Bayes-Klassifikator für diskrete Variablen

Prof. Dr. Jörg Frochte

Maschinelles Lernen

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

Satz von Bayes

Eine bedingte Wahrscheinlichkeit dafür, dass A eintritt unter der Bedingung, dass B bereits eingetreten ist, notiert man als $P(A \mid B)$. Mit dieser Notation gilt:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \tag{1}$$

Satz von der totalen Wahrscheinlichkeit

Seien $B_1, B_2, \ldots$ Ereignisse mit $P(B_j) > 0$ für alle j und

$$\bigcup_{j=1}^{\infty} B_j = \Omega,$$

also alle möglichen Ergebnisse – aufgepasst, es geht nicht um Ereignisse denn dann wäre von $\mathcal{P}(\Omega)$ die Rede – sind in den B_j enthalten, dann gilt:

$$P(A) = \sum_{j=1}^{\infty} P\left(A \mid B_j\right) \cdot P\left(B_j\right) \tag{2}$$

Satz von Bayes + Satz von der totalen Wahrscheinlichkeit

Modifizierte Darstellung des Satzes von Bayes

Wenn wir nun den Satz von der totalen Wahrscheinlichkeit (2) auf $P(B)$ im Nenner des Satzes von Bayes (1) anwenden, erhalten wir:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P\left(B \mid A_i\right) \cdot P(A_i)}{\sum_{j=1}^{N} P\left(B \mid A_j\right) \cdot P(A_j)} \tag{3}$$

Wichtig

Denken Sie daran, dass dafür natürlich alle Voraussetzungen des Satzes von der totalen Wahrscheinlichkeit verfüllt sein müssen.

- Ausgangspunkt für unseren ersten Klassifikator ist die Formel (3).

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{j=1}^{N} P(B \mid A_j) \cdot P(A_j)}$$

- Diese passen wir in der Notation ein wenig an. x soll jetzt der **Vektor unserer Merkmale** sein und i die Zugehörigkeit zur Klasse Nr. i.

$$P(i \mid x) = \frac{P(x \mid i) \cdot P(i)}{\sum_{j=1}^{N} P(x \mid j) \cdot P(j)}$$

- Mit einer zunehmenden Anzahl an Merkmalen ist man nicht mehr in der Lage so etwas wie $P(x \mid i)$ auszurechnen.
- Hintergrund ist die Kombinatorik, die einfach zu viele Möglichkeiten ergibt die einzelnen Merkmale zu kombinieren.
- Neben Rechenleistung ist der primäre Grund, dass der Umfang unserer Datenbank immer größer werden muss, um diese Wahrscheinlichkeiten noch ermitteln zu können. Man behilft sich mit einer Annahme bzgl. der stochastischen Unabhängigkeit.

Beispiel für den kombinatorischen Effekt

- Nehmen wir an, wir haben nur zwei Merkmale um Frauen(w) und Männer(m) (zwei Klassen) zu unterscheiden und diese sind dann sogar nur in wenigen diskreten Größen zusammengefasst:

$$\text{Körpergröße} = \{\text{small}, \text{medium}, \text{large}\} \qquad \text{Kurzhaarschnitt} = \{\text{yes}, \text{no}\}$$

- Der Merkmalsvektor besteht entsprechend aus zwei Einträgen $x = (\text{Körpergröße}, \text{Kurzhaarschnitt})$.
- Um $P(x \mid i)$ zu berechnen müssten wir schon jetzt viele Kombinationen auswerten:

$$P(\ (\text{small}, \text{yes}) \mid w)\ , P(\ (\text{medium}, \text{yes}) \mid w)\ ,\ P(\ (\text{large}, \text{yes}) \mid w)$$
$$P(\ (\text{small}, \text{no}) \mid w)\ , P(\ (\text{medium}, \text{no}) \mid w)\ ,\ P(\ (\text{large}, \text{no}) \mid w) \ldots$$

- Wir kommen also in diesem einfachen Beispiel kombinatorisch auf $2 \cdot 2 \cdot 3 = 12$ Terme die für die Berechnung des Nennes gebildet werden müssen.

Wichtig

Wenn zwei Ergebnisse A und B **stochastisch unabhängig** sind, gilt folgende Regel:

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

- Ein Beispiel ist das Würfeln mit zwei Würfeln.
- Die Chance, dass beide Würfel gleichzeitig eine 6 zeigen, ist:

$$P(\text{rot} = 6 \cap \text{blau} = 6) = P(\text{rot} = 6) \cdot P(\text{blau} = 6) = 1/6 \cdot 1/6 = 1/36$$

- Nehmen wir an, dass die Merkmale $x^{(k)}$ mit $k = 1, \ldots, m$ in dem Merkmalsvektor unabhängig sind, erhalten wir:

$$P(x \mid i) = \prod_{k=1}^{m} P(x^{(k)} \mid i)$$

- Das setzen wir nun ein:

$$P(i \mid x) = \frac{\prod_{k=1}^{m} P(x^{(k)} \mid i) \cdot P(i)}{\sum_{j=1}^{N} P(j) \cdot \prod_{k=1}^{m} P(x^{(k)} \mid j)}$$

Diese zunächst etwas ehrfurchtgebietende Formel mit Summen und Produktzeichen ist nun wirklich **unser Klassifikator**, den wir umsetzen können.

- Die Frage ist nur, wie valide unsere Annahme ist, dass die Merkmale stochastisch unabhängig sind.
- Tatsächlich wird diese in der Regel falsch sein und das hat etwas mit dem Unterschied zwischen Korrelation und Kausalität zu tun.
- Die meisten Dinge korrelieren dann doch irgendwie – das wäre das Gegenteil von stochastisch unabhängig.

Korrelation und Kausalität

- Ein Grund ist der wichtige Unterschied zwischen **Korrelation** und **Kausalität**.
- Das Problem ist, dass es viele Korrelationen zwischen zwei Größen gibt, denen kein Kausalzusammenhang (Ursache $\rightarrow$ Wirkung) zugrunde liegt.
- Der deutsche Ausdruck *Scheinkorrelation* ist da etwas unglücklich, denn es ist tatsächlich mathematisch eine Korrelation, die uns bei der Formel oben Probleme macht.

Beispiel Schokoladenkonsum

Es gab eine nachgewiesene Korrelation zwischen Schokoladenkonsum in einem Land und ein anderes die Anzahl der Nobelpreisträger. Der Grund ist, dass so ziemlich jedes Merkmal, das sich mit dem Einkommen oder Reichtum in einem Land erhöht, auch mit der Anzahl der Nobelpreisträger korreliert. Irgendwie kostet Forschung dann eben doch Geld.

- In der Praxis leidet die Qualität der Vorhersage unserer Formel mit dem Grad und der Häufigkeit der Verletzung der stochastisch unabhängig. Es ist keine *Alles oder Nichts*-Sache.
- Wir lernen später noch Techniken kennen um zu untersuchen ob zwei Merkmale korrelieren.

Naiver Bayes-Klassifikator – unser erster Klassifikator

- Grundlegenden Formel unseres Bayes-Klassifikator:

$$P(i \mid x) = \frac{\prod_{k=1}^{m} P(x^{(k)} \mid i) \cdot P(i)}{\sum_{j=1}^{N} P(j) \cdot \prod_{k=1}^{m} P(x^{(k)} \mid j)} = \frac{P(i) \cdot \prod_{k=1}^{m} P(x^{(k)} \mid i)}{\sum_{j=1}^{N} P(j) \cdot \prod_{k=1}^{m} P(x^{(k)} \mid j)}$$

- x ist der Vektor der Merkmale und i die Klassenzugehörigkeit, die wir vorhersagen wollen. m die Anzahl der Merkmale und N die der Klassen.

Hinweis

Die Annahme, dass die Merkmale unabhängig sind ist i.A. **falsch**. Trotzdem liefert der naive Bayes-Klassifikator in der Praxis oft recht gute Ergebnisse, wenn die Korrelation nicht zu ausgeprägt ist.

Diskrete und kontinuierliche (Zufalls-)Variablen

Diskrete Variablen

Diskrete Variablen haben **endlich viele Ausprägungen** (wir vernachlässigen die Möglichkeit abzählbar unendlich viele Ausprägungen).

Beispiel Beaufort-Skala

Die Beaufort-Skala für Windstärken geht von *Windstill* (0) bis *Orkan* (12). Es handelt sich um 12 diskrete Werte die Sie ggf. in einer Datenbank vorfinden.

Kontinuierliche Variablen

Bei kontinuierlichen Variablen ist zwischen zwei Werten $a < b$ auch jeder Zwischenwert im Intervall $[a, b]$ möglich. Entsprechend gilt $[a, b] \subset \mathbb{R}$.

Beispiel Temperatur

Die Temperatur ist eine kontinuierliche Variable, da zwischen zwei Temperaturen theoretisch jede andere vorkommen. Praktisch wird dies durch die Messgenauigkeit begrenzt.

Skalenniveaus

- In der praktischen Umsetzung rechnen wir beim maschinellen Lernen immer mit Float-Datentypen.
- Unsere Datenquellen enthalten, aber Werte aus völlig unterschiedlichen Skalenniveaus die konvertiert werden.

Skala	math. Operationen	Messbare Eigenschaften	Beispiel
Nominal	$= / \neq$	Häufigkeit	ja/nein
Ordinal	$= / \neq ; < / >$	Häufigkeit, Reihenfolge	Dienstränge
Intervall	$= / \neq ; < / >, + / -$	Häufigkeit, Reihenfolge, Abstand	Datum
Rational	$= / \neq ; < / >, + / -, * / :$	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Temperatur (K)

- Einträge aus der Nominalskala kommen häufig vor, dennoch darf man mit Ihnen eigentlich nicht rechnen.
- Bei manchen Algorithmen kommt man mit geringeren Anforderungen an die nötigen Operationen aus und bei anderen muss man sich klar machen, das man sich gerade in der Praxis auf dünnem Eis bewegt.

Skalenniveaus

- In der praktischen Umsetzung rechnen wir beim maschinellen Lernen immer mit Float-Datentypen.
- Unsere Datenquellen enthalten, aber Werte aus völlig unterschiedlichen Skalenniveaus die konvertiert werden.

Skala	math. Operationen	Messbare Eigenschaften	Beispiel
Nominal	$= / \neq$	Häufigkeit	ja/nein
Ordinal	$= / \neq ; < / >$	Häufigkeit, Reihenfolge	Dienstränge
Intervall	$= / \neq ; < / >, + / -$	Häufigkeit, Reihenfolge, Abstand	Datum
Rational	$= / \neq ; < / >, + / -, * / :$	Häufigkeit, Reihenfolge, Abstand, Nullpunkt	Temperatur (K)

- Einträge aus der Nominalskala kommen häufig vor, dennoch darf man mit Ihnen eigentlich nicht rechnen.
- Bei manchen Algorithmen kommt man mit geringeren Anforderungen an die nötigen Operationen aus und bei anderen muss man sich klar machen, das man sich gerade in der Praxis auf dünnem Eis bewegt.

Beispiel Bayes-Klassifikator für diskrete Merkmale (1/2)

- Um nun die Wahrscheinlichkeit dafür zu kennen, dass (l, n) zur einen oder anderen Klasse gehört berechnen wir nun noch den Zähler.
- Dazu berechnen wir zunächst den Nenner der in $P(m|(l,n)\ )$ und $P(w|(l,n)\ )$ gleich ist.
- $P(w)$ ist nach Tabelle die Wahrscheinlichkeit für w also $6/11$ und $P(m)$ ist entsprechend $5/11$.
- $P(x^{(k)} \mid j)$ ist dann z. B. die Wahrscheinlichkeit, dass eine Person sehr groß (l) ist, wenn es sich gleichzeitig um einen Mann handelt: $P(l|m) = 2/5$.
- Damit ergibt sich:

$$\sum_{j=1}^{N} P(j) \cdot \prod_{k=1}^{m} P(x^{(k)} \mid j) = \underbrace{\frac{6}{11} \cdot \frac{1}{6} \cdot \frac{4}{6}}_{=2/33} + \underbrace{\frac{5}{11} \cdot \frac{2}{5} \cdot \frac{1}{5}}_{2/55} = \frac{16}{165}$$

$$P(i \mid x) = \frac{P(i) \cdot \prod_{k=1}^{m} P(x^{(k)}|i)}{\sum_{j=1}^{N} P(j) \cdot \prod_{k=1}^{m} P(x^{(k)}|j)}$$

Nr.	Größe	Kurzhaarschnitt	Geschlecht
1	m	y	m
2	s	n	w
3	l	y	m
4	s	n	w
5	l	n	w
6	s	y	w
7	s	y	m
8	m	y	w
9	m	n	w
10	l	y	m
11	m	n	m

Beispiel für diskrete Merkmale (2/2)

- Die beiden Summanden sind dabei jeweils bereits die Nenner für die Wahrscheinlichkeitsberechnung, dass auf Basis der Tabelle ein Objekt mit den Merkmalen (l, n) zu Klasse w oder m gehört:

$$P(w|(l,n)) = \frac{\left(\frac{2}{33}\right)}{\left(\frac{16}{165}\right)} = \frac{5}{8} = 62.5\% \ \checkmark$$

$$P(m|(l,n)) = \frac{\left(\frac{2}{55}\right)}{\left(\frac{16}{165}\right)} = \frac{3}{8} = 37.5\%$$

- Wie man sieht würde der Klassifikator mit einer recht großen Unsicherheit zu w tendieren.

Nr.	Größe	Kurzhaarschnitt	Geschlecht
1	m	y	m
2	s	n	w
3	l	y	m
4	s	n	w
5	l	n	w
6	s	y	w
7	s	y	m
8	m	y	w
9	m	n	w
10	l	y	m
11	m	n	m

$P(w) = 6/11,\ P(m) = 5/11,\ P(y|w) = 2/6,$
$P(n|w) = 4/6,\ P(m|w) = 2/6,\ P(s|w) = 3/6,$
$P(l|w) = 1/6,\ P(y|m) = 4/5,\ P(n|m) = 1/5,$
$P(m|m) = 1/5,\ P(s|m) = 1/5,\ P(l|m) = 2/5$