat{\Sigma} = \hat{\hat{U}}\hat{\hat{S}}\hat{\hat{U}}^T$ with sorted eigenvalues $\hat{\lambda_j}$ in $\hat{\hat{S}}$

Compute the eigenvectors:

$$\hat{\Sigma}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda_{1,2}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} 14/3 & 6 \\ 6 & 9 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda_{1,2}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$\overset{\lambda_1}{\Rightarrow} \quad \begin{matrix} 14/3\, x + 6y = \lambda_1 x \\ 6x + 9y = \lambda_1 y \end{matrix} \quad \Rightarrow \text{1}^{\text{st}} \text{ (normed) eigenvector:} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix}$$

$$\text{Eigenvalues: } diag\left(\hat{\hat{S}}\right) = \begin{pmatrix} 13{,}21 & 0 \\ 0 & 0{,}45 \end{pmatrix}$$

$$\text{Eigenvectors: } \hat{\hat{U}} = \begin{pmatrix} 0{,}57 & 0{,}82 \\ 0{,}82 & -0{,}57 \end{pmatrix}$$

Assignment 10-3

iv. Reduce to one-dimensional space. For this purpose remove the second eigenvector and form the transformation matrix U:

$$U = \begin{pmatrix} 0{,}57 & 0 \\ 0{,}82 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix}$$

Now transform the data with

$$Y = \tilde{X} \cdot U = \begin{pmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix} = (-4{,}18 \;\; -3{,}6 \;\; -3{,}03 \;\; 3{,}03 \;\; 3{,}6 \;\; 4{,}18)^T$$

Assignment 10-3

iv. Reduce to one-dimensional space. For this purpose remove the second eigenvector and form the transformation matrix U:

We can now try to reconstruct the original data matrix with

$$\hat{Z} = \mu + Y \cdot U^T = \mu + \tilde{X} \cdot U \cdot U^T$$

$$\hat{Z} = \begin{pmatrix} 1{,}6 & -0{,}42 \\ 1{,}93 & 0{,}05 \\ 2{,}26 & 0{,}52 \\ 5{,}74 & 5{,}48 \\ 6{,}07 & 5{,}95 \\ 6{,}40 & 6{,}42 \end{pmatrix}$$

Assignment 10-3

v. As we have already reduced to the one-dimensional space (here we did that by eliminiating the second principal component), the reconstruction does not imply the information of the second pc: