Sentiment Analysis

Idee und Motivation

Ein Programm soll entschieden, ob sich sein Verfasser positiv oder negativ zu einer Sache äussert.

- Wie wird eine neues Produkt von den Konsumenten wahrgenommen?
- Wie kommt ein neuer Kinofilm bei den Zuschauern an?
- Welche politischen Meinungen sind laut Internet „mehrheitsfähig“?

Einordnung

Sentiment Analysis (Opinion Mining) ist verwandt mit anderen Klassifikationsaufgaben:

- Spam-Erkennung von E-Mails
- Kategorisierung von Textdokumenten
- Authorship attribution (war hat die Melodie zu „In My Life“ geschrieben?)

Terminologie

Bei der vorliegenden Aufgabe handelt es sich um ein *Klassifikationsproblem*. Da die Klassifikation durch *Trainingsbeispiele* „erlernt“ wird, handelt es sich um ein Verfahren des *Supervised Learning*.

Die Trainingsbeispiele bestehen aus Paaren

$$(d_1, c_1), (d_2, c_2), \ldots, (d_N, c_N)$$

wobei d_1, d_2, …, d_N die Dokumente und c_1, c_2, …, c_N die zugehörigen Klassen (*Labels*) bezeichnen.

Die i-te Klasse habe den Wert $c_i = 0$, wenn das Dokument einen Sachverhalt negativ beurteilt und den Wert $c_i = 1$, wenn es eine positive Meinung ausdrückt.

Bemerkungen

- Die Klassifizierung der Trainingsbeispiele muss im Voraus durch menschliche Beurteilung erfolgen (*Gold labels*).
- Im Falle der Sentiment Analysis wäre auch eine Einteilung der Dokumente in mehr als zwei Kategorien möglich. Oft wird eine dritte Klasse für neutrale Dokumente verwendet. Auch weitere Klassen für verschiedene Grade von Zustimmung oder Ablehnung sind denkbar.

Beispiel

Nr. i	Dokument d_i	Klasse c_i
1	schlechter film	0
2	schlechter plot	0
3	guter hauptdarsteller	1
4	wunderbarer film	1
5	guter film	1

Tabelle 1: Trainingsbeispiele

Das Modell

Für ein Dokument d und die Klassen $c \in \{0, 1\}$ suchen wir:

$$\hat{c} = \underset{c \in \{0,1\}}{\operatorname{argmax}} P(c|d) \tag{1}$$

Die Wahrscheinlichkeiten $P(c|d)$ lassen sich nur schwer aus den relativen Häufigkeiten der Trainingsdokumente schätzen, da diese im Allgemeinen zu verschieden sind.

Die Formel von Bayes

Für Ereignisse A, $B \subset \Omega$ mit $P(A) > 0$ gilt:

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)} \tag{2}$$

Einsetzen der Formel von Bayes in den Ausdruck (1) zur Klassifikation des Dokuments d ergibt:

$$\begin{aligned} \hat{c} &= \underset{c \in \{0,1\}}{\operatorname{argmax}} P(c|d) = \underset{c \in \{0,1\}}{\operatorname{argmax}} \frac{P(c)P(d|c)}{P(d)} \\ &= \underset{c \in \{0,1\}}{\operatorname{argmax}} P(c)P(d|c) \end{aligned} \tag{3}$$

Das letzte Gleichheitszeichen lässt sich damit begründen, dass die Wahrscheinlichkeit im Nenner unabhängig von der Klasse c ist.

$P(c)$ drückt aus, mit welcher Wahrscheinlichkeit ein Dokument zur Klasse c gehört, ohne dass der Inhalt des Dokuments d berücksichtigt wird (*A priori-Wahrscheinlichkeit*).

$P(d|c)$ bezeichnet die Wahrscheinlichkeit, mit der das Dokument d von der Klasse c „erzeugt“ wird (*Likelihood*).

Multimengen

Eine Multimenge ist eine Menge, bei der jedem Element der Menge zusätzlich eine Häufigkeit (≥ 0) zugeordnet wird. Wie bei (Multi-)Mengen üblich, ist die Reihenfolge der Elemente nicht von Bedeutung.

Fassen wir alle Wörter in einer Menge von Dokumenten D in willkürlicher Reihenfolge zu einer Menge $W = \{w_1, w_2, \ldots, w_n\}$ (Wörterbuch) zusammen, so lassen sich die einzelnen Dokumente als Multimenge (*bag-of-words*) darstellen.

Beispiel (Fortsetzung)

Wort	d_1	d_2	d_3	d_4	d_5
film	1	0	0	1	1
guter	0	0	1	0	1
hauptdarsteller	0	0	1	0	0
plot	0	1	0	0	0
schlechter	1	1	0	0	0
wunderbarer	0	0	0	1	0

Tabelle 2: Bag-of-words-Darstellung der Trainingsbeispiele

Starke Annahmen

Besteht also ein Dokument d aus den Häufigkeiten der n Wörter w_1, w_2, …, w_n, so erhalten wir:

$$\begin{aligned} P(d|c) &\approx P(w_1, w_2, \ldots, w_n|c) \\ &\approx P(w_1|c) \cdot P(w_2|c) \cdot \ldots \cdot P(w_n|c) \end{aligned} \tag{4}$$

In der ersten Umformung wird das Dokument vereinfachend durch eine Multimenge dargestellt. In der zweiten haben wir die „naive“ Annahme getroffen haben, dass jedes Wort unabhängig von den anderen im Dokument vorkommt (vorausgesetzt, dass es sich in der Klasse c befindet).

Dieser Vereinfachung und der Bayesschen Formel verdankt das Verfahren seinen Namen: *Naive Bayes*. Die vollständige Formel lautet dann:

$$\hat{c} = \operatorname*{argmax}_{c \in \{0,1\}} P(c|d) \approx \operatorname*{argmax}_{c \in \{0,1\}} P(c) \prod_{i=1}^{n} P(w_i|c) \tag{5}$$

Die Wahrscheinlichkeiten auf der linken Seite von (5) lassen sich aufgrund relativer Häufigkeiten in der Trainingsmenge schätzen:

$$P(c) \approx \frac{\text{Anzahl Trainingsdokumente in Klasse } c}{\text{Anzahl Trainingsdokumente insgesamt}} \tag{6}$$

$$\begin{aligned} P(w_i|c) = \frac{P(w_i, c)}{P(c)} &\approx \frac{\text{Anzahl}(w_i, c)}{\sum_{w_i \in W} \text{Anzahl}(w_i, c)} \\ &= \frac{\text{Häufigkeit von Wort } w_i \text{ in Klasse } c}{\text{Summe der Häufigkeiten aller Wörter in Klasse } c} \end{aligned} \tag{7}$$

Beispiel (Fortsetzung)

	$c = 0$	$c = 1$
Anzahl Dokumente	2	3
Wahrscheinlichkeit $P(c)$	2/5	3/5

Tabelle 3: geschätzte A priori-Wahrscheinlichkeiten

Wort	$c = 0$		$c = 1$	
film	1	(1/4)	2	(2/6)
guter	0	(0/4)	2	(2/6)
hauptdarsteller	0	(0/4)	1	(1/6)
plot	1	(1/4)	0	(0/6)
schlechter	2	(2/4)	0	(0/6)
wunderbarer	0	(0/4)	1	(1/6)
Summe	4	(4/4)	6	(6/6)

Tabelle 4: geschätzte Likelihoods

Klassifizierung

Um die Zugehörigkeit eines (noch unbesehenen) Dokuments d zu einer der Klassen zu ermitteln, müssen wir „nur“ die Formel (5) anwenden.

Beispiel (Fortsetzung)

$d =$ schlechter hauptdarsteller

$$\begin{aligned} P(0|d) &= P(0) \cdot P(\text{schlechter}|0) \cdot P(\text{hauptdarsteller}|0) \\ &= \frac{2}{5} \cdot \frac{2}{4} \cdot \frac{0}{4} = 0 \end{aligned}$$

$$\begin{aligned} P(1|d) &= P(1) \cdot P(\text{schlechter}|1) \cdot P(\text{hauptdarsteller}|1) \\ &= \frac{3}{5} \cdot \frac{0}{6} \cdot \frac{1}{6} = 0 \end{aligned}$$

Hoppla! Das Dokument enthält Wörter, die nicht in beiden Klassen vorkommen.

Add-one-smoothing

Eine Lösung für das obige Problem besteht darin, jedem Wort in jeder Klasse ein Vorkommen zu „schenken“ (*Pseudozähler*). Wenn man anschliessend korrekt normalisiert, bleiben die Grössenverhältnisse innerhalb der Klassen unverändert.

Beispiel (Fortsetzung)

Wort	$c=0$		$c=1$	
film	$1+1$	$(2/10)$	$2+1$	$(3/12)$
guter	$0+1$	$(1/10)$	$2+1$	$(3/12)$
hauptdarsteller	$0+1$	$(1/10)$	$1+1$	$(2/12)$
langweilig	$1+1$	$(2/10)$	$0+1$	$(1/12)$
schlechter	$2+1$	$(3/10)$	$0+1$	$(1/12)$
wunderbarer	$0+1$	$(1/10)$	$1+1$	$(2/12)$
Summe	$4+6$	$(10/10)$	$6+6$	$(12/12)$

Tabelle 5: geschätzte Likelihoods mit Add-one-smoothing

$d =$ schlechter hauptdarsteller

$$\begin{aligned} P(0|d) &= P(0) \cdot P(\text{schlechter}|0) \cdot P(\text{hauptdarsteller}|0) \\ &= \frac{2}{5} \cdot \frac{3}{10} \cdot \frac{1}{10} = 1.2 \cdot 10^{-2} \end{aligned}$$

$$\begin{aligned} P(1|d) &= P(1) \cdot P(\text{schlechter}|1) \cdot P(\text{hauptdarsteller}|1) \\ &= \frac{3}{5} \cdot \frac{1}{12} \cdot \frac{2}{12} \approx 8.3 \cdot 10^{-3} \end{aligned}$$

Da sich für $c = 0$ die grössere Wahrscheinlichkeit ergibt, klassifizieren wir d „negativ“

Bemerkungen

- Durch das Multiplizieren vieler kleiner Zahlen entsteht das Problem des *Underflows*. Dieses lässt sich durch Logarithmieren der Gleichung (5) entschärfen.

$$\begin{aligned} \hat{c} &= \operatorname*{argmax}_{c \in \{0,1\}} \ln P(c|d) \\ &\approx \operatorname*{argmax}_{c \in \{0,1\}} \ln \left[P(c) \prod_{i=1}^{n} P(w_i|c) \right] \\ &= \operatorname*{argmax}_{c \in \{0,1\}} \left[\ln P(c) + \sum_{i=1}^{n} \ln P(w_i|c) \right] \end{aligned} \tag{8}$$

  Die Monotonie der Logarithmusfunktion erhält das Maximum.

- Trifft am in einem zu klassifizierenden Dokument auf ein Wort, das nicht in den Trainingsdaten vorkommt, so wird es bei der Berechnung der Wahrscheinlichkeiten ignoriert.

- Es kann auch sinnvoll sein, die in einem Dokument mehrfach vorkommenden Wörter nur einmal zu zählen.

- Werden Verneinungen erkannt (`"kein guter Film"`), lassen sich die damit verbundenen Probleme durch Bildung negativer Wortvarianten entschärfen (`"NEG_guter"`).

- Professionelle Sentiment Analysis-Systeme ...
  - berücksichtigen Wortarten,
  - verwenden N-Gramme (Monogramme, Bigramme, Trigramme),
  - führen Stammformreduktion (Stemming) durch,
  - benutzen gelabelte Wortlisten (sentiwordnet.isti.cnr.it).