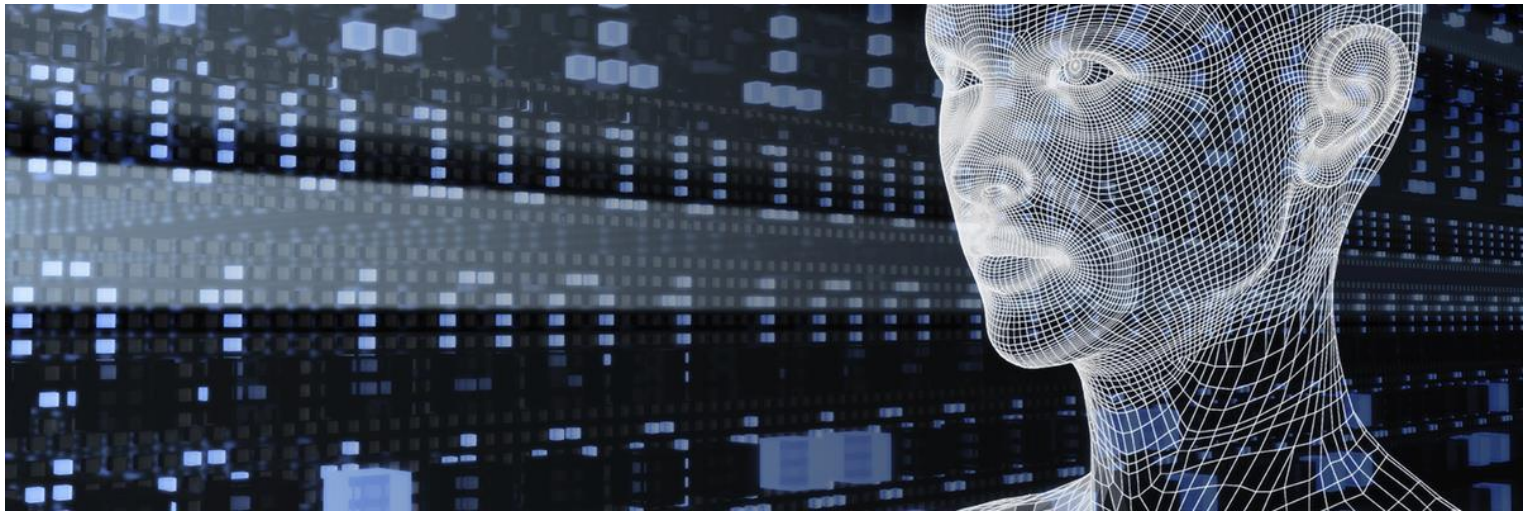
Künstliche Intelligenz

Vorlesung 10: Schließen mit Bayes Netzen



Schließen mit Unsicherheit

1. Tweety ist ein Pinguin
2. Pinguine sind Vögel
3. Vögel können fliegen

Als Wissensbasis:

pinguin(tweety)
pinguin(x) → vogel(x)
vogel(x) → fliegen(x)

- Es läßt sich ableiten: fliegen(tweety)

Schließen mit Unsicherheit

Neuer Versuch: Pinguine können nicht fliegen

- pinguin(x))→¬ fliegen(x)

Es läßt sich ableiten: ¬ fliegen(tweety)

Aber: Es läßt sich auch ableiten: fliegen(tweety)

➢Die Wissensbasis ist widersprüchlich.

➢die Logik ist monoton:

- neues Wissen kann altes nicht ungültig machen

Wahrscheinlichkeitslogik

Unsicherheit: 99% aller Vögel können fliegen

Unvollständigkeit: Agent hat unvollständige Informationen über den Zustand der Welt (Realzeitentscheidungen)

Heuristische Suche

Schließen mit unsicherem und unvollständigem Wissen

Andere Formalismen zur Modellierung von Unsicherheit

- nichtmonotone Logiken
- Defaultlogik
- Dempster-Schäfer-Theorie: ordnet einer logischen Aussage A eine Glaubensfunktion (engl. belief function) Bel (A) zu
- Fuzzy-Logik: Regelungstechnik

Schließen mit bedingten Wahrscheinlichkeiten

- bedingte Wahrscheinlichkeiten statt Implikation (materiale Implikation)
- subjektive Wahrscheinlichkeiten
- Wahrscheinlichkeitstheorie sehr gut fundiert
- Schließen mit unsicherem und unvollständigem Wissen
- Methode der maximalen Entropie (MaxEnt)
- Bayes-Netze

Wiederholung: Rechnen mit Wahrscheinlichkeiten

Definition

Sei Ω die zu einem Versuch gehörende endliche Menge von **Ereignissen**. Jedes Ereignis $\omega \in \Omega$ steht für einen möglichen Ausgang des Versuchs. Schließen sich die Ereignisse $w_i \in \Omega$ gegenseitig aus, decken aber alle möglichen Ausgänge des Versuchs ab, so werden diese **Elementarereignisse** genannt.

Wiederholung: Rechnen mit Wahrscheinlichkeiten

Definition

Sei $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ endlich. Es sei kein Elementarereignis bevorzugt, d.h. man setzt eine Symmetrie bezüglich der Häufigkeit des Auftretens aller Elementarereignisse voraus. Die **Wahrscheinlichkeit** $P(A)$ des Ereignisses A ist dann

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für } A \text{ günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}$$

Wiederholung: Rechnen mit Wahrscheinlichkeiten

Definition

Für zwei Ereignisse A und B ist die Wahrscheinlichkeit $P(A|B)$ für A unter der Bedingung B **(bedingte Wahrscheinlichkeit)** definiert durch

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

Wiederholung: Rechnen mit Wahrscheinlichkeiten

Definition

Gilt für zwei Ereignisse A und B

$$P(A|B) = P(A),$$

so nennt man diese Ereignisse unabhängig.

Satz

Für unabhängige Ereignisse A und B folgt aus der Definition

$$P(A \wedge B) = P(A) \cdot P(B).$$

Wiederholung: Rechnen mit Wahrscheinlichkeiten

Produktregel: $P(A \wedge B) = P(A|B)P(B)$

Kettenregel:

$$
\begin{aligned}
&\boldsymbol{P}(X_1, \ldots, X_n) \\
&= \boldsymbol{P}(X_n|X_1, \ldots, X_{n-1}) \cdot \boldsymbol{P}(X_1, \ldots, X_{n-1}) \\
&= \boldsymbol{P}(X_n|X_1, \ldots, X_{n-1}) \cdot \boldsymbol{P}(X_{n-1}|X_1, \ldots, X_{n-2}) \cdot \boldsymbol{P}(X_1, \ldots, X_{n-2}) \\
&= \boldsymbol{P}(X_n|X_1, \ldots, X_{n-1}) \cdot \boldsymbol{P}(X_{n-1}|X_1, \ldots, X_{n-2}) \cdot \ldots \cdot \boldsymbol{P}(X_2|X_1) \cdot \boldsymbol{P}(X_1) \\
&= \prod_{i=1}^{n} \boldsymbol{P}(X_i|X_1 \ldots, X_{i-1}),
\end{aligned}
$$

Die Bayes-Formel

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad \text{sowie} \quad P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

Bayes-Formel:

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

Bayes-Regel: Beispiel

- Sehr zuverlässige Einbruchmeldeanlage
- Meldet die Einbrüche mit 99% Sicherheit
- Also mit hohe Sicherheit: Wenn Alarm dann Einbruch!

Bayes-Regel: Beispiel

- Sehr zuverlässige Einbruchmeldeanlage
- Meldet die Einbrüche mit 99% Sicherheit
- Also mit hohe Sicherheit: Wenn Alarm dann Einbruch!
- NEIN!

$$P(A|E) = 0.99, \quad P(A) = 0.1, \quad P(E) = 0.001$$

$$P(E|A) = \frac{P(A|E)P(E)}{P(A)} = \frac{0.99 \cdot 0.001}{0.1} = 0.01$$

Schließen mit Bayes Netzen

- d Variablen $X_1, \ldots, X_d$ mit je n Werten
- Wahrscheinlichkeitsverteilung hat $n^d - 1$ Werte.
- in d. Praxis: Verteilung enthält viel Redundanz.

Unabhängige Variablen

$P(X_1, \ldots, X_d) = P(X_1)\ P(X_2) \ldots P(X_d)$:

bedingte Wahrscheinlichkeiten werden trivial:

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

- nur noch $(n-1) \cdot d$ Variablen!

Das Alarm-Beispiel

Wissensbasis:

$$P(J|Al) = 0.90 \qquad P(M|Al) = 0.70$$
$$P(J|\neg Al) = 0.05 \qquad P(M|\neg Al) = 0.01$$

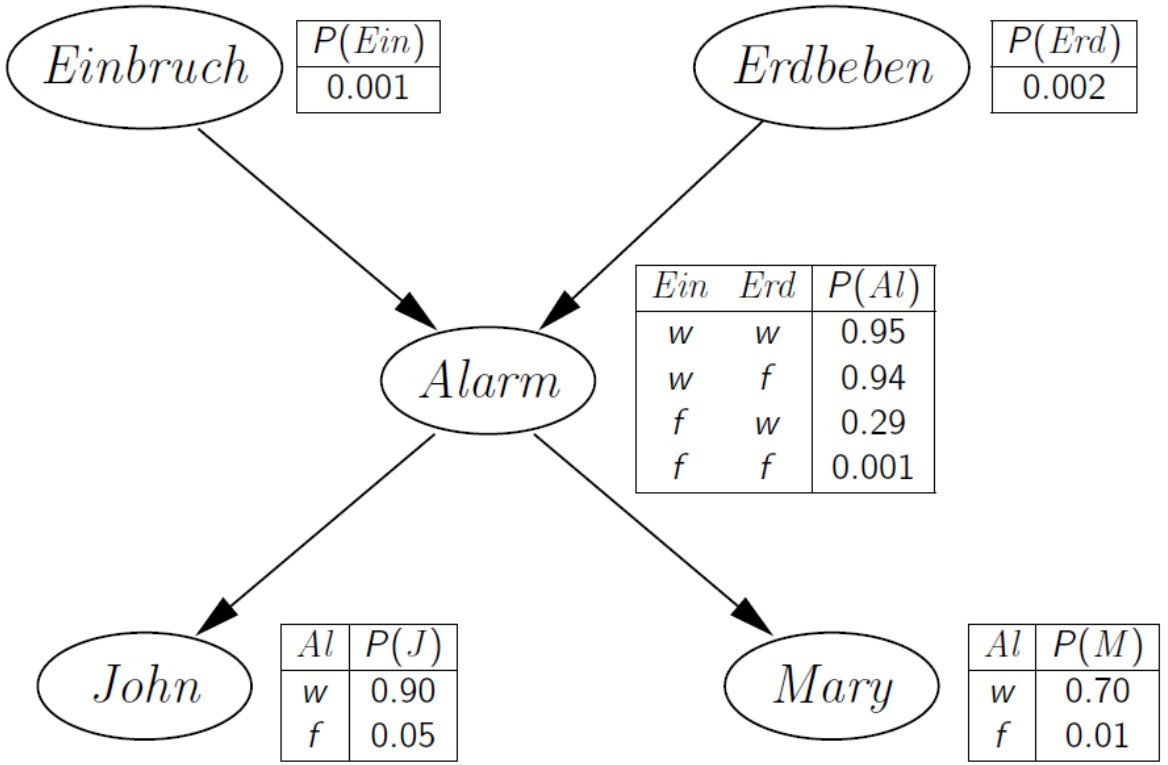
$$\begin{aligned} P(Al|Ein, Erd) &= 0.95 \\ P(Al|Ein, \neg Erd) &= 0.94 \\ P(Al|\neg Ein, Erd) &= 0.29 \\ P(Al|\neg Ein, \neg Erd) &= 0.001, \end{aligned}$$

$P(Erd) = 0.002.$

Anfragen: $P(Ein|J \vee M)$ $P(J|Ein)$ $P(M|Ein)$

J	„John ruft an"	M	„Mary ruft an"	Al	„Alarmsirene ertönt
Ein	„Einbruch"	Erd	„Erdbeben"		

Graphische Darstellung des Wissens als Bayes-Netz



Bedingte Unabhängigkeit

Definition

Zwei Variablen A und B heißen **bedingt unabhängig**, gegeben C, wenn

$$\boldsymbol{P}(A, B|C) = \boldsymbol{P}(A|C) \cdot \boldsymbol{P}(B|C).$$

Beispiele:

$$\boldsymbol{P}(J, M|Al) = \boldsymbol{P}(J|Al) \cdot \boldsymbol{P}(M|Al)$$

$$\boldsymbol{P}(J, Ein|Al) = \boldsymbol{P}(J|Al) \cdot \boldsymbol{P}(Ein|Al)$$

Satz

Folgende Gleichungen sind paarweise äquivalent, das heißt jede einzelne dieser Gleichungen beschreibt die bedingte Unabhängigkeit der Variablen A und B gegeben C.

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

$$P(A|B, C) = P(A|C)$$

$$P(B|A, C) = P(B|C)$$

Praktische Anwendung von Bayes-Netze

$$P(J \mid Ein) = \frac{P(J, Ein)}{P(Ein)} = \frac{P(J, Ein, Al) + P(J, Ein, \neg Al)}{P(Ein)}$$

$$\boldsymbol{P}(J, Ein, Al) = \boldsymbol{P}(J|Ein, Al)\,\boldsymbol{P}(Al|Ein)\,\boldsymbol{P}(Ein) = \boldsymbol{P}(J|Al)\,\boldsymbol{P}(Al|Ein)\,\boldsymbol{P}(Ein)$$

$$\begin{aligned} P(J \mid Ein) &= \frac{P(J \mid Al)\,P(Al \mid Ein)\,P(Ein) + P(J \mid \neg Al)\,P(\neg Al \mid Ein)\,P(Ein)}{P(Ein)} \\ &= P(J \mid Al)\,P(Al \mid Ein) + P(J \mid \neg Al)\,P(\neg Al \mid Ein). \end{aligned}$$

$$\begin{aligned} P(Al \mid Ein) &= \frac{P(Al, Ein)}{P(Ein)} = \frac{P(Al, Ein, Erd) + P(Al, Ein, \neg Erd)}{P(Ein)} \\ &= \frac{P(Al \mid Ein, Erd)\,P(Ein)\,P(Erd) + P(Al|Ein, \neg Erd)\,P(Ein)\,P(\neg Erd)}{P(Ein)} \\ &= P(Al \mid Ein, Erd)\,P(Erd) + P(Al \mid Ein, \neg Erd)\,P(\neg Erd) \\ &= 0{,}95 \cdot 0{,}002 + 0{,}94 \cdot 0{,}998 = 0{,}94 \end{aligned}$$

Praktische Anwendung von Bayes-Netze

$$P(J \mid Ein) = 0{,}9 \cdot 0{,}94 + 0{,}05 \cdot 0{,}06 = 0{,}849$$

Analog berechnet man $P(M|Ein) = 0{,}659$.

Wir wissen nun also, dass John bei etwa 85% aller Einbrüche anruft und Mary bei etwa 66% aller Einbrüche. Die Wahrscheinlichkeit, dass beide anrufen, ergibt sich aufgrund der bedingten Unabhängigkeit zu

$$\begin{aligned} P(J, M \mid Ein) &= P(J, M \mid Al)\, P(Al \mid Ein) + P(J, M \mid \neg Al)\, P(\neg Al \mid Ein)) \\ &= P(J \mid Al)\, P(M \mid Al)\, P(Al \mid Ein) + P(J \mid \neg Al)\, P(M \mid \neg Al)\, P(\neg Al \mid Ein) \\ &= 0{,}9 \cdot 0{,}7 \cdot 0{,}94 + 0{,}05 \cdot 0{,}01 \cdot 0{,}06 = 0{,}5922. \end{aligned}$$

Interessanter ist aber die Wahrscheinlichkeit für einen Anruf von John oder Mary

$$\begin{aligned} P(J \vee M \mid Ein) &= P(\neg(\neg J, \neg M) \mid Ein) = 1 - P(\neg J, \neg M \mid Ein) \\ &= 1 - [P(\neg J \mid Al)\, P(\neg M \mid Al)\, P(Al \mid Ein) \\ &\qquad + P(\neg J \mid \neg Al)\, P(\neg M \mid \neg Al)\, P(\neg Al \mid Ein)] \\ &= 1 - [0{,}1 \cdot 0{,}3 \cdot 0{,}94 + 0{,}95 \cdot 0{,}99 \cdot 0{,}06] = 1 - 0{,}085 = 0{,}915. \end{aligned}$$

Praktische Anwendung von Bayes-Netze

- Bob bekommt also etwa 92% aller Einbrüche gemeldet. Um nun P(*Ein*|*J*) zu berechnen, wenden wir die Bayes-Formel an:

$$P(Ein \mid J) = \frac{P(J \mid Ein)\,P(Ein)}{P(J)} = \frac{0{,}849 \cdot 0{,}001}{0{,}052} = 0{,}016$$

- Offenbar haben nur etwa 1,6% aller Anrufe von John einen Einbruch als Ursache.
- Da die Wahrscheinlichkeit für Fehlalarme bei Mary fünfmal geringer ist als bei John, erhalten wir mit P(*Ein* |*M*) =0,056 eine wesentlich höhere Sicherheit bei einem Anruf von Mary.
- Wirkliche Sorgen um sein Eigenheim sollte sich Bob aber erst dann machen, wenn beide anrufen, denn P(*Ein* | *J*,*M*) = 0,284.

Praktische Anwendung von Bayes-Netze

- Einschieben einer Variable: $P(J \mid Ein) = P(J \mid Al)\,P(Al \mid Ein) + P(J \mid \neg Al)\,P(\neg Al \mid Ein)$

Konditionierung:

$$P(A|B) = \sum_c P(A|B, C = c)P(C = c|B).$$

Entwicklung von Bayes-Netzen

Bei den Variablen $v_1, \ldots, v_n$ mit jeweils $|v_1|, \ldots, |v_n|$ verschiedenen Werten hat die Verteilung insgesamt

$$\prod_{i=1}^{n} |v_i| - 1$$

unabhängige Einträge.
Alarm-Beispiel: $2^5 - 1 = 31$ unabhängige Einträge.
Bayes-Netz: Für einen Knoten v_i mit den k_i Elternknoten $e_{i1}, \ldots, e_{ik_i}$ besitzt die zugehörige CPT

$$(|v_i| - 1) \prod_{j=1}^{k_i} |e_{ij}|$$

Einträge.

Entwicklung von Bayes-Netzen

Alle CPTs zusammen:

$$\sum_{i=1}^{n}(|v_i| - 1)\prod_{j=1}^{k_i}|e_{ij}|$$

Alarm-Beispiel:

$$2 + 2 + 4 + 1 + 1 = 10$$

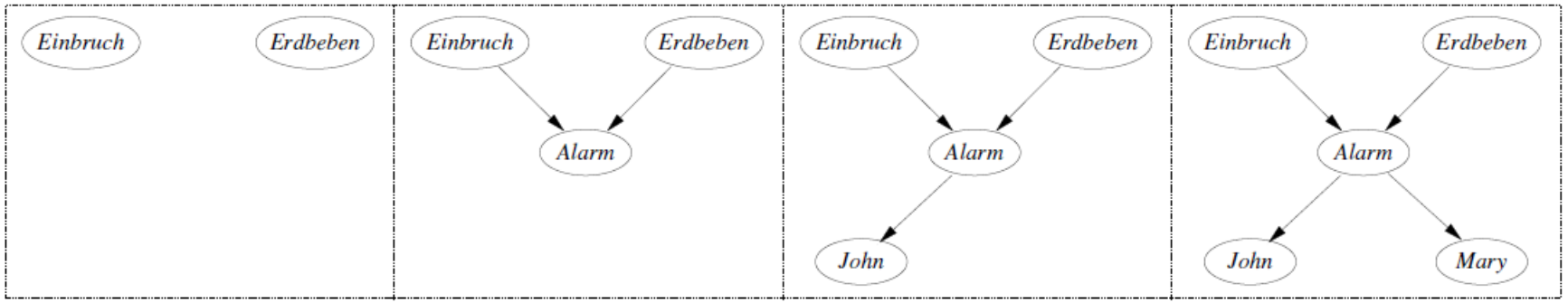
Kausalität und Netzstruktur

Aufbau eines Bayes-Netzes:

1. Entwurf der Netzwerkstruktur
   meist manuell
2. Eintragen der Wahrscheinlichkeiten in die CPTs
   meist automatisch

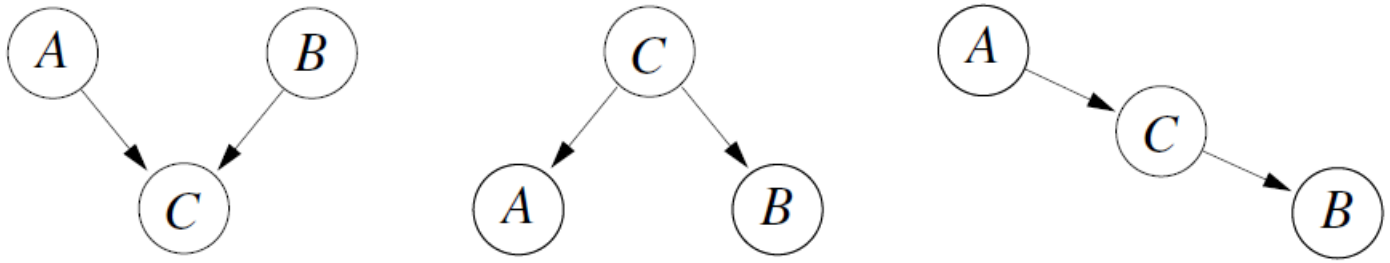
- Ursachen *Einbruch* und *Erdbeben*
- Symptome *John* und *Mary*
- Alarm: nicht beobachtbare Variable
- Kausalität beachten: von Ursache zu Wirkung vorgehen

Kausalität und Netzstruktur



Schrittweiser Aufbau des Alarm-Netzes unter Beachtung der Kausalität.

Semantik von Bayes-Netzen



Zwischen zwei Knoten A und B wird keine Kante eingetragen, wenn sie unabhängig (links) oder bedingt unabhängig sind (Mitte, rechts).

Forderungen:

- Bayes-Netz hat keine Zyklen
- keine Variable hat einen Nachfolger mit kleinerer Nummer

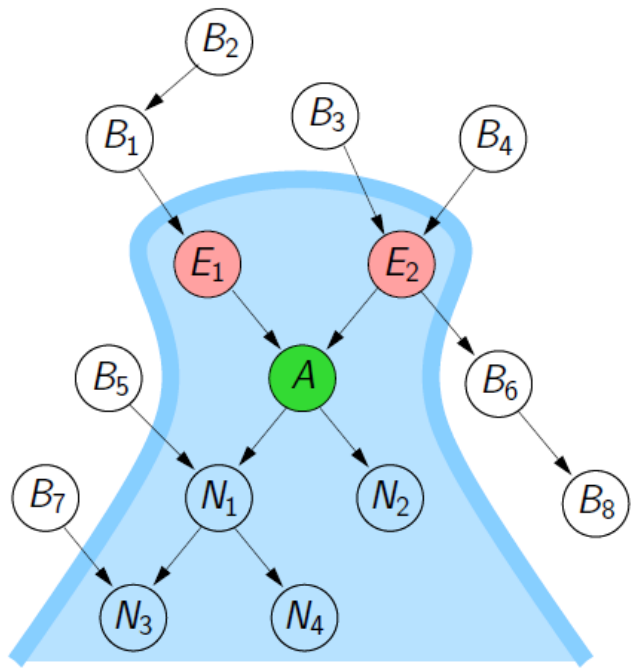
Es gilt

$$\boldsymbol{P}(X_n|X_1,\ldots,X_{n-1}) = \boldsymbol{P}(X_n|\text{Eltern}(X_n)).$$

Satz

Ein Knoten in einem Bayes-Netz ist bedingt unabhängig von allen Nicht-Nachfolgern, gegeben seine Eltern.

Sind die Elternknoten E_1 und E_2 gegeben, so sind alle Nichtnachfolger $B_1, \ldots B_8$ unabhängig von A.



[i]

Kettenregel für Bayes-Netze:

$$\boldsymbol{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \boldsymbol{P}(X_i | X_1 \ldots, X_{i-1}) = \prod_{i=1}^{n} \boldsymbol{P}(X_i | \text{Eltern}(X_i))$$

Damit gilt Gleichung

$$\boldsymbol{P}(J, Ein, Al) = \boldsymbol{P}(J|Al)\boldsymbol{P}(Al|Ein)\boldsymbol{P}(Ein)$$

Die wichtigsten Begriffe und Grundlagen von Bayes-Netzen sind nun bekannt und wir können diese zusammenfassen

Definition

Ein Bayes-Netz ist definiert durch:

- Eine Menge von Variablen und einer Menge von gerichteten Kanten zwischen diesen Variablen.
- Jede Variable hat endlich viele mögliche Werte.
- Die Variablen zusammen mit den Kanten stellen einen gerichteten azyklischen Graphen (engl. directed acyclic graph, DAG) dar. Ein DAG ist ein gerichteter Graph ohne Zyklen, das heißt ohne Pfade der Form $(A, \ldots, A)$.
- Zu jeder Variablen A ist die CPT, das heißt die Tabelle der bedingten Wahrscheinlichkeiten $\boldsymbol{P}(A|\text{Eltern}(A))$, angegeben.

Bayes-Formel: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Marginalisierung:

$P(B) = P(A, B) + P(\neg A, B) = P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)$

Konditionierung: $P(A|B) = \sum_c P(A|B, C = c)P(C = c|B)$

Eine Variable in einem Bayes-Netz ist bedingt unabhängig von allen Nicht-Nachfolge-Variablen, gegeben ihre Eltern-Variablen. Wenn $X_1, \ldots, X_{n-1}$ keine Nachfolger von X_n sind, gilt $\boldsymbol{P}(X_n|X_1, \ldots, X_{n-1}) = \boldsymbol{P}(X_n|\text{Eltern}(X_n))$. Diese Bedingung muss beim Aufbau eines Netzes beachtet werden.

Beim Aufbau eines Bayes-Netzes sollten die Variablen in Sinne der Kausalität angeordnet werden. Zuerst die Ursachen, dann die verdeckten Variablen und zuletzt die Diagnosevariablen.

Naive Bayes Klassifikation

- **Naive Bayes ist ein einfacher, aber effektiver und häufig verwendeter Klassifikator für maschinelles Lernen.**
- **Es ist ein probabilistischer Klassifikator, der die Maximum A-Posteriori Regel nutzt.**
- **Es kann auch mit einem einfachen Bayes'schen Netzwerk dargestellt werden.**
- **Naive Bayes-Klassifizierer waren häufig für die Textklassifizierung benutzt und sind eine klassische Lösung für Probleme wie Spam-Erkennung.**

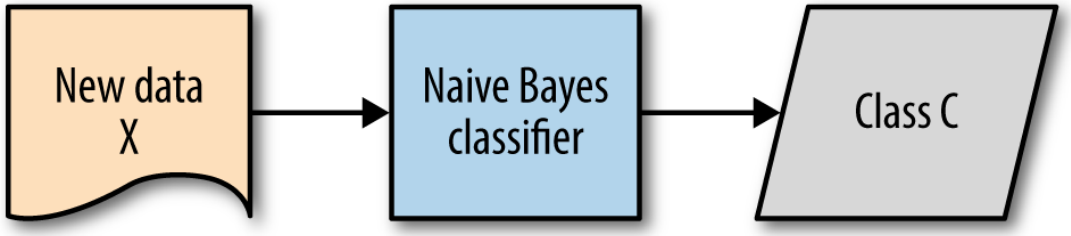
Das Modell:

- Ziel eines jeden probabilistischen Klassifizierer mit Merkmalen $x_0,\ldots,x_n$ und Klassen $c_0,\ldots, c_k$ ist um die Wahrscheinlichkeit des Auftretens der Merkmale in jeder Klasse zu bestimmen und die wahrscheinlichste Klasse zurückzugeben
- Für jede Klasse berechnen wir folgende bedingte Wahrscheinlichkeit: $P(c_i \mid x_0, \ldots, x_n)$.
- Wir benutzen dafür die **Bayes Regel**:

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

Naive Bayes Klassifikation

- Naive Bayes wird, im Vergleich zu anderen gängigen Klassifizierungsmethoden, nur sehr wenig explizit trainiert.
- Die einzige Arbeit, die vor der Vorhersage erledigt werden muss, besteht darin, die Parameter für die einzelnen Wahrscheinlichkeitsverteilungen der Merkmale zu finden, was typischerweise schnell und deterministisch erledigt werden kann.
- Naive Bayes-Klassifikatoren können eine gute Leistung erbringen, auch bei höherdimensionalen Datenpunkten und / oder einer großen Anzahl von Datenpunkten.



Naive Bayes Klassifikation

Klassifikation:

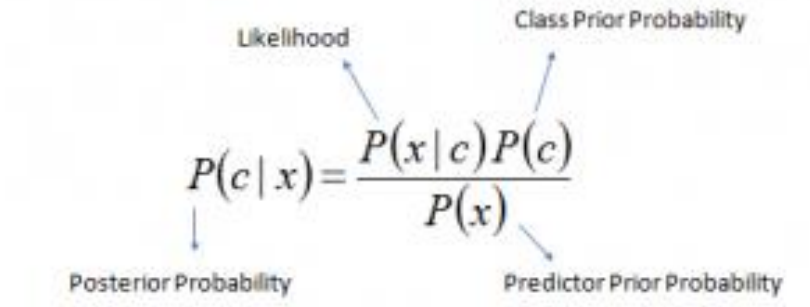
- Nachdem wir nun die Möglichkeit haben, die Wahrscheinlichkeit zu schätzen, dass ein bestimmter Datenpunkt in eine bestimmte Klasse fällt, müssen wir dies verwenden können, um Klassifizierungen zu erstellen.
- Naive Bayes handhabt dies auf sehr einfache Weise: Wähle das c_i mit der größten Wahrscheinlichkeit aus.

$$y = \underset{c_i}{argmax}\ P(c_i) \prod_{j=1}^{n} P(x_j|c_i)$$

- Diese Regel ist die **Maximum A Posteriori** Entscheidungsregel.
- Die Erklärung für diesen Namen beruht auf die Bayes Regel: wir benutzen nur P(B|A) und P(A), d.h. die Wahrscheinlichkeit und die vorherige Bedingungen.
- Falls wir nur P(B|A) benutzen, d.h. die bedingte Wahrscheinlichkeit, die Entscheidungsregel heißt **Maximum Likelihood** .

Naive Bayes Klassifikation: Beispiel

- Beachten Sie Folgendes: *Der Naive Bayes-Klassifikator ist ein Lernenverfahren, bei dem nur gezählt wird, wie oft jedes Merkmal mit jeder Klasse zusammen auftritt*



$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* c given *predictor* (*features*).
- $P(c)$ is the probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

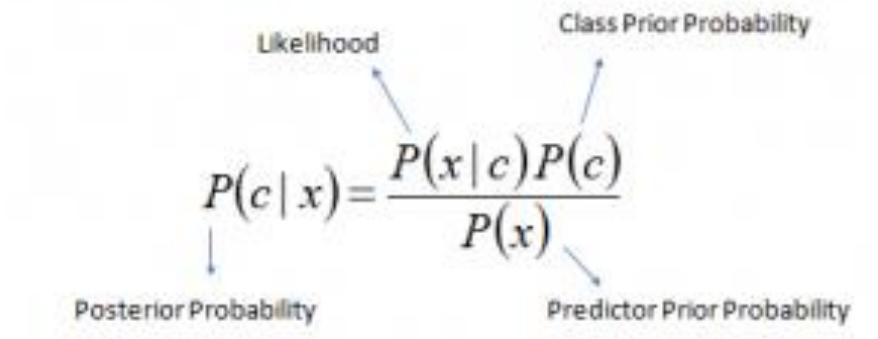
Naive Bayes Klassifikation: Beispiel

Fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

- 50% der Früchte sind Bananen
- 30% sind Orangen
- 20% sind andere Früchte
- Basierend auf unsere training Menge können wir folgendes sagen:
  - Aus 500 Bananen, 400 (0.8) sind vom Typ Long, 350 (0.7) sind Sweet und 450 (0.9) sind Yellow
  - Aus 300 Orangen, 0 sind Long, 150 (0.5) sind Sweet und 300 (1) sind Yellow
  - Aus der restlichen Menge von 200 Früchte, 100 (0.5) sind Long, 150 (0.75) sind Sweet und 50 (0.25) sind Yellow

Naive Bayes Klassifikation: Beispiel

- Gegeben die Merkmale einer Frucht, müssen wir die entsprechende Klasse bestimmen.
- Wenn uns gesagt wird, dass die zusätzliche Frucht lang, süß und gelb ist, können wir sie anhand der folgenden Formeln und Werte klassifizieren, unabhängig davon, ob es sich um eine Banane, eine Orange oder eine andere Frucht handelt.
- Diejenige Furcht mit der höchsten Wahrscheinlichkeit (Punktzahl) ist der Gewinner.



$$P(c|\mathbf{X}) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Naive Bayes Klassifikation: Beispiel

Banane:

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = \frac{P\left(\frac{Long}{Banana}\right) \times P\left(\frac{Sweet}{Banana}\right) \times P\left(\frac{Yellow}{Banana}\right) \times P(Banana)}{P(Long)\, P(Sweet)\, P(Yellow)}$$

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = \frac{(0.8) \times (0.7) \times (0.9) \times (0.5)}{0.25 \times 0.33 \times 0.41}$$

$$P\left(\frac{Banana}{Long, Sweet, Yellow}\right) = 0.252$$

Naive Bayes Classification: Example

Orange:

$$P(\frac{Orange}{Long, Sweet, Yellow}) = 0$$

Naive Bayes Classification: Example

Eine andere Frucht:

$$P(\frac{Other}{Long, Sweet, Yellow}) = \frac{P(\frac{Long}{Other}) \times P(\frac{Sweet}{Other}) \times P(\frac{Yellow}{Other}) \times P(Other)}{P(Long)\,P(Sweet)\,P(Yellow)}$$

$$P(\frac{Other}{Long, Sweet, Yellow}) = \frac{(0.5) \times (0.75) \times (0.25) \times (0.2)}{0.25 \times 0.33 \times 0.41}$$

$$P(\frac{Other}{Long, Sweet, Yellow}) = 0.01875$$

Naive Bayes Klassifikation: Beispiel

- In diesem Fall können wir basierend auf der hohen Punktzahl (0,252 für die Banane) annehmen, dass diese lange, süße und gelbe Frucht tatsächlich eine Banane ist.