Spam-Filterung mit Naive-Bayes-Klassifikator

Team 7: Jannik Weisser & Leon Jerke

Beschreibung des Problems

Im digitalen Zeitalter ist die Flut an E-Mails enorm, und nicht alle davon sind erwünscht oder sogar sicher. Spam-E-Mails, die oft unerwünschte Werbung oder sogar schädliche Inhalte enthalten, sind ein signifikantes Problem für Nutzer und E-Mail-Dienstanbieter. Um diese unerwünschten E-Mails effektiv zu filtern und die Nutzererfahrung zu verbessern, setzen viele E-Mail-Dienste auf maschinelles Lernen, insbesondere auf Naive-Bayes-Klassifikatoren. Diese Klassifikatoren nutzen bedingte Wahrscheinlichkeiten, um zu bestimmen, ob eine eingehende E-Mail Spam ist oder nicht, basierend auf dem Vorkommen bestimmter Schlüsselwörter in der Nachricht.

Formalisierung des Problems

Der Naive-Bayes-Klassifikator basiert auf dem Bayes'schen Theorem, welches bedingte Wahrscheinlichkeiten verwendet, um die Wahrscheinlichkeit einer Hypothese (hier: eine E-Mail ist Spam) auf Basis von vorliegenden Beweisen (hier: Wörter in der E-Mail) zu aktualisieren.

Gegebene Wahrscheinlichkeiten:

1. **Prior-Wahrscheinlichkeit** $P(Spam)$: Die Wahrscheinlichkeit, dass eine beliebige eingehende E-Mail Spam ist, bevor der Inhalt betrachtet wird.
2. **Wahrscheinlichkeiten für das Auftreten bestimmter Wörter in Spam- und Nicht-Spam-E-Mails**:

   $P(Wort|Spam)$: Wahrscheinlichkeit, dass ein bestimmtes Wort in einer Spam-E-Mail erscheint.

   $P(Wort|\neg Spam)$: Wahrscheinlichkeit, dass dasselbe Wort in einer Nicht-Spam-E-Mail erscheint.

Berechnung nach Bayes'schem Theorem:

Das Bayes'sche Theorem wird genutzt, um die Posterior-Wahrscheinlichkeit $P(Spam|Wörter)$ zu berechnen, die die Wahrscheinlichkeit angibt, dass eine E-Mail Spam ist, gegeben die Wörter, die sie enthält:

$$P(Spam|Wörter) = \frac{P(Wörter|Spam) \times P(Spam)}{P(Wörter)}$$

Hierbei ist $P(Wörter|Spam)$ das Produkt der Wahrscheinlichkeiten jedes Wortes, gegeben dass die E-Mail Spam ist, unter der Annahme der Unabhängigkeit der Wörter (daher "naive").

Marginal Likelihood $P(Wörter)$:

Die Gesamtwahrscheinlichkeit $P(Wörter)$, dass die Wörter unabhängig von der Klassifikation auftreten, wird wie folgt berechnet:

$$P(Wörter) = P(Wörter|Spam) \times P(Spam) + P(Wörter|\neg Spam) \times P(\neg Spam)$$

Diese Berechnungen ermöglichen es, jede eingehende E-Mail auf der Basis ihres Inhalts effektiv zu klassifizieren und tragen dazu bei, das Spam-Problem zu minimieren.

Beispiel 1:

Gegebene Daten (fiktiv):

- **Prior-Wahrscheinlichkeit, dass eine zufällige E-Mail Spam ist** $P(Spam)$: 20%
- **Prior-Wahrscheinlichkeit, dass eine zufällige E-Mail kein Spam ist** $P(\neg Spam)$: 80%

Wir nehmen an, dass bestimmte Schlüsselwörter wie "kostenlos", "Gewinn" und "Sonderangebot" Indikatoren für Spam sein könnten. Wir verwenden folgende fiktive Wahrscheinlichkeiten:

- **Wahrscheinlichkeit des Wortes "kostenlos" in Spam-E-Mails** $P(kostenlos|Spam)$: 30%
- **Wahrscheinlichkeit des Wortes "kostenlos" in Nicht-Spam-E-Mails** $P(kostenlos|\neg Spam)$: 5%

Ziel:

Berechnung der Wahrscheinlichkeit, dass eine E-Mail Spam ist, gegeben das Wort "kostenlos" erscheint.

Formeln:

1. **Bayessches Theorem**:

   $$P(Spam|kostenlos) = \frac{P(kostenlos|Spam) \times P(Spam)}{P(kostenlos)}$$

2. **Marginal Likelihood** $P(kostenlos)$:

   $$P(kostenlos) = P(kostenlos|Spam) \times P(Spam) + P(kostenlos|\neg Spam) \times P(\neg Spam)$$

Berechnungen:

1. Zuerst berechnen wir $P(kostenlos)$:

   $$P(kostenlos) = 0.3 \times 0.2 + 0.05 \times 0.8 = 0.1$$

2. Dann nutzen wir das Bayessche Theorem, um $P(Spam|kostenlos)$ zu berechnen:

   $P(Spam|kostenlos) = \frac{0.3 \times 0.2}{0.1} = 0.6$

Basierend auf den Berechnungen haben wir die folgenden Ergebnisse:

- Die **Gesamtwahrscheinlichkeit, dass das Wort "kostenlos" in einer E-Mail erscheint** ($P(kostenlos)$), beträgt 10%.
- Die **Wahrscheinlichkeit, dass eine E-Mail Spam ist, gegeben dass das Wort "kostenlos" erscheint** ($P(Spam|kostenlos)$), beträgt 60%.

Beispiel 2:

Gegeben sind die folgenden Emails, die bereits nach Spam oder kein Spam klassifiziert sind:

Nr.	Email	Spam / !Spam
1	send us your password	Spam
2	send us your review	kein Spam
3	review your password	kein Spam
4	review us	Spam
5	send your password	Spam
6	send us your account	Spam

Es soll nun basierend auf diesen Daten, die Wahrscheinlichkeit berechnet werden, dass eine Email mit den Worten "Change your password" Spam ist oder nicht.

Berechnung:

- **Prior-Wahrscheinlichkeit, dass eine zufällige E-Mail Spam ist** $P(Spam)$:$\frac{4}{6}$
- **Prior-Wahrscheinlichkeit, dass eine zufällige E-Mail kein Spam ist** $P(\neg Spam)$: $\frac{2}{6}$

Bestimmen der Wortfrequenz:

Nr.	Wort	Wortfrequenz in Spam	Wortfrequenz in ¬ Spam
1	password	2/4	1/2
2	review	1/4	2/2
3	send	3/4	1/2
4	us	3/4	1/2
5	your	3/4	2/2
6	account	1/4	0/2

Laplace-Smoothing: Es sollte eine Glättung vorgenommen werden, um das Szenario zu vermeiden, dass ein Wort in den Spam-Trainingsbeispielen nicht vorkommt, wohl aber in den $\neg Spam$-Trainingsbeispielen oder umgekehrt. Hierfür wird einfach zu jeder Wortanzahl +1 hinzugefügt und zum Nenner wird +2 hinzugefügt.

Nr.	Wort	Wortfrequenz in Spam	Wortfrequenz in ¬ Spam
1	password	3/6	2/4
2	review	2/6	3/4
3	send	4/6	2/4
4	us	4/6	2/4
5	your	4/6	3/4
6	account	2/6	1/4

Als nächstes berechnen wir die Wahrscheinlichkeit dafür, dass "Change your password" in Spam Mails und die Wahrscheinlichkeit dass es in Nicht-Spam Mails vorkommt. Da "Change" nicht in unserem Vokabular vorkommt, kann es einfach weggelassen werden und wir betrachten nur "your password".

- **Spam:**

  $P(your|Spam) = \frac{4}{6}$

  $P(password|Spam) = \frac{3}{6}$

  $P(your\,password|Spam) = P(your|Spam) * P(password|Spam) = \frac{4}{6} * \frac{3}{6} = \frac{1}{3}$

- **Nicht Spam:**

  $P(your|\neg Spam) = \frac{3}{4}$

  $P(password|\neg Spam) = \frac{2}{4}$

  $P(your\ password|\neg Spam) = P(your|\neg Spam) * P(password|\neg Spam) = \frac{3}{4} * \frac{2}{4} = \frac{3}{8}$

Berechnen der Gesamtwahrscheinlichkeit:

1. Marginal Likelyhood:

   $P(your\,password) = P(your\,password|Spam) * P(Spam) + P(your\,password|\neg Spam) * P(\neg Spam) = \frac{1}{3} * \frac{4}{6} + \frac{3}{8} * \frac{2}{6} = \frac{2}{9} + \frac{1}{8} \approx 0.347$

2. Bayessches Theorem:

   $P(Spam|your\ password) = \frac{P(your\ password|Spam) * P(Spam)}{P(your\ password)} = \frac{\frac{1}{3} * \frac{4}{6}}{0.347} = 0.64$

   $P(\neg Spam|your\ password) = \frac{P(your\ password|\neg Spam) * P(\neg Spam)}{P(your\ password)} = \frac{\frac{3}{8} * \frac{2}{4}}{0.347} = 0.54$

Basierend auf den Berechnungen haben wir die folgenden Ergebnisse:

- Die **Gesamtwahrscheinlichkeit, dass die Wörter "your password" in einer E-Mail erscheinen** ($P(your\ password)$), beträgt 34,7%.
- Die **Wahrscheinlichkeit, dass eine E-Mail Spam ist, gegeben dass die Wörter "your password" erscheinen** ($P(Spam|your\ password)$), beträgt 64%.