# Data Mining Tutorial

## Evaluation von Algorithmen

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2014-06-27 — KDD Übung

# Evaluation von Clusterings

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 |

# Evaluation von Clusterings

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

# Evaluation von Clusterings

## Set Matching

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

"Precision": Summe Maxima in Zeilen:

# Evaluation von Clusterings

## Set Matching

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

"Precision": Summe Maxima in Zeilen:

$(13 + 10 + 15 + 8)/71 \approx 0.6479$

# Evaluation von Clusterings

## Set Matching

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

"Precision": Summe Maxima in Zeilen:

$(13 + 10 + 15 + 8)/71 \approx 0.6479$

"Recall": Summe Maxima in Spalten:

# Evaluation von Clusterings

## Set Matching

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

"Precision": Summe Maxima in Zeilen:

$(13 + 10 + 15 + 8)/71 \approx 0.6479$

"Recall": Summe Maxima in Spalten:

$(15 + 10 + 11 + 8 + 13)/71 \approx 0.8028$

# Evaluation von Clusterings

## Set Matching

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

"Precision": Summe Maxima in Zeilen:

$(13 + 10 + 15 + 8)/71 \approx 0.6479$

"Recall": Summe Maxima in Spalten:

$(15 + 10 + 11 + 8 + 13)/71 \approx 0.8028$

F-Measure: $\approx 0.7171$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$13/27$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$13/27 + 10/14$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$13/27 + 10/14 + 15/19$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$13/27 + 10/14 + 15/19 + 8/11$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$

Mittlere Precision Spalten:

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$

Mittlere Precision Spalten:

$15/17$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$

Mittlere Precision Spalten:

$15/17 + 10/11$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$$

Mittlere Precision Spalten:

$$15/17 + 10/11 + 11/17$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$$

Mittlere Precision Spalten:

$$15/17 + 10/11 + 11/17 + 8/11$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$$

Mittlere Precision Spalten:

$$15/17 + 10/11 + 11/17 + 8/11 + 13/15$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$$

Mittlere Precision Spalten:

$$(15/17 + 10/11 + 11/17 + 8/11 + 13/15)/5 \approx 0.8065$$

# Evaluation von Clusterings

## Average Precision

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

Mittlere Precision Zeilen:

$(13/27 + 10/14 + 15/19 + 8/11)/4 \approx 0.6781$

Mittlere Precision Spalten:

$(15/17 + 10/11 + 11/17 + 8/11 + 13/15)/5 \approx 0.8065$

F-Measure: $\approx 0.7368$

# Evaluation von Clusterings

## Average F-Measure

Was ist das F-Measure einer Zelle?

# Evaluation von Clusterings

## Average F-Measure

Was ist das F-Measure einer Zelle?

Precision: $c_{ij}/|A_i|$
Recall: $c_{ij}/|B_j|$

# Evaluation von Clusterings

## Average F-Measure

Was ist das F-Measure einer Zelle?

Precision: $c_{ij}/|A_i|$
Recall: $c_{ij}/|B_j|$

F-Measure daraus:

$$F_1(c_{ij}, A_i, B_j) = \frac{2 \cdot \frac{c_{ij}}{|A_i|} \cdot \frac{c_{ij}}{|B_j|}}{\frac{c_{ij}}{|A_i|} + \frac{c_{ij}}{|B_j|}}$$

# Evaluation von Clusterings

## Average F-Measure

Was ist das F-Measure einer Zelle?

Precision: $c_{ij}/|A_i|$
Recall: $c_{ij}/|B_j|$

F-Measure daraus:

$$F_1(c_{ij}, A_i, B_j) = \frac{2 \cdot \frac{c_{ij}}{|A_i| \cdot |B_j|}}{\frac{|B_j| + |A_i|}{|A_i| \cdot |B_j|}}$$

# Evaluation von Clusterings

## Average F-Measure

Was ist das F-Measure einer Zelle?

Precision: $c_{ij}/|A_i|$
Recall: $c_{ij}/|B_j|$

F-Measure daraus:

$$F_1(c_{ij}, A_i, B_j) = \frac{2 \cdot c_{ij}}{|B_j| + |A_i|}$$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | $0$ | $\frac{26}{42}$ |
| $B_2$ | $0$ | $\frac{20}{25}$ | $0$ | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | $0$ | $\frac{8}{36}$ | $0$ | $0$ |
| $B_4$ | $0$ | $0$ | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4$$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

Spaltenmaxima:

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | $0$ | $\frac{26}{42}$ |
| $B_2$ | $0$ | $\frac{20}{25}$ | $0$ | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | $0$ | $\frac{8}{36}$ | $0$ | $0$ |
| $B_4$ | $0$ | $0$ | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

Spaltenmaxima:

$$(5/6 + 4/5 + 1/2 + 8/11 + 13/21)/5$$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

Spaltenmaxima:

$$(5/6 + 4/5 + 1/2 + 8/11 + 13/21)/5 \approx 0.6959$$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | 0 | $\frac{26}{42}$ |
| $B_2$ | 0 | $\frac{20}{25}$ | 0 | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | 0 | $\frac{8}{36}$ | 0 | 0 |
| $B_4$ | 0 | 0 | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

Spaltenmaxima:

$$(5/6 + 4/5 + 1/2 + 8/11 + 13/21)/5 \approx 0.6959$$

Mittelwert daraus: $\approx 0.7091$

# Evaluation von Clusterings

## Average F-Measure

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| $B_1$ | $\frac{4}{44}$ | $\frac{2}{38}$ | $\frac{22}{44}$ | $0$ | $\frac{26}{42}$ |
| $B_2$ | $0$ | $\frac{20}{25}$ | $0$ | $\frac{6}{25}$ | $\frac{2}{29}$ |
| $B_3$ | $\frac{30}{36}$ | $0$ | $\frac{8}{36}$ | $0$ | $0$ |
| $B_4$ | $0$ | $0$ | $\frac{4}{28}$ | $\frac{16}{22}$ | $\frac{2}{26}$ |

Zeilenmaxima:

$$(13/21 + 4/5 + 5/6 + 8/11)/4 \approx 0.7222$$

Spaltenmaxima:

$$(5/6 + 4/5 + 1/2 + 8/11 + 13/21)/5 \approx 0.6959$$

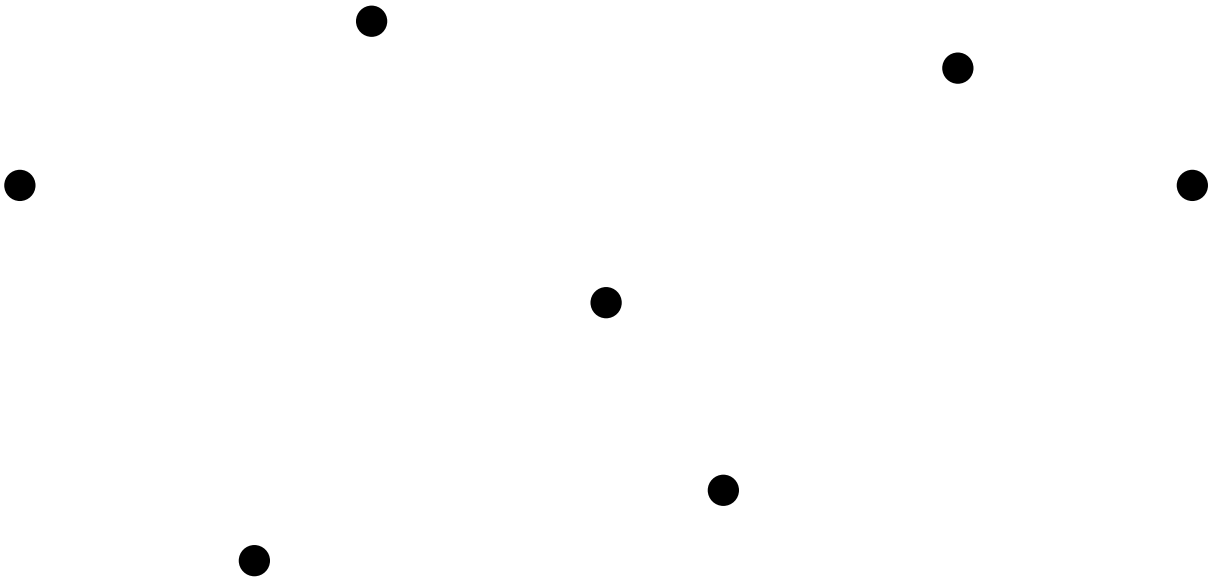
Mittelwert daraus: $\approx 0.7091$

F-Measure daraus: $\approx 0.7088$

# Evaluation von Clusterings

## Pair Counting

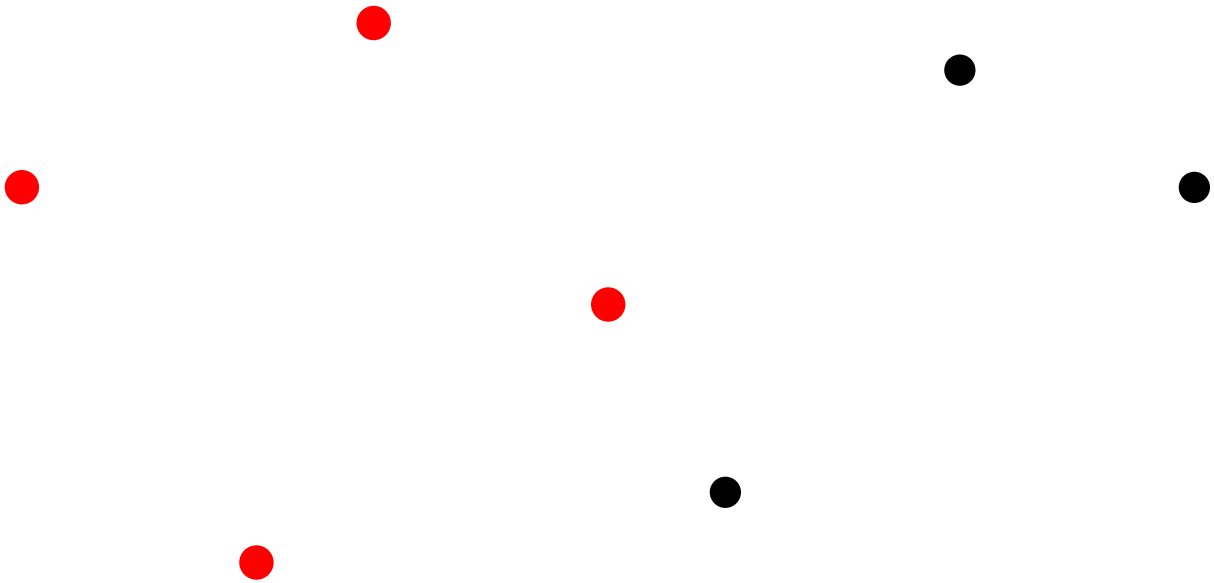
Intuition:



# Evaluation von Clusterings

## Pair Counting

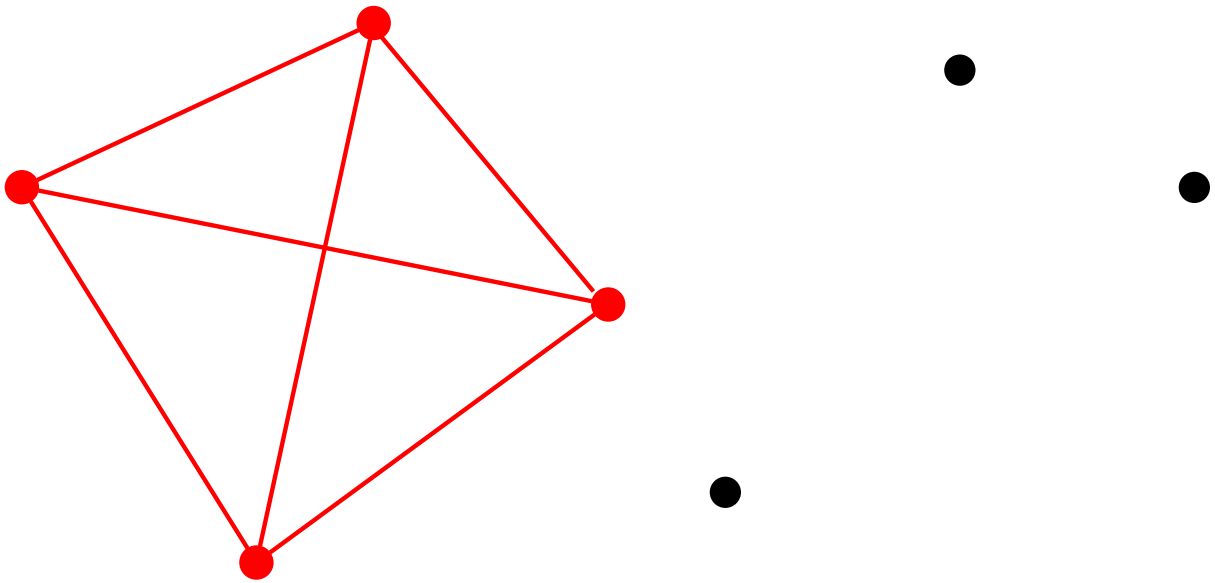
Intuition:



# Evaluation von Clusterings

## Pair Counting

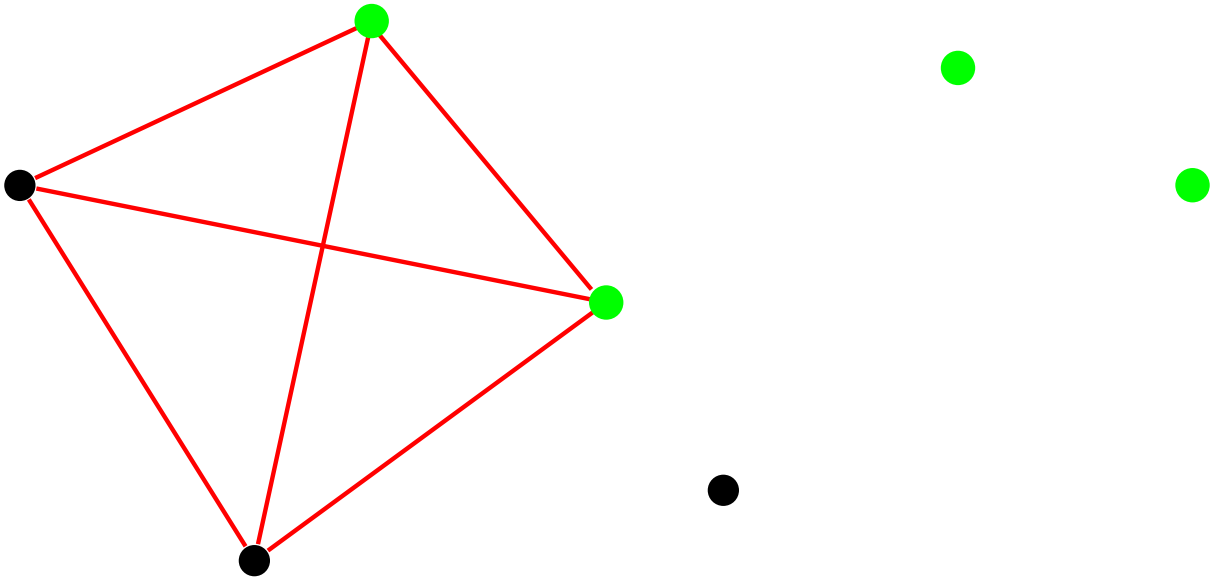
Intuition:



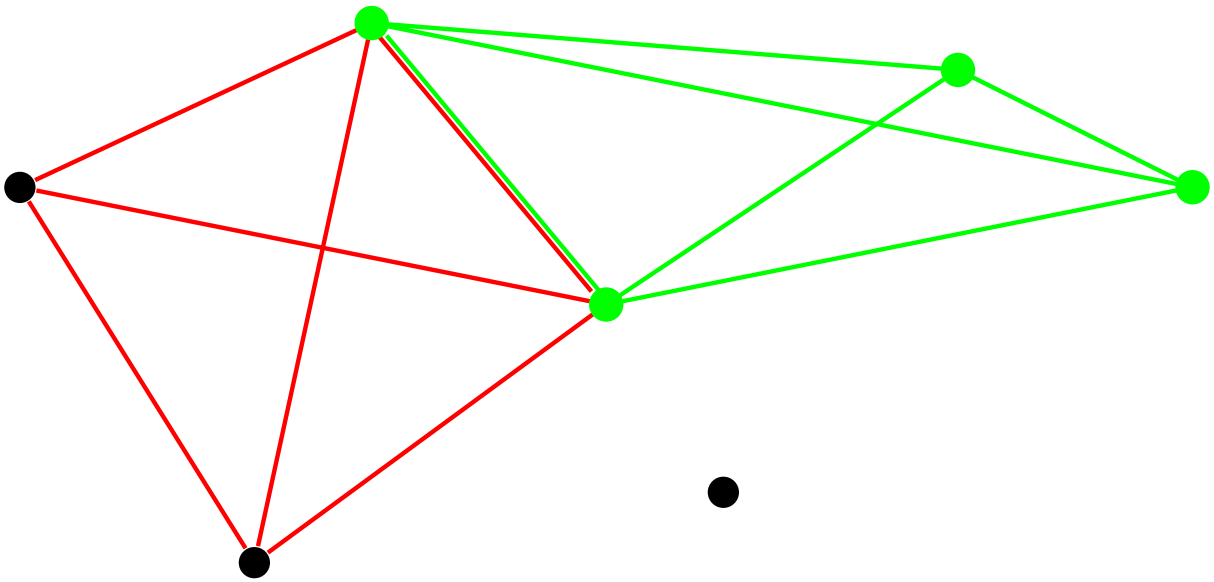
# Evaluation von Clusterings

## Pair Counting

Intuition:



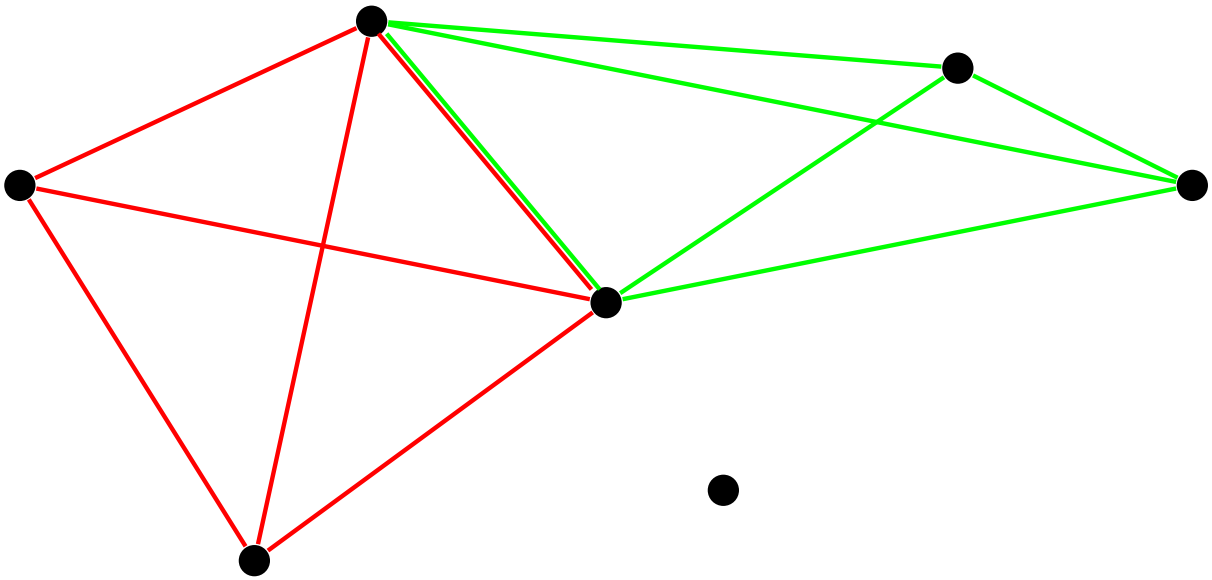
Intuition:



# Evaluation von Clusterings

## Pair Counting

Intuition:



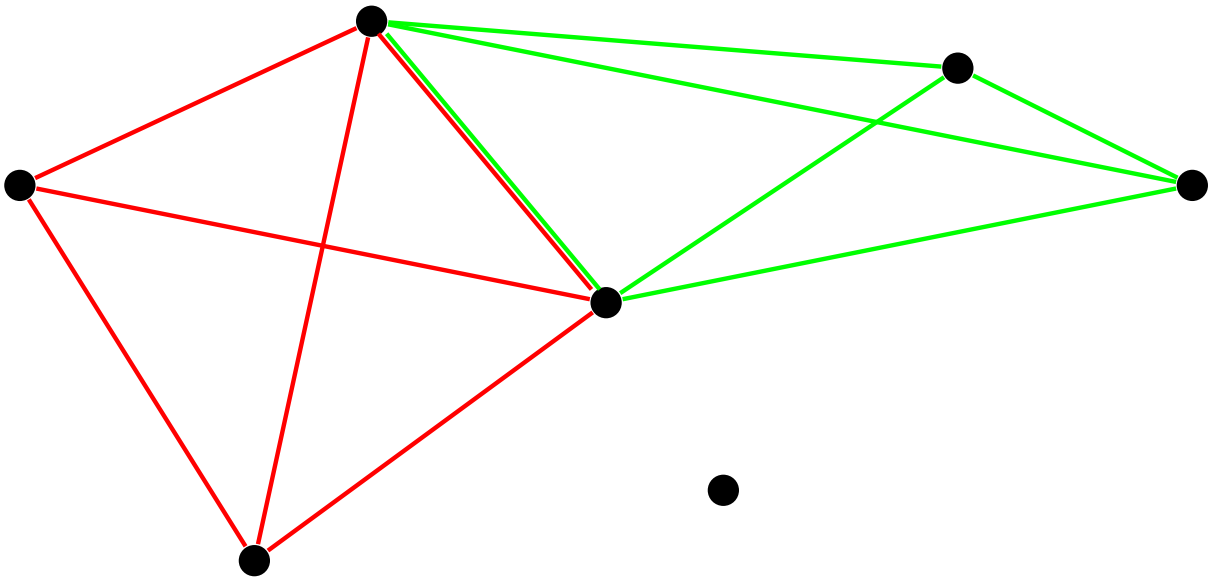
Cluster haben 1 gemeinsames Paar

Geringe Überlappung $\approx$ wenig Paare gemeinsam

# Evaluation von Clusterings

## Pair Counting

Intuition:



Cluster haben 1 gemeinsames Paar

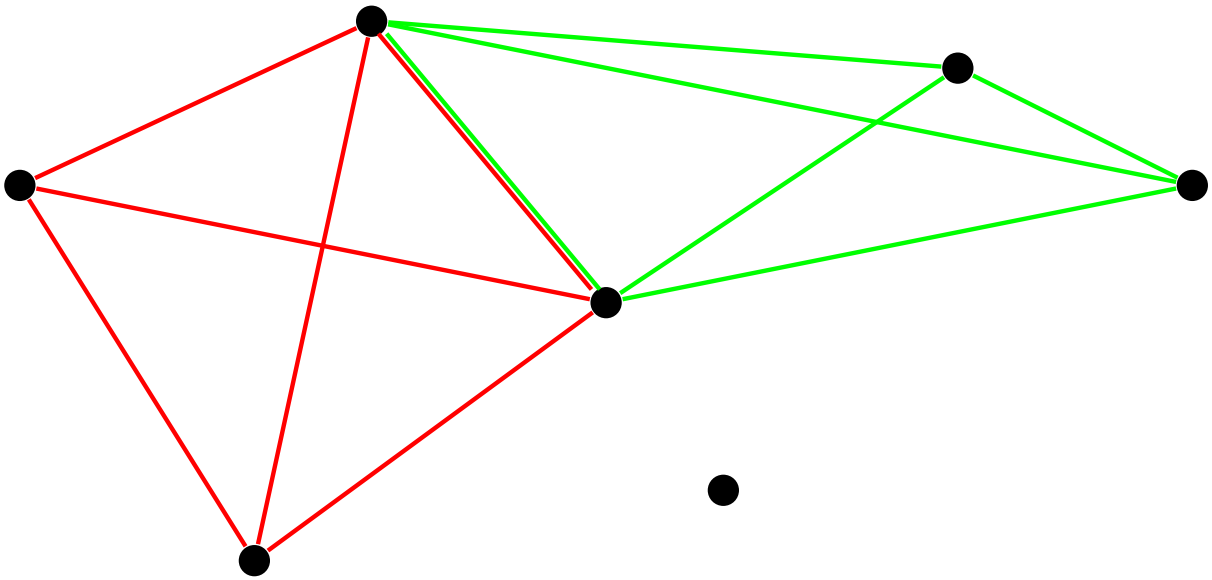
Geringe Überlappung $\approx$ wenig Paare gemeinsam

Anzahl Paare in einer Menge von $n$ Elementen:

# Evaluation von Clusterings

## Pair Counting

Intuition:



Cluster haben 1 gemeinsames Paar

Geringe Überlappung $\approx$ wenig Paare gemeinsam

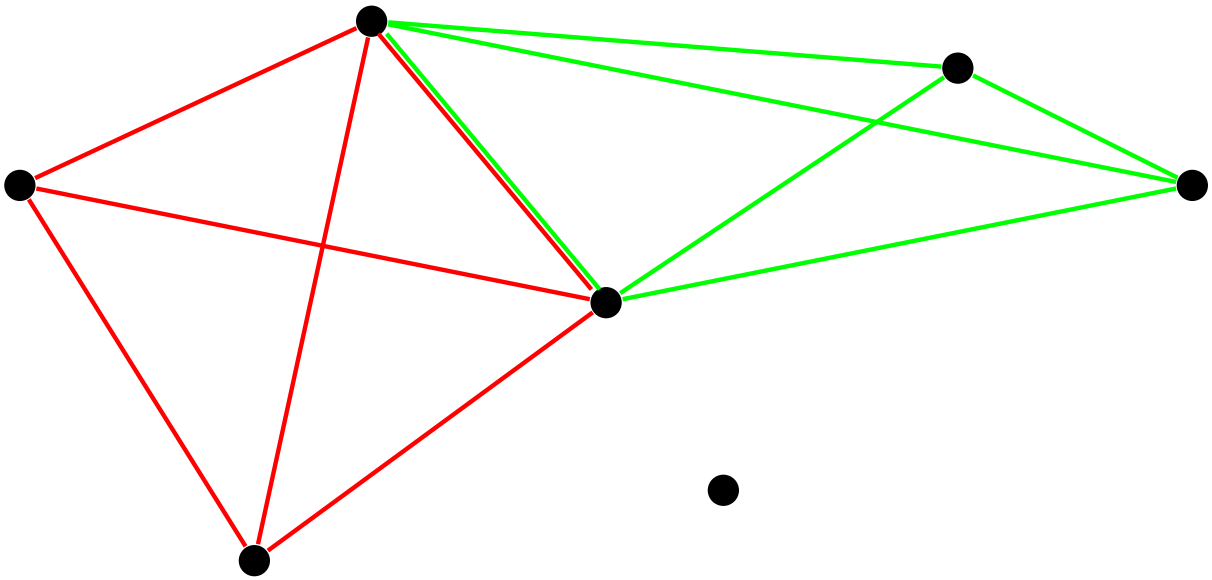
Anzahl Paare in einer Menge von $n$ Elementen:

$\binom{n}{2}$

# Evaluation von Clusterings

## Pair Counting

Intuition:



Cluster haben 1 gemeinsames Paar

Geringe Überlappung $\approx$ wenig Paare gemeinsam

Anzahl Paare in einer Menge von $n$ Elementen:

$\binom{n}{2} = \frac{n(n-1)}{2}$ ("jedes mit jedem anderen, aber nicht doppelt")

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Achtung: Nicht mehr die Summen in der letzten Spalte/Zeile!

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

Summe Zeilen $\Sigma_j \binom{|B_j|}{2} =$

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

Summe Zeilen $\sum_j \binom{|B_j|}{2} = 668$ (“Paare in $B$”, $a + c$)

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

Summe Zeilen $\Sigma_j \binom{|B_j|}{2} = 668$ (“Paare in $B$”, $a + c$)

Summe Spalten $\Sigma_i \binom{|A_i|}{2} =$

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

Summe Zeilen $\Sigma_j \binom{|B_j|}{2} = 668$ (“Paare in $B$”, $a + c$)

Summe Spalten $\Sigma_i \binom{|A_i|}{2} = 487$ (“Paare in $A$”, $a + b$)

# Evaluation von Clusterings

## Pair Counting

| | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\|B_j\|$ |
|---|---|---|---|---|---|---|
| $B_1$ | 2 | 1 | 11 | 0 | 13 | 27 |
| $B_2$ | 0 | 10 | 0 | 3 | 1 | 14 |
| $B_3$ | 15 | 0 | 4 | 0 | 0 | 19 |
| $B_4$ | 0 | 0 | 2 | 8 | 1 | 11 |
| $\|A_i\|$ | 17 | 11 | 17 | 11 | 15 | 71 |

| $\binom{c_{ij}}{2}$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $\binom{\|B_j\|}{2}$ |
|---|---|---|---|---|---|---|
| $B_1$ | 1 | 0 | 55 | 0 | 78 | 351 |
| $B_2$ | 0 | 45 | 0 | 3 | 0 | 91 |
| $B_3$ | 105 | 0 | 6 | 0 | 0 | 171 |
| $B_4$ | 0 | 0 | 1 | 28 | 0 | 55 |
| $\binom{\|A_j\|}{2}$ | 136 | 55 | 136 | 55 | 105 | |

Summe Konfusionsmatrix:

$\sum_{i,j} \binom{c_{ij}}{2} = 322$ (“Übereinstimmungen”, $a$)

Summe Zeilen $\sum_j \binom{|B_j|}{2} = 668$ (“Paare in $B$”, $a + c$)

Summe Spalten $\sum_i \binom{|A_i|}{2} = 487$ (“Paare in $A$”, $a + b$)

Gesamtzahl (möglicher) Paare: $\binom{n}{2} = 2485$ ($M$)

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$

$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision:

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$

$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c)$

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall:

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b)$

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus:

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index:

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d}$

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$

Adjusted Rand:

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M}$

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M} \approx 130.91$

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M} \approx 130.91$
$ARI = \frac{a-E}{\frac{(a+c)+(a+b)}{2}-E}$

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M} \approx 130.91$
$ARI = \frac{a-E}{\frac{(a+c)+(a+b)}{2}-E} \approx 0.4279$

Jaccard Index:

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M} \approx 130.91$
$ARI = \frac{a-E}{\frac{(a+c)+(a+b)}{2}-E} \approx 0.4279$

Jaccard Index: $J = \frac{a}{a+b+c}$

# Evaluation von Clusterings

## Pair Counting

D.h. $a = 322$, $a + c = 668$, $a + b = 487$, $M = 2485$
$d = M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652$

Precision: $a/(a + c) \approx 0.66119$
Recall: $a/(a + b) \approx 0.48204$
F-Measure daraus: $\approx 0.55758$

Rand Index: $\frac{a+d}{a+b+c+d} = \frac{322+1652}{2485} \approx 0.7944$
Adjusted Rand: Erwartungswert: $E = \frac{(a+c)\cdot(a+b)}{M} \approx 130.91$
$ARI = \frac{a-E}{\frac{(a+c)+(a+b)}{2}-E} \approx 0.4279$

Jaccard Index: $J = \frac{a}{a+b+c} = \frac{322}{487+668-322} \approx 0.3866$

# Evaluation von Clusterings

## Zusammenfassung

| Name | Preciosion and Recall | | F / Wert |
|---|---|---|---|
| Set Matching | 0.8028 | 0.6479 | 0.7171 |
| Average Prec. | 0.8065 | 0.6781 | 0.7368 |
| Cell-F-Measure | 0.7222 | 0.6959 | 0.7088 |
| Pair-Counting | 0.6612 | 0.4820 | 0.5576 |
| Rand | | | 0.7944 |
| ARI | | | 0.4279 |
| Jaccard | | | 0.3866 |

# Evaluation von Clusterings

## Zusammenfassung

| Name | Preciosion and Recall | | F / Wert |
|---|---|---|---|
| Set Matching | 0.8028 | 0.6479 | 0.7171 |
| Average Prec. | 0.8065 | 0.6781 | 0.7368 |
| Cell-F-Measure | 0.7222 | 0.6959 | 0.7088 |
| Pair-Counting | 0.6612 | 0.4820 | 0.5576 |
| Rand | | | 0.7944 |
| ARI | | | 0.4279 |
| Jaccard | | | 0.3866 |

Es gibt nicht “das” Maß aller Dinge. Populär sind vor allem: ARI und Pair-Counting-F-Measure.

# Evaluation von Clusterings

## Zusammenfassung

| Name | Preciosion and Recall | | F / Wert |
|---|---|---|---|
| Set Matching | 0.8028 | 0.6479 | 0.7171 |
| Average Prec. | 0.8065 | 0.6781 | 0.7368 |
| Cell-F-Measure | 0.7222 | 0.6959 | 0.7088 |
| Pair-Counting | 0.6612 | 0.4820 | 0.5576 |
| Rand | | | 0.7944 |
| ARI | | | 0.4279 |
| Jaccard | | | 0.3866 |

Es gibt nicht "das" Maß aller Dinge. Populär sind vor allem: ARI und Pair-Counting-F-Measure.
Absolute Werte sind nicht besonders aussagekräftig, insbesondere wenn sie nicht "corrected for chance" sind.

# Evaluation von Clusterings

## Zusammenfassung

| Name | Preciosion and Recall | | F / Wert |
|---|---|---|---|
| Set Matching | 0.8028 | 0.6479 | 0.7171 |
| Average Prec. | 0.8065 | 0.6781 | 0.7368 |
| Cell-F-Measure | 0.7222 | 0.6959 | 0.7088 |
| Pair-Counting | 0.6612 | 0.4820 | 0.5576 |
| Rand | | | 0.7944 |
| ARI | | | 0.4279 |
| Jaccard | | | 0.3866 |

Es gibt nicht "das" Maß aller Dinge. Populär sind vor allem: ARI und Pair-Counting-F-Measure.

Absolute Werte sind nicht besonders aussagekräftig, insbesondere wenn sie nicht "corrected for chance" sind.

Aber: Aussagen wie $B_1$ ist ähnlicher zu $A$ als $B_2$ zu $A$ sind meist dennoch möglich (und i.d.R. bei allen Indizes gleich).

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile

# Evaluation

## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung
  - Misst die Ähnlichkeit Methode $\leftrightarrow$ Maß?

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung
  - Misst die Ähnlichkeit Methode $\leftrightarrow$ Maß?
- Manuelle Evaluierung durch Experten

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung
  - Misst die Ähnlichkeit Methode $\leftrightarrow$ Maß?
- Manuelle Evaluierung durch Experten
  - Aufwendig und subjektiv

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung
  - Misst die Ähnlichkeit Methode $\leftrightarrow$ Maß?
- Manuelle Evaluierung durch Experten
  - Aufwendig und subjektiv
- Indirekte Evaluierung?

# Evaluation
## Fazit

Eine gute Evaluierung von “unsupervised” Methoden ist *überraschend* schwierig!

- Es gibt zahlreiche Maße, ohne klare Vorteile
- Vergleich mit bekannten Labels (“supervised”)
  - Bestraft echt neues Wissen!
- Interne Evaluierung
  - Misst die Ähnlichkeit Methode $\leftrightarrow$ Maß?
- Manuelle Evaluierung durch Experten
  - Aufwendig und subjektiv
- Indirekte Evaluierung?
  - Bewertung durch Verbesserung der Ergebnisse anderer Methoden?

# Evaluation von Outlier-Detection-Algorithmen

Nicht ganz einfach:

- Unbalanciertes Problem
- Scores oft nicht aussagekräftig
- Labels i.d.R. nicht vollständig (weitere, “uninteressante” Ausreißer)
- i.d.R. dennoch “supervised” Evaluation
- besser eigentlich: Validieren durch Experten!

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Precision@2:

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Precision@2: jeweils $1/2$. Recall: $1/2$. F-Measure: $1/2$.

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Precision@2: jeweils $1/2$. Recall: $1/2$. F-Measure: $1/2$.
Sehr grobes Maß! $0/2$, $1/2$ oder $2/2$ richtig!

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}(1/1 +$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}(1/1 + 1/2 +$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + \right.$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right)$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

$$AveP(S_2, 4) := \frac{1}{4}\left(\right.$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

$$AveP(S_2, 4) := \frac{1}{4}\left(0/1 +\right.$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

$$AveP(S_2, 4) := \frac{1}{4}\left(0/1 + 1/2 +\right.$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

$$AveP(S_2, 4) := \frac{1}{4}\left(0/1 + 1/2 + 2/3 +\right.$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

$$AveP(S_2, 4) := \frac{1}{4}\left(0/1 + 1/2 + 2/3 + 2/4\right)$$

# Evaluation von Outlier-Detection-Algorithmen

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

Average Precision für $k = [1 \ldots 4]$:

$$AveP(S_1, 4) := \frac{1}{4}\left(1/1 + 1/2 + 1/3 + 2/4\right) = 7/12$$

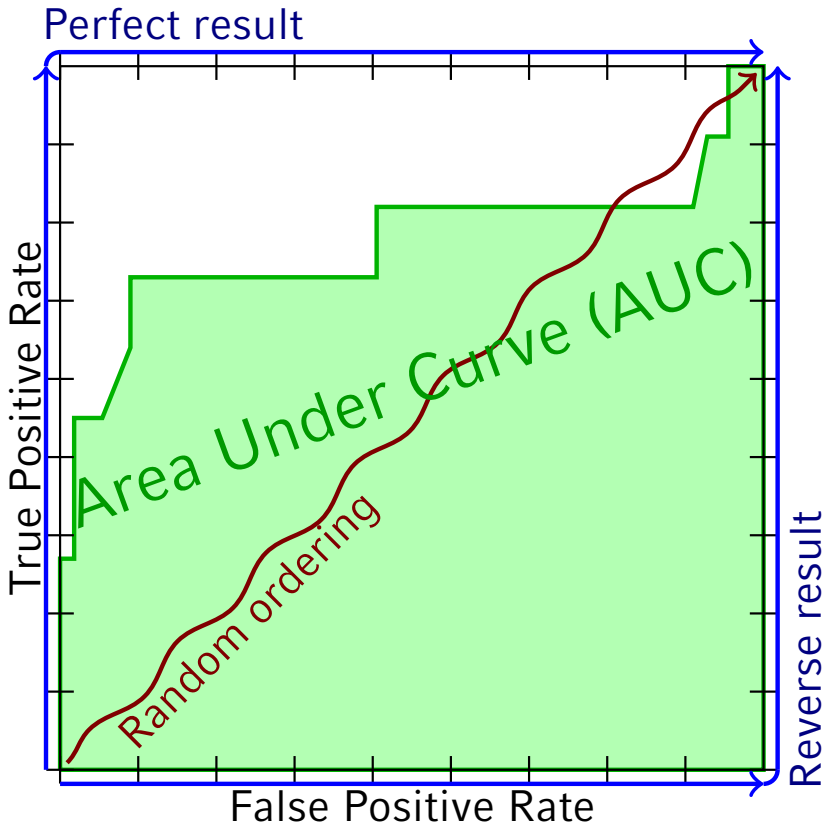
$$AveP(S_2, 4) := \frac{1}{4}\left(0/1 + 1/2 + 2/3 + 2/4\right) = 5/12$$

# ROC-Kurven

ROC-Kurven: False-Positive-Rate vs. True-Positive-Rate für *alle* $k$! Implizit gewichtet nach Klassengröße, da $x$-Achse: $|I|$ Stufen, $y$-Achse: $|O|$ Stufen!

# ROC-Kurven

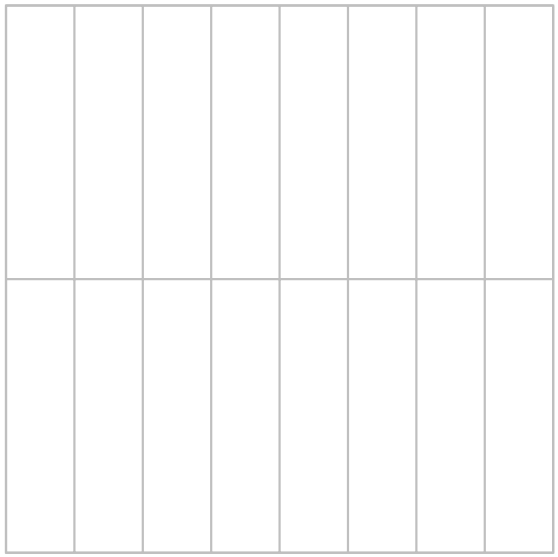
ROC-Kurven: False-Positive-Rate vs. True-Positive-Rate für *alle* $k$! Implizit gewichtet nach Klassengröße, da $x$-Achse: $|I|$ Stufen, $y$-Achse: $|O|$ Stufen!



# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

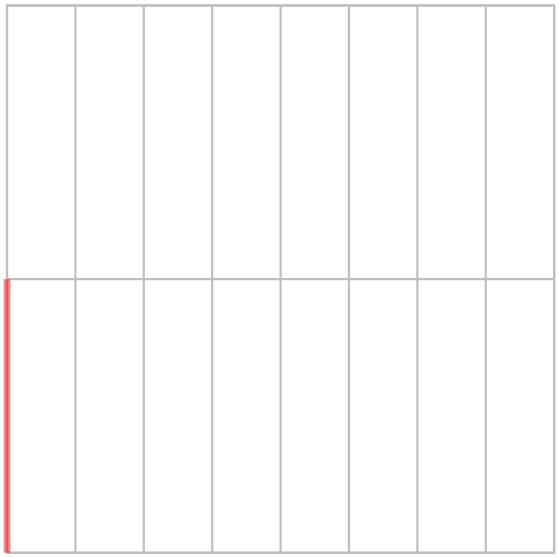
| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |



# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |



# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

# ROC-Kurven

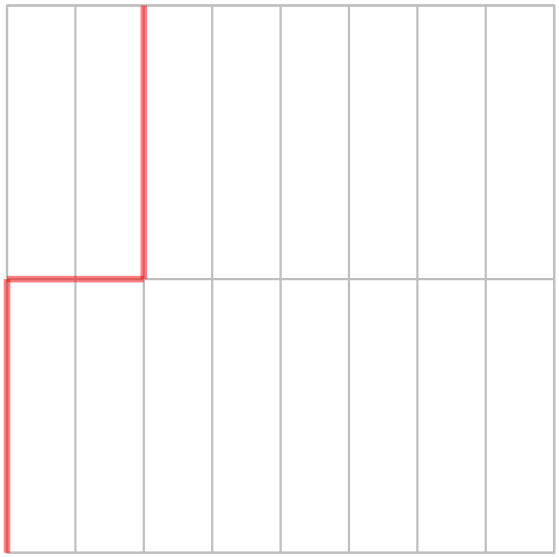
Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

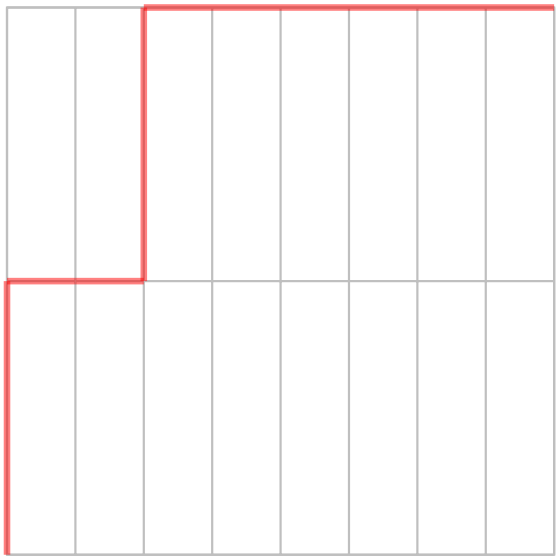
| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |



# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

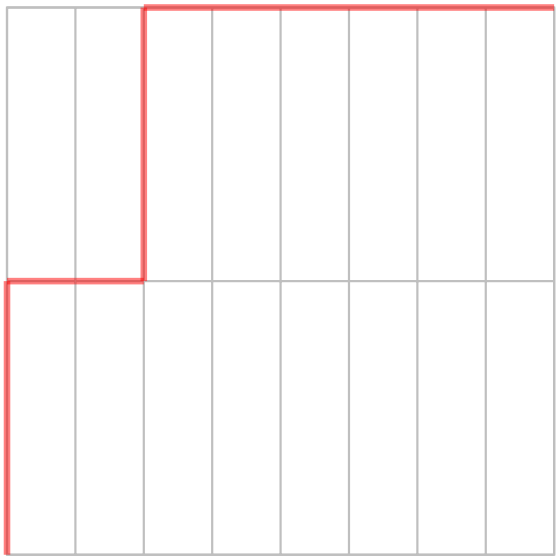
| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |



# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

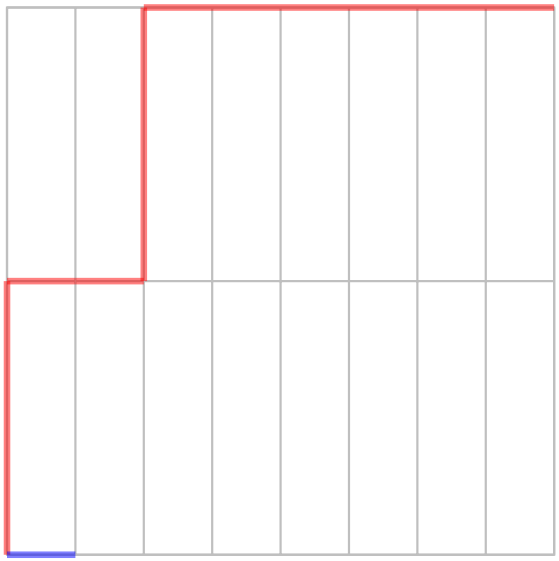


$$AUC(ROC(S_1)) = 14/16 = 0.875$$

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

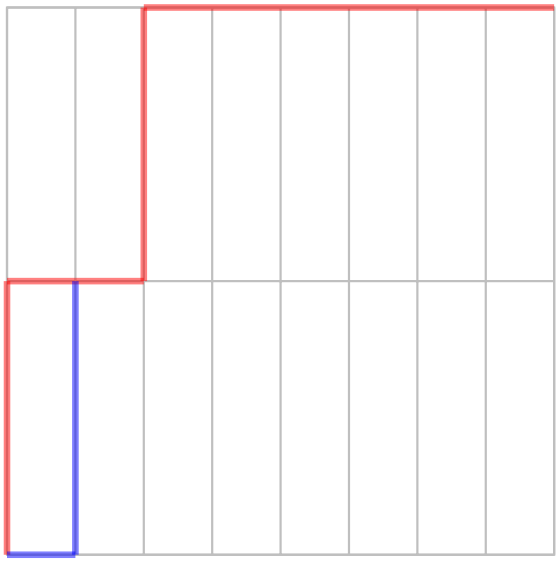


$$AUC(ROC(S_1)) = 14/16 = 0.875$$

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

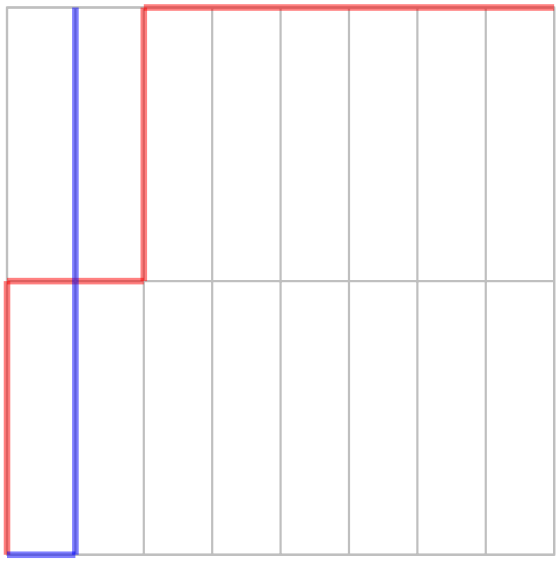


$$AUC(ROC(S_1)) = 14/16 = 0.875$$

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

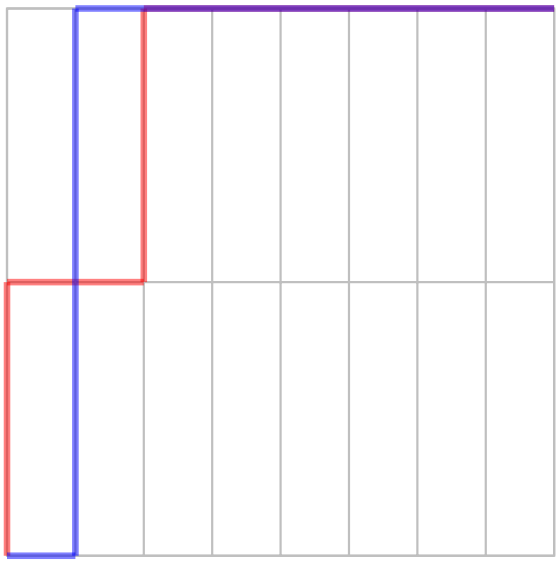


$$AUC(ROC(S_1)) = 14/16 = 0.875$$

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |

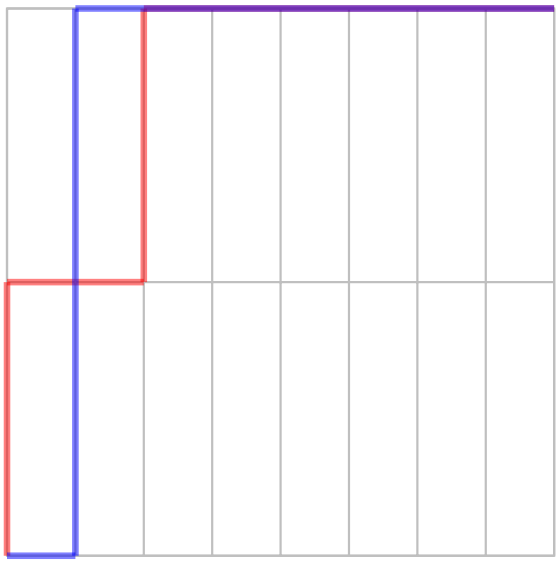


$$AUC(ROC(S_1)) = 14/16 = 0.875$$

# ROC-Kurven

Sortiert nach $S_1$ bzw. $S_2$:

| Label | + | − | − | + | − | − | − | − | − | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | 3.0 | 2.0 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1.1 | 1.0 | 0.9 |
| Label | − | + | + | − | − | − | − | − | − | − |
| $S_2$ | .91 | .90 | .89 | .81 | .80 | .80 | .50 | .50 | .20 | .10 |



$$AUC(ROC(S_1)) = 14/16 = 0.875$$

$$AUC(ROC(S_2)) = 14/16 = 0.875$$

# Kostenbasierte Bewertung



Idee: Scores sollten nicht nur als Rangfolge mit einfließen.

# Kostenbasierte Bewertung

Idee: Scores sollten nicht nur als Rangfolge mit einfließen.

Wenn die Scores zwischen 0 (inlier) und 1 (outlier) liegen, kann man den “Fehler” messen.

# Kostenbasierte Bewertung

Idee: Scores sollten nicht nur als Rangfolge mit einfließen.

Wenn die Scores zwischen 0 (inlier) und 1 (outlier) liegen, kann man den "Fehler" messen.

Gewichtung notwendig, da $|O| \ll |I|$.
Heuristik: $\frac{1}{2}$ Gewicht auf Inlier verteilen,
$\frac{1}{2}$ Gewicht auf die Outlier.

Publiziert in: :-)
H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. "Interpreting and Unifying Outlier Scores". In: *Proceedings of the 11th SIAM International Conference on Data Mining (SDM), Mesa, AZ*. 2011, pp. 13–24

# Kostenbasierte Bewertung



Lineare Skalierung auf $[0; 1]$:

| Label | $-$ | $-$ | $-$ | $-$ | $+$ | $-$ | $-$ | $-$ | $+$ | $-$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $n(S_1)$ | .0476 | .0952 | .0952 | .1904 | 1 | .5238 | .2857 | 0 | .2381 | .1429 |
| $n(S_2)$ | .8642 | .8642 | 0 | .8765 | .9753 | .4938 | .4938 | 1 | .9877 | .1235 |

# Kostenbasierte Bewertung



Lineare Skalierung auf $[0;1]$:

| Label | $-$ | $-$ | $-$ | $-$ | $+$ | $-$ | $-$ | $-$ | $+$ | $-$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $n(S_1)$ | .0476 | .0952 | .0952 | .1904 | 1 | .5238 | .2857 | 0 | .2381 | .1429 |
| $n(S_2)$ | .8642 | .8642 | 0 | .8765 | .9753 | .4938 | .4938 | 1 | .9877 | .1235 |

Kosten Inlier: $\frac{1}{|I|}\sum_{p\in I} \quad n(S_1(p))$

Kosten Outlier: $\frac{1}{|O|}\sum_{p\in O} 1 - n(S_1(p))$

# Kostenbasierte Bewertung

Lineare Skalierung auf $[0; 1]$:

| Label | − | − | − | − | + | − | − | − | + | − |
|---|---|---|---|---|---|---|---|---|---|---|
| $n(S_1)$ | .0476 | .0952 | .0952 | .1904 | 1 | .5238 | .2857 | 0 | .2381 | .1429 |
| $n(S_2)$ | .8642 | .8642 | 0 | .8765 | .9753 | .4938 | .4938 | 1 | .9877 | .1235 |

Kosten Inlier: $\frac{1}{|I|} \sum_{p \in I} \quad n(S_1(p))$

Kosten Outlier: $\frac{1}{|O|} \sum_{p \in O} 1 - n(S_1(p))$

| | Kosten Inlier | Kosten Outlier | Mittelwert |
|---|---|---|---|
| $S_1$ | .1726 | .3810 | .2768 |
| $S_2$ | .5895 | .0185 | .3040 |

# Kostenbasierte Bewertung

Lineare Skalierung auf $[0; 1]$:

| Label | $-$ | $-$ | $-$ | $-$ | $+$ | $-$ | $-$ | $-$ | $+$ | $-$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $n(S_1)$ | .0476 | .0952 | .0952 | .1904 | 1 | .5238 | .2857 | 0 | .2381 | .1429 |
| $n(S_2)$ | .8642 | .8642 | 0 | .8765 | .9753 | .4938 | .4938 | 1 | .9877 | .1235 |

Kosten Inlier: $\frac{1}{|I|} \sum_{p \in I} \quad n(S_1(p))$

Kosten Outlier: $\frac{1}{|O|} \sum_{p \in O} 1 - n(S_1(p))$

| | Kosten Inlier | Kosten Outlier | Mittelwert |
|---|---|---|---|
| $S_1$ | .1726 | .3810 | .2768 |
| $S_2$ | .5895 | .0185 | .3040 |

$S_2$ deutlich geringere Kosten auf den Ausreißern, aber dafür auch mehr “false positives”.