Universität Stuttgart
Studiengang Simulation Technology

SimTech
Cluster of Excellence

Bayes'sches Spamfilter

Referenten:

- Daniel Fink
- Marcel Messer

Betreuer:

- Michael Sinsbeck

Einführung

E-Mail

- 1971 wurde erste E-Mail versandt
  - Ray Tomlinson gilt als Erfinder der E-Mail
- Werbung entdeckte die E-Mail
- Statistik besagt heute 70% Spam und 30% Ham
- Spamfilter bietet Lösung

Einführung

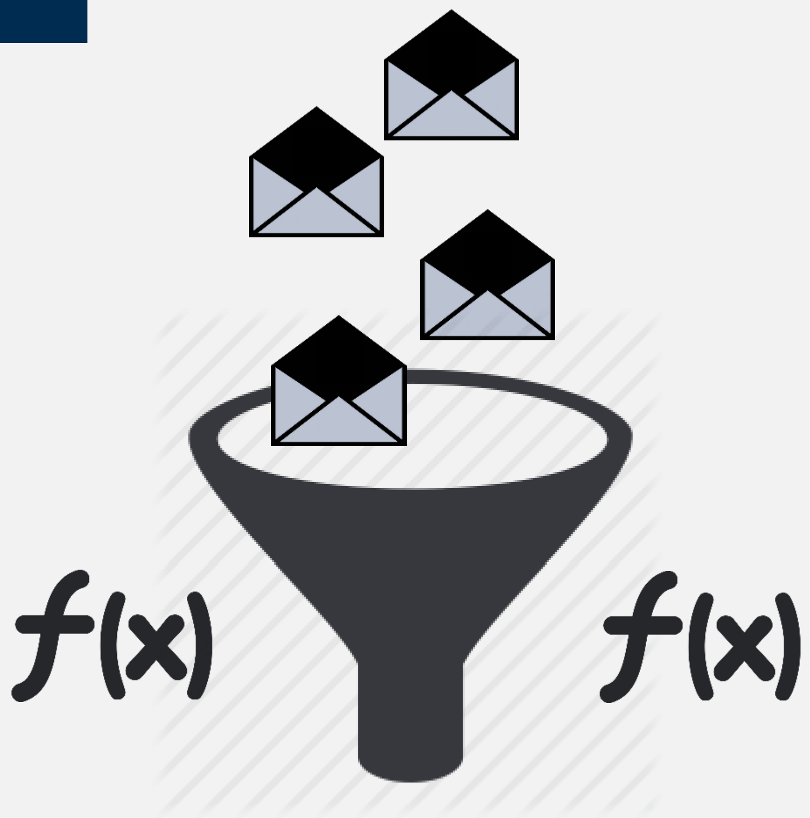
Anforderungen an den Spamfilter

- Effizienz
- Schnelligkeit
- Zuverlässigkeit
- Anpassungsfähigkeit

Interaktiver Spamfilter

Übersicht

- E-Mails schreiben
- Mathematische Grundlagen
- Bayes‘sches Spamfilter
- Erweiterungen



Interaktiver Spamfilter

E-Mail schreiben

- Jeder hat jetzt die Chance eine E-Mail zu schreiben
- Gemäß den Vorgaben auf der Rückseite des Handouts
  - future-campus.de oder QR-Code

Mathematische Grundlagen

Mathematische Grundlagen

Definition 2.1 (Endlicher Wahrscheinlichkeitsraum)

Ein endlicher **Wahrscheinlichkeitsraum** ist ein Paar (Ω, P) mit einer endlichen nichtleeren Menge Ω (dem **Grundraum**) und einer Abbildung $P: \mathcal{P}(\Omega) \to \mathbb{R}$ (der **Wahrscheinlichkeitsverteilung**) mit folgenden Eigenschaften:

- Nichtnegativität $\forall A \subset \Omega : \; P(A) \geq 0$
- Normiertheit $P(\Omega) = 1$
- Additivität $\forall A, B \subset \Omega, A \cap B = \emptyset : \; P(A + B) = P(A) + P(B)$

Mathematische Grundlagen

Definition 2.2 (Bedingte Wahrscheinlichkeit)

Sei (Ω, P) ein endlicher **Wahrscheinlichkeitsraum** und $A, B \subset \Omega$ **Ereignisse** mit $P(B) > 0$. Dann heißt

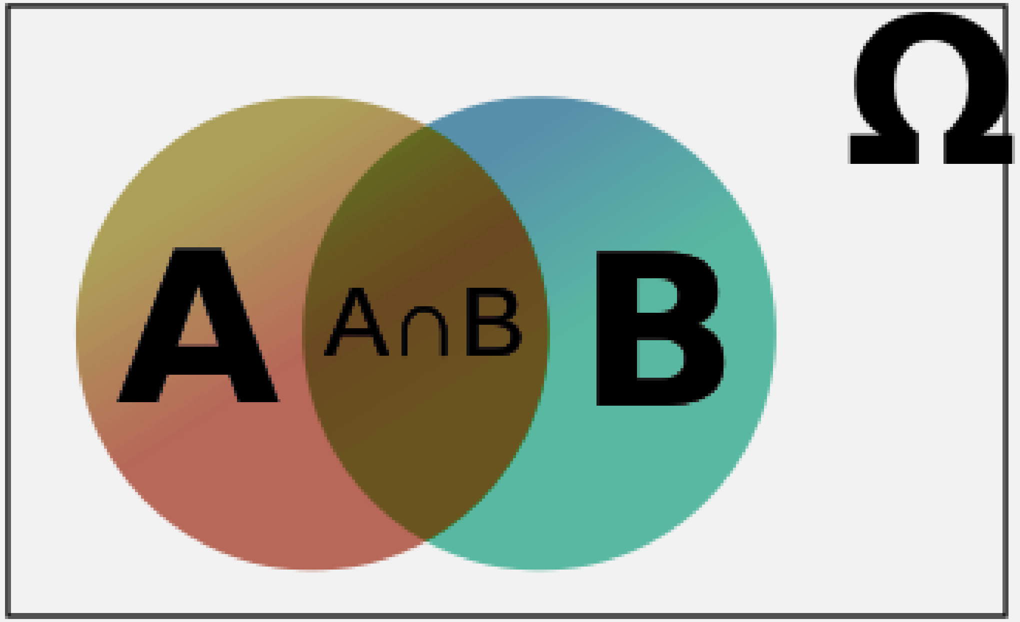
$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

bedingte Wahrscheinlichkeit von A unter der Bedingung B.

Mathematische Grundlagen

Das Flächenverhältnis illustriert

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



Mathematische Grundlagen

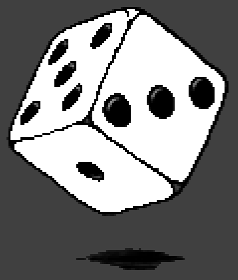
Definition 2.3 (Stochastische Unabhängigkeit)

In einem endlichen **Wahrscheinlichkeitsraum** (Ω, P) heißen zwei **Ereignisse** $A, B \in \Omega$ genau dann **stochastisch unabhängig**, wenn

$$P(A \cap B) = P(A)P(B)$$

Ein Beispiel zur stochastischen Unabhängigkeit

Ein Würfel wird einmal geworfen...



Mathematische Grundlagen

$$\left.\begin{array}{ll} \boldsymbol{A} \coloneqq \text{„Gerade Augenzahl“} & = \{\boldsymbol{2}, \boldsymbol{4}, \boldsymbol{6}\} \\ \boldsymbol{B} \coloneqq \text{„Augenzahl durch 3 teilbar“} & = \{\boldsymbol{3}, \boldsymbol{6}\} \end{array}\right\} \quad \boldsymbol{A} \cap \boldsymbol{B} = \{\boldsymbol{6}\}$$

$$\left.\begin{array}{l} P(A) = \frac{3}{6} = \frac{1}{2} \\[1em] P(B) = \frac{2}{6} = \frac{1}{3} \\[1em] P(A \cap B) = \frac{1}{6} \end{array}\right] \quad {\color{red}P(A)P(B)} = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} = {\color{red}P(A \cap B)}$$

Mathematische Grundlagen

Lemma 2.4

Sei (Ω, P) ein endlicher **Wahrscheinlichkeitsraum** und $A, B, C \in \Omega$ drei **Ereignisse** mit A und B stochastisch unabhängig. Dann gilt

$$P(A \cap B|C) = P(A|C)P(B|C)$$

Mathematische Grundlagen

Satz 2.5 (Satz von der totalen Warscheinlichkeit)

Sei (Ω, P) ein endlicher **Wahrscheinlichkeitsraum** und $B_1, B_2, \dots, B_n \subset \Omega$ paarweise **disjunkte Ereignisse**, die eine **Partition** von Ω bilden, also

$$\Omega = \bigcup_{i=1}^{n} B_i$$

Sei weiter $P(B_i) > 0$ für alle $1 \leq i \leq n$. Dann gilt für jedes Ereignis $\mathrm{A} \subset \Omega$

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Mathematische Grundlagen

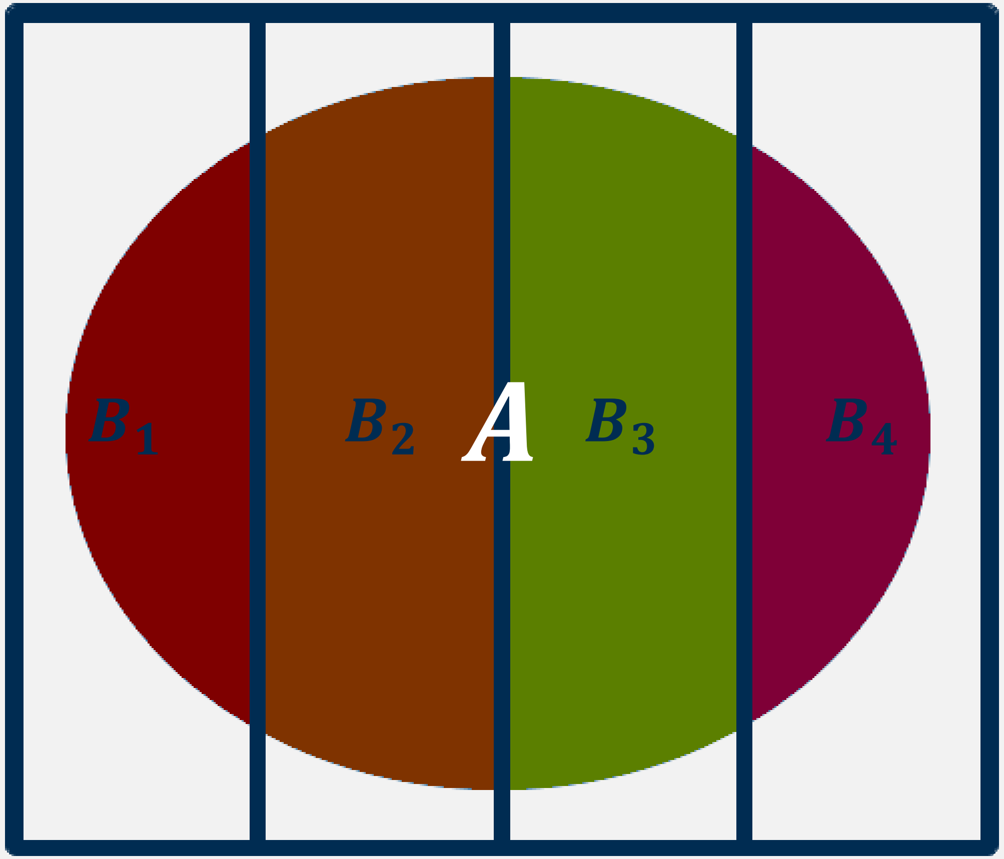
Illustration zur Berechnung der totalen Wahrscheinlichkeit

$$A = (A \cap B_1) \cup (A \cap B_1) \cup (A \cap B_1) \cup (A \cap B_1)$$

$$P(A) = P(A \cap B_1) + P(A \cap B_1) + P(A \cap B_1) + P(A \cap B_1)$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) + P(A|B_4)P(B_4)$$

$$P(A) = \sum_{i=1}^{4} P(A|B_i)P(B_i)$$



Mathematische Grundlagen

Satz 2.6 (Satz von Bayes)

Sei (Ω, P) ein endlicher **Wahrscheinlichkeitsraum** und $A, B \subset \Omega$ **Ereignisse** mit $P(B) > 0$. Dann gilt

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

Mathematische Grundlagen

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

„Umdrehen“ der bedingten Wahrscheinlichkeit

benötigt die Wahrscheinlichkeiten von A, B

Ein Beispiel zum Satz von Bayes

„Eine Schülerin fährt in **70%** der Schultage mit dem **Bus**. In **80%** dieser Fälle kommt sie **pünktlich** zur Schule. **Durchschnittlich** kommt sie aber nur an **60%** der Schultage **pünktlich** an.“

Heute kommt die Schülerin pünktlich zur Schule.
Mit welcher Wahrscheinlichkeit hat sie den Bus benutzt?

Mathematische Grundlagen

$\mathbf{B}$ ≔ „Die Schülerin fährt mit dem Bus“

$\boldsymbol{P}$ ≔ „Die Schülerin kommt pünktlich an“

„Eine Schülerin fährt in 70% der Schultage mit dem Bus“ $\Longrightarrow \boldsymbol{P(B) = 70\%}$

„In 80% dieser Fälle kommt sie pünktlich zur Schule“ $\Longrightarrow \boldsymbol{P(P|B) = 80\%}$

„ Durchschnittlich kommt sie aber nur an 60% der Schultage pünktlich an“ $\Longrightarrow \boldsymbol{P(P) = 60\%}$

Mathematische Grundlagen

$$P(B) = 70\%$$
$$P(P|B) = 80\%$$
$$P(P) = 60\%$$

$\mathbf{B}$ ≔ „Die Schülerin fährt mit dem Bus“

$\boldsymbol{P}$ ≔ „Die Schülerin kommt pünktlich an“

Gesucht ist also die Wahrscheinlichkeit für „Bus“ unter der Bedingung „Pünktlichkeit“ …

… oder anders ausgedrückt, gesucht ist $\boldsymbol{P(B|P)}$.

Mathematische Grundlagen

$$P(B) = 70\%$$
$$P(P|B) = 80\%$$
$$P(P) = 60\%$$

$\mathbf{B}$:= „Die Schülerin fährt mit dem Bus“

P := „Die Schülerin kommt pünktlich an“

Der Satz von Bayes liefert uns

$$P(B|P) = \frac{P(P|B)P(B)}{P(P)} = \frac{0{,}8 \cdot 0{,}7}{0{,}6} \approx 93\%$$

Die Problemstellung

Wie kann man möglichst effizient gewünschte E-Mails (Ham) von unerwünschten E-Mails (Spam) trennen?

Das Bayes‘sche Spamfilter

Bayes‘sches Spamfilter

- Einzelne charakteristische Wörter (Ereignis) → Spam-Wahrscheinlichkeit
  - Gegeben: $\boldsymbol{P(wort|S)}$ und $\boldsymbol{P(wort|H)}$
  - Gesucht: $\boldsymbol{P(S|wort)}$
- „Umdrehen“ der Wahrscheinlichkeiten → Satz von Bayes

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

Ein Beispiel

Durch Auszählen von 100 Ham und 100 Spam E-Mails haben wir für die Wörter „haben“, „online“ und „ich“ die folgenden Zahlen erhalten:

	Ham	Spam
„haben“	30	7
„online“	3	8
„ich“	25	27

Bayes‘sches Spamfilter

Je 100 E-Mails	Ham	Spam
„haben“	30	7
„online“	3	8
„ich“	25	27

Wir erhalten also die folgenden Wahrscheinlichkeiten

$$\boldsymbol{P(haben|H) = 30\%} \qquad \boldsymbol{P(haben|S) = 7\%}$$

$$\boldsymbol{P(online|H) = 3\%} \qquad \boldsymbol{P(online|S) = 8\%}$$

$$\boldsymbol{P(ich|H) = 25\%} \qquad \boldsymbol{P(ich|S) = 27\%}$$

Datenbank

Bayes‘sches Spamfilter

Werden neue E-Mails vom Benutzer als Ham oder Spam markiert → Update in der Datenbank

Je 100 E-Mails	Ham	Spam
„haben“	30	7
„online“	3	8
„ich“	25	27

Bayes‘sches Spamfilter

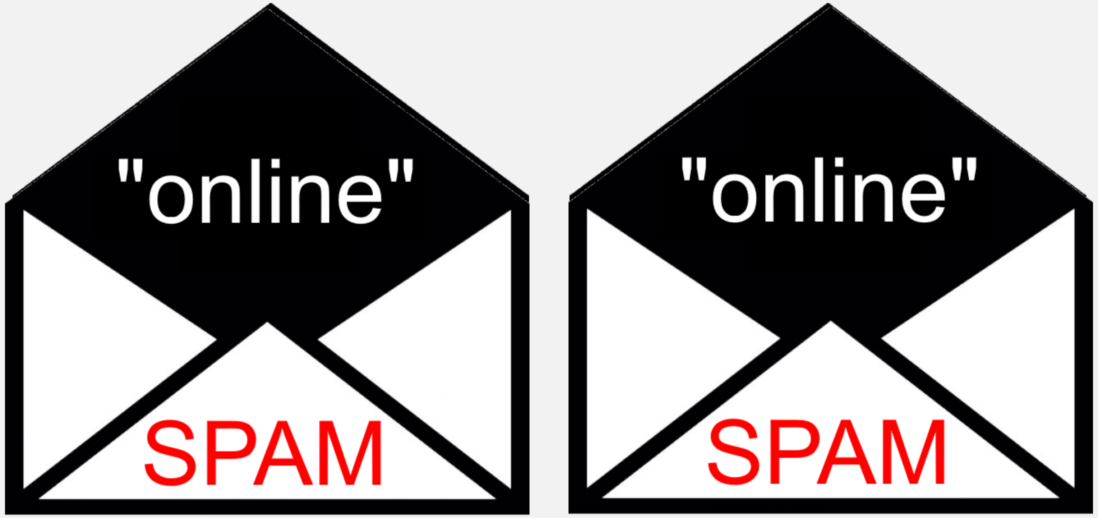
Je 100 E-Mails	Ham	Spam
„haben“	33	7
„online“	3	8
„ich“	25	27

Werden neue E-Mails vom Benutzer als Ham oder Spam markiert → Update in der Datenbank

Bayes‘sches Spamfilter

Werden neue E-Mails vom Benutzer als Ham oder Spam markiert → Update in der Datenbank

Je 100 E-Mails	Ham	Spam
„haben“	33	7
„online“	3	10
„ich“	25	27



Bayes‘sches Spamfilter

Je 100 E-Mails	Ham	Spam
„haben“	33	7
„online“	3	10
„ich“	25	27

Werden neue E-Mails vom Benutzer als Ham oder Spam markiert → Update in der Datenbank

→ Lernprozess

Bayes‘sches Spamfilter

- Einzelne charakteristische Wörter (Ereignis) → Spam-Wahrscheinlichkeit
  - Gegeben: $P(wort|S)$ und $P(wort|H)$ ✓
  - Gesucht: $P(S|wort)$

Bayes‘sches Spamfilter

Lemma 3.1 (Spam-Wahrscheinlichkeit für ein Ereignis)

Sei Ω die Menge aller Wörter (Ereignisse) einer E-Mail und $wort \subset \Omega$. Die **Spam-Wahrscheinlichkeit** ist dann gegeben durch

$$P(S|wort) = \frac{P(wort|S)}{P(S)P(wort|S) + P(H)P(wort|H)} P(S)$$

Bayes‘sches Spamfilter

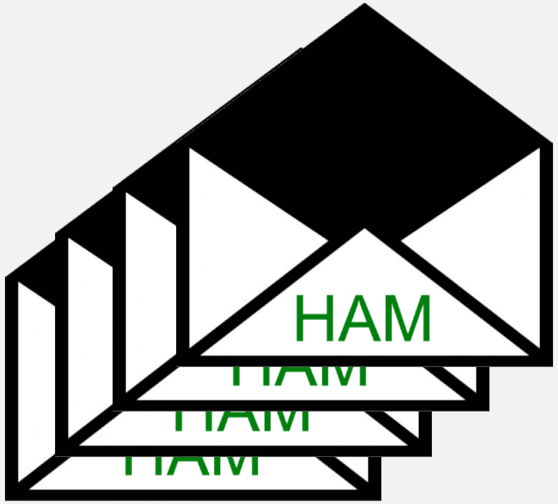
Lemma 3.1 (Spam-Wahrscheinlichkeit für ein Ereignis)

Sei Ω die Menge aller Wörter (Ereignisse) einer E-Mail und $wort \subset \Omega$. Die **Spam-Wahrscheinlichkeit** ist dann gegeben durch

$$P(S|wort) = \frac{P(wort|S)}{P(S)P(wort|S) + P(H)P(wort|H)} P(S)$$

Bemerkung: Für die Wahrscheinlichkeiten $P(S)$ und $P(H)$ müssen geeignete Annahmen getroffen werden.

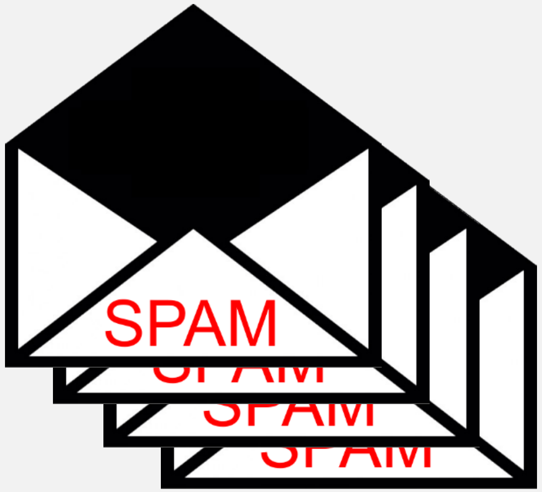
Bayes‘sches Spamfilter



Eine mögliche Annahme wäre

$$P(S) = P(H) = \mathbf{50\%}$$

Gerechtfertigt?



Bayes‘sches Spamfilter

Annahme 3.2

Die betrachteten Worte treten voneinander **stochastisch unabhängig** auf, d.h.

$$P(wort_1 \cap wort_2|S) = P(wort_1|S)P(wort_2|S)$$

Bayes'sches Spamfilter

Satz 3.3 (Spam-Wahrscheinlichkeit für zwei Ereignisse)

Sei Ω die Menge aller Wörter (Ereignisse) einer E-Mail. Für zwei Ereignisse $wort_1, wort_2 \subset \Omega$ ist die **Spam-Wahrscheinlichkeit** dann gegeben durch

$$P(S|wort_1 \cap wort_2) = \frac{P(wort_1|S)P(wort_2|S)}{P(S)P(wort_1|S)P(wort_2|S) + P(H)P(wort_2|H)P(wort_2|H)} P(S)$$

Bayes‘sches Spamfilter

Theorem 3.4 (Spam-Wahrscheinlichkeit für endlich viele Ereignisse)

Sei Ω die Menge aller Wörter (Ereignisse) einer E-Mail. Für jede endliche Teilmenge von Wörtern $\{wort_i\}_{i=1}^{n}$ ist die **Spam-Wahrscheinlichkeit** gegeben durch

$$P\left(S \middle| \bigcap_{i=1}^{n} wort_i\right) = \frac{\prod_{i=1}^{n} P(wort_i|S)}{P(S)\prod_{i=1}^{n} P(wort_i|S) + P(H)\prod_{i=1}^{n} P(wort_i|H)} P(S)$$

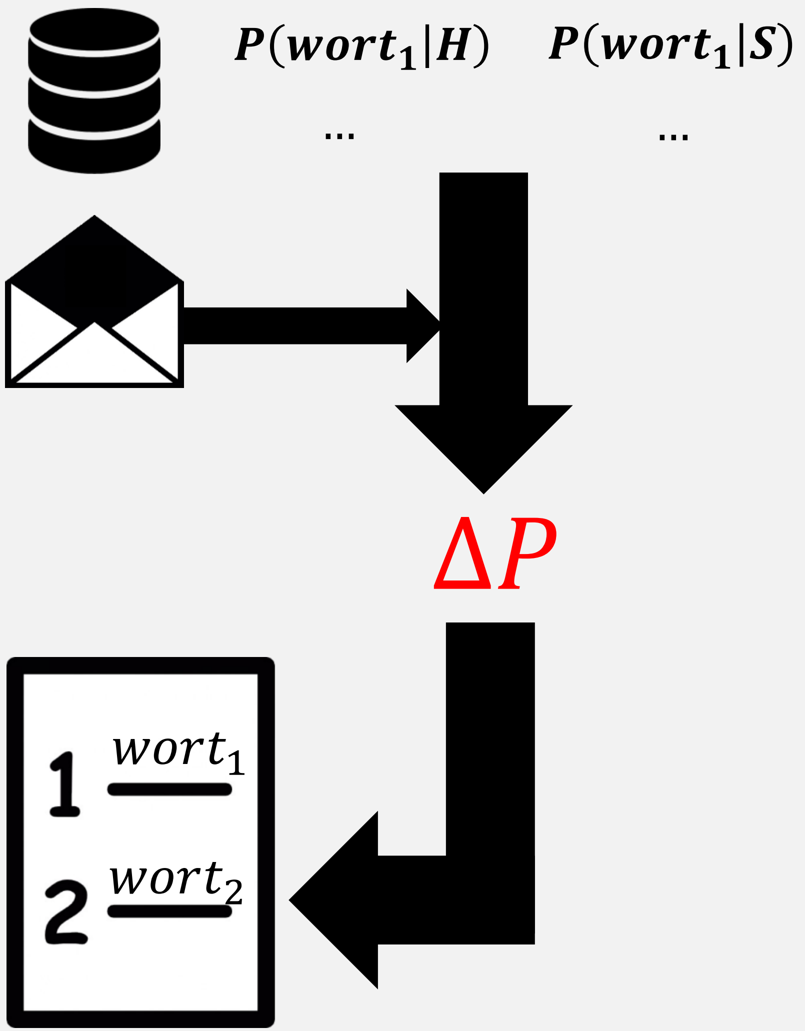
Bayes‘sches Spamfilter

- Einzelne charakteristische Wörter (Ereignis) → Spam-Wahrscheinlichkeit
  - Gegeben: $\boldsymbol{P(wort|S)}$ und $\boldsymbol{P(wort|H)}$ ✓
  - Gesucht: $\boldsymbol{P(S|wort)}$ ✓

Bayes‘sches Spamfilter

Signifikante Wörter

- Datenbank mit bedingter Wahrscheinlichkeit
- Analysieren der Wörter aus E-Mail
  - Abgleichen dieser mit Datenbank
- Berechnung der Beträge
  - Abspeichern in separater Liste
- Auswahl der ersten n Wörter



Unser Beispiel

Durch Auszählen von 100 Ham und 100 Spam E-Mails haben wir für die Wörter „haben“, „online“ und „ich“ die folgenden Zahlen erhalten:

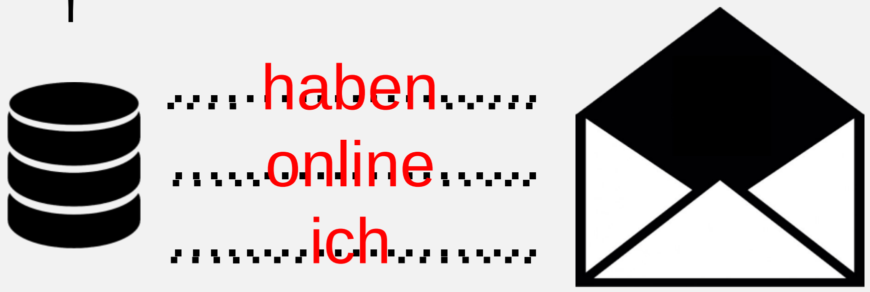
	Ham	Spam
„haben“	30	7
„online“	3	8
„ich“	25	27

Bayes‘sches Spamfilter

Wir haben die folgenden Wahrscheinlichkeiten

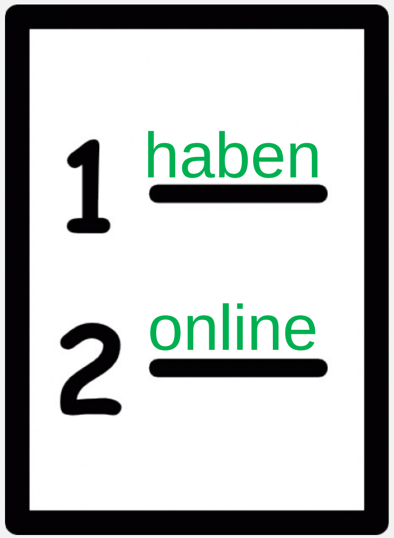
$$P(haben|H) = 30\% \qquad P(haben|S) = 7\%$$
$$P(online|H) = 3\% \qquad P(online|S) = 8\%$$
$$P(ich|H) = 25\% \qquad P(ich|S) = 27\%$$

... ...



$$\Delta P(haben) = 23$$
$$\Delta P(online) = 5$$
$$\Delta P(ich) = 2$$

Signifikante Wörter



Bayes‘sches Spamfilter

Wir betrachten also nur noch die folgenden Wahrscheinlichkeiten

$\boldsymbol{P(haben|H) = 30\%}$ $\boldsymbol{P(haben|S) = 7\%}$

$\boldsymbol{P(online|H) = 3\%}$ $\boldsymbol{P(online|S) = 8\%}$

und erhalten zusammen mit der Annahme $\boldsymbol{P(H) = P(S) = 50\%}$

$$P(S|haben \cap online) = \frac{P(haben|S)P(online|S)}{P(S)P(haben|S)P(online|S) + P(H)P(haben|H)P(online|H)}P(S)$$

$$= \frac{0{,}07 \cdot 0{,}08}{0{,}5 \cdot 0{,}07 \cdot 0{,}08 + 0{,}5 \cdot 0{,}30 \cdot 0{,}03} \cdot 0{,}5 \approx 38\%$$

Bayes‘sches Spamfilter

Signifikante Wörter

- Nur wichtige Wörter betrachten

Konsequenz

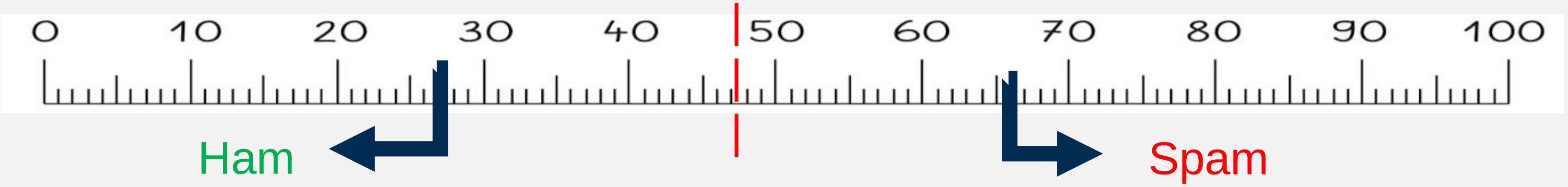
- Wahrscheinlichkeit dennoch ausschlagkräftig

Bayes'sches Spamfilter

Klassifizierungsgrenze

- Ab wie viel Prozent E-Mail als Spam erkannt wird

$P(S|\cap_{i=1}^{n} wort_i)$ **Klassifzierungsgrenze**



Konsequenz

- Je höher Klassifizierung desto früher wird aussortiert

Bayes‘sches Spamfilter

P(S)

- Grundannahme
- $P(H) = 1 - P(S)$
- $P(S|\bigcap_{i=1}^{n} wort_i) = \dfrac{\prod_{i=1}^{n}\left(P(wort_i|S)\right) P(S)}{P(S)\prod_{i=1}^{n} P(wort_i|S) + (1-P(S))\prod_{i=1}^{n} P(wort_i|H)}$

Konsequenz

- Wahrscheinlichkeit Spam erhöht sich, wenn P(S) größer wird.

Bayes‘sches Spamfilter

Problemstellung

- Was passiert mit E-Mails in, denen keine ausschlaggebenden signifikanten Wörter vorkommen?

Lösung

- Man lässt das Programm mit dieser E-Mail lernen!

→ Spamfilter wird dadurch auf den Benutzer trainiert

Bayes‘sches Spamfilter

Lernprozess mit entsprechender E-Mail

- Bedingte Wahrscheinlichkeit werden Aufgrund der E-Mail neu berechnet oder angelegt

Konsequenz

- Spamfilter erkennt nun Wörter und berechnet Wahrscheinlichkeit

→ E-Mail kann nun korrekt klassifiziert werden

Erweiterungen

Erweiterungen

Erkenntnis

- Parameter nicht selektiv genug!

Lösung

- Anhänge mitbetrachten und nach Dateiendung deklarieren

Erweiterungen

Anhänge

- Deklarieren der Anhänge nach Dateiendung
- Parameter P(S)
- Parameter Klassifizierung

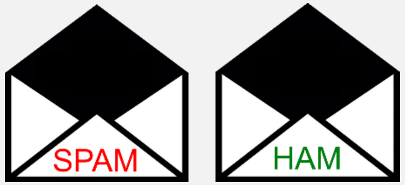
$P(S|\cap_{i=1}^{n} format_i)$

< Klassifizierung

$P(S|\cap_{i=1}^{n} wort_i)$

> Klassifizierung

P(E-Mail)



Konsequenz

- Gefährliche Inhalte werden vorab über Anhang aussortiert.

Erweiterungen

Zweite E-Mail schreiben

- Ziel: Spamfilter mit einer Spam E-Mail zu umgehen
  - future-campus.de oder QR-Code

Analyse der Spam E-Mails

Quellen

- Illustration zur bedingten Wahrscheinlichkeit
  - https://de.wikipedia.org/wiki/Bedingte_Wahrscheinlichkeit
- Beispiel zum Satz von Bayes
  - http://www.mathebibel.de/satz-von-bayes
- Theorie des Satzes von Bayes
  - https://de.wikipedia.org/wiki/Bayessches_Filter
- Beispiel zur stochastischen Unabhängigkeit
  - https://de.serlo.org/mathe/stochastik/uebersicht-aller-artikel-zur-stochastik/unabhaengigkeit-von-ereignissen

Quellen

- Artikel zum Bayes'schen Spamfilter
  - http://www.math.kit.edu/ianm4/~ritterbusch/seite/spam/de
- Wikipedia zum Bayes'schen Spamfilter
  - https://de.wikipedia.org/wiki/Bayessches_Filter
- Skript zur Vorlesung „Numerische und Stochastische Grundlagen"
  - Zu finden unter Ilias, Dirk Pflüger, Stefan Zimmer, 8 Februar 2017