

**6. Konferenz für SAS-Anwender in Forschung und Entwicklung**

**28. Februar – 1. März 2002**

**Universität Dortmund**

# **Diskriminanzanalyse mit binären Daten**

Bernd Jäger<sup>1</sup>, Michael Wodny<sup>1</sup>, Karl-Ernst Biebler<sup>1</sup>,  
Paul Eberhard Rudolph<sup>2</sup>, Karen Mathies<sup>3</sup>

<sup>1</sup> Institut für Biometrie und Medizinische Informatik, Ernst-Moritz-Arndt-Universität Greifswald

<sup>2</sup> Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere, Dummerstorf

<sup>3</sup> Klinik und Poliklinik für Innere Medizin A, Ernst-Moritz-Arndt-Universität Greifswald

## Abstände zweier Vektoren im $\mathbb{R}^n$

Euklidischer Abstand

$$d_E^2(x, y) = \sum_{i=1}^n (x_i - y_i)^2$$

Mahalanobisabstand  $d_M$

$$d_M(x, y) = (x - y) \cdot COV^{-1} \cdot (x - y)^T,$$

wobei  $COV^{-1}$  die Inverse der empirischen Kovarianzmatrix ist.

## Abstände für Binärdaten

Für  $x = (0,0,0,1,1,0,1,0) \in \{0,1\}^8$  und

$y = (0,1,0,1,1,0,0,0) \in \{0,1\}^8$

		$y_i$		
		1	0	$\Sigma$
$x_i$	1	$\alpha = 2$	$\beta = 1$	3
	0	$\gamma = 1$	$\delta = 4$	5
	$\Sigma$	3	5	$r = 8$

## Simple-Matching-Distance

$$d_{SM}(x, y) = 1 - \frac{\alpha + \delta}{n} = \frac{1}{4}$$

## Jaccard-Abstand (Tanimoto-Abstand)

$$d_j(x, y) = 1 - \frac{\alpha}{\alpha + \beta + \gamma} = \frac{5}{7}$$

## Jaccard-Metrik ist reichhaltiger als Simple-Matching-Distance

$$x = (1,1,0,0,0,1,1,1) \in \{0,1\}^8$$

$$y = (1,1,0,0,0,0,0,0) \in \{0,1\}^8$$

$$z = (0,0,1,0,0,1,1,1) \in \{0,1\}^8$$

$$d_{SM}(x, y) = 1 - \frac{2 + 3}{8} = \frac{3}{8}$$

$$d_j(x, y) = 1 - \frac{2}{2 + 3 + 0} = \frac{3}{5}$$

$$d_{SM}(x, z) = 1 - \frac{3 + 2}{8} = \frac{3}{8}$$

$$d_j(x, z) = 1 - \frac{3}{3 + 2 + 1} = \frac{1}{2}$$

# Jaccard-Metrik ist reichhaltiger als Simple-Matching-Distance

Simple-Matching-Distance

$$d_{SM}(x, y) = 1 - \frac{\alpha + \delta}{n}$$

$\alpha$  von 0 bis  $n$   
 $\delta$  von 0 bis  $(n - \alpha)$

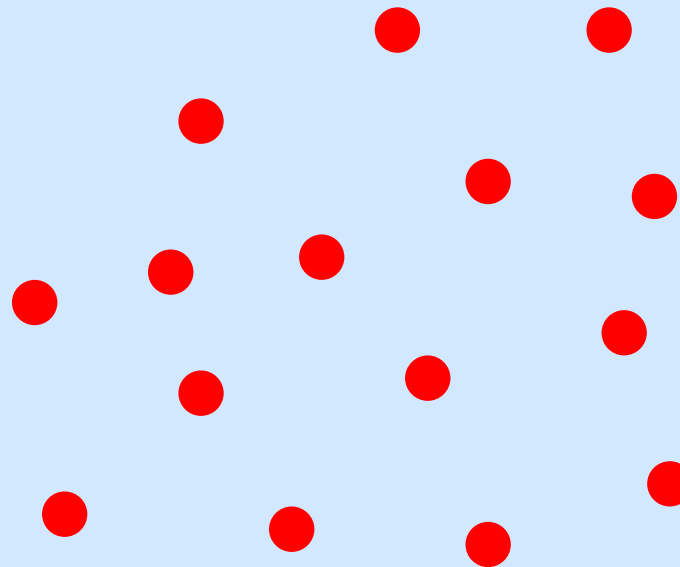
Jaccard-Abstand (Tanimoto-Abstand)

$$d_j(x, y) = 1 - \frac{\alpha}{\alpha + \beta + \gamma}$$

$\alpha$  von 0 bis  $n$   
 $\beta$  von 0 bis  $(n - \alpha)$   
 $\gamma$  von 0 bis  $(n - \alpha - \beta)$

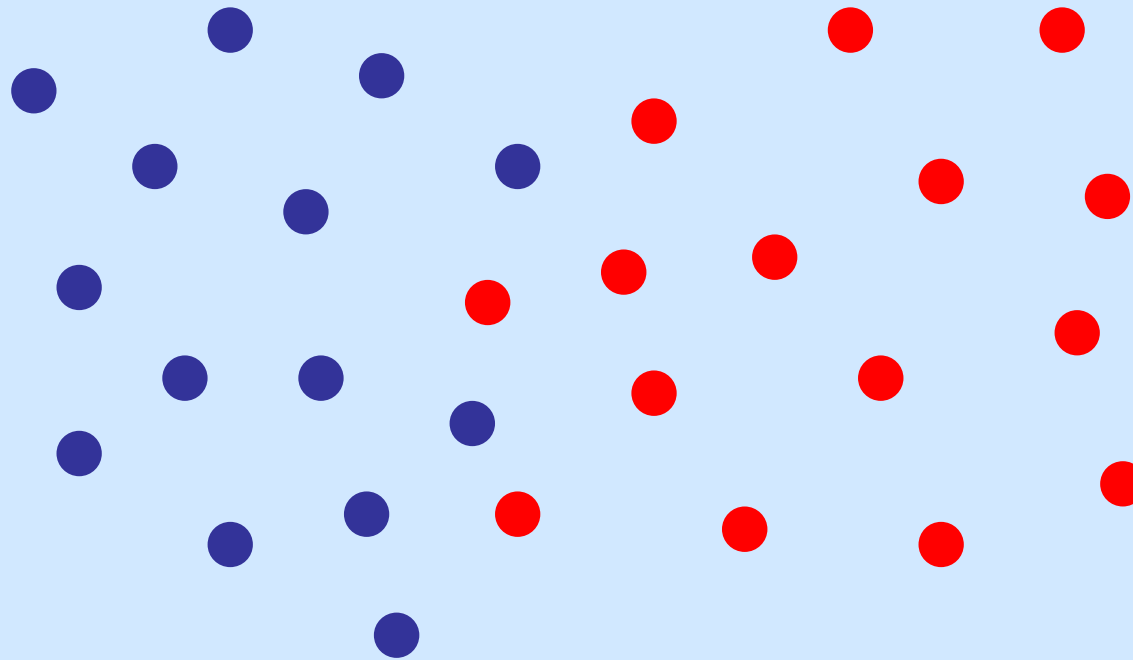
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

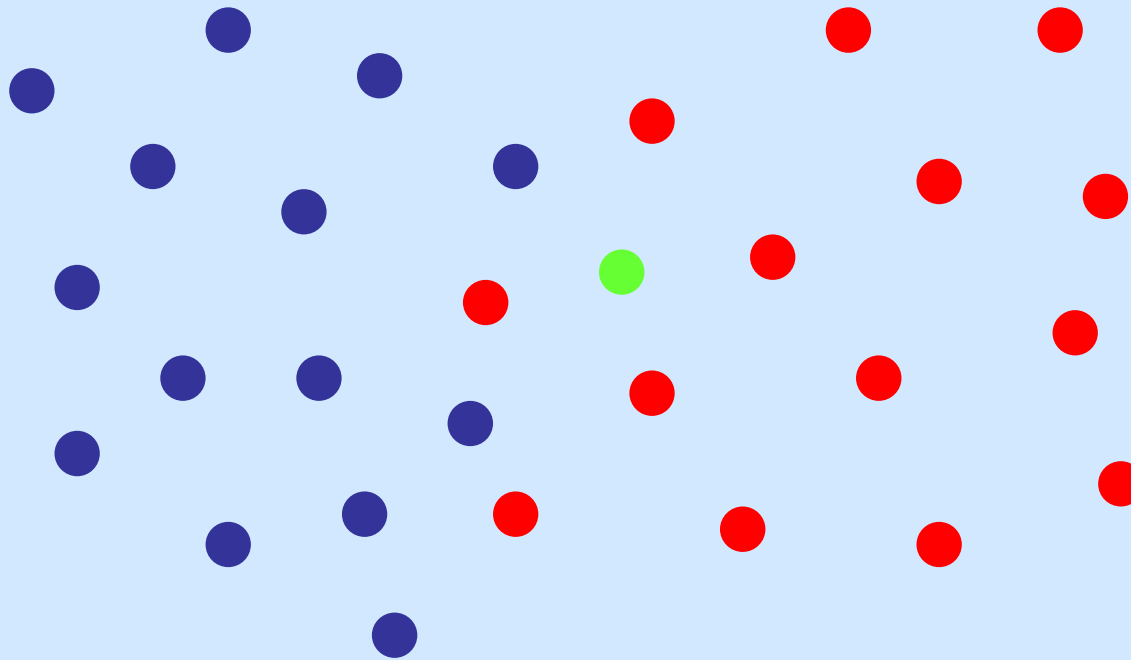
**n=5**





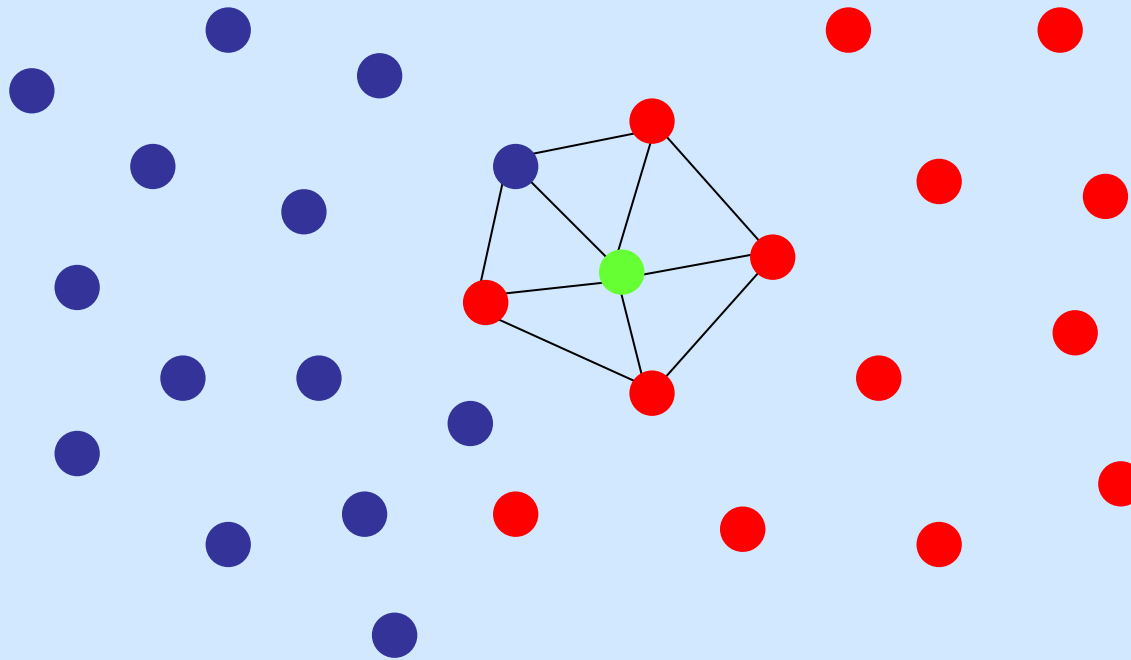
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



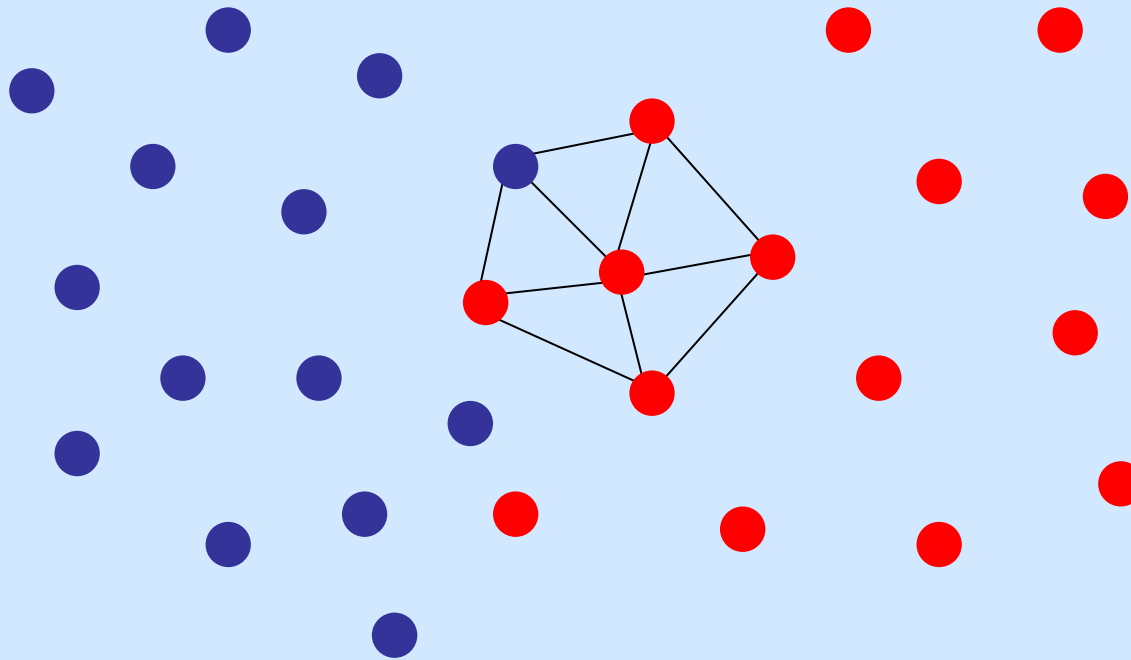
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



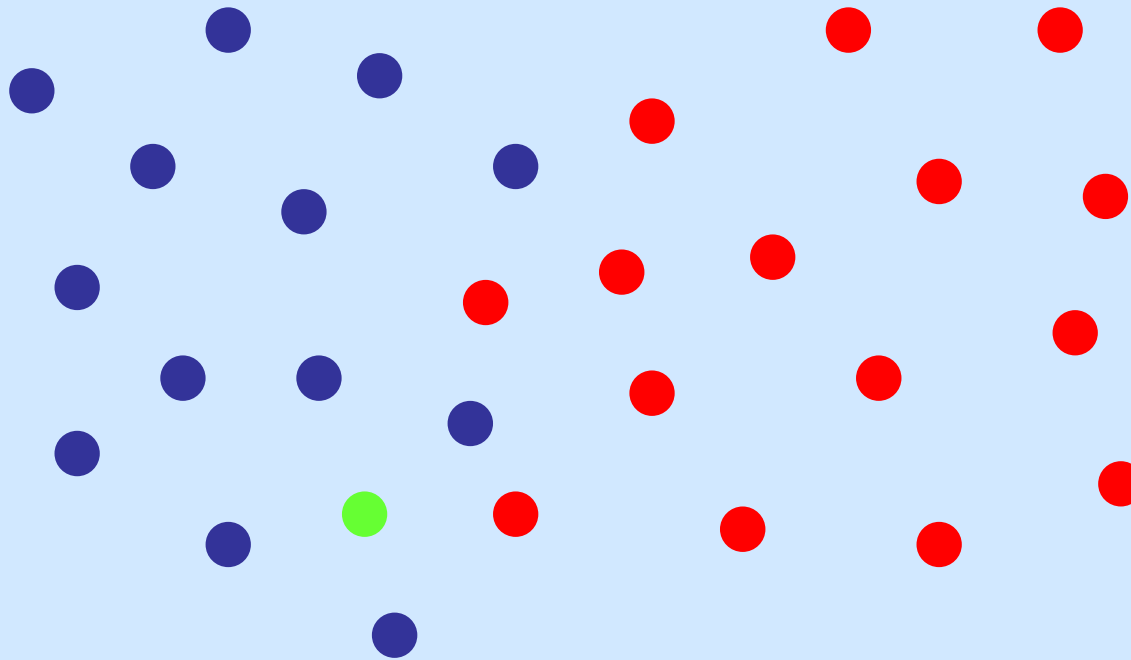
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



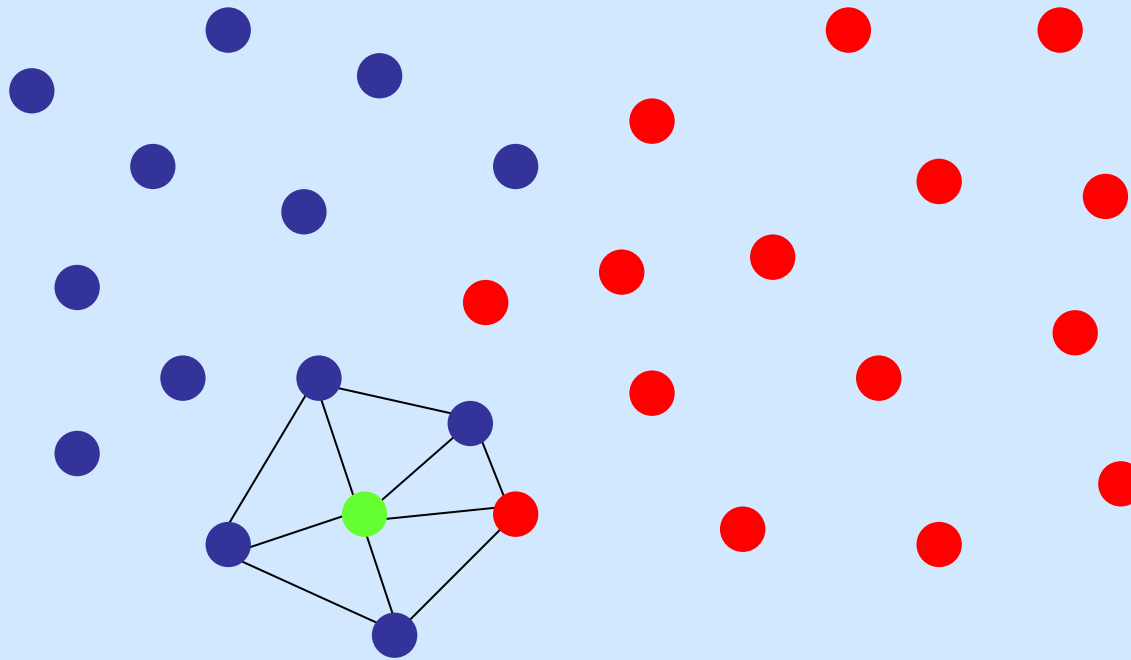
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



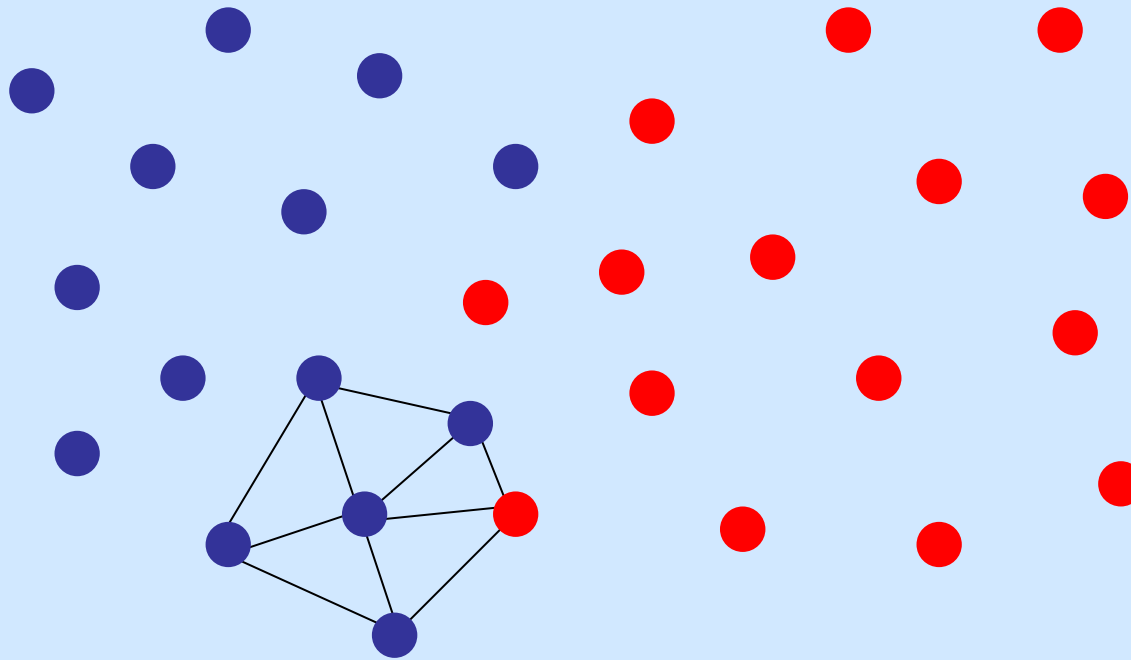
**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**



**n-nächste-Nachbarn-Regel  
(n-nearest-neighbor-rule)**

**n=5**

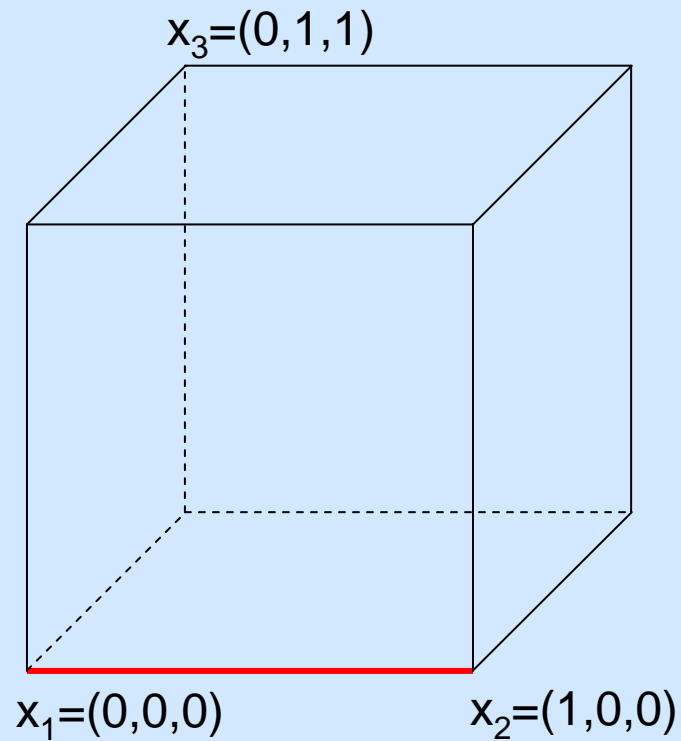


**Euklidischer Abstand**

$$d_E(x_1, x_2) = 1$$

**simple-matching distance**

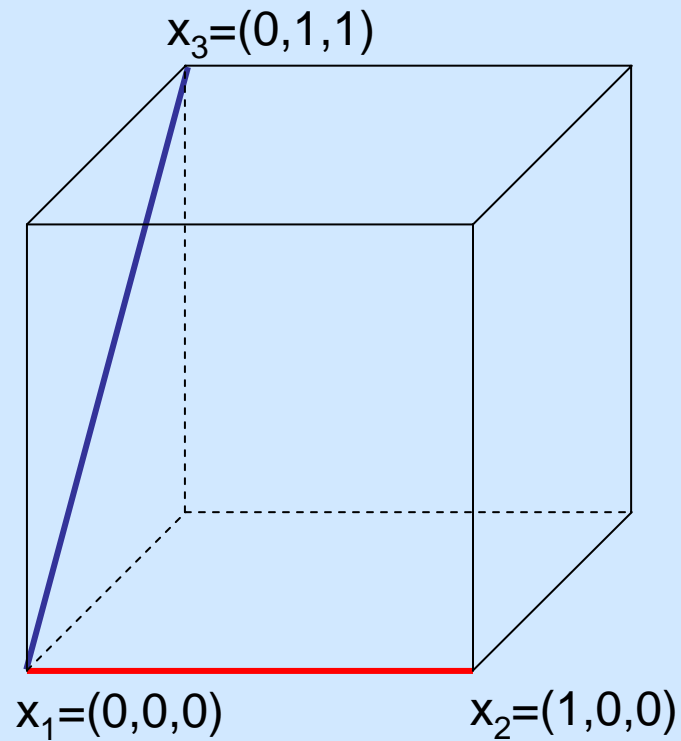
$$d_{SM}(x_1, x_2) = 1/3$$



## Euklidischer Abstand

$$d_E(x_1, x_2) = 1$$

$$d_E(x_1, x_3) = \sqrt{2}$$



## simple-matching distance

$$d_{SM}(x_1, x_2) = 1/3$$

$$d_{SM}(x_1, x_3) = 2/3$$

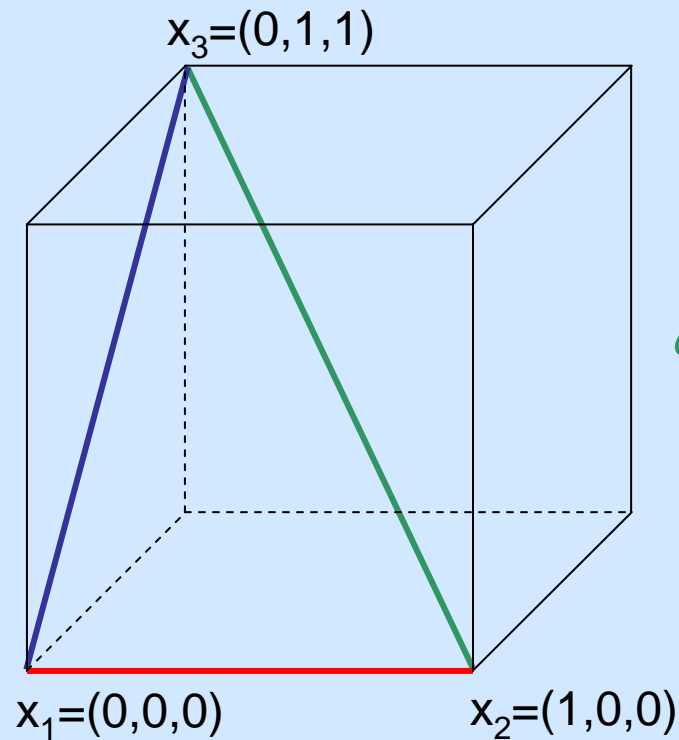


## Euklidischer Abstand

$$d_E(x_1, x_2) = 1$$

$$d_E(x_1, x_3) = \sqrt{2}$$

$$d_E(x_2, x_3) = \sqrt{3}$$



## simple-matching distance






$$d_{SM}(x_1, x_2) = 1/3$$

$$d_{SM}(x_1, x_3) = 2/3$$

$$d_{SM}(x_2, x_3) = 1$$

## Variation der 5 nächsten-Nachbarn-Methode

1.Fall: Zuordnung bis zum vierten Nachbarn entschieden

	1.N.	2.N.	3.N.	4.N	5.N.
	1/50	1/50	1/50	2/50	2/50
					

## Variation der 5 nächsten-Nachbarn-Methode

1.Fall: Zuordnung bis zum vierten Nachbarn entschieden

	1.N.	2.N.	3.N.	4.N	5.N.
●	1/50	1/50	1/50	2/50	2/50
	●	●	●	●	

Entscheidung nach der 5 nächsten-Nachbarn-Methode

## Variation der 5 nächsten-Nachbarn-Methode

2.Fall: Zuordnung bis zum vierten Nachbarn nicht entschieden

	1.N.	2.N.	3.N.	4.N.	5.N.	6.N.	7.N.
	1/50	1/50	2/50	3/50			
							

## Variation der 5 nächsten-Nachbarn-Methode

2.Fall: Zuordnung bis zum vierten Nachbarn nicht entschieden

	1.N.	2.N.	3.N.	4.N.	5.N.	6.N.	7.N.
	1/50	1/50	2/50	3/50	3/50	3/50	4/50
							

## Variation der 5 nächsten-Nachbarn-Methode

2.Fall: Zuordnung bis zum vierten Nachbarn nicht entschieden

	1.N.	2.N.	3.N.	4.N.	5.N.	6.N.	7.N.
●	1/50	1/50	2/50	3/50	3/50	3/50	4/50
	●	●	●	●	●	●	

Entscheidung nach der 6 nächsten-Nachbarn-Methode

## Reklassifikation nach der Lachenbruch-Methode

Jeder Datensatz  $x$  wird einzeln aus der Datei entfernt und bezüglich der Restdatei (Lernstichprobe) klassifiziert. Anschließend wird der Datensatz wieder hinzugefügt.

		in		
		Gruppe 1	Gruppe 2	
aus	Gruppe 1	$n_{11}$	$n_{12}$	$n_{1\bullet}$
	Gruppe 2	$n_{21}$	$n_{22}$	$n_{2\bullet}$
		$n_{\bullet 1}$	$n_{\bullet 2}$	$n$

$$\frac{n_{11} + n_{22}}{n}$$

Maß für die **Richtigklassifikation**

$$\frac{n_{12} + n_{21}}{n}$$

Maß für die **Falschklassifikation**

## Reklassifikation beim abbauenden Verfahren

Anzahl	Reduzierte Variable	Morbus Wegner		Kontrollgruppe	
		richtig	falsch	richtig	falsch
58		12	15	14	14
57	Neubau	14	13	19	9
56	Infektionsanzahl	15	12	20	8
55	Myk_IgA	15	12	21	7
54	Blut 0	16	11	23	5
53	Land	18	9	24	4
52	CMV_IgM	19	8	24	4
51	Toxo_IgM	19	8	24	4
50	PBV_IgG	19	8	25	3



## Reklassifikation beim abbauenden Verfahren

Anzahl	Reduzierte Variable	Morbus Wegner		Kontrollgruppe	
		richtig	falsch	richtig	falsch
32	Katze	20	7	26	2
31	Hund	21	6	26	2
30	Röteln	22	5	25	3
29	Parfüme	22	5	25	3
28	Hydrocarbon	22	5	26	2
27	Altbau	22	5	25	3
26	Scharlach	22	5	24	4
25	Blut A	22	5	25	3
24	CMV_PCR	23	4	24	4

## Reklassifikation beim abbauenden Verfahren

Anzahl	Reduzierte Variable	Morbus Wegner		Kontrollgruppe	
		richtig	falsch	richtig	falsch
11	Pilze	24	3	23	5
10	Masern	25	2	23	5
9	Blut Rhesus	25	2	24	4
8	Blut B	24	3	24	7
7	Tonsillenektomie	24	3	22	6
6	andere Säuger	24	3	21	7
5	Grippe	24	3	23	5
4	Metalle	22	5	25	3
3	Allergien	20	7	23	5

Die Diskriminanzanalyse für binäre Daten und das abbauende Verfahren sind als Makro-Programme unter Verwendung von Prozeduren aus SAS/STAT und PROC IML geschrieben.

Interessenten können die Programme anfordern unter:

[bjaeager@biometrie.uni-greifswald.de](mailto:bjaeager@biometrie.uni-greifswald.de)