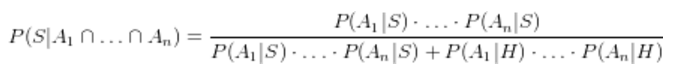
Bayes-Spam-Filter

# Idee & Funktionsweise

1. Eine Menge von Spam-Emails analysieren und für jedes Wort Buchhaltung führen, wie oft es vorkommt.
2. Das gleiche mit Ham-Emails (also nicht-Spam-Mails) machen.
3. Nun kann für jedes Wort die bedingte Wahrscheinlichkeit berechnet werden, mit der es sich um Spam oder Ham handelt.
   1. Beispiel: Das Wort «money» kommt in 30 von 100 Spam Mails vor, also ist die bedingte Wahrscheinlichkeit P(money | Spam) = 0.3
   2. Das gleiche Wort kommt nur in 25 von 100 Ham-Mails vor, also ist die bedingte Wahrscheinlichkeit P(money | Ham) = 0.15
4. Jetzt wird eine beliebige Email eingelesen und für jedes Wort die oben beschriebene Wahrscheinlichkeit berechnet.
5. Danach werden die einzelnen Wahrscheinlichkeiten in die nachfolgend beschriebene Formel eingefügt und es resultiert die Wahrscheinlichkeit P(Spam | Wort A kommt vor, Wort B kommt vor, Wort C kommt vor, ...), also die Wahrscheinlichkeit, ob es sich bei der Email um Spam handelt.

# Verwendete Formel



Herleitung siehe: http://www.math.kit.edu/ianm4/~ritterbusch/seite/spam/de

# Vorteile

* Die Formel und damit auch die Implementierung ist sehr simpel.
* Es ist nur einmal zu Beginn ein grosser Analyseaufwand erforderlich. Die gelernten Daten können danach beliebig oft für Berechnungen verwendet oder auch erweitert werden.

# Nachteile

* Bei der Herleitung der Formel wird angenommen, dass Wörter in Spam und Ham stochastisch unabhängig voneinander auftreten, was grundsätzlich nicht der Fall ist (z.B. «Freundliche» und «Grüsse» treten sehr häufig miteinander auf, sind also nicht unabhängig).
* Je grösser die Email, desto mehr Wahrscheinlichkeiten werden mitenander multipliziert 🡪 enorm kleine Zahlen entstehen im Nenner und Zähler der Formel 🡪 Gefahr für Underflow.

# Grobstruktur des Programms

