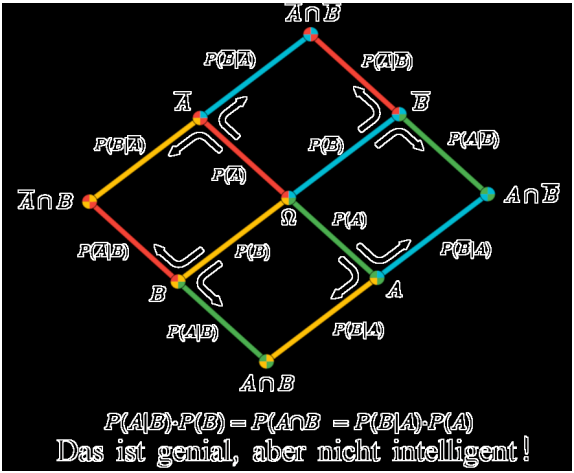
KI – der neue Hype, doch ML hilft seit Jahren gegen SPAM

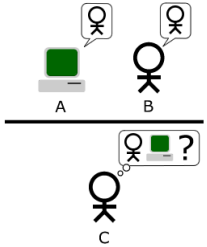


KI künstliche Intelligenz

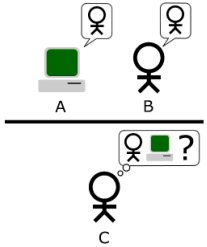
Was ist Intelligenz?

Was ist Intelligenz?

Was ist Intelligenz?
Turingtest



Was ist Intelligenz?
Turingtest



Lovelace Test 2.0
Beispiel: Komponiere ein Stück in Moll über die Liebe.

Maschinelles Lernen

Ziel ist es, dass Muster, Regel, Gesetzmässigkeiten aus den Trainingsdaten gelernt werden.

Bayes-Spamfilter

Thomas Bayes

war ein englischer Mathematiker, Statistiker, Philosoph und presbyterianischer Pfarrer. Nach ihm ist der Satz von Bayes benannt, der in der Wahrscheinlichkeitsrechnung große Bedeutung hat.

https://de.wikipedia.org/wiki/Thomas_Bayes

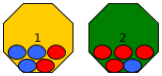
Bayes-Satz

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

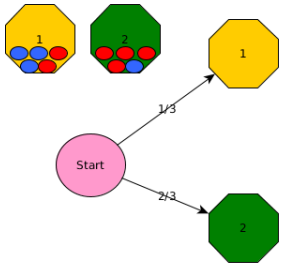
$P(A|B)$ die bedingte Wahrscheinlichkeit A unter der Bedignung das B eingetreten ist
$P(B|A)$ die bedingte Wahrscheinlichkeit B unter der Bedignung das A eingetreten ist
$P(A)$ die A-priori-Wahscheinlichkeit A $P(B)$ die B-priori-Wahscheinlichkeit A

Aufgabe: Wir ziehen 1 Ball aus einen von 2 Säcken.
Wir würfeln, bei 1 & 2 Kiste 1, beim Rest Kiste 2

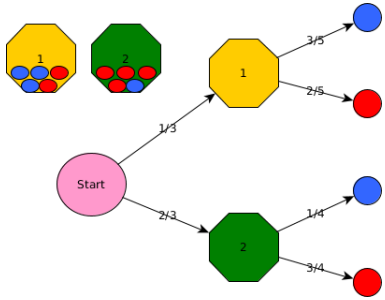
Aufgabe: Wir ziehen 1 Ball aus einen von 2 Säcken.
Wir würfeln, bei 1 & 2 Kiste 1, beim Rest Kiste 2



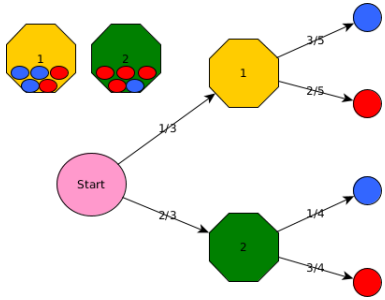
Aufgabe: Wir ziehen 1 Ball aus einen von 2 Säcken.
Wir würfeln, bei 1 & 2 Kiste 1, beim Rest Kiste 2



Aufgabe: Wir ziehen 1 Ball aus einen von 2 Säcken.
Wir würfeln, bei 1 & 2 Kiste 1, beim Rest Kiste 2



Aufgabe: Wir ziehen 1 Ball aus einen von 2 Säcken.
Wir würfeln, bei 1 & 2 Kiste 1, beim Rest Kiste 2



$$P(Kiste1|Blau) = \frac{\frac{1}{3} \cdot \frac{3}{5}}{\frac{1}{3} \cdot \frac{3}{5} + \frac{2}{3} \cdot \frac{1}{4}} = \frac{6}{11}$$

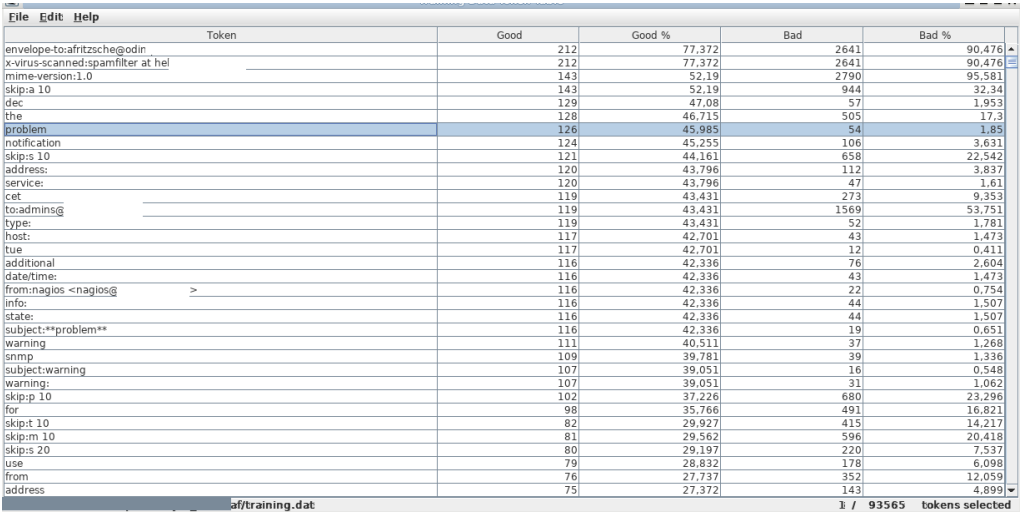
$$P(Kiste1|Blau) = \frac{6}{11}$$

Neue Aufgabe

Wir wollen wissen, ob eine Email Spam ist.

Wir haben E-Mails vorher schon eingestuft in Spam & Ham.

Alle Wörter aus dem Training sind Spam und Ham zugeordnet.



File Edit Help

Token	Good	Good %	Bad	Bad %
envelope-to:afritzsche@odin	212	77,372	2641	90,476
x-virus-scanned:spamfilter at hel	212	77,372	2641	90,476
mime-version:1.0	143	52,19	2790	95,581
skip:a 10	143	52,19	944	32,34
dec	129	47,08	57	1,953
the	128	46,715	505	17,3
problem	126	45,985	54	1,85
notification	124	45,255	106	3,631
skip:s 10	121	44,161	658	22,542
address:	120	43,796	112	3,837
service:	120	43,796	47	1,61
cet	119	43,431	273	9,353
to:admins@	119	43,431	1569	53,751
type:	119	43,431	52	1,781
host:	117	42,701	43	1,473
tue	117	42,701	12	0,411
additional	116	42,336	76	2,604
date/time:	116	42,336	43	1,473
from:nagios <nagios@ >	116	42,336	22	0,754
info:	116	42,336	44	1,507
state:	116	42,336	44	1,507
subject:**problem**	116	42,336	19	0,651
warning	111	40,511	37	1,268
snmp	109	39,781	39	1,336
subject:warning	107	39,051	16	0,548
warning:	107	39,051	31	1,062
skip:p 10	102	37,226	680	23,296
for	98	35,766	491	16,821
skip:t 10	82	29,927	415	14,217
skip:m 10	81	29,562	596	20,418
skip:s 20	80	29,197	220	7,537
use	79	28,832	178	6,098
from	76	27,737	352	12,059
address	75	27,372	143	4,899

af/training.dat 1 / 93565 tokens selected

http://bayesjunktool.mozdev.org

problem	126	45,985	54	1,85

$$P(spam) = \frac{2919}{2919 + 274} = 0,0858 \; P(ham) = \frac{274}{2919 + 274} = 0,914$$

$$P(Problem|spam) = \frac{54}{2919} = 0.0185 \; P(Problem|ham) = \frac{126}{274} = 0.45985$$

$$P(spam|Problem) = \frac{P(Problem|spam) \cdot P(spam)}{P(Problem)}$$

$$= \frac{0.185 \cdot 0.0858}{0.0858 \cdot 0.0185 + 0.914 \cdot 0.45985} = .0376 = 3,76\%$$

Email besteht aus $w_1, \ldots, w_n$ und sind unabhäning

$$P(Email|spam) = P(w_1 \cap \cdots \cap w_n|spam) = P(w_1|spam) \cdots P(w_n|spam)$$

$$Q = \frac{P(spam|Email)}{P(ham|Email)} = \frac{P(w_1|spam) \cdots P(w_n|spam)}{P(w_1|ham) \cdots P(w_n|ham)}$$

Email besteht aus $w_1, \ldots, w_n$ und sind unabhäning

$$P(Email|spam) = P(w_1 \cap \cdots \cap w_n|spam) = P(w_1|spam) \cdots P(w_n|spam)$$

$$Q = \frac{P(spam|Email)}{P(ham|Email)} = \frac{P(w_1|spam) \cdots P(w_n|spam)}{P(w_1|ham) \cdots P(w_n|ham)}$$

Token	Good	Good %	Bad	Bad %
mime-version:1.0	143	52,19	2790	95,581
content-type/type:multipart/alternative	55	20,073	1733	59,37
.	46	16,788	1275	43,679
com	34	12,409	1263	43,268
charset:utf-8	24	8,759	1222	41,864

Sind die Wörter unabhängig?

Email besteht aus $w_1, \ldots, w_n$ und sind unabhäning

$$P(Email|spam) = P(w_1 \cap \cdots \cap w_n|spam) = P(w_1|spam) \cdots P(w_n|spam)$$

$$Q = \frac{P(spam|Email)}{P(ham|Email)} = \frac{P(w_1|spam) \cdots P(w_n|spam)}{P(w_1|ham) \cdots P(w_n|ham)}$$

Token	Good	Good %	Bad	Bad %
mime-version:1.0	143	52,19	2790	95,581
content-type/type:multipart/alternative	55	20,073	1733	59,37
.	46	16,788	1275	43,679
com	34	12,409	1263	43,268
charset:utf-8	24	8,759	1222	41,864

Sind die Wörter unabhängig?

Eher nein, funktioniert aber denoch sehr gut.

$P(spam|Email) > P(ham|Email)$ oder $P(spam|Email) < P(ham|Email)$ für Entscheidung

Email besteht aus $w_1, \ldots, w_n$ und sind unabhäning

$$P(Email|spam) = P(w_1 \cap \cdots \cap w_n|spam) = P(w_1|spam) \cdots P(w_n|spam)$$

$$Q = \frac{P(spam|Email)}{P(ham|Email)} = \frac{P(w_1|spam) \cdots P(w_n|spam)}{P(w_1|ham) \cdots P(w_n|ham)}$$

Token	Good	Good %	Bad	Bad %
mime-version:1.0	143	52,19	2790	95,581
content-type/type:multipart/alternative	55	20,073	1733	59,37
.	46	16,788	1275	43,679
com	34	12,409	1263	43,268
charset:utf-8	24	8,759	1222	41,864

Sind die Wörter unabhängig?

Eher nein, funktioniert aber denoch sehr gut.

$P(spam|Email) > P(ham|Email)$ oder $P(spam|Email) < P(ham|Email)$ für Entscheidung

Praxis: $w_1 \ldots w_{10}$

Email besteht aus $w_1, \ldots, w_n$ und sind unabhäning

$$P(Email|spam) = P(w_1 \cap \cdots \cap w_n|spam) = P(w_1|spam) \cdots P(w_n|spam)$$

$$Q = \frac{P(spam|Email)}{P(ham|Email)} = \frac{P(w_1|spam) \cdots P(w_n|spam)}{P(w_1|ham) \cdots P(w_n|ham)}$$

Token	Good	Good %	Bad	Bad %
mime-version:1.0	143	52,19	2790	95,581
content-type/type:multipart/alternative	55	20,073	1733	59,37
.	46	16,788	1275	43,679
com	34	12,409	1263	43,268
charset:utf-8	24	8,759	1222	41,864

Sind die Wörter unabhängig?

Eher nein, funktioniert aber denoch sehr gut.

$P(spam|Email) > P(ham|Email)$ oder $P(spam|Email) < P(ham|Email)$ für Entscheidung

Praxis: $w_1 \ldots w_{10}$

$w_{spam_1} \ldots w_{spam_5}$ $w_{ham_1} \ldots w_{ham_5}$

mit größter Wahrscheinlichkeit in Ham bzw. Spam

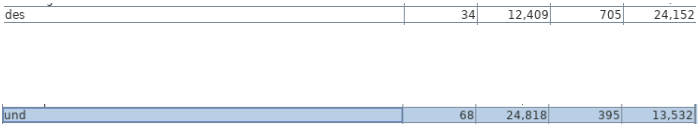
Sehr kurze und inhaltsleere Funktionswörter sind am häufigsten.

Top 20 Deutscher Wortschatz

der, die, und, in, den, von, zu, das, mit, sich, des, auf, für, ist, im, dem, nicht, ein, Die, eine

Top 20 SAP Wortschatz

die, Sie, der, und, in, werden, den, für, das, im, können, wird, zu, eine, auf, des, %N%, Die, ist, mit



Markow

Basis Hidden Markov Mode von Andrei Andrejewitsch Markow
Idee ähnliche Wortketten

Markow

Basis Hidden Markov Mode von Andrei Andrejewitsch Markow
Idee ähnliche Wortketten

Bayes

$$P = 0,5 + \frac{P_{gut} - P_{schlecht}}{P_{gut} + P_{schlecht} + 1}$$

Markow

Wichtung 2^{2N}

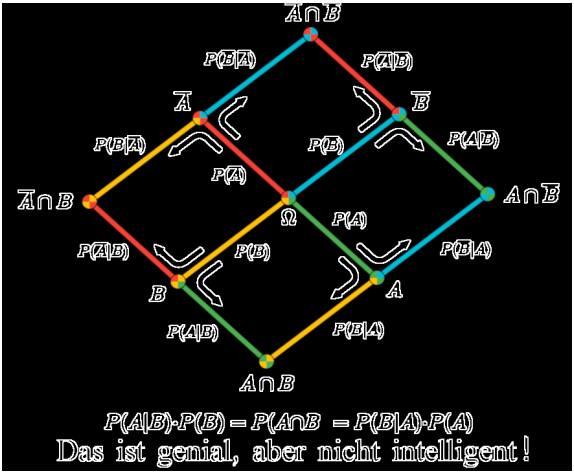
N = Anzahl der übereinstimmenden Wörter -1

$$P = 0,5 + \frac{(P_{gut} - P_{schlecht}) \cdot Wichtung}{(P_{gut} + P_{schlecht} + 1) \cdot Wichtung_{max}}$$

Wenn diese Nachricht nicht korrekt

Kette	W	N	A
wenn	1	0	1
wenn Nachricht	4	1	2
wenn diese ... nicht	16	2	3
wenn diese ... nicht korrekt	64	3	4

Danke


$\bar{A} \cap B$
$P(B|\bar{A})$
$P(\bar{A}|B)$
$\bar{A}$
$\bar{B}$
$P(\bar{B}|\bar{A})$
$P(\bar{A}|\bar{B})$
$P(\bar{A})$
$P(B)$
$P(A|\bar{B})$
$\bar{A} \cap B$
Ω
$A \cap \bar{B}$
$P(A)$
$P(\bar{B})$
$P(\bar{B}|A)$
B
A
$P(A|B)$
$P(B|A)$
$A \cap B$
$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$
Das ist genial, aber nicht intelligent !