

Data Mining Tutorial

Kernel, Regression, Evaluation

Erich Schubert, Arthur Zimek

Ludwig-Maximilians-Universität München

2013-07-05 — KDD Übung

“Kernel” kann mehrdeutig sein! Unterscheidet zwischen:

- ▶ Kernel function (diese Aufgabe)
- ▶ Kernel density function (in der Statistik)
- ▶ Kernel matrix (oftmals: eine vorberechnete Distanz-Matrix – z.B. mit Kernelfunktion)
- ▶ Positiv (semi-) definite Matrix A in $d(x, x) := x^T A x \geq 0$

“Kernel” kann mehrdeutig sein! Unterscheidet zwischen:

- ▶ Kernel function (diese Aufgabe)
- ▶ Kernel density function (in der Statistik)
- ▶ Kernel matrix (oftmals: eine vorberechnete Distanz-Matrix – z.B. mit Kernelfunktion)
- ▶ Positiv (semi-) definite Matrix A in $d(x, x) := x^T A x \geq 0$

Positiv definite Matrix $A \Rightarrow x^T A y$ ist eine Kernel-Funktion.

“Kernel” kann mehrdeutig sein! Unterscheidet zwischen:

- ▶ Kernel function (diese Aufgabe)
- ▶ Kernel density function (in der Statistik)
- ▶ Kernel matrix (oftmals: eine vorberechnete Distanz-Matrix – z.B. mit Kernelfunktion)
- ▶ Positiv (semi-) definite Matrix A in $d(x, x) := x^T A x \geq 0$

Positiv definite Matrix $A \Rightarrow x^T A y$ ist eine Kernel-Funktion.

Aber nicht jede Kernelfunktion ist als positiv definite *Matrix* repräsentierbar!

Positiv semi-definit $\leftrightarrow$ Generalisierte Skalarprodukte

Standardskalarprodukt: $\langle x, y \rangle = \sum_i x_i y_i$

Generalisiertes Skalarprodukt: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Positiv semi-definit $\leftrightarrow$ Generalisierte Skalarprodukte

Standardskalarprodukt: $\langle x, y \rangle = \sum_i x_i y_i$

Generalisiertes Skalarprodukt: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E so dass $x^T \cdot E \cdot y = \langle x, y \rangle$?

$$\langle x, y \rangle = \sum_i \sum_j e_{ij} \cdot x_i \cdot y_j$$

Positiv semi-definit $\leftrightarrow$ Generalisierte Skalarprodukte

Standardskalarprodukt: $\langle x, y \rangle = \sum_i x_i y_i$

Generalisiertes Skalarprodukt: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E so dass $x^T \cdot E \cdot y = \langle x, y \rangle$?

$$\langle x, y \rangle = \sum_i \sum_j e_{ij} \cdot x_i \cdot y_j$$

$$e_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Positiv semi-definit $\leftrightarrow$ Generalisierte Skalarprodukte

Standardskalarprodukt: $\langle x, y \rangle = \sum_i x_i y_i$

Generalisiertes Skalarprodukt: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E so dass $x^T \cdot E \cdot y = \langle x, y \rangle$?

$$\langle x, y \rangle = \sum_i \sum_j e_{ij} \cdot x_i \cdot y_j$$

$$e_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Das ist die Einheitsmatrix!

Beweise für ein paar Kernel-Funktionen:

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i$$

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2$$

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2 \geq 0 \text{ offensichtlich}$$

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2 \geq 0 \text{ offensichtlich}$$

$$A) k_1(x, y) = 1$$

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2 \geq 0 \text{ offensichtlich}$$

$$A) k_1(x, y) = 1 = c^+ \text{ für nicht-negative Konstante } c^+ \in \mathbb{R}^{0+}$$

Beweise für ein paar Kernel-Funktionen:

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

$$0) k_0(x, y) = \langle x, y \rangle = x^T \cdot y$$

$$k_0(x, x) = \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2 \geq 0 \text{ offensichtlich}$$

$$A) k_1(x, y) = 1 = c^+ \text{ für nicht-negative Konstante } c^+ \in \mathbb{R}^{0+}$$

$$k_1(x, x) = c^+ \geq 0 \text{ trivial.}$$

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\text{}} \cdot \underbrace{k_0(x, x)}_{\text{}}$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k_0(x, x)}_{\geq 0} \geq 0$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k(x, x)}_{\geq 0} \geq 0$$

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k(x, x)}_{\geq 0} \geq 0$$

$$\text{C) } k_3(x, y) = 3 \cdot x^T \cdot y + 5$$

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k(x, x)}_{\geq 0} \geq 0$$

$$\text{C) } k_3(x, y) = 3 \cdot x^T \cdot y + 5 = c^+ \cdot k_0(x, y) + d^+$$

Beweise für ein paar Kernel-Funktionen:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$

$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k(x, x)}_{\geq 0} \geq 0$$

$$\text{C) } k_3(x, y) = 3 \cdot x^T \cdot y + 5 = c^+ \cdot k_0(x, y) + d^+$$

Genauso! Im allgemeinen: ein beliebiges Polynom aus nicht-negativen Faktoren und positiv semi-definiten Kernel-Funktionen ist wiederum positiv semi-definit!

Beispiel: $2k_0(x, y) \cdot k_1(x, y) + k_0(x, y)^2 + k_1(x, y)^2 + 7$

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Lineare Regression $\Leftrightarrow$ beste Ausgleichsgerade!

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Optimale Steigung: $\text{Cov}(X, Y) / \text{Var}(X)$

(Kleinste-Quadrate-Schätzer, siehe math. Statistik)

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Optimale Steigung: $\text{Cov}(X, Y) / \text{Var}(X)$

(Kleinste-Quadrate-Schätzer, siehe math. Statistik)

Optimaler Achsenabschnitt: $\bar{y} = \alpha \bar{x} + \beta$

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Optimale Steigung: $\text{Cov}(X, Y) / \text{Var}(X)$

(Kleinste-Quadrate-Schätzer, siehe math. Statistik)

Optimaler Achsenabschnitt: $\beta = \bar{y} - \alpha \bar{x}$

Für Mittelwerte $\bar{x}$, $\bar{y}$ und α .

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Optimale Steigung: $\text{Cov}(X, Y) / \text{Var}(X)$

(Kleinste-Quadrate-Schätzer, siehe math. Statistik)

Optimaler Achsenabschnitt: $\beta = \bar{y} - \alpha \bar{x}$

Für Mittelwerte $\bar{x}$, $\bar{y}$ und α .

Also als Erstes: Mittelwerte berechnen, Daten zentrieren.

Eingabedaten:

x	3	8	9	13	3	6	11	21	1	16
y	30	57	64	72	36	43	59	90	20	83

Allgemeine Gerade: $y = \alpha \cdot x + \beta$.

Benötigt also: α (Steigung), β (Achsenabschnitt).

Optimale Steigung: $\text{Cov}(X, Y) / \text{Var}(X)$

(Kleinste-Quadrate-Schätzer, siehe math. Statistik)

Optimaler Achsenabschnitt: $\beta = \bar{y} - \alpha \bar{x}$

Für Mittelwerte $\bar{x}$, $\bar{y}$ und α .

Also als Erstes: Mittelwerte berechnen, Daten zentrieren.

Mittelwerte: $\bar{x} = 9.1$, $\bar{y} = 55.4$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\frac{1}{n-1} \sum (x - \bar{x})(y - \bar{y})}{\frac{1}{n-1} \sum (x - \bar{x})(x - \bar{x})}$$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})(x - \bar{x})} = \frac{1269.6}{358.9} \approx 3.53747562 = \alpha$$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})(x - \bar{x})} = \frac{1269.6}{358.9} \approx 3.53747562 = \alpha$$

$$\beta = \bar{y} - \alpha \bar{x} \approx 23.2089719$$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61
$\alpha x + \beta \approx$		33.82	51.51	55.05	69.20	33.82	44.43	62.12	97.50	26.74	79.81

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})(x - \bar{x})} = \frac{1269.6}{358.9} \approx 3.53747562 = \alpha$$

$$\beta = \bar{y} - \alpha \bar{x} \approx 23.2089719$$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61
$\alpha x + \beta \approx$		33.82	51.51	55.05	69.20	33.82	44.43	62.12	97.50	26.74	79.81
reales y		30.	57.	64.	72.	36.	43.	59.	90.	20.	83.
$(\alpha x + \beta - y)^2 \approx$		14.60	30.15	80.17	7.86	4.75	2.06	9.74	56.19	45.51	10.19

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})(x - \bar{x})} = \frac{1269.6}{358.9} \approx 3.53747562 = \alpha$$

$$\beta = \bar{y} - \alpha \bar{x} \approx 23.2089719$$

Zentrierte Daten:

$\bar{x} = 9.1$	$x - \bar{x}$	-6.1	-1.1	-0.1	3.9	-6.1	-3.1	1.9	11.9	-8.1	6.9
$\bar{y} = 55.4$	$y - \bar{y}$	-25.4	1.6	8.6	16.6	-19.4	-12.4	3.6	34.6	-35.4	27.6
$(x - \bar{x}) \cdot (y - \bar{y})$		154.94	-1.76	-0.86	64.74	118.34	38.44	6.84	411.74	286.74	190.44
$(x - \bar{x})^2$		37.21	1.21	0.01	15.21	37.21	9.61	3.61	141.61	65.61	47.61
$\alpha x + \beta \approx$		33.82	51.51	55.05	69.20	33.82	44.43	62.12	97.50	26.74	79.81
reales y		30.	57.	64.	72.	36.	43.	59.	90.	20.	83.
$(\alpha x + \beta - y)^2 \approx$		14.60	30.15	80.17	7.86	4.75	2.06	9.74	56.19	45.51	10.19

$$\frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})(x - \bar{x})} = \frac{1269.6}{358.9} \approx 3.53747562 = \alpha$$

$$\beta = \bar{y} - \alpha \bar{x} \approx 23.2089719$$

Summe quadratischer Fehler: ≈ 261.220

Wurzel aus mittlerem quadratischen Fehler: ≈ 5.11097

Prognose:

Formel: $\hat{y} = 3.53747562 \cdot x + 23.2089719$

Prognose:

$$\text{Formel: } \hat{y} = 3.53747562 \cdot x + 23.2089719$$

A) $x = 20$

Prognose:

$$\text{Formel: } \hat{y} = 3.53747562 \cdot x + 23.2089719$$

A) $x = 20 \Rightarrow \hat{y} \approx 93.9585$

B) $x = 8$

Prognose:

$$\text{Formel: } \hat{y} = 3.53747562 \cdot x + 23.2089719$$

A) $x = 20 \Rightarrow \hat{y} \approx 93.9585$

B) $x = 8 \Rightarrow \hat{y} \approx 51.5088$

C) $x = 11$

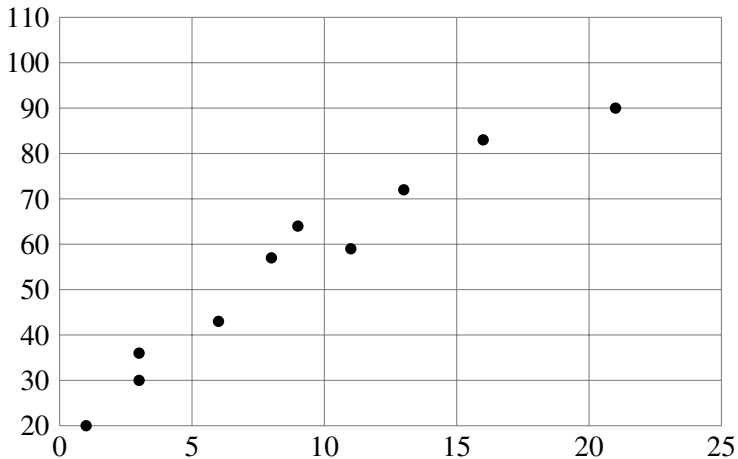
Prognose:

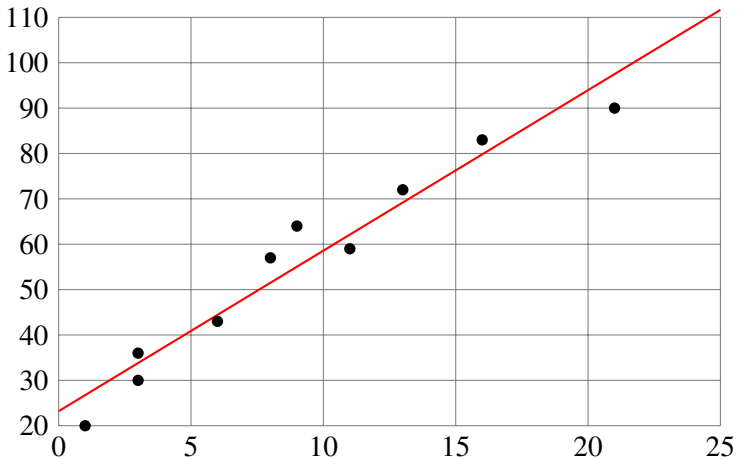
Formel: $\hat{y} = 3.53747562 \cdot x + 23.2089719$

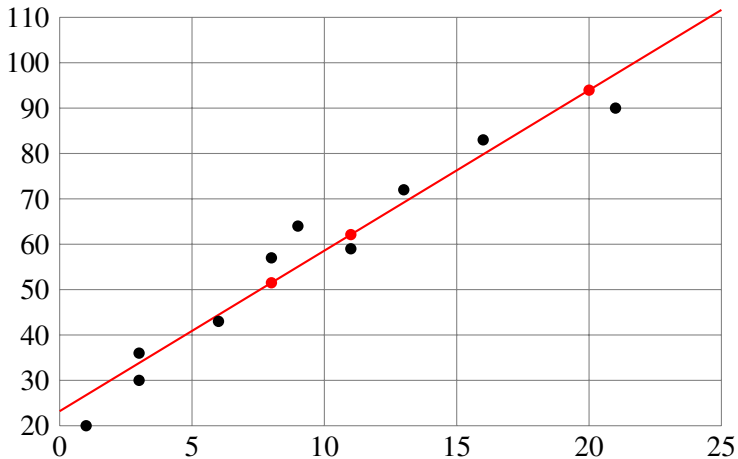
A) $x = 20 \Rightarrow \hat{y} \approx 93.9585$

B) $x = 8 \Rightarrow \hat{y} \approx 51.5088$

C) $x = 11 \Rightarrow \hat{y} \approx 62.1212$







```
import numpy as np
x = np.array([ 3,  8,  9, 13,  3,  6, 11, 21,  1, 16])
y = np.array([30, 57, 64, 72, 36, 43, 59, 90, 20, 83])

x.mean(), y.mean()
# (9.0999999999999996, 55.399999999999999)

x-x.mean()
# array([-6.1, -1.1, -0.1,  3.9, -6.1, ...,

(x-x.mean())*(y-y.mean())
# array([ 154.94, -1.76, -0.86,  64.74, ...,

(x-x.mean())**2
# array([ 3.72100000e+01,  1.21000000e+00, ...,

sum((x-x.mean())*(y-y.mean())), sum((x-x.mean())**2)
# (1269.6000000000001, 358.90000000000003)
```

```
alpha=sum((x-x.mean())*(y-y.mean()))/sum((x-x.mean())**2)
beta =y.mean() - alpha * x.mean()
alpha, beta
# (3.5374756199498467, 23.208971858456394)
```

```
alpha * x + beta
# array([ 33.82139872,  51.50877682,  55.04625244, ...
```

```
(alpha * x + beta - y) ** 2
# array([ 14.60308816,  30.15353203,  80.1695954 , ...
```

```
sum(((alpha * x + beta - y) ** 2))
# 261.22095291167477
```

```
np.sqrt(sum(((alpha * x + beta - y) ** 2))/len(x))
# 5.1109779192604101
```

```
np.array([20, 8, 11]) * alpha + beta
# array([ 93.95848426,  51.50877682,  62.12120368])
```

(The first two lines are the actual regression, the last is prediction!)

Es gibt zahlreiche Varianten von Regression!
Hier: ein optimaler KQ-Schätzer bekannt.
Oftmals: numerische Suche nach lokalem Maximum!

- ▶ quadratischer Fehler vs linearer Fehler
- ▶ Polynome statt linearen Funktionen
- ▶ Regularisierung, um Overfitting zu vermeiden
- ▶ RANSAC – RANdom SAmple Consensus, robuster gegen Ausreißer
- ▶ Gradient descent für differenzierbare Funktionen
- ▶ u.v.m.

Mehr in der Vorlesung “Maschinelles Lernen”, denn Regression braucht Trainingsdaten, und “lernt”!

	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

	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_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:

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

	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:

	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:

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

	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: ≈ 0.7171

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

	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:

$$13/27 + 10/14$$

	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:

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

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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:

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

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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:

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

	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:

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

	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: ≈ 0.7368

Was ist das F-Measure einer Zelle?

Was ist das F-Measure einer Zelle?

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

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

Was ist das F-Measure einer Zelle?

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

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

F-Measure daraus:

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

Was ist das F-Measure einer Zelle?

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

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

F-Measure daraus:

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

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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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:

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

	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:

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

	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:

	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:

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

	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:

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

	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: ≈ 0.7091

	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:

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

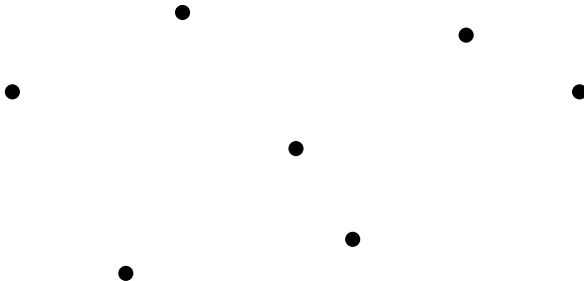
Spaltenmaxima:

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

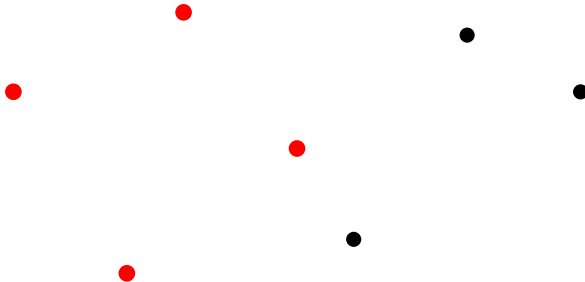
Mittelwert daraus: ≈ 0.7091

F-Measure daraus: ≈ 0.7088

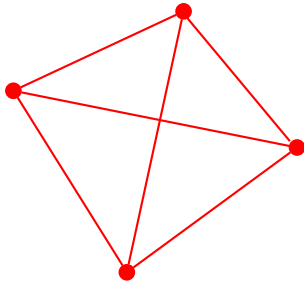
Intuition:



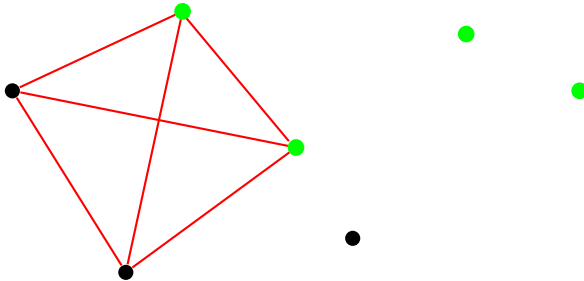
Intuition:



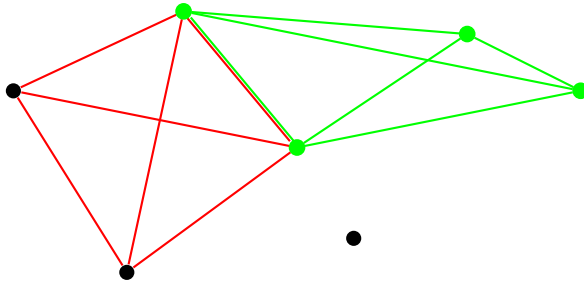
Intuition:



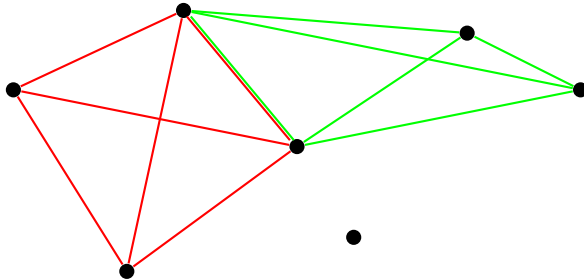
Intuition:



Intuition:

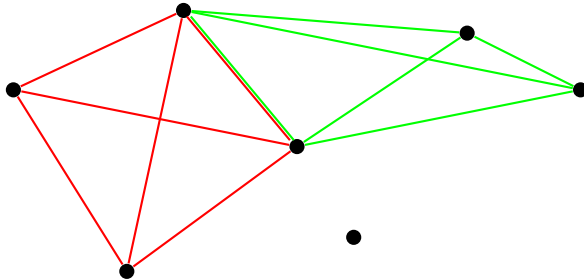


Intuition:



Cluster haben 1 gemeinsames Paar
Geringe Überlappung $\approx$ wenig Paare gemeinsam

Intuition:

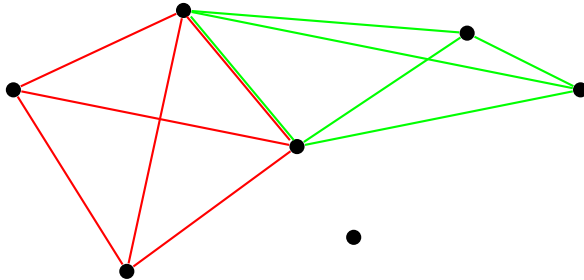


Cluster haben 1 gemeinsames Paar

Geringe Überlappung $\approx$ wenig Paare gemeinsam

Anzahl Paare in einer Menge von n Elementen:

Intuition:



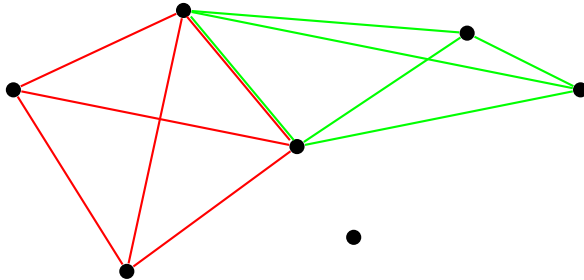
Cluster haben 1 gemeinsames Paar

Geringe Überlappung $\approx$ wenig Paare gemeinsam

Anzahl Paare in einer Menge von n Elementen:

$$\binom{n}{2}$$

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")

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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

Data Mining
Tutorial

E. Schubert,
A. Zimek

Aufgabe 10-1

Aufgabe 10-2

Aufgabe 10-3

	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:

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

	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:

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

$$\text{Summe Zeilen } \sum_j \binom{|B_j|}{2} =$$

	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:

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

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

	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
$\binom{ A_i }{2}$	136	55	136	55	105	

Summe Konfusionsmatrix:

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

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

$$\text{Summe Spalten } \sum_i \binom{|A_i|}{2} =$$

	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
$\binom{ A_i }{2}$	136	55	136	55	105	

Summe Konfusionsmatrix:

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

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

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

	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:

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

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

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

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

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

$$\begin{aligned} \text{D.h. } a &= 322, a + c = 668, a + b = 487, M = 2485 \\ d &= M - (a + b + c) = 2485 - (668 + 487 - 322) = 1652 \end{aligned}$$

Precision:

$$\text{D.h. } a = 322, a + c = 668, a + b = 487, M = 2485$$

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

$$\text{Precision: } a/(a + c)$$

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: $a/(a + b)$

$$\text{D.h. } a = 322, a + c = 668, a + b = 487, M = 2485$$

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

$$\text{Precision: } a/(a + c) \approx 0.66119$$

$$\text{Recall: } a/(a + b) \approx 0.48204$$

F-Measure daraus:

$$\text{D.h. } a = 322, a + c = 668, a + b = 487, M = 2485$$

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

$$\text{Precision: } a/(a + c) \approx 0.66119$$

$$\text{Recall: } a/(a + b) \approx 0.48204$$

$$\text{F-Measure daraus: } \approx 0.55758$$

Rand Index:

$$\text{D.h. } a = 322, a + c = 668, a + b = 487, M = 2485$$

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

$$\text{Precision: } a/(a + c) \approx 0.66119$$

$$\text{Recall: } a/(a + b) \approx 0.48204$$

$$\text{F-Measure daraus: } \approx 0.55758$$

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

$$\text{D.h. } a = 322, a + c = 668, a + b = 487, M = 2485$$

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

$$\text{Precision: } a/(a + c) \approx 0.66119$$

$$\text{Recall: } a/(a + b) \approx 0.48204$$

$$\text{F-Measure daraus: } \approx 0.55758$$

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

Adjusted Rand:

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: ≈ 0.55758

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

Adjusted Rand: Erwartungswert: $E = \frac{(a+c) \cdot (a+b)}{M}$

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: ≈ 0.55758

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

Adjusted Rand: Erwartungswert: $E = \frac{(a+c) \cdot (a+b)}{M} \approx 130.91$

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: ≈ 0.55758

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

Adjusted Rand: Erwartungswert: $E = \frac{(a+c) \cdot (a+b)}{M} \approx 130.91$

$$ARI = \frac{a - E}{\frac{(a+c) + (a+b)}{2} - E}$$

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: ≈ 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:

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: ≈ 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: $J = \frac{a}{a+b+c}$

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

Name	Precision 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

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.

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.

Name	Precision 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).

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

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

- ▶ Es gibt zahlreiche Maße, ohne klare Vorteile

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

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

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!

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

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?

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

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

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?

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?