

Data Mining Tutorial

Session 4: Classification

Erich Schubert, Eirini Ntoutsis

Ludwig-Maximilians-Universität München

2012-05-31 — KDD class tutorial

Build the confusion matrix:

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A				
B				
C				
K_i				

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	
B	2	2	1	
C	1	1	3	
K_i				

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4		
2		
3		

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2		
3		

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3		

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

$$\text{Precision}(K, A) = 4/7$$

$$\text{Precision}(K, B) = 2/3$$

$$\text{Precision}(K, C) = 3/5$$

$$\text{Precision}(K, i) = \frac{|\{o \in K_i \mid K(o) = C(o)\}|}{|K_i|} = \frac{|TP_i|}{|TP_i| + |FP_i|}$$

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

$$\begin{aligned}
 \text{Precision}(K, A) &= 4/7 & \text{Recall}(K, A) &= 4/5 \\
 \text{Precision}(K, B) &= 2/3 & \text{Recall}(K, B) &= 2/5 \\
 \text{Precision}(K, C) &= 3/5 & \text{Recall}(K, C) &= 3/5
 \end{aligned}$$

$$\text{Recall}(K, i) = \frac{|\{o \in C_i \mid K(o) = C(o)\}|}{|C_i|} = \frac{|TP_i|}{|TP_i| + |FN_i|}$$

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

$$\text{Precision}(K, A) = 4/7$$

$$\text{Recall}(K, A) = 4/5$$

$$F_1(K, A) = 2/3$$

$$\text{Precision}(K, B) = 2/3$$

$$\text{Recall}(K, B) = 2/5$$

$$F_1(K, B) = 1/2$$

$$\text{Precision}(K, C) = 3/5$$

$$\text{Recall}(K, C) = 3/5$$

$$F_1(K, C) = 3/5$$

$$F_1(K, i) = \frac{2 \cdot \text{Recall}(K, i) \cdot \text{Precision}(K, i)}{\text{Recall}(K, i) + \text{Precision}(K, i)}$$

$$\left(\text{not general: } = \frac{2|TP_i|}{2|TP_i| + |FP_i| + |FN_i|} \right)$$

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

Micro Average F_1 :

$$|TP| = 4 + 2 + 3 = 9$$

$$|FP| = 3 + 1 + 2 = 6$$

$$|FN| = 1 + 3 + 2 = 6$$

Precision: $9/15$

Recall: $9/15$

Micro Average F_1 : $9/15 = 0.6$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

Micro Average F_1 :

$$|TP| = 4 + 2 + 3 = 9$$

$$\text{Precision: } 9/15$$

$$|FP| = 3 + 1 + 2 = 6$$

$$\text{Recall: } 9/15$$

$$|FN| = 1 + 3 + 2 = 6$$

$$\text{Micro Average } F_1: 9/15 = 0.6$$

Notice: Precision = Recall = F_1

Not just by chance: sum of diagonal / sum total

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

Macro Average F_1 :

average precision: $1/3(4/7 + 2/3 + 3/5) \approx 0.613$

average recall: $1/3(4/5 + 2/5 + 3/5) = 0.6$

Macro Average $F_1 \approx \frac{2 \cdot 0.6 \cdot 0.613}{0.6 + 0.613} = 0.606.$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Build the confusion matrix:

	A	B	C	C_i
A	4	0	1	5
B	2	2	1	5
C	1	1	3	5
K_i	7	3	5	15

$ TP $	$ FP $	$ FN $
4	3	1
2	1	3
3	2	2

Micro average: all instances are weighted equally.

Macro average: all classes are weighted equally.

Consider the following scenario:

1% of object are in class "interesting",

99% are in class "uninteresting".

Which measure is then more useful?

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

weather	snow	ski?	weather	snow	ski?
sunny	< 50	no	snow	< 50	no
rainy	< 50	no	sunny	≥ 50	yes
rainy	≥ 50	no	snow	≥ 50	yes
snow	≥ 50	yes	rainy	< 50	yes

A priori probabilities and conditional probabilities:

weather	snow	ski?	weather	snow	ski?
sunny	< 50	no	snow	< 50	no
rainy	< 50	no	sunny	≥ 50	yes
rainy	≥ 50	no	snow	≥ 50	yes
snow	≥ 50	yes	rainy	< 50	yes

A priori probabilities and conditional probabilities:

$$P(ski) = 1/2$$

$$P(\neg ski) = 1/2$$

weather	snow	ski?	weather	snow	ski?
sunny	< 50	no	snow	< 50	no
rainy	< 50	no	sunny	≥ 50	yes
rainy	≥ 50	no	snow	≥ 50	yes
snow	≥ 50	yes	rainy	< 50	yes

A priori probabilities and conditional probabilities:

$$P(\text{weather} = \text{sunny} | \text{ski}) = 1/4$$

$$P(\text{weather} = \text{snow} | \text{ski}) = 2/4$$

$$P(\text{weather} = \text{rainy} | \text{ski}) = 1/4$$

$$P(\text{weather} = \text{sunny} | \neg \text{ski}) = 1/4$$

$$P(\text{weather} = \text{snow} | \neg \text{ski}) = 1/4$$

$$P(\text{weather} = \text{rainy} | \neg \text{ski}) = 2/4$$

weather	snow	ski?	weather	snow	ski?
sunny	< 50	no	snow	< 50	no
rainy	< 50	no	sunny	≥ 50	yes
rainy	≥ 50	no	snow	≥ 50	yes
snow	≥ 50	yes	rainy	< 50	yes

A priori probabilities and conditional probabilities:

$$P(\text{snow} \geq 50 | \text{ski}) = \frac{3}{4}$$

$$P(\text{snow} < 50 | \text{ski}) = \frac{1}{4}$$

$$P(\text{snow} \geq 50 | \neg \text{ski}) = \frac{1}{4}$$

$$P(\text{snow} < 50 | \neg \text{ski}) = \frac{3}{4}$$

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

		weather			snow	
	a priori	sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

A) $\text{weather} = \text{sunny}, \text{snow} \geq 50$

$$\begin{aligned}
 & P(\text{ski} | \text{weather} = \text{sunny}, \text{snow} \geq 50) \\
 &= \frac{P(\text{weather} = \text{sunny} | \text{ski}) \cdot P(\text{snow} \geq 50 | \text{ski}) \cdot P(\text{ski})}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)} \\
 &= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)} \\
 &= \frac{\frac{3}{32}}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)}
 \end{aligned}$$

		weather			snow	
	a priori	sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

A) $\text{weather} = \text{sunny}, \text{snow} \geq 50$

$$\begin{aligned}
 & P(\neg \text{ski} | \text{weather} = \text{sunny}, \text{snow} \geq 50) \\
 &= \frac{P(\text{weather} = \text{sunny} | \neg \text{ski}) \cdot P(\text{snow} \geq 50 | \neg \text{ski}) \cdot P(\neg \text{ski})}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)} \\
 &= \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)} \\
 &= \frac{\frac{1}{32}}{P(\text{weather} = \text{sunny}, \text{snow} \geq 50)}
 \end{aligned}$$

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	1/2	1/4	2/4	1/4	3/4	1/4
$\neg$ ski	1/2	1/4	1/4	2/4	1/4	3/4

A) $\text{weather} = \text{sunny}, \text{snow} \geq 50$

$$P(\text{ski} | \text{weather} = \text{sunny}, \text{snow} \geq 50) = \frac{\frac{3}{32}}{P(\dots)}$$

$$P(\neg \text{ski} | \text{weather} = \text{sunny}, \text{snow} \geq 50) = \frac{\frac{1}{32}}{P(\dots)}$$

$\Rightarrow$ Ski

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

B) $\text{weather} = \text{rainy}, \text{snow} < 50$

$$\begin{aligned}
 & P(\text{ski} | \text{weather} = \text{rainy}, \text{snow} < 50) \\
 = & \frac{P(\text{weather} = \text{rainy} | \text{ski}) \cdot P(\text{snow} < 50 | \text{ski}) \cdot P(\text{ski})}{P(\text{weather} = \text{rainy}, \text{snow} < 50)} \\
 = & \frac{\frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{rainy}, \text{snow} < 50)} \\
 = & \frac{\frac{1}{32}}{P(\text{weather} = \text{rainy}, \text{snow} < 50)}
 \end{aligned}$$

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	1/2	1/4	2/4	1/4	3/4	1/4
$\neg$ ski	1/2	1/4	1/4	2/4	1/4	3/4

B) $\text{weather} = \text{rainy}, \text{snow} < 50$

$$\begin{aligned}
 & P(\neg \text{ski} | \text{weather} = \text{rainy}, \text{snow} < 50) \\
 &= \frac{P(\text{weather} = \text{rainy} | \neg \text{ski}) \cdot P(\text{snow} < 50 | \neg \text{ski}) \cdot P(\neg \text{ski})}{P(\text{weather} = \text{rainy}, \text{snow} < 50)} \\
 &= \frac{\frac{2}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{rainy}, \text{snow} < 50)} \\
 &= \frac{\frac{6}{32}}{P(\text{weather} = \text{rainy}, \text{snow} < 50)}
 \end{aligned}$$

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	1/2	1/4	2/4	1/4	3/4	1/4
$\neg$ ski	1/2	1/4	1/4	2/4	1/4	3/4

B) weather=rainy, snow < 50

$$P(\text{ski} | \text{weather} = \text{rainy}, \text{snow} < 50) = \frac{\frac{1}{32}}{P(\dots)}$$

$$P(\neg \text{ski} | \text{weather} = \text{rainy}, \text{snow} < 50) = \frac{\frac{6}{32}}{P(\dots)}$$

$\Rightarrow$ do not ski

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

C) $\text{weather} = \text{snow}, \text{snow} < 50$

$$\begin{aligned}
 & P(\text{ski} | \text{weather} = \text{snow}, \text{snow} < 50) \\
 = & \frac{P(\text{weather} = \text{snow} | \text{ski}) \cdot P(\text{snow} < 50 | \text{ski}) \cdot P(\text{ski})}{P(\text{weather} = \text{snow}, \text{snow} < 50)} \\
 = & \frac{\frac{2}{4} \cdot \frac{1}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{snow}, \text{snow} < 50)} \\
 = & \frac{\frac{2}{32}}{P(\text{weather} = \text{snow}, \text{snow} < 50)}
 \end{aligned}$$

	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	$1/2$	$1/4$	$2/4$	$1/4$	$3/4$	$1/4$
$\neg$ ski	$1/2$	$1/4$	$1/4$	$2/4$	$1/4$	$3/4$

C) $\text{weather} = \text{snow}, \text{snow} < 50$

$$\begin{aligned}
 & P(\neg \text{ski} | \text{weather} = \text{snow}, \text{snow} < 50) \\
 &= \frac{P(\text{weather} = \text{snow} | \neg \text{ski}) \cdot P(\text{snow} < 50 | \neg \text{ski}) \cdot P(\neg \text{ski})}{P(\text{weather} = \text{snow}, \text{snow} < 50)} \\
 &= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{2}}{P(\text{weather} = \text{snow}, \text{snow} < 50)} \\
 &= \frac{\frac{3}{32}}{P(\text{weather} = \text{snow}, \text{snow} < 50)}
 \end{aligned}$$

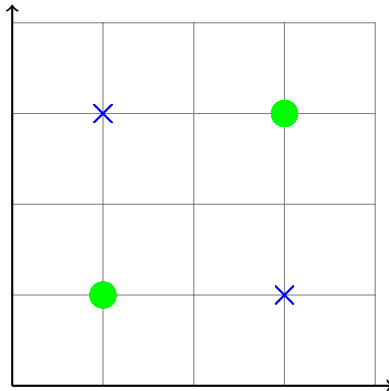
	a priori	weather			snow	
		sunny	snow	rainy	≥ 50	< 50
ski	1/2	1/4	2/4	1/4	3/4	1/4
$\neg$ ski	1/2	1/4	1/4	2/4	1/4	3/4

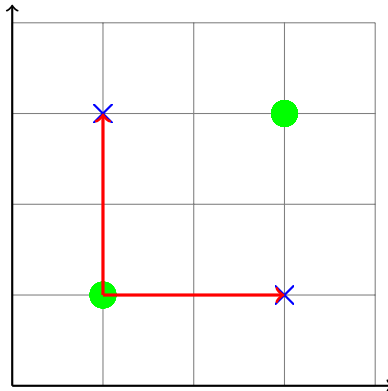
C) weather=snow, snow < 50

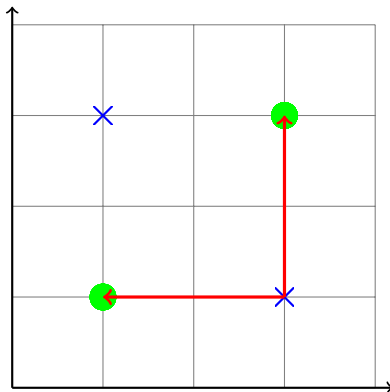
$$P(\text{ski} | \text{weather} = \text{snow}, \text{snow} < 50) = \frac{\frac{2}{32}}{P(\dots)}$$

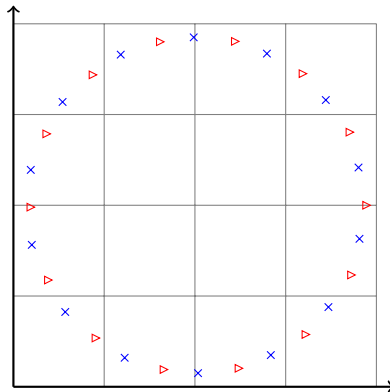
$$P(\neg \text{ski} | \text{weather} = \text{snow}, \text{snow} < 50) = \frac{\frac{3}{32}}{P(\dots)}$$

$\Rightarrow$ do not ski









Optimal classifier (only for random labels):
always assign the majority label.
Expected error rate?

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Optimal classifier (only for random labels):
always assign the majority label.
Expected error rate?

Since $|A| = |B| = |D|/2$, the error rate will be 50%.

Leave one out validation:
Expected error rate?

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Leave one out validation:
Expected error rate?

The wrong class always becomes majority since we leave out only the test object.

Expected error will be 100%!
This is obviously too pessimistic.

Bootstrap by sampling with replacement: every object is left out with a probability of $(1 - \frac{1}{n})^n \approx 0.368$ i.e. only about 63.2% of the objects are used for training. (In 10-fold cross validation, 90% of the data is used!) Regular error estimation would be pessimistic, as it contains duplicates.

Bootstrap by sampling with replacement: every object is left out with a probability of $(1 - \frac{1}{n})^n \approx 0.368$ i.e. only about 63.2% of the objects are used for training. (In 10-fold cross validation, 90% of the data is used!) Regular error estimation would be pessimistic, as it contains duplicates.

Common practise: also include observed classification error (on training data) during evaluation:

$$\begin{aligned} \text{error rate} &= 0.632 \cdot \text{Error on test set} \\ &\quad + 0.368 \cdot \text{Error on training set} \end{aligned}$$

This will be repeated multiple times (with different samples) and averaged.

The error rate of the constant classifier is $\approx 50\%$.
(If we average enough iterations)

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

The error rate of the constant classifier is $\approx 50\%$.
(If we average enough iterations)

New best classifier for the training set: “memorize”!
On the training set the “memorize” approach can achieve a
precision of 100%!

The error rate of the constant classifier is $\approx 50\%$.
(If we average enough iterations)

New best classifier for the training set: “memorize”!
On the training set the “memorize” approach can achieve a
precision of 100%!

Then:

$$\text{error rate} = 0.632 \cdot 50\% + 0.368 \cdot 0\% = 31.6\%$$

which is a too optimistic estimation.

Nearest Neighbor Classification

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

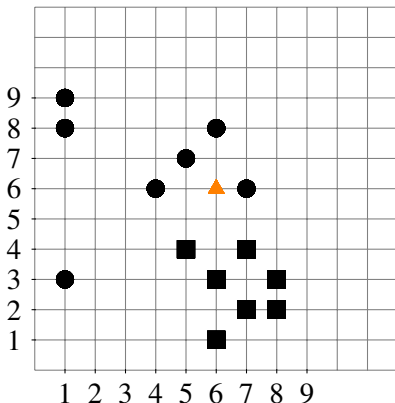
Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2



Nearest Neighbor Classification

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

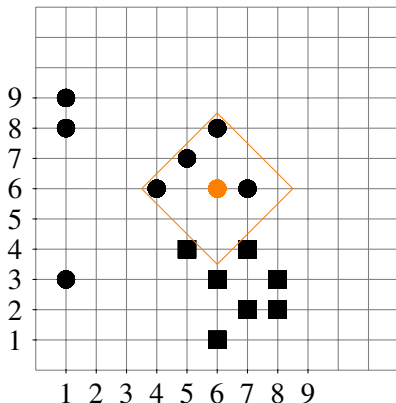
Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2



Nearest Neighbor Classification

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

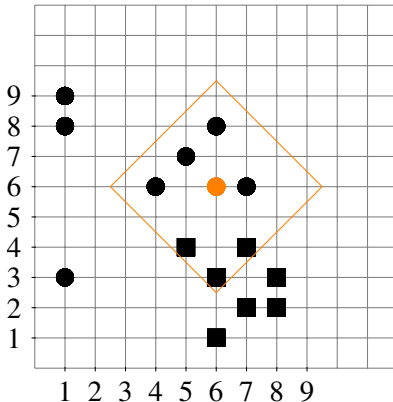
Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2



Nearest Neighbor Classification

Data Mining
Tutorial

E. Schubert,
E. Ntoutsi

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

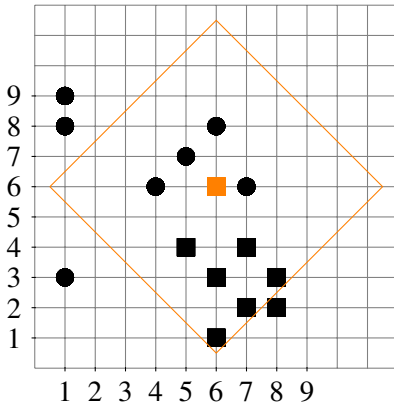
Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2



Recall: when splitting T on attribute A into partitions $T_1 \dots T_m$:

$$\text{entropy}(T) = - \sum_{i=1}^k p_i \cdot \log p_i$$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_{i=1}^m \frac{|T_i|}{|T|} \text{entropy}(T_i)$$

Full data set:

$$\text{entropy}(T) = 1, \text{ since } p(R = \text{low}) = \frac{1}{2} = p(R = \text{high})$$

Information gain for time attribute: Entropy for T_1

1-2 years: $T_1 = \text{Person } 1,4,6$

$$p(R = \text{low}) = \frac{1}{3}$$

$$p(R = \text{high}) = \frac{2}{3}$$

$$\text{entropy}(T_1) = - \sum_{i=1,2} p_i \log p_i$$

$$= - \left(\frac{1}{3} \log \frac{1}{3} + \frac{2}{3} \log \frac{2}{3} \right)$$

$$\approx 0.918$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for time attribute: Entropy for T_2
2-7 years: $T_2 = \text{Person } 2,7,8$

$$p(R = \text{low}) = \frac{2}{3}$$

$$p(R = \text{high}) = \frac{1}{3}$$

$$\begin{aligned} \text{entropy}(T_2) &= \text{entropy}(T_1) \\ &\approx 0.918 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for time attribute: Entropy for T_3
 > 7 years: $T_3 = \text{Person } 3,5$

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\begin{aligned} \text{entropy}(T_3) &= - \left(\frac{1}{2} \log \frac{1}{2} \right) \cdot 2 \\ &= 1 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for time attribute.

$$\begin{aligned}
 & \text{information-gain}(T, \text{Time}) \\
 &= \text{entropy}(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} \text{entropy}(T_i) \\
 &= 1 - \left(\frac{3}{8} \cdot 0.918 + \frac{3}{8} \cdot 0.918 + \frac{2}{8} \cdot 1 \right) \\
 &\approx 0.06
 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for gender attribute: Entropy for T_1
m: $T_1 = \text{Person } 1,2,5,6,8$

$$p(R = \text{low}) = \frac{2}{5}$$

$$p(R = \text{high}) = \frac{3}{5}$$

$$\text{entropy}(T_1) \approx 0.971$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for gender attribute: Entropy for T_2
w: $T_2 = \text{Person } 3,4,7$

$$p(R = \text{low}) = \frac{2}{3}$$

$$p(R = \text{high}) = \frac{1}{3}$$

$$\text{entropy}(T_2) \approx 0.918$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for gender attribute.

$$\begin{aligned} & \text{information-gain}(T, \text{Gender}) \\ &= \text{entropy}(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} \text{entropy}(T_i) \\ &= 1 - \left(\frac{5}{8} \cdot 0.971 + \frac{3}{8} \cdot 0.918 \right) \\ &\approx 0.05 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for area attribute: Entropy for T_1
urban: $T_1 = \text{Person } 1,7,8$

$$p(R = \text{low}) = 1$$

$$p(R = \text{high}) = 0$$

$$\text{entropy}(T_1) = 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for area attribute: Entropy for T_2

Rural: $T_2 = \text{Person } 2,3,4,5,6$

$$p(R = \text{low}) = \frac{1}{5}$$

$$p(R = \text{high}) = \frac{4}{5}$$

$$\text{entropy}(T_2) \approx 0.722$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

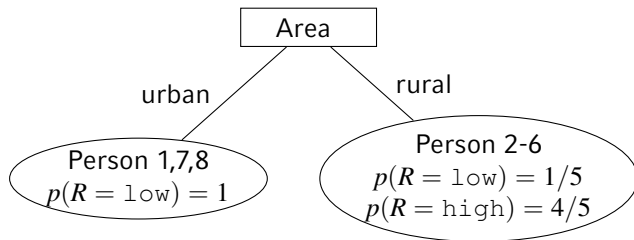
Aufgabe 5-3

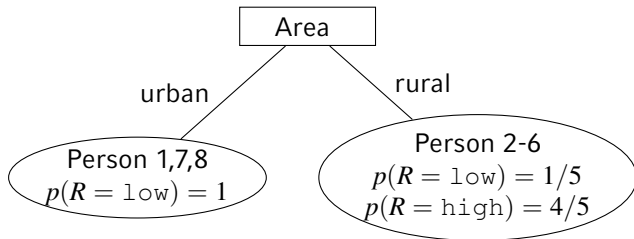
Aufgabe 6-2

Information gain for gender attribute.

$$\begin{aligned} & \text{information-gain}(T, \text{Area}) \\ &= 1 - \left(0 + \frac{5}{8} \cdot 0.722 \right) \\ &\approx 0.55 \end{aligned}$$

Attribute Area has the highest gain.





Right branch:

$$\text{entropy}(T) = - \left(\frac{1}{5} \log \frac{1}{5} + \frac{4}{5} \log \frac{4}{5} \right) \approx 0.722$$

Information gain for time attribute: Entropy for T_1
1-2 years: $T_1 = \text{Person 4,6}$

$$p(R = \text{high}) = 1$$

$$\text{entropy}(T_1) = 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for time attribute: Entropy for T_2
2-7 years: $T_2 = \text{Person 2}$

$$p(R = \text{high}) = 1$$

$$\text{entropy}(T_2) = 0$$

Information gain for time attribute: Entropy for T_3
 > 7 years: $T_3 = \text{Person } 3,5$

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\begin{aligned} \text{entropy}(T_3) &= - \left(\frac{1}{2} \log \frac{1}{2} \right) \cdot 2 \\ &= 1 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for time attribute.

$$\begin{aligned}
 & \textit{information-gain}(T, \textit{Time}) \\
 &= \textit{entropy}(T) - \sum_{i=1,2,3} \frac{|T_i|}{|T|} \textit{entropy}(T_i) \\
 &= 0.722 - \left(\frac{2}{5} \cdot 0 + \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot 1 \right) \\
 &\approx 0.322
 \end{aligned}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for gender attribute: Entropy for T_1
m: $T_1 = \text{Person } 2,5,6$

$$p(R = \text{high}) = 1$$

$$\text{entropy}(T_1) = 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Information gain for gender attribute: Entropy for T_2
w: $T_2 = \text{Person 3,4}$

$$p(R = \text{low}) = \frac{1}{2}$$

$$p(R = \text{high}) = \frac{1}{2}$$

$$\text{entropy}(T_2) = 1$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

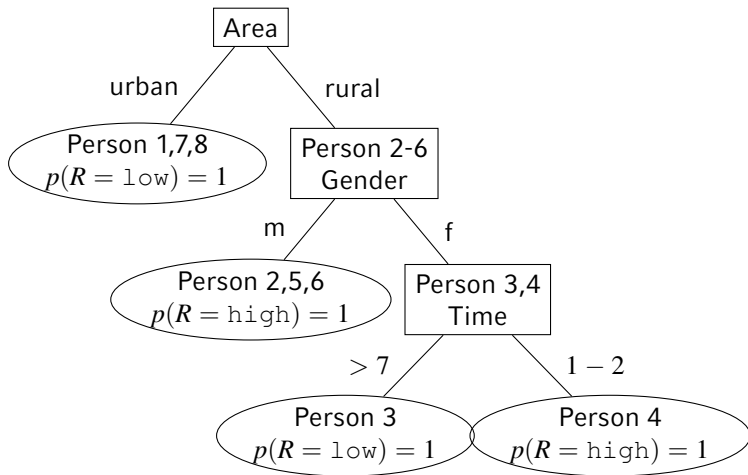
Aufgabe 5-3

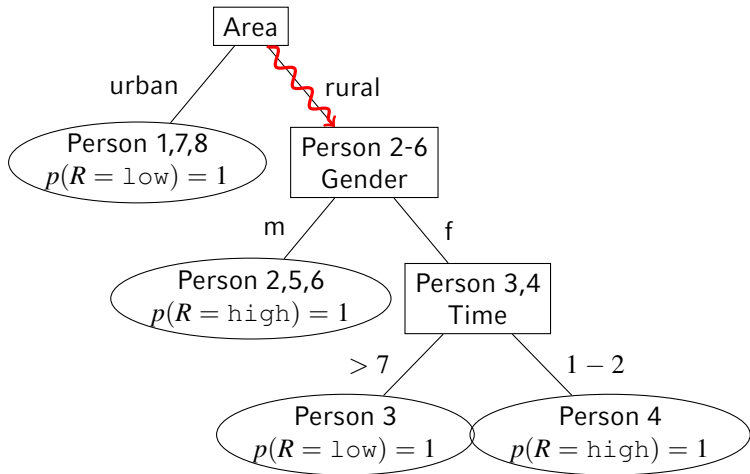
Aufgabe 6-2

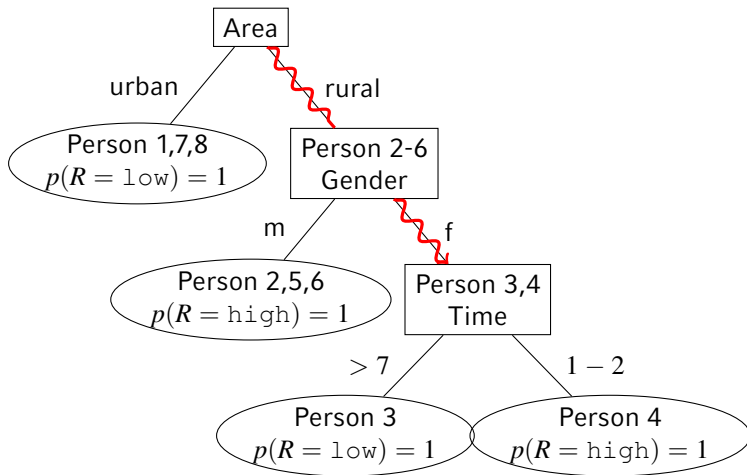
Information gain for gender attribute.

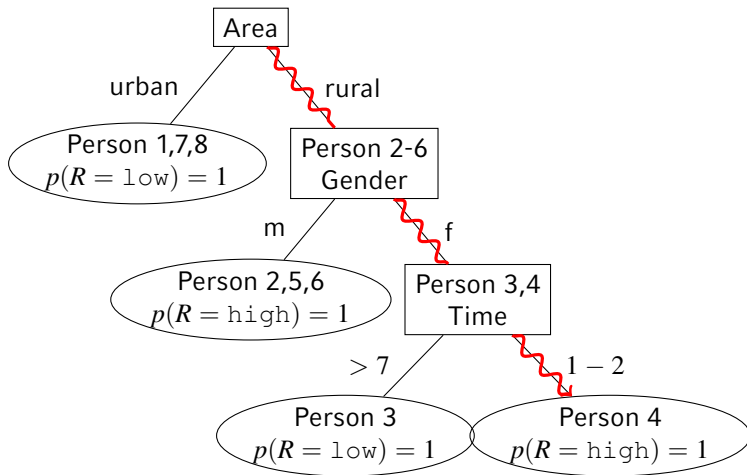
$$\begin{aligned}
 & \text{information-gain}(T, \text{Gender}) \\
 &= \text{entropy}(T) - \sum_{i=1,2} \frac{|T_i|}{|T|} \text{entropy}(T_i) \\
 &= 0.722 - \left(\frac{3}{5} \cdot 0 + \frac{2}{5} \cdot 1 \right) \\
 &\approx 0.322
 \end{aligned}$$

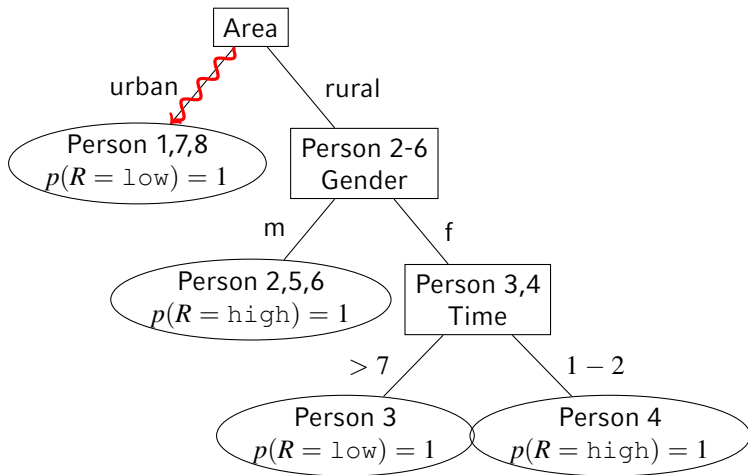
Same gain for both. Choose any.











A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k}$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k$$

Since $p_i = p_i(T_j^A)$, $\text{entropy}(T_j^A) = \text{entropy}(T) = \log k$

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k$$

Since $p_i = p_i(T_j^A)$, $\text{entropy}(T_j^A) = \text{entropy}(T) = \log k$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_j \frac{|T_j^A|}{|T|} \cdot \text{entropy}(T_j^A)$$

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k$$

Since $p_i = p_i(T_j^A)$, $\text{entropy}(T_j^A) = \text{entropy}(T) = \log k$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_j \frac{|T_j^A|}{|T|} \cdot \text{entropy}(T_j^A)$$

$$= \log k - m_A \frac{1}{m_A} \log k = 0$$

A) Uniform distribution on classes and values: $\forall_i p_i = \frac{1}{k}$

$$-\sum_i p_i \log p_i = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k$$

Since $p_i = p_i(T_j^A)$, $\text{entropy}(T_j^A) = \text{entropy}(T) = \log k$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_j \frac{|T_j^A|}{|T|} \cdot \text{entropy}(T_j^A)$$

$$= \log k - m_A \frac{1}{m_A} \log k = 0$$

As expected, splitting on this attribute yields no gain.

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_j \frac{|T_j^A|}{|T|} \cdot \text{entropy}(T_j^A)$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

$$\dots - \frac{1}{|T|} \left(\sum_j^{m_A} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'}) + |T_{m_A+1}^{A'}| \cdot \text{entropy}(T_{m_A+1}^{A'}) \right)$$

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

$$\dots - \frac{1}{|T|} \left(\sum_j^{m_A} |T_i^{A'}| \cdot \underbrace{\text{entropy}(T_i^{A'})}_{=\text{entropy}(T_i^A) \leq \log k} + |T_{m_A+1}^{A'}| \cdot \text{entropy}(T_{m_A+1}^{A'}) \right)$$

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

$$\dots - \frac{1}{|T|} \left(\sum_j^{m_A} |T_i^{A'}| \cdot \underbrace{\text{entropy}(T_i^{A'})}_{=\text{entropy}(T_i^A) \leq \log k} + |T_{m_A+1}^{A'}| \cdot \underbrace{\text{entropy}(T_{m_A+1}^{A'})}_{=\log k} \right)$$

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

$$\dots - \frac{1}{|T|} \left(\sum_j^{m_A} |T_i^{A'}| \cdot \underbrace{\text{entropy}(T_i^{A'})}_{=\text{entropy}(T_i^A) \leq \log k} + |T_{m_A+1}^{A'}| \cdot \underbrace{\text{entropy}(T_{m_A+1}^{A'})}_{=\log k} \right)$$

Gain cannot improve, $\log k$ is the maximal entropy!

B) Additional uniform attribute:

$$\text{information-gain}(T, A) = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A} |T_i^A| \cdot \text{entropy}(T_i^A)$$

$$\text{information-gain}(T, A') = \text{entropy}(T) - \frac{1}{|T|} \sum_j^{m_A+1} |T_i^{A'}| \cdot \text{entropy}(T_i^{A'})$$

$$\dots - \frac{1}{|T|} \left(\sum_j^{m_A} |T_i^{A'}| \cdot \underbrace{\text{entropy}(T_i^{A'})}_{=\text{entropy}(T_i^A) \leq \log k} + |T_{m_A+1}^{A'}| \cdot \underbrace{\text{entropy}(T_{m_A+1}^{A'})}_{=\log k} \right)$$

Gain cannot improve, $\log k$ is the maximal entropy!
Therefore, a split on A is preferred to A' .

C) At most one instance per attribute value.

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0 \Rightarrow \forall_i \text{entropy}(T_i^A) = 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0 \Rightarrow \forall_i \text{entropy}(T_i^A) = 0$$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_i p_i \cdot 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0 \Rightarrow \forall_i \text{entropy}(T_i^A) = 0$$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_i p_i \cdot 0$$

Best choice – maximum information gain!

Single split: every branch is pure and the tree complete.

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0 \Rightarrow \forall_i \text{entropy}(T_i^A) = 0$$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_i p_i \cdot 0$$

Best choice – maximum information gain!

Single split: every branch is pure and the tree complete.

Danger of overfitting! There is no *training* error, but the tree is just memorizing the data, and will not generalize to unseen data. True error will likely be much higher.

C) At most one instance per attribute value.

$$\exists_i \Rightarrow p_i = 1, \forall_{j \neq i} p_j = 0 \Rightarrow \forall_i \text{entropy}(T_i^A) = 0$$

$$\text{information-gain}(T, A) = \text{entropy}(T) - \sum_i p_i \cdot \text{entropy}(T_i^A)$$

Best choice – maximum information gain!

Single split: every branch is pure and the tree complete.

Danger of overfitting! There is no *training* error, but the tree is just memorizing the data, and will not generalize to unseen data. True error will likely be much higher.
Example: split by unique record ID.

“Kernel” can be confusing. Distinguish:

- ▶ Kernel function (this section)
- ▶ Kernel density function (in probability distributions)
- ▶ Kernel matrix (often: a precomputed distance matrix)
- ▶ positive (semi-) definite matrix in $d(x, x) := x^T A x \geq 0$

“Kernel” can be confusing. Distinguish:

- ▶ Kernel function (this section)
- ▶ Kernel density function (in probability distributions)
- ▶ Kernel matrix (often: a precomputed distance matrix)
- ▶ positive (semi-) definite matrix in $d(x, x) := x^T A x \geq 0$

Positive definite matrix $A \Rightarrow x^T A y$ is a kernel function.

“Kernel” can be confusing. Distinguish:

- ▶ Kernel function (this section)
- ▶ Kernel density function (in probability distributions)
- ▶ Kernel matrix (often: a precomputed distance matrix)
- ▶ positive (semi-) definite matrix in $d(x, x) := x^T A x \geq 0$

Positive definite matrix $A \Rightarrow x^T A y$ is a kernel function.
Arbitrary kernel function $\nRightarrow$ representable as positive definite matrix

positive semi-definite $\leftrightarrow$ generalized dot products

Usual dot product: $\langle x, y \rangle = \sum_i x_i y_i$

Generalized dot product: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

positive semi-definite $\leftrightarrow$ generalized dot products

Usual dot product: $\langle x, y \rangle = \sum_i x_i y_i$

Generalized dot product: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E such that $x^T \cdot E \cdot y = \langle x, y \rangle$?

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

positive semi-definite $\leftrightarrow$ generalized dot products

Usual dot product: $\langle x, y \rangle = \sum_i x_i y_i$

Generalized dot product: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E such that $x^T \cdot E \cdot y = \langle x, y \rangle$?

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

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

positive semi-definite $\leftrightarrow$ generalized dot products

Usual dot product: $\langle x, y \rangle = \sum_i x_i y_i$

Generalized dot product: $\langle x, y \rangle_A = x^T \cdot A \cdot y$

Matrix E such that $x^T \cdot E \cdot y = \langle x, y \rangle$?

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

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

This is the unit matrix!

Proof for some kernel functions:

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

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

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

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

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

$$\begin{aligned} 0) \quad k_0(x, y) &= \langle x, y \rangle = x^T \cdot y \\ k_0(x, x) &= \langle x, x \rangle = \sum_i x_i x_i = \sum_i x_i^2 \end{aligned}$$

Proof for some kernel functions:

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

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

Proof for some kernel functions:

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

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

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

Proof for some kernel functions:

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

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

$$A) k_1(x, y) = 1 = c^+ \text{ for } c^+ \geq 0$$

Proof for some kernel functions:

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

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

$$A) k_1(x, y) = 1 = c^+ \text{ for } c^+ \geq 0$$

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

Proof for some kernel functions:

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

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

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

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

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

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$
$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k_0(x, x)}_{\geq 0} \geq 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

$$\text{B) } k_2(x, y) = 3 \cdot x^T \cdot y = c^+ \cdot k_0(x, y)$$
$$k_2(x, x) = \underbrace{c^+}_{\geq 0} \cdot \underbrace{k(x, y)}_{\geq 0} \geq 0$$

Aufgabe 3-3

Aufgabe 4-1

Aufgabe 4-2

Aufgabe 4-3

Aufgabe 5-1

Aufgabe 5-2

Aufgabe 5-3

Aufgabe 6-2

Proof for some kernel functions:

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

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

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

Proof for some kernel functions:

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

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

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

Proof for some kernel functions:

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

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

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

Same thing. More general: any polynomial built of non-negative factors and positive semi-definite kernel functions is positive semi-definite.

$$\text{Example: } 2k_0(x, y) \cdot k_1(x, y) + k_0(x, y)^2 + k_1(x, y)^2 + 7$$