Statistik zur Datenanalyse

Dr. Meike Wocken

Vorlesung 4

HS Bielefeld
Digitale Technologien (M.Sc.)
WiSe 2023/24

Meike.Wocken@codecentric.de

Bis jetzt: Supervised Verfahren

- Supervised Learning = Überwachtes Lernen

- Zielgröße ist bekannt
- Aufwendige Datenaquise

Unsupervised Learning

Clusterverfahren

- Unsupervised
- Deskriptiv
- Keine Rückschlüsse von Stichprobe auf Grundgesamtheit möglich

<u>Ziel:</u>

<u>innerhalb</u> eines Clusters Beobachtungen mit <u>starken Ähnlichkeiten</u> sammeln,

<u>außerhalb</u> eines Cluster Beobachtungen mit <u>großen Unterschieden</u> sammeln.

Clusterverfahren

Bereits bekannte Verfahren:

- Kmeans
- Hierarchisches Clustern

Jetzt:

<u>Modell-basiertes Clustern </u>(Finite Mixture models)

Hier im Bild:

- Zwei sich überlappende Normalverteilungen
- Jedes Cluster hat eine eigene Verteilung

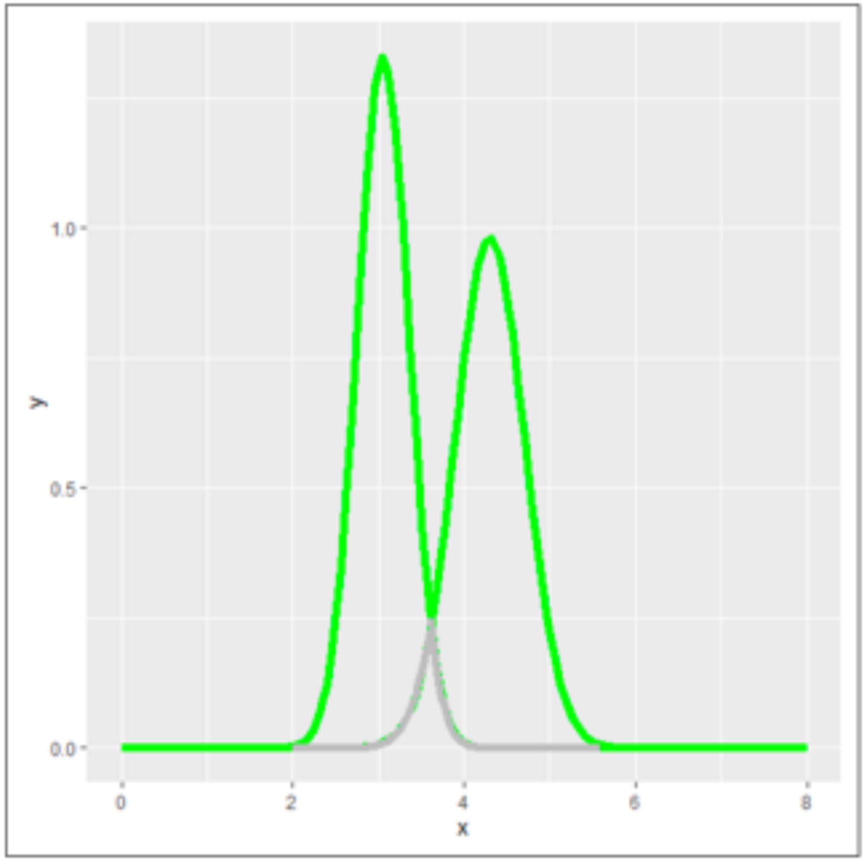


Abbildung 39: Mischmodell bestehend aus zwei Normalverteilungen.

Clusterverfahren

$$p(X) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

$\pi_k = p(z_k = 1)$ mit $\sum_{k=1}^{K} \pi_k = 1$

z_k ist eine latente (nicht beobachtbare) Variable. Für jede Beobachtung gilt, dass z_k für nur genau ein Cluster gleich 1 ist, sonst 0 (binär). Sie gibt an, zu welchem Cluster die jeweilige Beobachtung gehört.

Σ_k ist Varianz-Kovarianz-Matrix im multivariaten Fall, z.B.

$$\Sigma_k = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}, \mu_k = \begin{pmatrix} E(X) \\ E(Y) \end{pmatrix}$$

Log-Likelihood

- Für modell-basiertes Clustern sind nun die Vektoren π, μ und Matrix Σ zu bestimmen.
- Dafür wird die Log-Likelihood aufgestellt:

$$\ln p(x|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln(\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k))$$

- Maximierung der log-likelihood komplex, daher EM-Algorithmus, dafür wird Satz von Bayes verwendet.

Verwendung Satz von Bayes

Prior Information

$$\boxed{p(Z_k = 1|x)} = \frac{p(Z_k = 1)p(x|Z_k = 1)}{\sum_{j=1}^{K} p(Z_j = 1)p(x|Z_j = 1)} = \frac{\boxed{\pi_k} N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

Posterior Information:
Responsibility, das Cluster k die Beobachtung x beschreibt

EM Algorithmus:

Initialisierung: freie Wahl (oder kmeans, hierarchisches Clustern) von μ, Σ, π

Expectation: Ermittlung erwartete posterior WSK mit festgelegten Parametern

Maximization: Benutzung posterior WSK, um μ, Σ, π zu bestimmen, so dass Bedingung für Maximum der Log-Likelihood erfüllt sind

EM Algorithmus

Bestimmt durch Parameter

- Gemeinsame Verteilung $p(X, Z|\theta)$

Beobachtbar Latente Variable

- Ziel: maximiere Likelihood $p(X|\theta)$ w.r.t. θ

1) Initiale Parameter θ^{old} wählen

2) Expect/Evaluate: $p\left(X \middle| \theta^{old}\right)$

3) Maximize: $\theta^{New} = \underset{\theta}{max} L\left(\theta, \theta^{old}\right)$

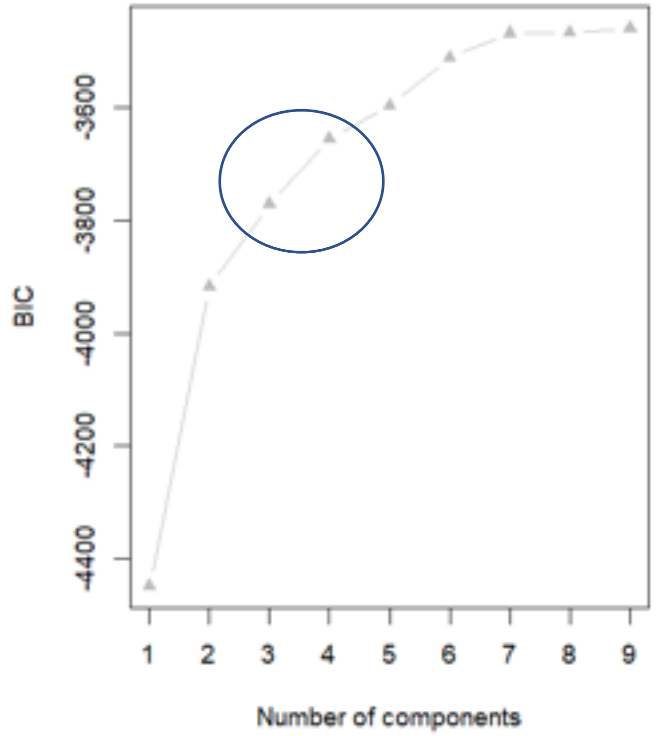
mit $L\left(\theta, \theta^{old}\right) = \sum_Z p(Z|X, \theta^{old}) \cdot \ln p(X, Z|\theta)$

Ergebnis EM Algorithmus

- Wir erhalten monotone Konvergenz gegen lokales Maximum, daher wichtig, mit unterschiedlichen Startwerten zu arbeiten
- EM Algorithmus konvergiert langsam, ist aber auch sehr flexibel
- Ergebnis: *softe* Clusterzuordnung durch Angabe einer Wahrscheinlichkeit einer Clusterzugehörigkeit

Wahl der Cluster-Anzahl K

- Verwendung Log-Likelihod (je größer, je besser)
- Ableitung Informationskriterium, z.B. BIC das für Anzahl geschätzter Parameter/Anzahl der Cluster korrigiert
- Grafische Darstellung der Werte über die Anzahl der Cluster
- Wahl der Clusteranzahl „am Knick“: nur noch geringere Verbesserung bei Hinzunahme weiterer Cluster
- In R: Scree-Plot mit TSS-within
- Standardisierung von Variablen (mean = 0, var =1) hilft manchmal, Cluster leichter zu ermitteln



Wahl nicht immer eindeutig

Anwendungsbereich für finite mixture models als Clusterverfahren

- Überlappende Cluster
- „softe“ Clusterzuordnung durch Wahrscheinlichkeiten

<u>Wichtig und allgemein gültig für Clusterverfahren:</u>

Clusterverfahren sind immer NUR beschreibend!

Auf Clustergrößen (Anzahl an Beobachtungen je Cluster) achten. Zu kleine Cluster zu haben macht u.U. keinen Sinn (z.B. nur eine Beobachtung in einem Cluster).

Aufgabe: siehe R Code

2. **cluster_example.csv**

   a) Lesen Sie den Datensatz cluster_example.csv ein. Stellen Sie die Daten grafisch dar.

   b) Ermitteln Sie die Cluster mit dem EM-Algorithmus. Welche Cluster-Anzahl würden Sie wählen?

Bayessche Netze

Kurze Wiederholung Wahrscheinlichkeiten

- UND Verknüpfung von Ereignissen $p(E1 \cdot E2) = p(E2|E1)p(E1)$
- Unabhängigkeit $p(E1 \cdot E2) = p(E2)p(E1)$
- Totale Wahrscheinlichkeit

$$P(A) = \sum_{j=1}^{n} P(E_j)p(A|E_j)$$

Mit $\sum_{j=1}^{n} P(E_j) = 1$ und für alle $j \neq k$ gilt $P(E_j \cdot E_k) = 0$ (gegenseitiges Ausschließen der Ereignisse)

- Satz von Bayes $P(E_j|A) = \frac{P(E_j)P(A|E_j)}{P(A)}$
- Grenzen von Naive Bayes Klassifikatoren: bedingte Unabhängigkeit sehr restriktiv
- Unsupervised Learning:
  - Vielzahl an Zielgrößen (z.B. Fehler im Drucker, Diagnosen)
  - Unsichere Informationen
  - Fehlende Informationen

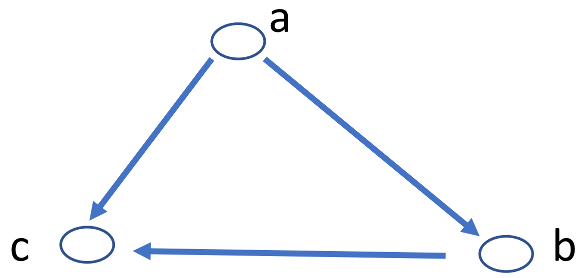
Bayessche Netze: Modelle zur Umsetzung von lernenden Systemen

- Was ist ein Bayessches Netz?
  - Gerichteter azyklischer Graph
  - Annotation dieses Graphen mit Wahrscheinlichkeiten, deren Kombination eine Verteilung definiert
  - Knoten sind Variablen
  - Kanten sind Wahrscheinlichkeitsbeziehungen

Gemeinsame Wahrscheinlichkeitsverteilung

$$p(a, b, c) = p(c|a, b) \cdot p(a \cdot b) = p(c|a, b) \cdot p(b|a) \cdot p(a)$$

(Kettenregel)



Zerlegung hätte auch anders aussehen können (z.B. $p(a|b, c) \cdot p(b|c) \cdot p(c)$), dann wäre es aber eine andere grafische Darstellung.

Gemeinsame Wahrscheinlichkeitsverteilung

Die gemeinsame Verteilung definiert durch einen Graphen (Bayessches Netz). Gemeinsame Wahrscheinlichkeitsverteilung ist gegeben durch Produkt über alle Knoten vom Graphen und den bedingten Verteilungen für jeden Knoten in Abhängigkeiten zu den Eltern des jeweiligen Knotens $X = \{x_1, \dots, x_K\}$

$$p(X) = \prod_{k=1}^{K} p(x_k | pa_k)$$

mit pa_k ist Set an Eltern (parents) für x_k. Der Graph hat K Knoten.

Anwendung Bayessche Netze

- Abbildung Expertenwissen
- Regelbasierte Systeme mit Unsicherheiten erweitern

Lernen mit Bayesschen Netzen

- Netztopologie/Struktur
- Parameter für jeden Knoten (bedingte Wahrscheinlichkeitstabelle für jede Variable)

Hier:

- Annahme diskrete Variablen
- Experte gibt Netzstruktur vor (sonst z.B. über Greedy Search)
- Trainingsdaten vorhanden

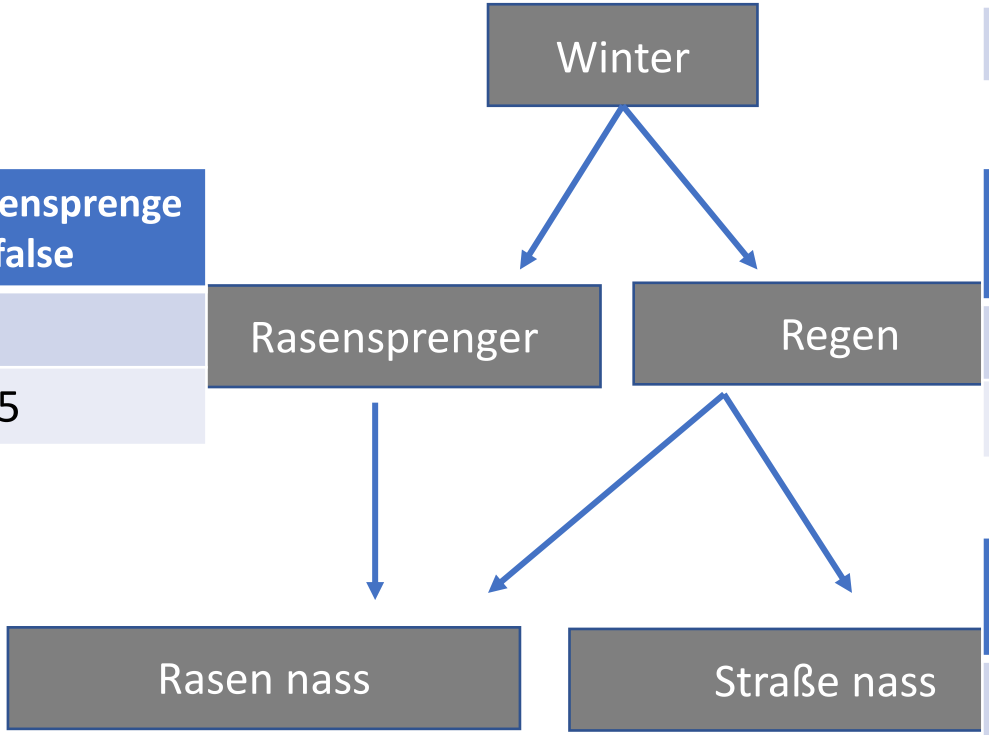
Dann Maximum-Likelihood Schätzung möglich:

$$\theta_{i,j,k} = p(x_i = k | pa_i = j) = \frac{N_{ijk}}{\sum_k N_{ijk}}$$

Beispiel

Winter = true	Winter = false
0,6	0,4

Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25

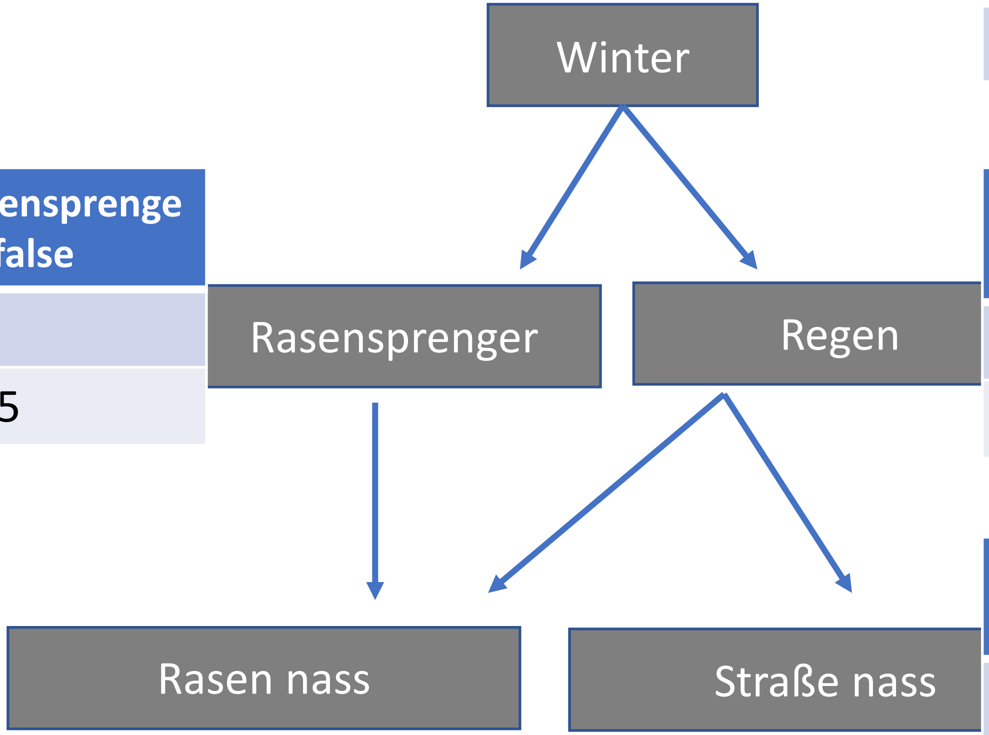


Winter	Regen = true	Regen = false
true	0,8	0,2
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Beispiel



Winter = true	Winter = false
0,6	0,4

Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25

Winter	Regen = true	Regen = false
true	0,8	0,8
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

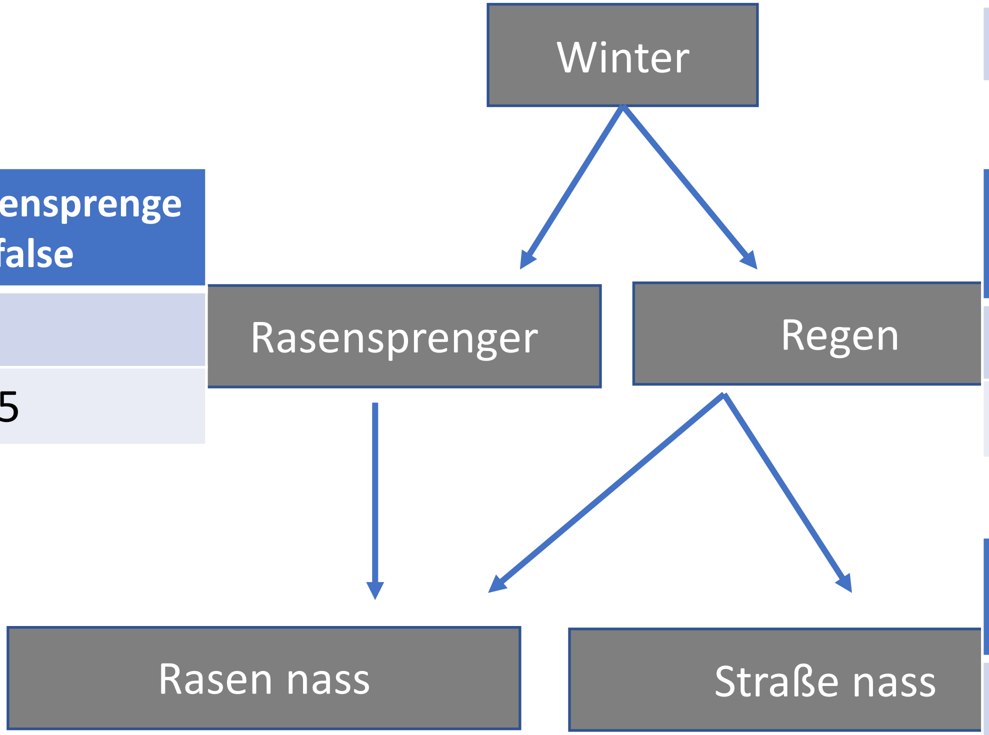
Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Gemeinsame Verteilung:
P(**Winter** und **kein Rasensprenger** und **Regen** und **Rasen nass** und **Straße nass**)?

Beispiel

Winter = true	Winter = false
0,6	0,4

Winter	Rasensprenger = true	Rasensprenger = false
true	0,2	0,8
false	0,75	0,25



Winter	Regen = true	Regen = false
true	0,8	0,8
false	0,1	0,9

Regen	Straße nass = true	Straße nass = false
true	0,7	0,3
false	0	1

Rasensprenger	Regen	Rasen nass = true	Rasen nass = false
true	true	0,95	0,05
true	false	0,9	0,1
false	true	0,8	0,2
false	false	0	1

Gemeinsame Verteilung:
P(**Winter** und **kein Rasensprenger** und **Regen** und **Rasen nass** und **Straße nass**) =
0,6 * 0,8 * 0,8 * 0,7 * 0,8 = 0,21504

D-Seperation

- Bedingte Unabhängigkeit ist wichtig in Bayesschen Netzen, da es Faktorisierung von Verteilungen vereinfacht.
- Z.B. a ist unabhängig von b gegeben c, wenn gilt

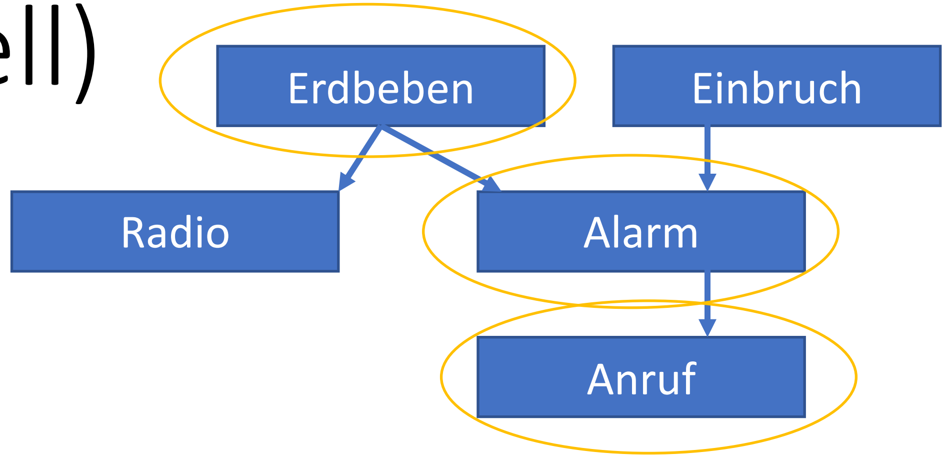
$$p(a|bc) = p(a|c)$$

d.h. a und b sind statistisch unabhängig gegeben c: $a \perp b \mid c$

$$p(ab|c) = p(a|bc) \cdot p(b|c) = p(a|c) \cdot p(b|c)$$

- Unabhängigkeit kann in grafischer Weise charakterisiert/abgelesen werden!

Head-to-tail (Sequentiell)

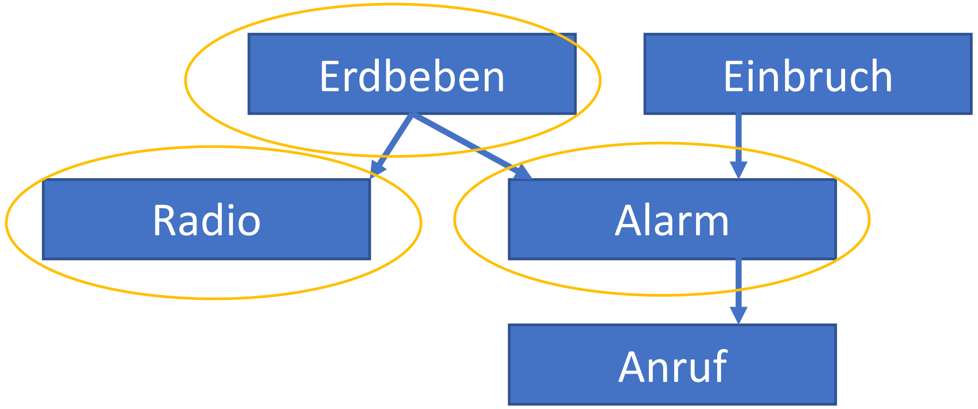


Unabhängigkeit von Knoten, wenn sie unterbrochen sind:

Knoten auf Pfad als Ventil

Wenn Alarm beobachtet, dann Erdbeben und Anruf unabhängig

Tail-to-tail



2 Fälle:

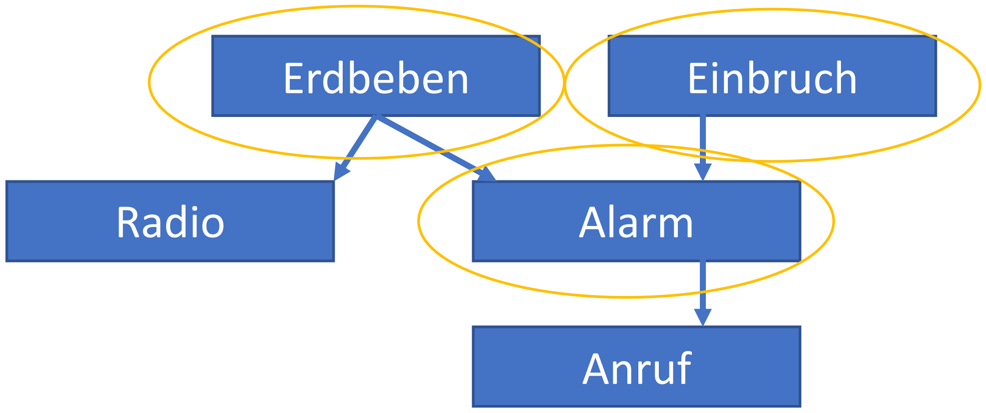
1. Wert von Erdbeben unbekannt

Alarm und Radio sind abhängig, Radionachricht über Erdbeben erhöht Wahrscheinlichkeit eines Alarms

2. Wert von Erdbeben bekannt

Radio und Alarm unabhängig

Head-to-head



2 Fälle:

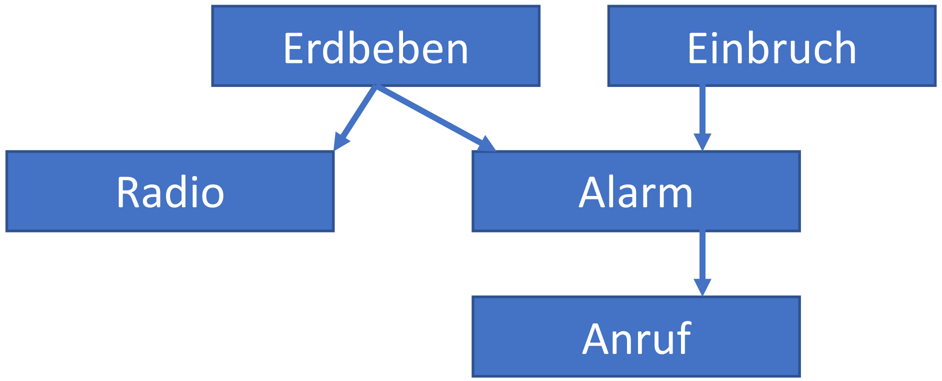
1. Wert von Alarm und Anruf unbekannt

Erdbeben und Einbruch sind unabhängig

2. Wert von Alarm bekannt

Abhängigkeit von Erdbeben und Einbruch, da Erdbeben die Wahrscheinlichkeit eines Einbruchs verringert

Explaining Away



- Erdbeben ist unbekannt
- Radio und Anruf sind abhängig (Tail-to-Tail)
- Es gilt dann:

$$P(\text{Einbruch}|\text{Anruf}) > P(\text{Einbruch}|\text{Anruf, Radio})$$

- Radiomeldung über Erdbeben erklärt einen Alarm, daher wird Einbruch unwahrscheinlicher. (Zahlenbeispiel dazu im Skript)

Inferenz

- Alarm, wie wahrscheinlich ist ein Einbruch?
  Einbruch E → Alarm A:

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

- Anruf, wie wahrscheinlich ist ein Einbruch?
  Einbruch E→ Alarm A → (Telefon-)Anruf T:

$$P(T) = P(E \cdot T) + P(kein\ E \cdot T)$$

$$P(E|T) = \frac{P(E \cdot T)}{P(T)}$$

Aufgabe Bayessche Netze

Wichtig! Ausgefüllte Variablen stehen für beobachtete Werte!

12.3 Aufgaben

Aufgabe 1: Beurteilen Sie die Bayesschen Netze in Abbildung 50 hinsichtlich der Abhängigkeiten der Variablen a und b.

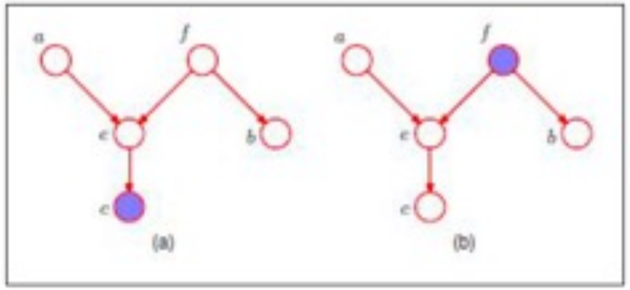


Abbildung 50: Beispiel Head to Head.

Lösung

12.3 Aufgaben

Aufgabe 1: Beurteilen Sie die Bayesschen Netze in Abbildung $\boxed{50}$ hinsichtlich der Abhängigkeiten der Variablen a und b.

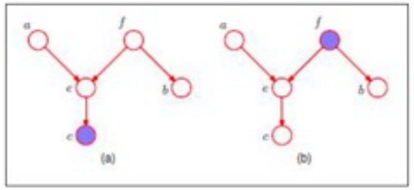


Abbildung 50: Beispiel Head to Head.

a)

Zwischen a und b sind:

f: tail-to-tail

e: head-to-head mit Nachfolger, der beobachtet wurde (c)

Damit sind a und b gegeben c Abhängig.

Lösung

12.3 Aufgaben

Aufgabe 1: Beurteilen Sie die Bayesschen Netze in Abbildung 50 hinsichtlich der Abhängigkeiten der Variablen a und b.

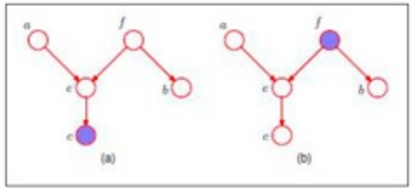


Abbildung 50: Beispiel Head to Head.

b)

Gleicher Fall wie a) nur f ist anstatt c beobachtet

a und b ist von f blockiert: tail-to-tail und f bekannt, daher a und b unabhängig bedingt f

Analog mit unbeobachtetem c:

e blockiert a und b, da es Head-to-Head ist, wo e und c (Nachfolger) nicht beobachtet werden