02. Juni 2017

Expectation Maximization

Problem:

Lernen der Parameter einer Verteilung aus Trainingsdaten $\mathcal{D}$ mit fehlenden oder schlechten Features. Die Daten werden eingeteilt in gute ($\mathcal{D}_g$) und schlechte ($\mathcal{D}_b$)

Beispiel: Punkte aus Gaußverteilung

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}, \mathcal{D}_b = x_{41}$$

Für eine Maximum-Likelihood-Schätzung der Verteilungsparameter $\boldsymbol{\theta}$ suchen wir das $\boldsymbol{\theta}$, das die likelihood $p\left(\mathcal{D}_g \mid \boldsymbol{\theta}\right)$ maximiert:

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} p\left(\mathcal{D}_g \mid \boldsymbol{\theta}\right) = \arg\max_{\boldsymbol{\theta}} \sum_{\mathcal{D}_b} p\left(\mathcal{D}_g, \mathcal{D}_b \mid \boldsymbol{\theta}\right)$$

Falls $\mathcal{D}_b$ kontinuierlich ist, muss natürlich bei der Marginalisierung die Summe durch ein Integral ersetzt werden.

In der Praxis ist die Berechnung von $\sum_{\mathcal{D}_b} p\left(\mathcal{D}_g, \mathcal{D}_b \mid \boldsymbol{\theta}\right)$ oft schwierig, oder nicht möglich, z.B. aufgrund der vielen Möglichkeiten für $\mathcal{D}_b$ (exponentiell in der Anzahl der Dimensionen). Es kann aber gezeigt werden, dass $p\left(\mathcal{D}_g, \mathcal{D}_b \mid \boldsymbol{\theta}\right)$ auch mit einer Hilfsfunktion $Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^i\right)$ optimiert werden kann.

Expectation Maximization

Wir stellen folgende Funktion auf:

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^i\right) = E_{\mathcal{D}_b|\mathcal{D}_g,\boldsymbol{\theta}^i}\left[\ln p\left(\mathcal{D}_g,\mathcal{D}_b \mid \boldsymbol{\theta}\right)\right]$$

$$= \sum_{\mathcal{D}_b} p\left(\mathcal{D}_b \mid \mathcal{D}_g,\boldsymbol{\theta}^i\right)\ln p\left(\mathcal{D}_g,\mathcal{D}_b \mid \boldsymbol{\theta}\right)$$

Dabei ist $\boldsymbol{\theta}^i$ die zur Zeit beste Schätzung für die Verteilungsparameter.

$\boldsymbol{\theta}$ ist ein Kandidatenvektor für eine verbesserte Schätzung.

$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^i\right)$ gibt uns den Erwartungswert für die log-Likelihood für die Daten an, wobei zu dessen Berechnung die schlechten Daten als mit der Verteilung $p\left(\mathbf{x} \mid \boldsymbol{\theta}^i\right)$ verteilt angenommen werden.

Expectation Maximization

Der Expectation Maximization Algorithmus läuft nun in zwei abwechselnd durchgeführten Schritten ab:

E-Step: Berechne $p\left(\mathcal{D}_b \mid \mathcal{D}_g, \boldsymbol{\theta}^i\right)$ und damit

$$Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^i\right) = E_{\mathcal{D}_b \mid \mathcal{D}_g, \boldsymbol{\theta}^i}\left[\ln p\left(\mathcal{D}_g, \mathcal{D}_b \mid \boldsymbol{\theta}\right)\right]$$

M-Step: Berechne $\boldsymbol{\theta}^{i+1}$ als $\arg\max_{\boldsymbol{\theta}} Q\left(\boldsymbol{\theta}; \boldsymbol{\theta}^i\right)$

Diese Vorgehensweise, bei der für die fehlenden Daten eine Wahrscheinlichkeitsverteilung angenommen wird, wird als Soft Expectation Maximization bezeichnet.

Expectation Maximization

Alternativ können statt einer Verteilung im E-Step auch die aktuell wahrscheinlichsten fehlenden Daten berechnet werden. Dann gilt:

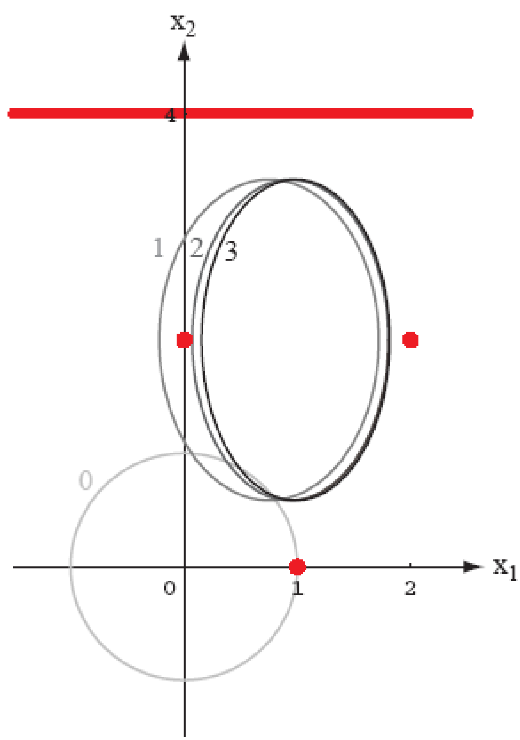
E-Step: Berechne $\mathcal{D}_{b_i} = \arg\max_{\mathcal{D}_b} p\left(\mathcal{D}_g, \mathcal{D}_b \mid \boldsymbol{\theta}^i\right)$

M-Step: Berechne $\boldsymbol{\theta}^{i+1}$ als $\arg\max_{\boldsymbol{\theta}} p\left(\mathcal{D}_g, \mathcal{D}_{b_i} \mid \boldsymbol{\theta}\right)$

Diese Vorgehensweise wird als Hard Expectation Maximization bezeichnet.

Beispiel: Finden einer 2D-Normalverteilung (achsenparallel) mit den angegebenen Daten:

$$\mathcal{D} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}$$



Expectation Maximization

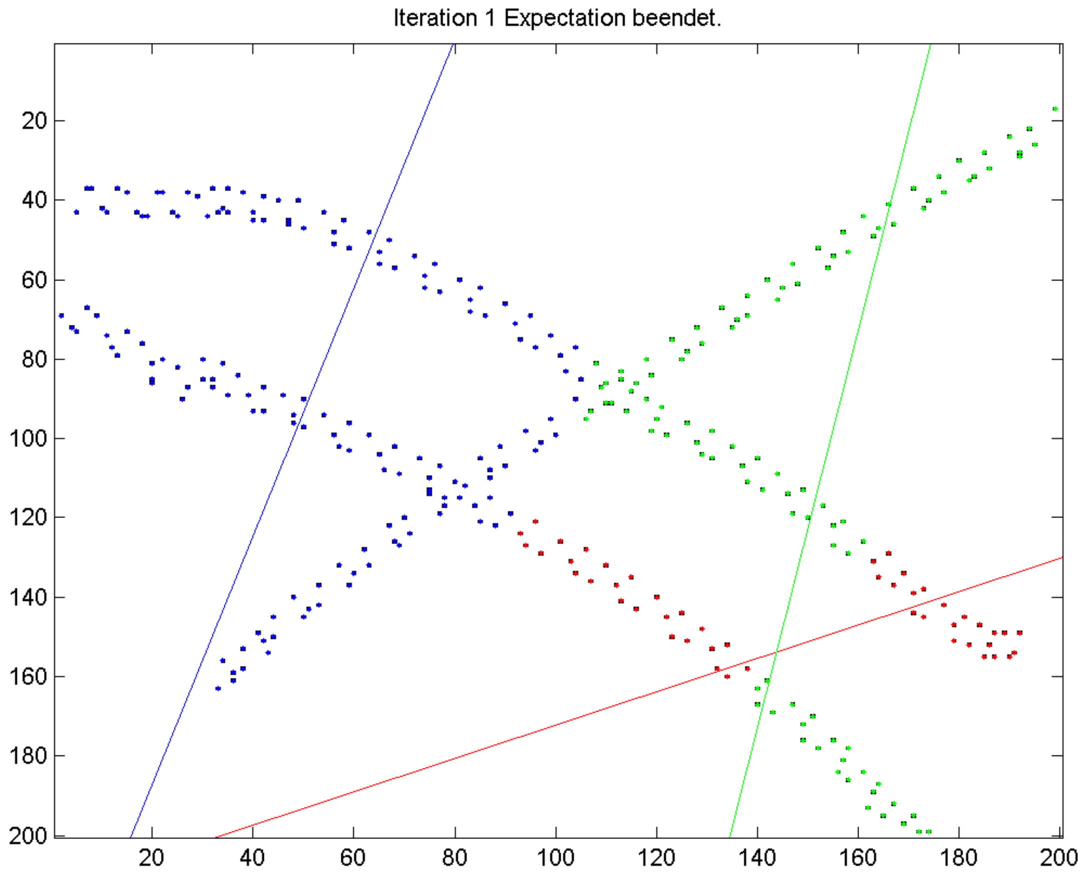
Generalized Expectation Maximization (GEM):

- Verbessere θ^i in jedem Schritt
- nicht unbedingt optimale Lösung
- Eine Version ist Hard EM, wo in jedem Schritt wahrscheinlichste Werte für die unbekannten Features geschätzt werden.
  - Beispiel [Egorova 02]:
    - Gesucht: Beste drei Geraden
    - Gegebene Daten: Punkte
    - Fehlende Daten: welcher Punkt gehört zu welcher Geraden

[Egorova 02] http://www.cims.nyu.edu/~fischer/SemBayNet/, Vortrag v. 5.12.02 (Link nicht mehr erreichbar)

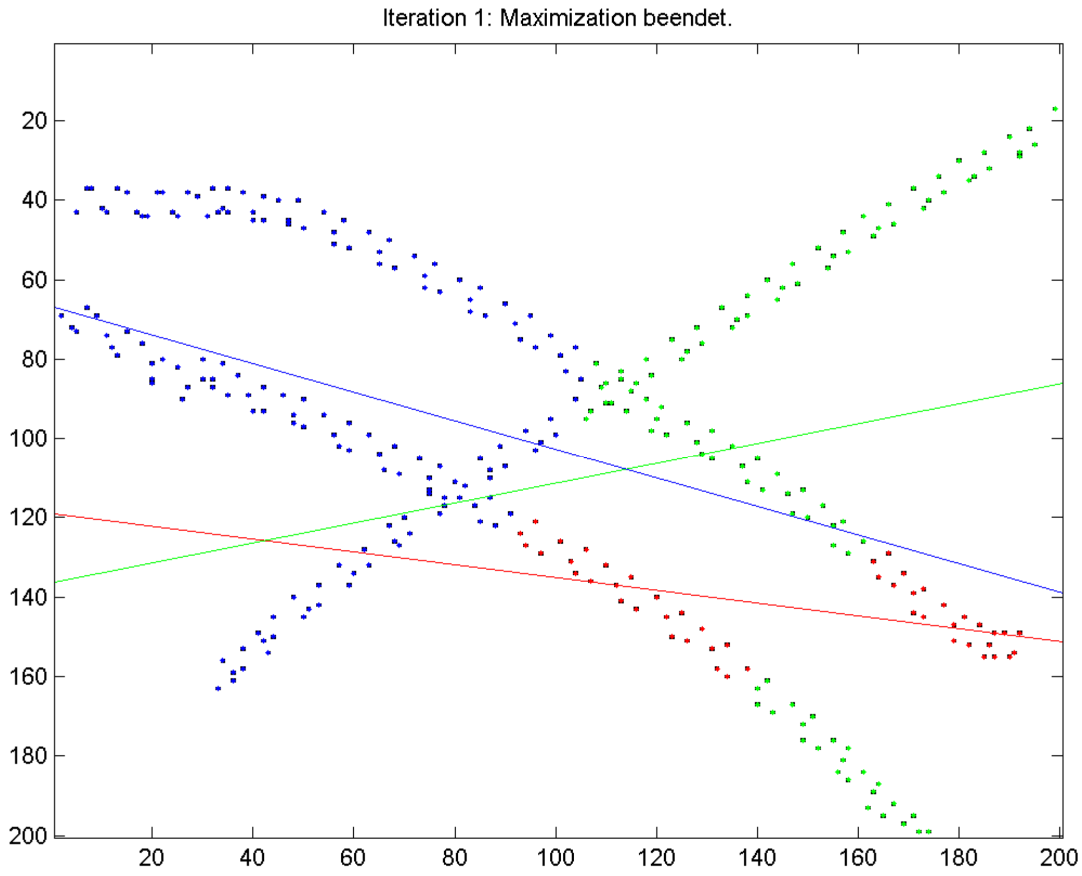
3 Geraden – schönes Beispiel

02. Juni 2017



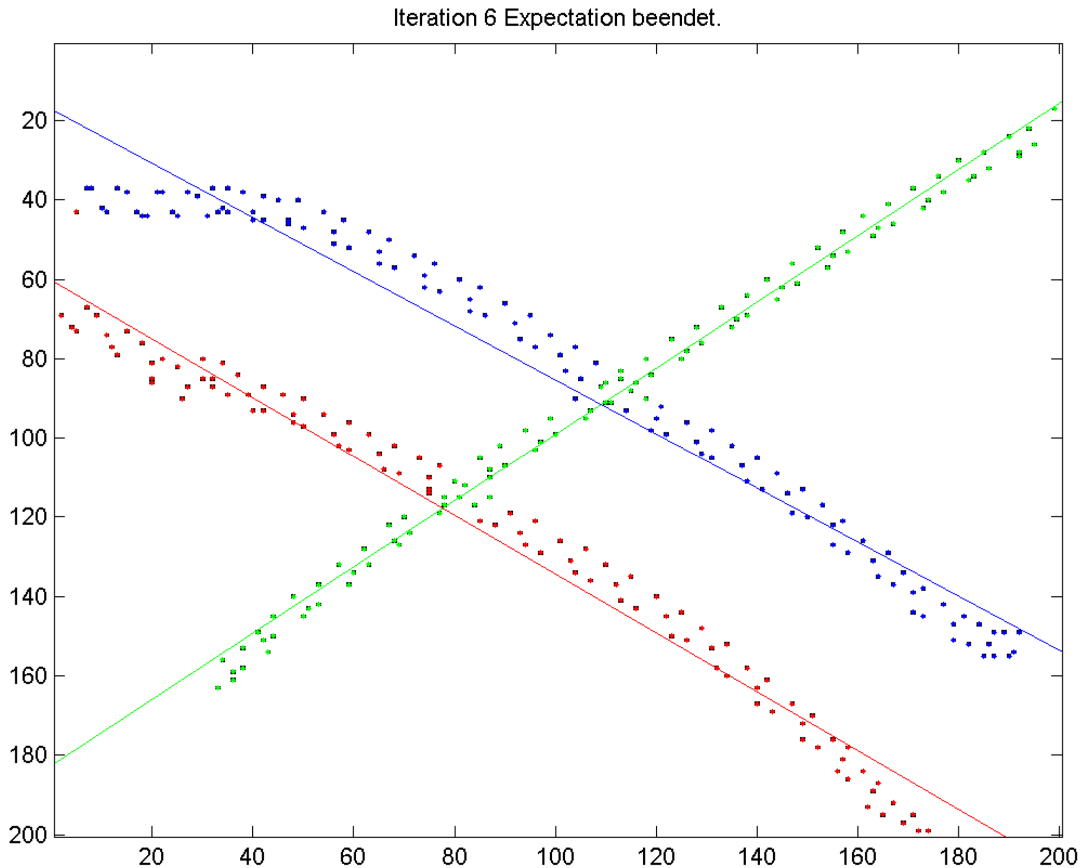
[Egorova 02]

3 Geraden – schönes Beispiel



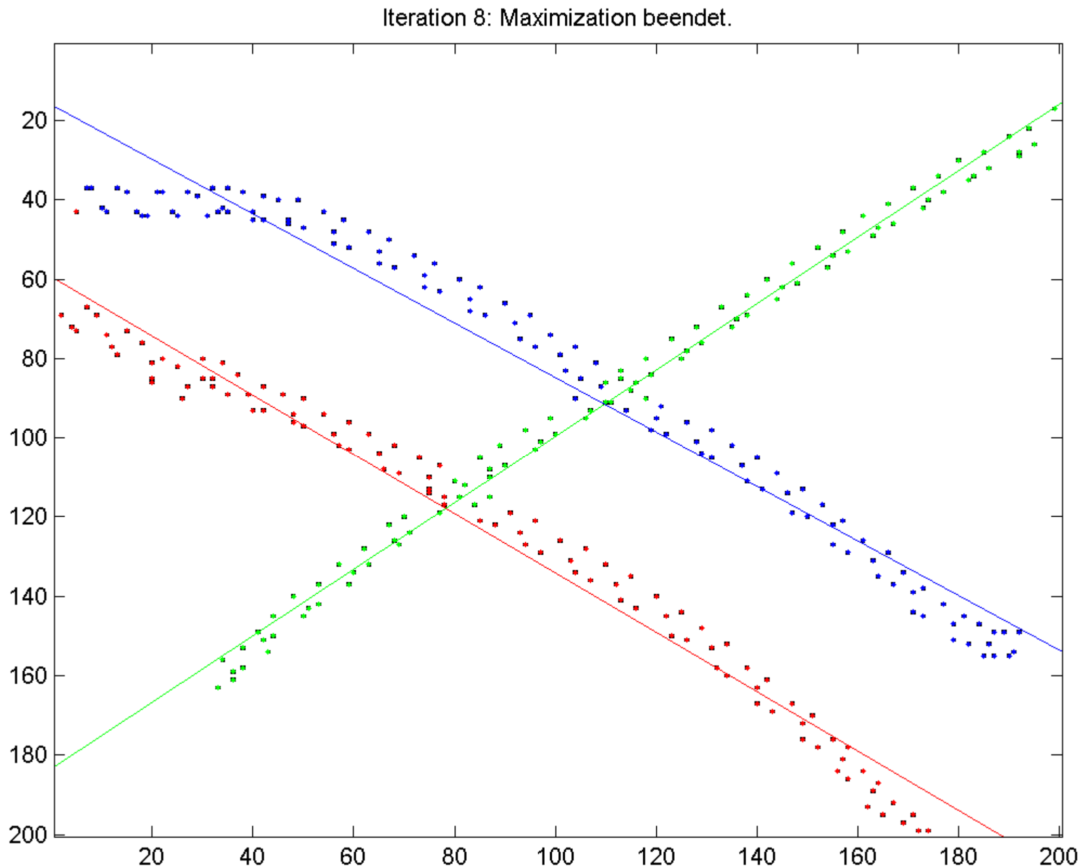
[Egorova 02]

3 Geraden – schönes Beispiel



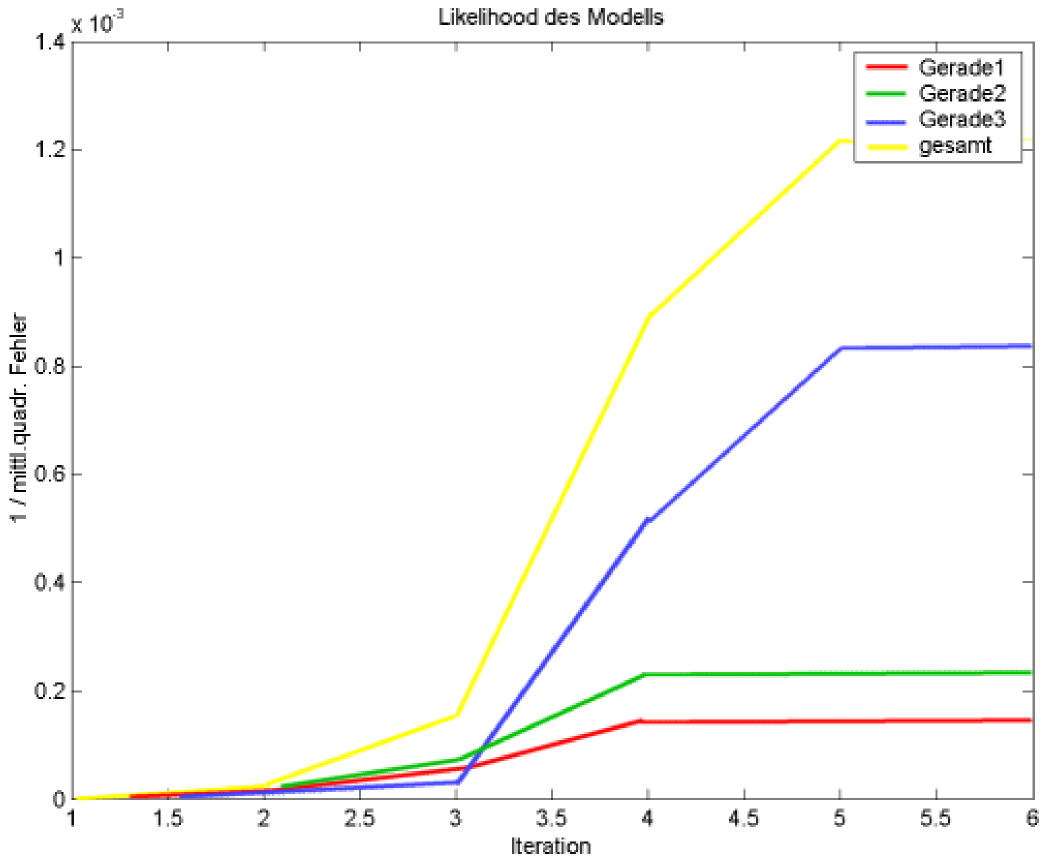
[Egorova 02]

3 Geraden – schönes Beispiel



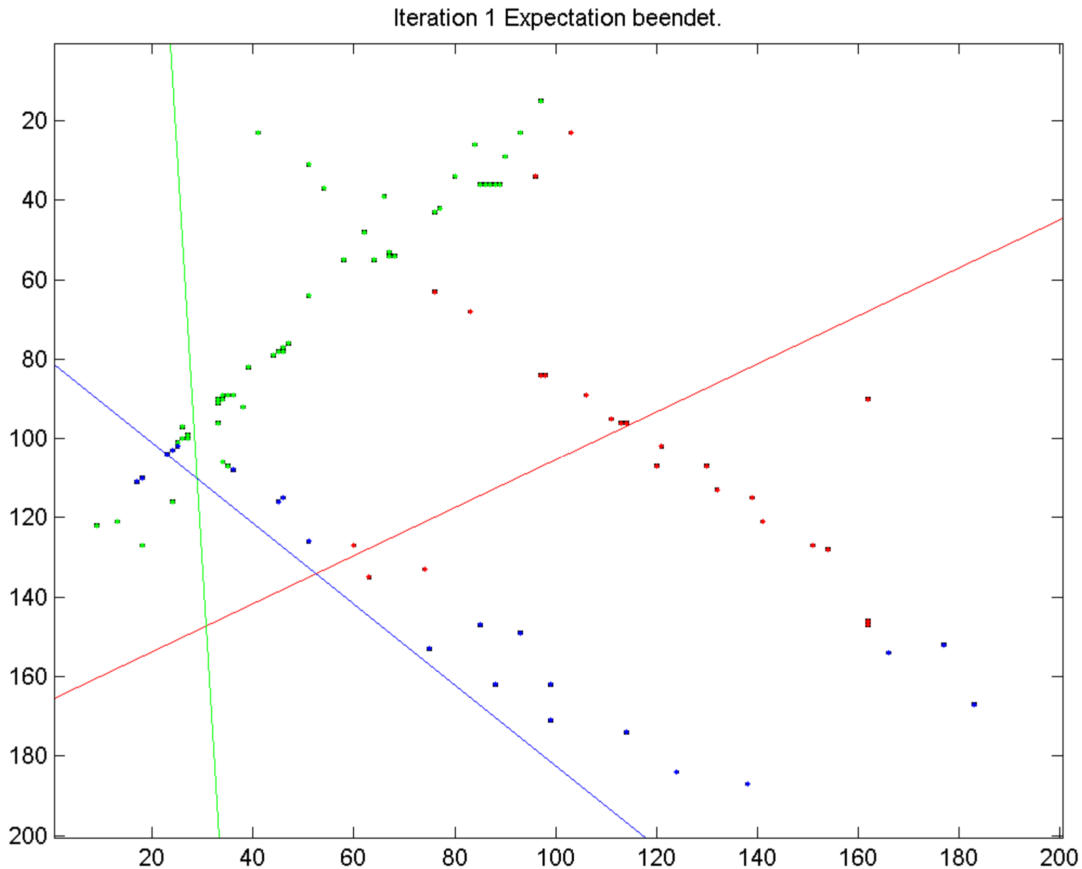
[Egorova 02]

3 Geraden – schönes Beispiel



[Egorova 02]

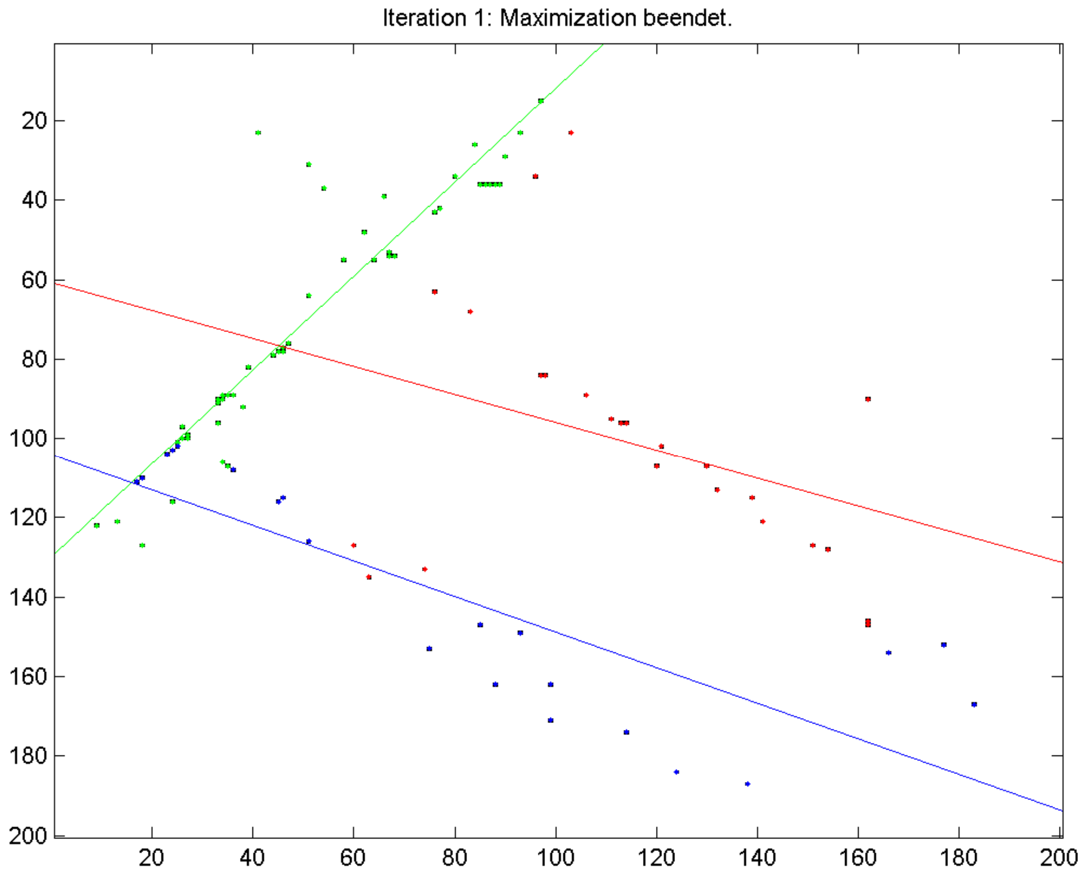
3 Geraden – noch ein Beispiel



[Egorova 02]

3 Geraden – noch ein Beispiel

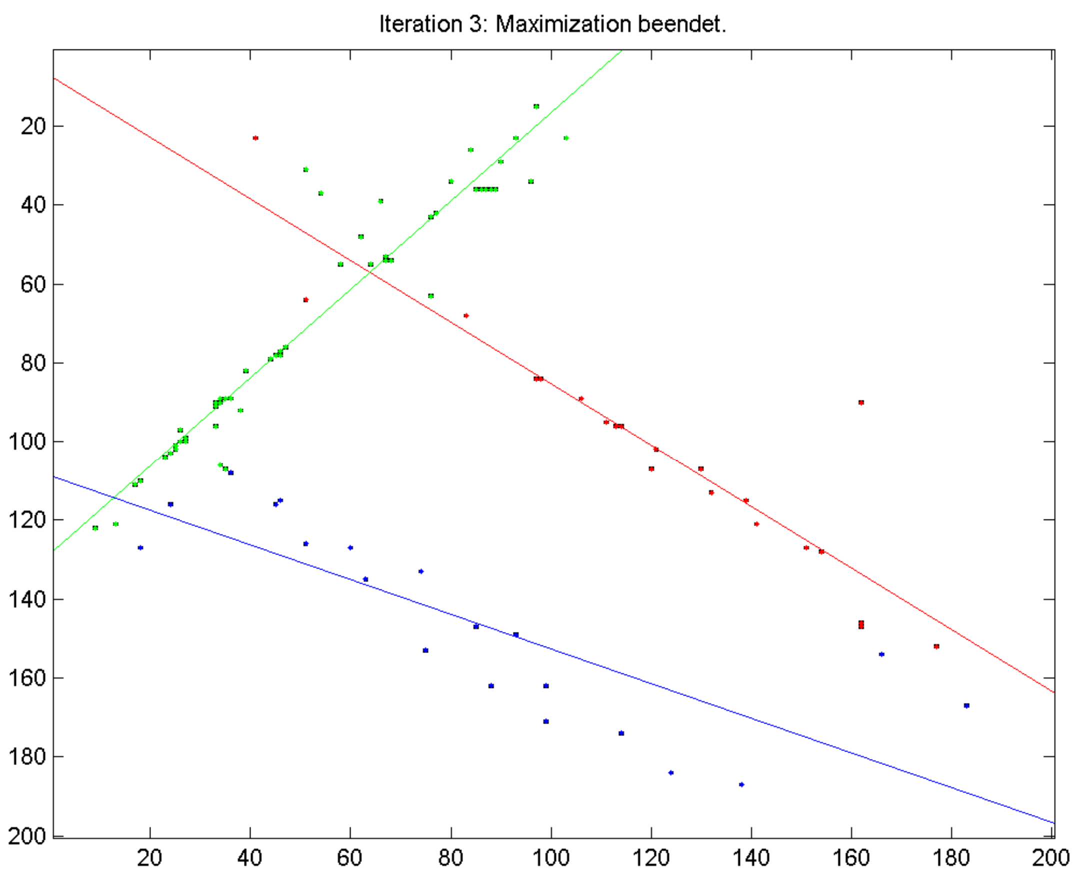
02. Juni 2017



[Egorova 02]

3 Geraden – noch ein Beispiel

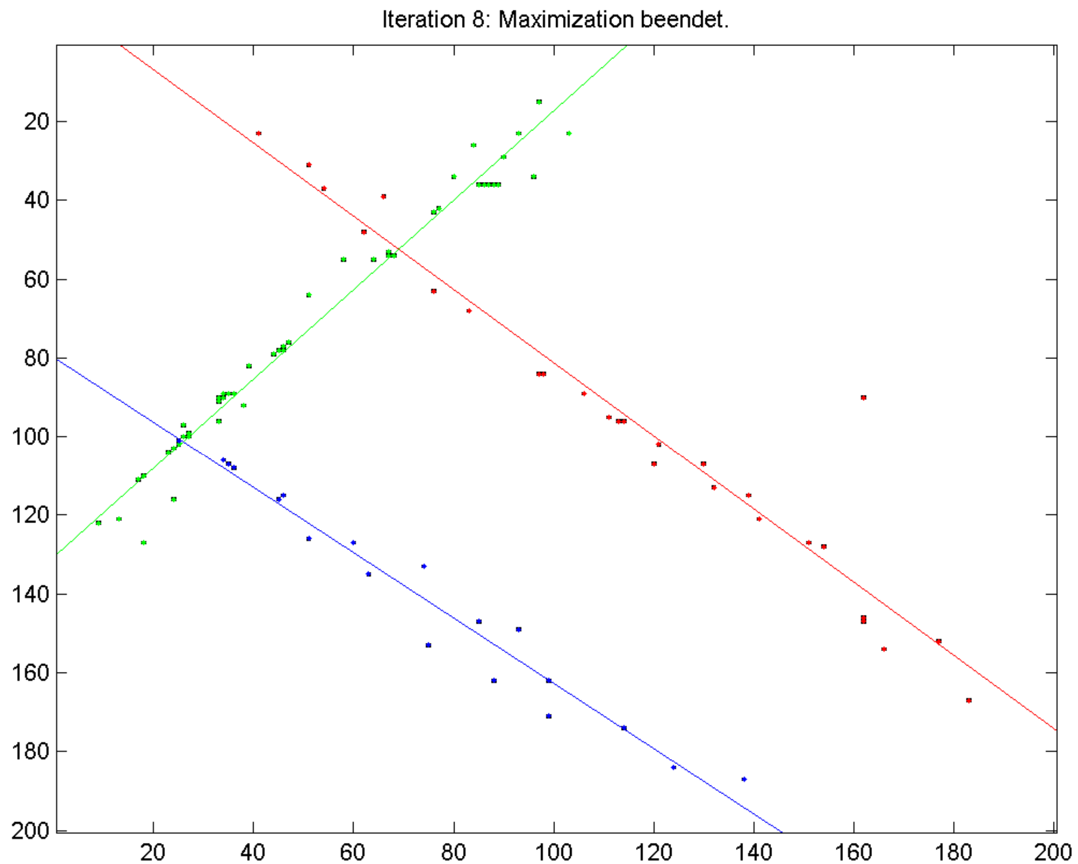
02. Juni 2017



[Egorova 02]

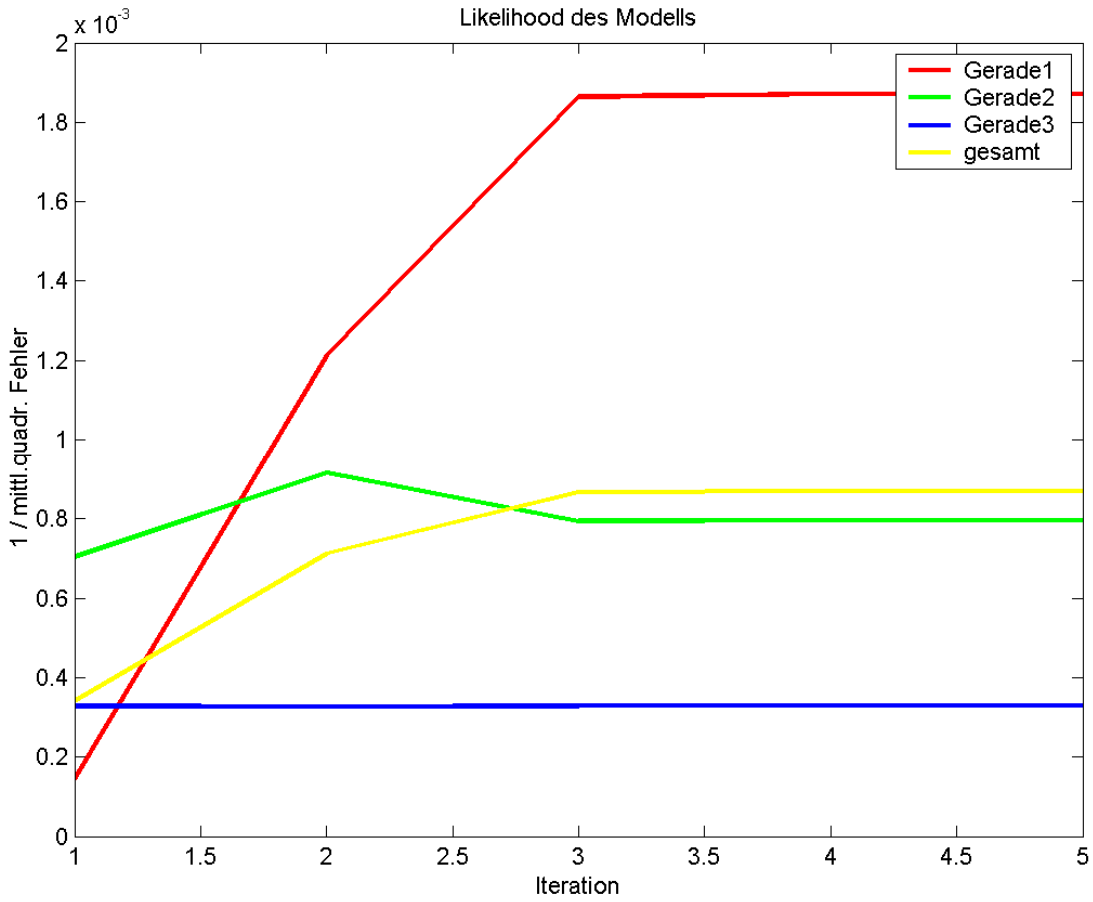
3 Geraden – noch ein Beispiel

02. Juni 2017



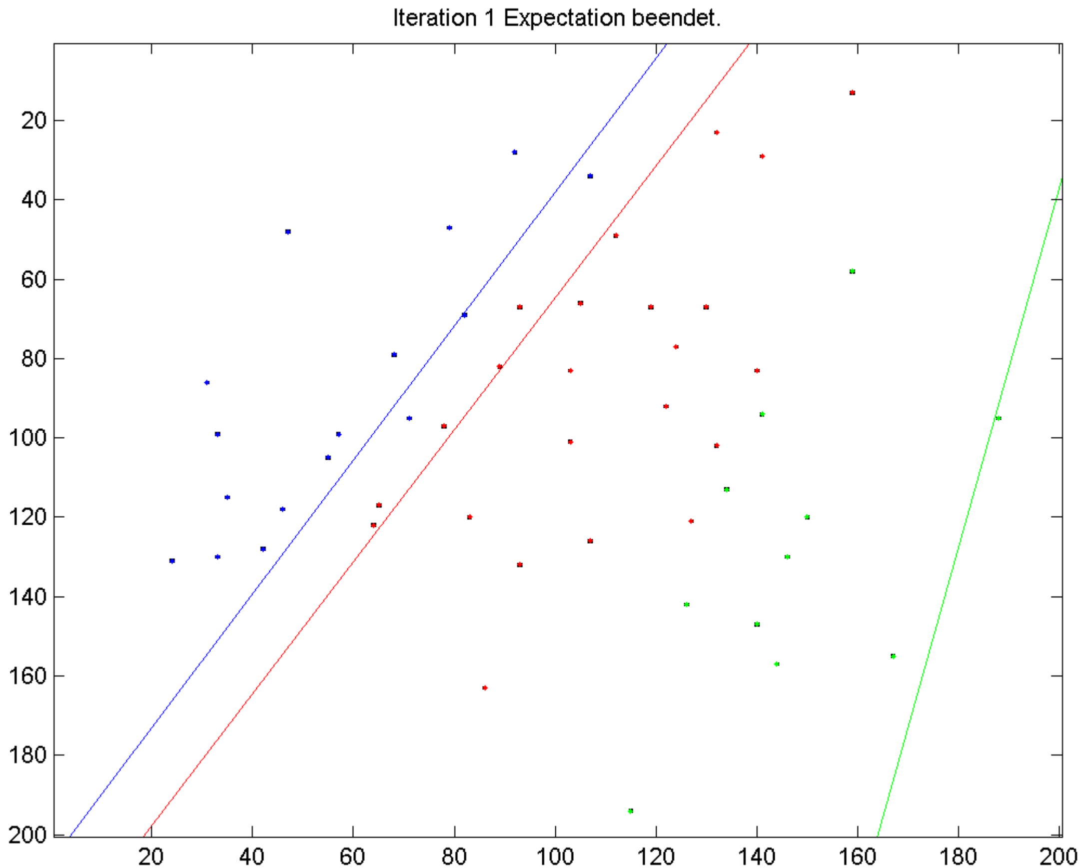
[Egorova 02]

3 Geraden – noch ein Beispiel



[Egorova 02]

drei Geraden – falsches Modell

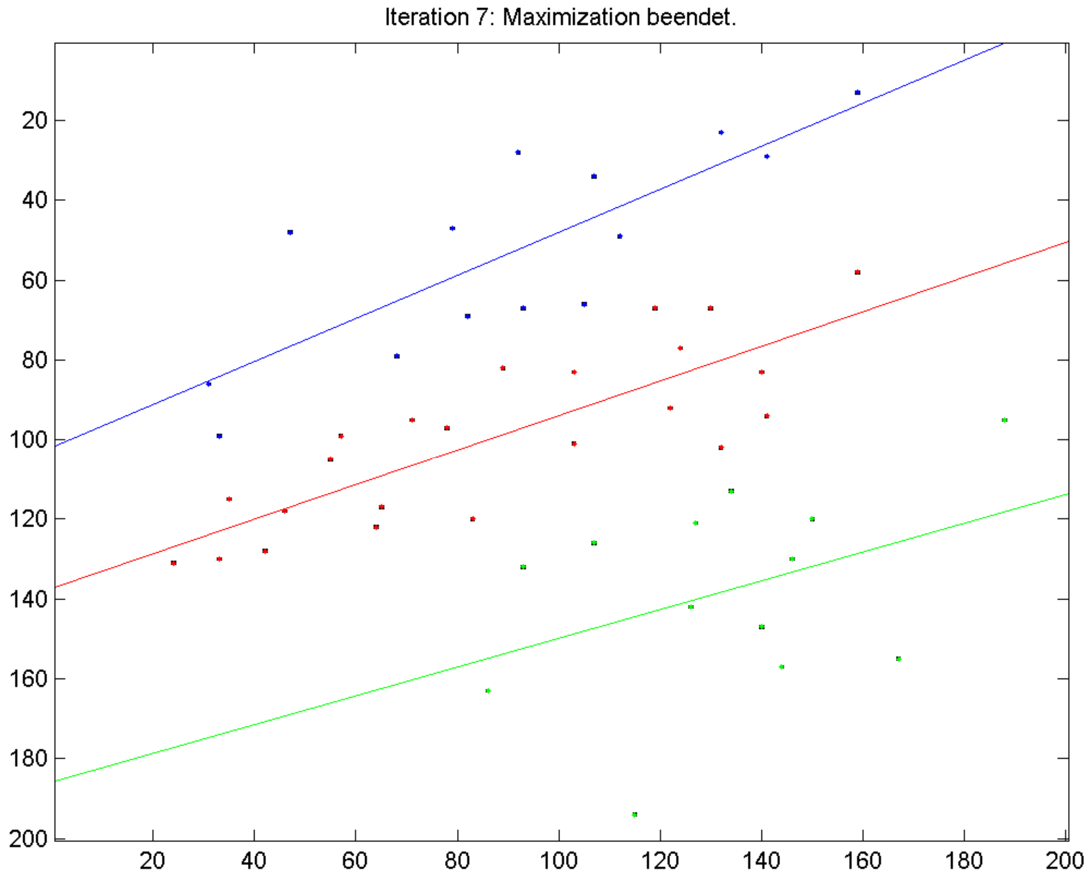


[Egorova 02]

drei Geraden – falsches Modell

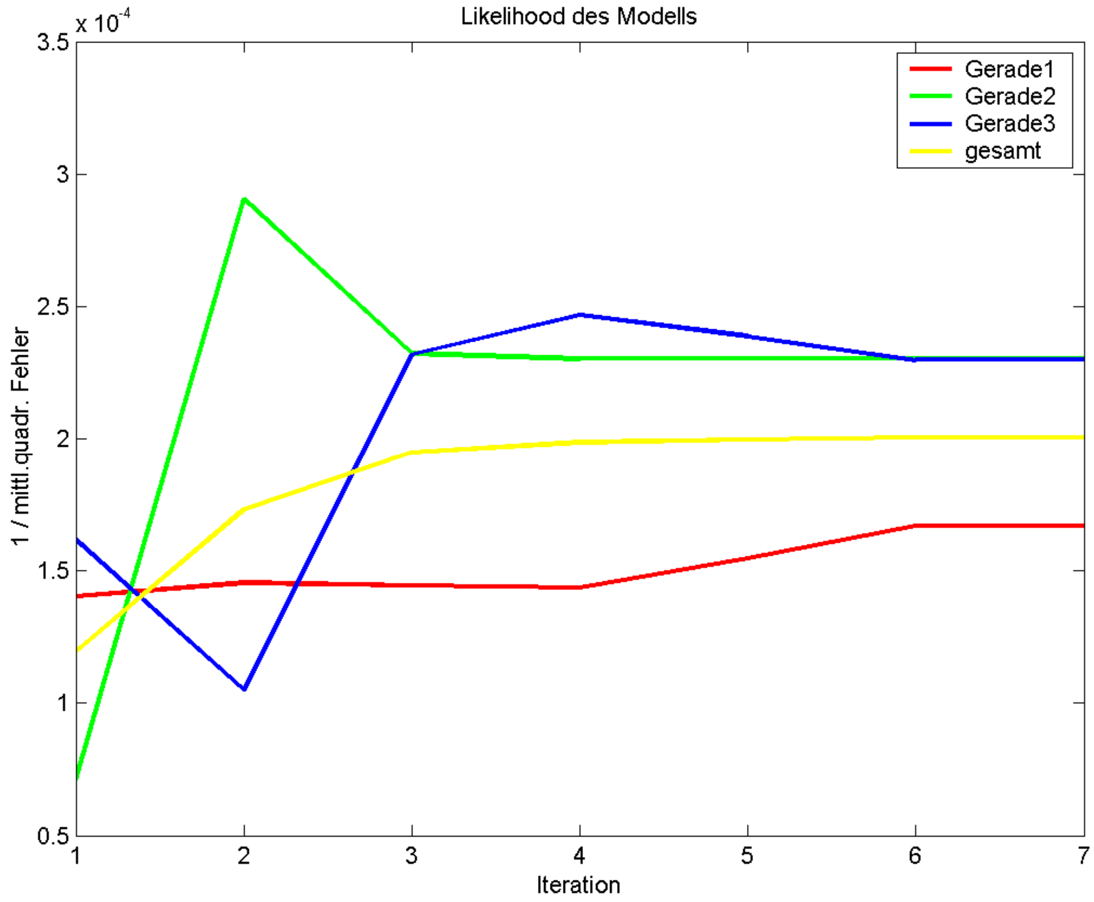
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

02. Juni 2017



[Egorova 02]

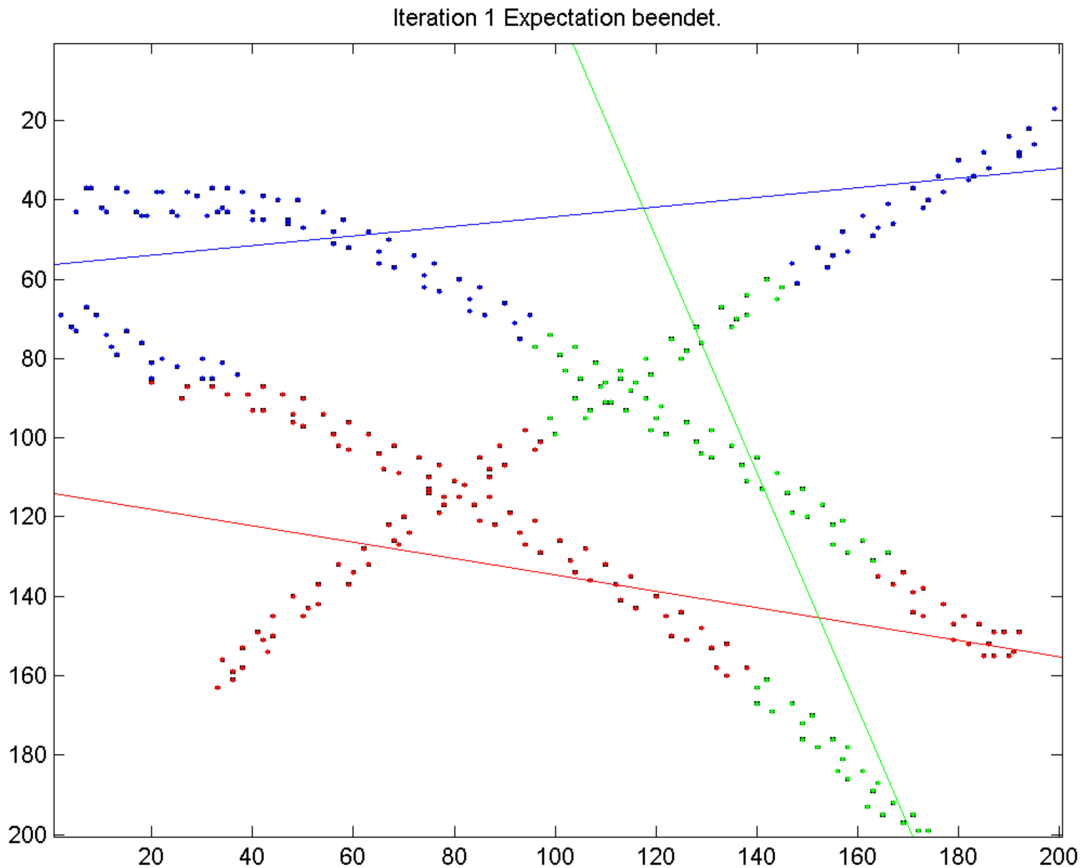
drei Geraden – falsches Modell



[Egorova 02]

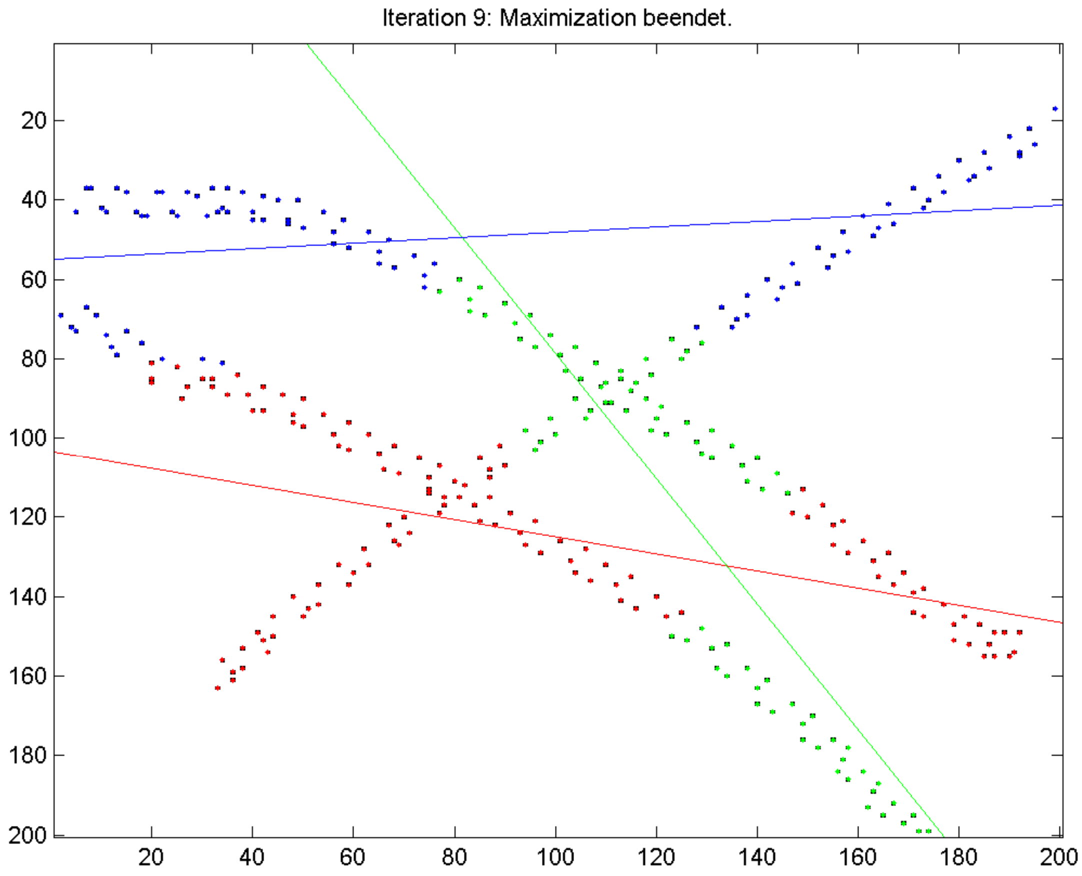
drei Geraden – falsches Maximum

02. Juni 2017



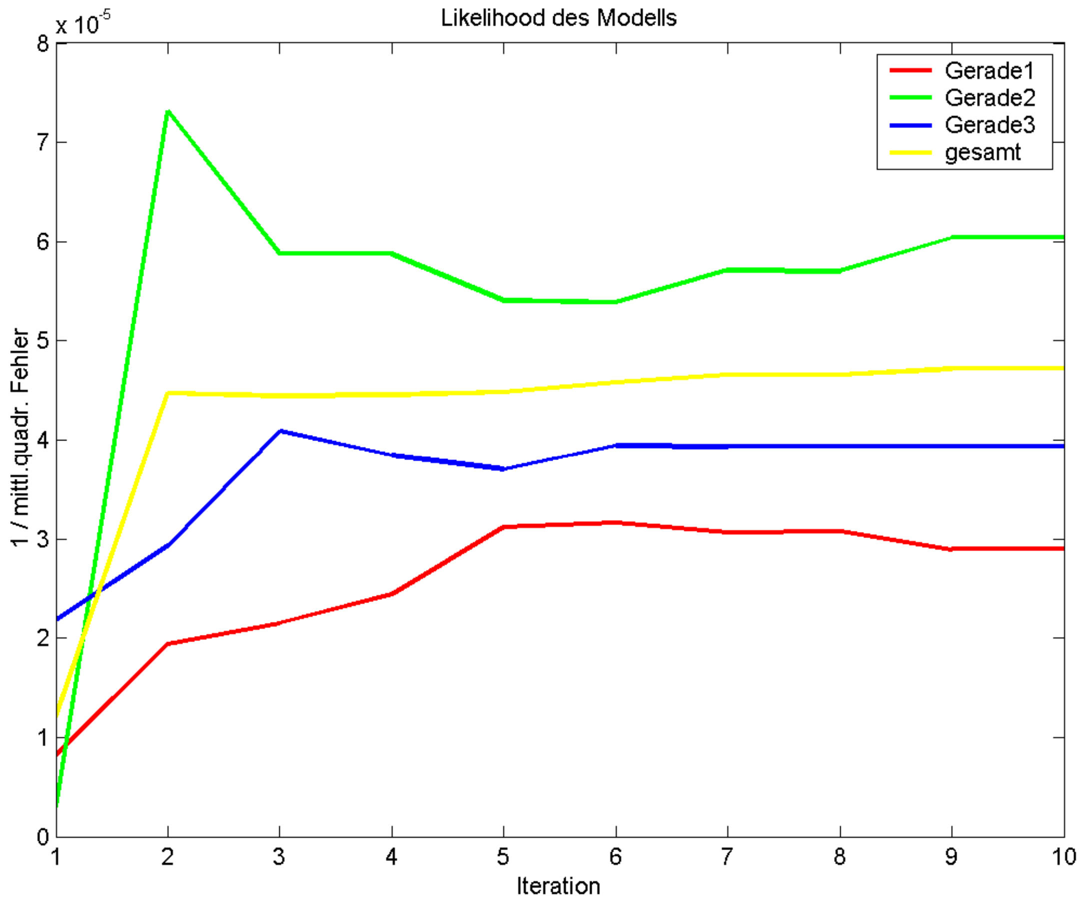
[Egorova 02]

drei Geraden – falsches Maximum



[Egorova 02]

drei Geraden – falsches Maximum



[Egorova 02]

Referenzen:

Wie immer:

R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, New York: John Wiley & Sons, 2001

Zusätzlich:

Wikipedia contributors, "Expectation–maximization algorithm," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/w/index.php?title=Expectation%E2%80%93maximization_algorithm (accessed June 25, 2013).

David McAllester, Expectation Maximization (EM), lecture note, TTIC 103 (CMSC 35420): Statistical Methods for Artificial Intelligence, Autumn 2007 http://ttic.uchicago.edu/~dmcallester/ttic101-07/lectures/em/em.pdf

Bayes‘sche Netze

Kausalnetze

Wissen wird durch Wahrscheinlichkeitsverteilungen repräsentiert. Manchmal sind kausale Abhängigkeiten bekannt. Dann bieten sich Bayesian belief nets an:

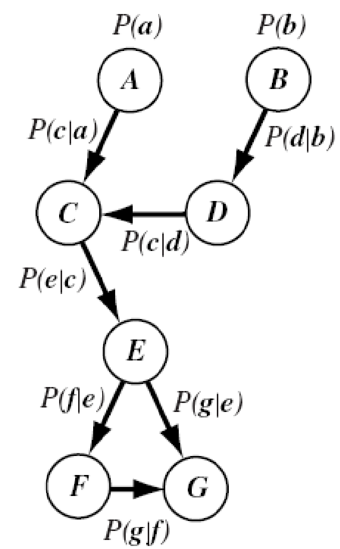
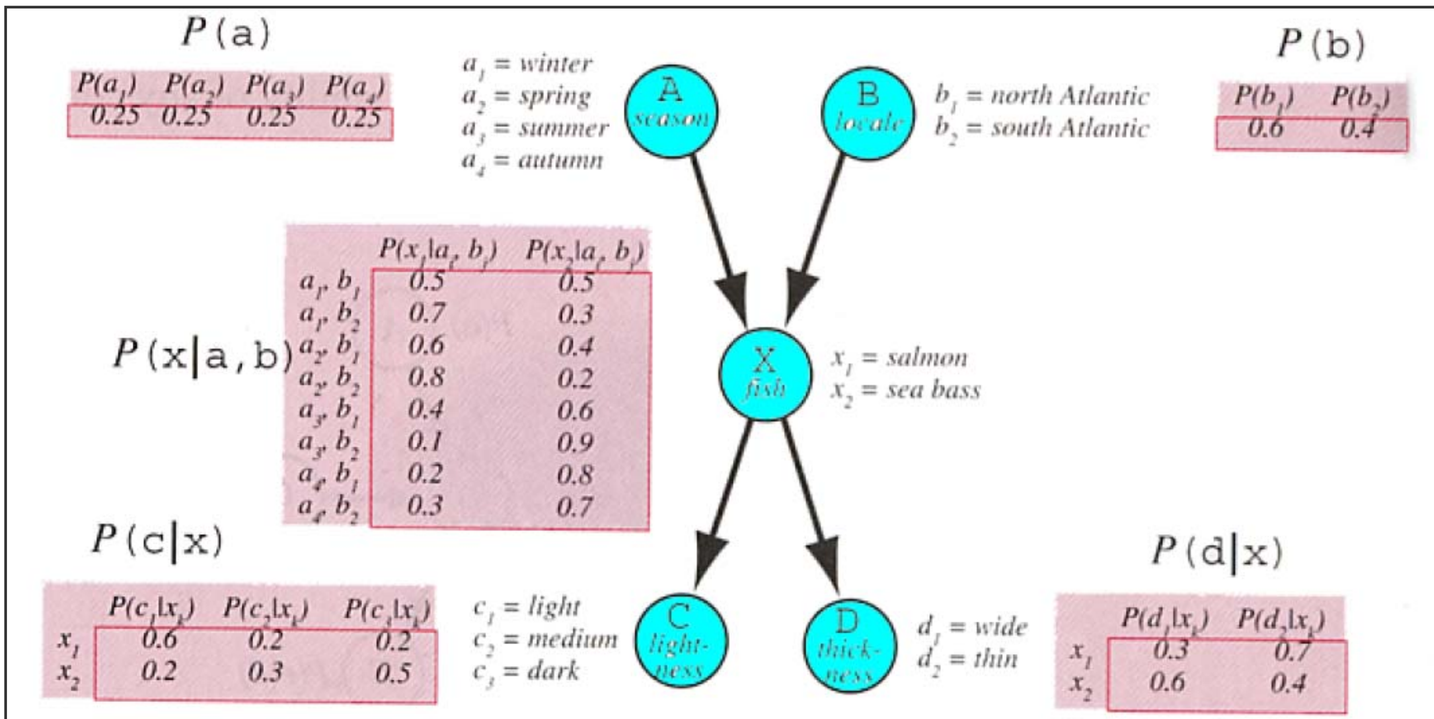


FIGURE 2.24. A belief network consists of nodes (labeled with uppercase bold letters) and their associated discrete states (in lowercase). Thus node **A** has states $a_1, a_2, \ldots,$ denoted simply **a**; node **B** has states $b_1, b_2, \ldots,$ denoted **b**, and so forth. The links between nodes represent conditional probabilities. For example, $P(\mathbf{c}|\mathbf{a})$ can be described by a matrix whose entries are $P(c_i|a_j)$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Außerdem brauchen wir natürlich die bedingten Wahrscheinlichkeiten, die im diskreten Fall als Tabellen gegeben sind:



Kausalnetze

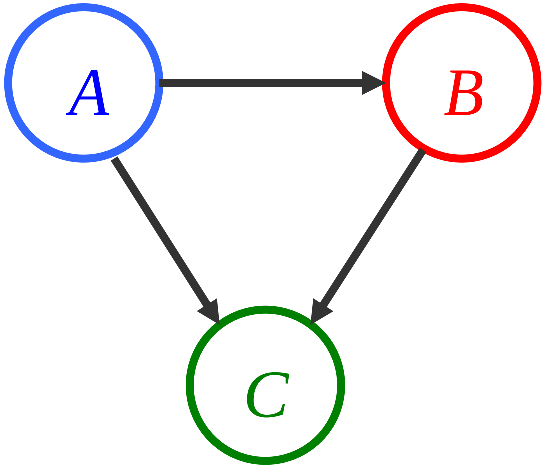
Wenn Abhängigkeit nicht bekannt, wird oft Unabhängigkeit angenommen (naive Bayes‘ rule, idiot Bayes‘ rule):

$$P\left(x \mid \mathbf{a}, \mathbf{b}\right) = P\left(x \mid \mathbf{a}\right) P\left(x \mid \mathbf{b}\right)$$

Nocheinmal grundsätzlich:

Produktregel

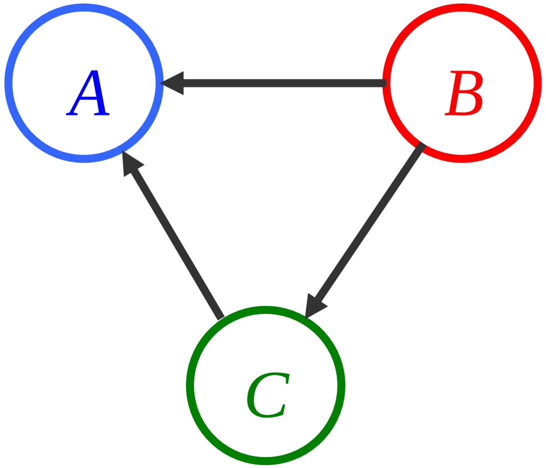
$$P(A,B,C) = P(A)P(B|A)P(C|A,B)$$



Kausalnetze

Andere, ebenso mögliche Faktorisierungen:

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

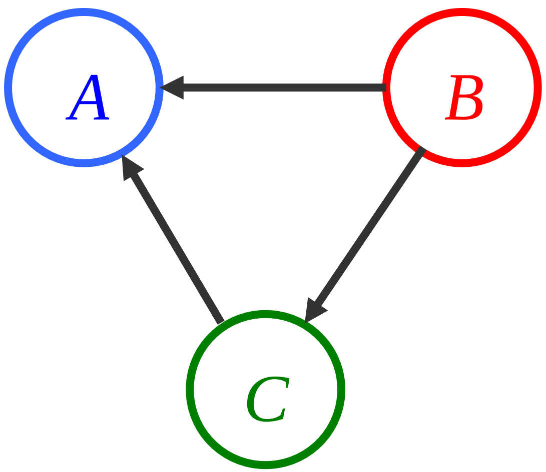


Kausalnetze

A – Fanggebiet, B – Fisch heute, C – Fisch gestern

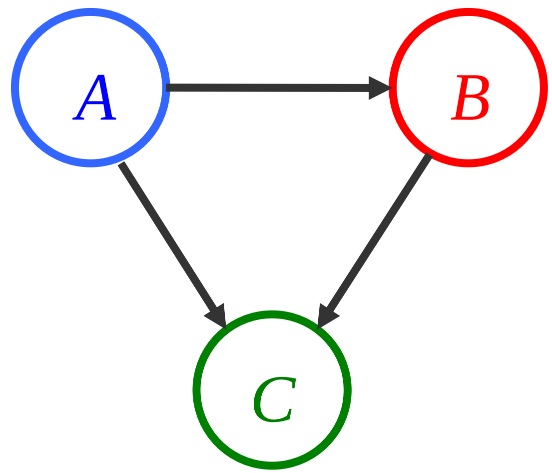
Welche Faktorisierung ist besser,

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



oder

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



?

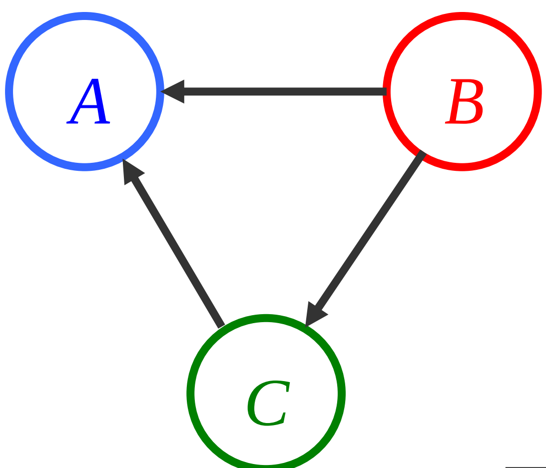
Kausalnetze

Repräsentation der Verbundwahrscheinlichkeit durch Tabellen. Faktorisierung 1:

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

B,C	Nord	Süd
L,L	54,6%	45,4%
L,B	59,5%	40,5%
B,L	59,5%	40,5%
B,B	64,3%	35,7%

Lachs	Barsch
44,5%	55,5%



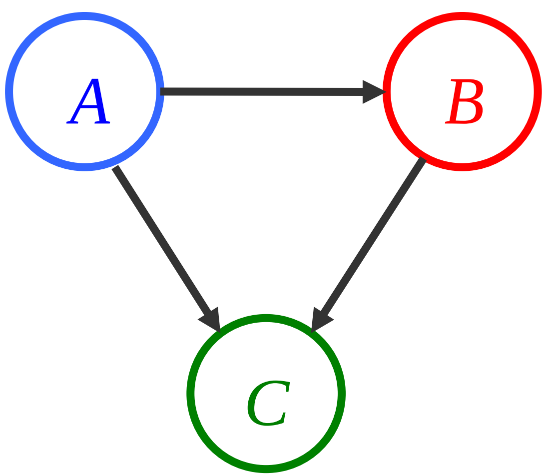
B	Lachs	Barsch
Lachs	44,6%	55,4%
Barsch	44,4%	55,6%

Achtung: $P(B) \neq P(C \mid B)$

Repräsentation der Verbundwahrscheinlichkeit durch Tabellen. Faktorisierung 2:

Nord	Süd
60%	40%

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



A	Lachs	Barsch
Nord	42,5%	57,5%
Süd	47,5%	52,5%

A,B	Lachs	Barsch
Nord,L	42,5%	57,5%
Nord,B	42,5%	57,5%
Süd,L	47,5%	52,5%
Süd,B	47,5%	52,5%

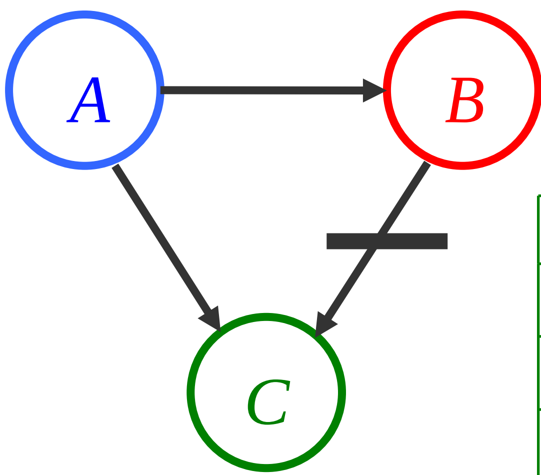
Kausalnetze

Repräsentation der Verbundwahrscheinlichkeit durch Tabellen. Faktorisierung 2:

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

$$\rightarrow \quad P(A,B,C)=P(A)P(B\mid A)P(C\mid A)$$

Nord	Süd
60%	40%



A	Lachs	Barsch
Nord	42,5%	57,5%
Süd	47,5%	52,5%

A,B	Lachs	Barsch
Nord,L	42,5%	57,5%
Nord,B	42,5%	57,5%
Süd,L	47,5%	52,5%
Süd,B	47,5%	52,5%

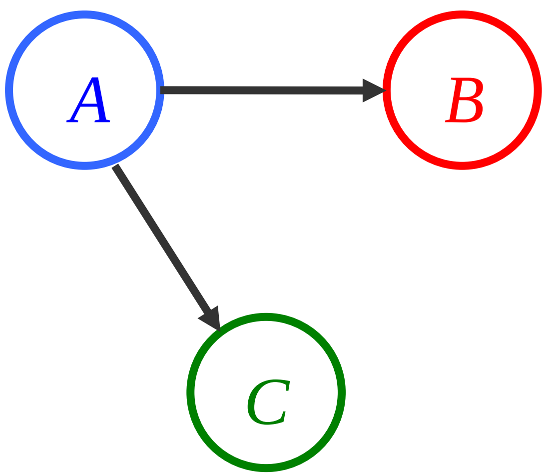
A	Lachs	Barsch
Nord	42,5%	57,5%
Süd	47,5%	52,5%

Kausalnetze

Repräsentation der Verbundwahrscheinlichkeit durch Tabellen. Faktorisierung 2:

Nord	Süd
60%	40%

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



Faktorisierung 2 ist besser, nur so kann Pfeil weggelassen werden, und außerdem gilt:

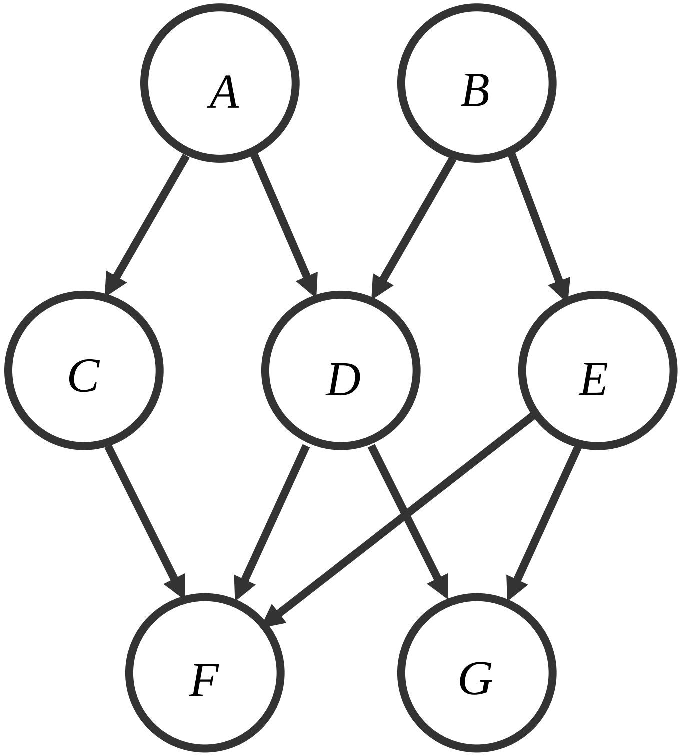
$$P(B \mid A) = P(C \mid A)$$

A	Lachs	Barsch
Nord	42,5%	57,5%
Süd	47,5%	52,5%

=

A	Lachs	Barsch
Nord	42,5%	57,5%
Süd	47,5%	52,5%

Umwandlung von Graph in Faktorisierung



$$P(A,B,C,D,E,F,G)$$

$$= \prod_{X \in \{A,B,C,D,E,F,G\}} P(X \mid \text{Eltern}(X))$$

$$= P(A)P(B)$$
$$*P(C \mid A)P(D \mid A,B)P(E \mid B)$$
$$*P(F \mid C,D,E)P(G \mid D,E)$$

Conditioning, Marginalisierung, Inferenz

Angenommen, ich kenne die Verbundwahrscheinlichkeit:

- Wie berechne ich Wahrscheinlichkeit für einzelne Zufallsvariable?
- Was lerne ich, wenn ich den Zustand einer Zufallsvariablen erfahre, wie verwerte ich dieses Wissen?

Conditioning, Marginalisierung, Inferenz

Wie berechne ich Wahrscheinlichkeit für einzelne Zufallsvariable, wenn $\operatorname{prob}(A, B, C)$ bekannt ist?

- Durch Marginalisierung über alle anderen unbekannten Zufallsvariablen:

$$\operatorname{prob}(A) = \sum_{b_i:\text{ alle möglichen Werte von } B} \left(\sum_{c_k:\text{ alle möglichen Werte von C}} \operatorname{prob}(A, b_i, c_k) \right)$$

Conditioning, Marginalisierung, Inferenz

Was lerne ich, wenn ich den Zustand einer Zufallsvariablen erfahre ($\mathrm{prob}(A,B,C)$ bekannt)?

Anwort: Conditioning bzw. Inferenz

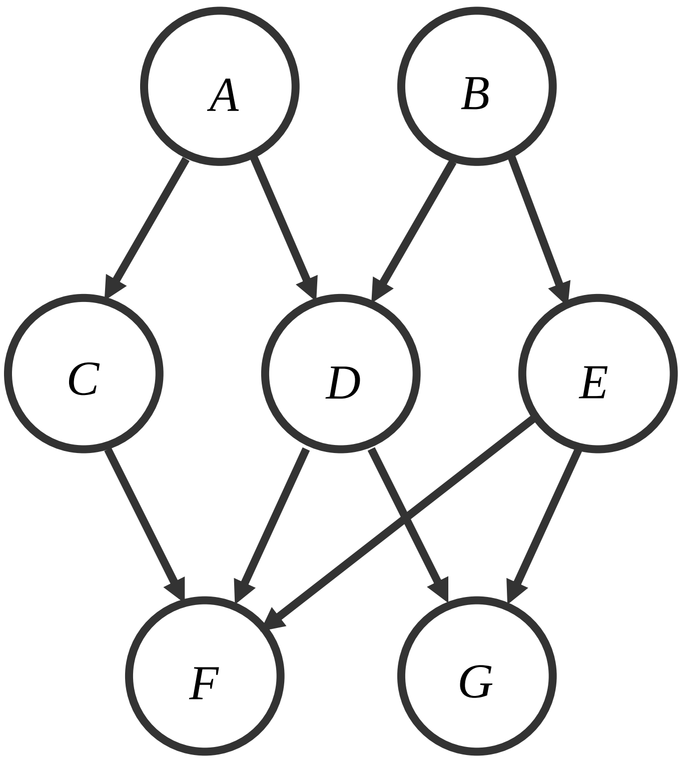
Conditioning:

- Es interessieren nur entsprechende Zeilen/Ebenen… aus der Tabelle. Beispiel: es wird bekannt, dass $A = a_r$. Dann wird Verteilung zu:

$$\boxed{\mathrm{prob}(B,C \mid A = a_r) = \frac{\mathrm{prob}(B,C,A = a_r)}{\mathrm{prob}(A = a_r)}}$$

Conditioning, Marginalisierung, Inferenz

Wenn Verteilung als Bayesnetz gegeben, ist Conditioning einfach, wenn Zustand von Wurzelknoten bekannt wird, da bedingte Wahrscheinlichkeiten gegeben sind:



$$P(A,B,C,D,E,F,G)$$

$$= \prod_{X\in\{A,B,C,D,E,F,G\}} P(X \mid \mathrm{Eltern}(X))$$

$$= P(A)P(B)$$
$$*P(C \mid A)P(D \mid A,B)P(E \mid B)$$
$$*P(F \mid C,D,E)P(G \mid D,E)$$

Conditioning, Marginalisierung, Inferenz

Conditioning und Marginalisierung sieht dann z.B. so aus (wenn $\text{prob}(B \mid a_r)$ gesucht ist:

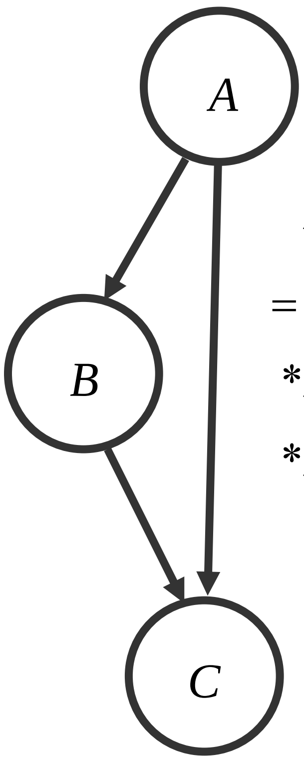
$$\boxed{\text{prob}(B \mid a_r) = \sum_{c_k:\ \text{alle möglichen Werte von C}} \text{prob}(B, c_k \mid a_r)}$$

Inferenz

02. Juni 2017

Schließen vom Zustand der Kindknoten auf Zustand bzw. Wahrscheinlichkeiten für Elternknoten wird Inferenz genannt.

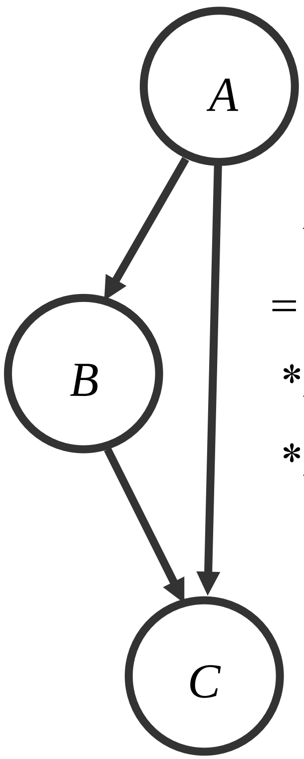
Beispiel: Zustand von C bekannt ($C = c_r$).



$$
\begin{aligned}
&P(A,B,C) \\
&= P(A) \\
&*P(B \mid A) \\
&*P(C \mid A,B)
\end{aligned}
$$

Inferenz

Wir wollen z.B. $P(A \mid C = c_r)$ berechnen.
Das geschieht durch Einsetzen von *c* und Marginalisierung:



$$P(A,B,C)$$
$$= P(A)$$
$$*P(B \mid A)$$
$$*P(C \mid A,B)$$

$$\boxed{\operatorname{prob}(A, c_r) = \sum_{b_k:\ \text{alle möglichen Werte von B}} \operatorname{prob}(A, B = b_k, C = c_r)}$$

Anschließend muss normiert werden:

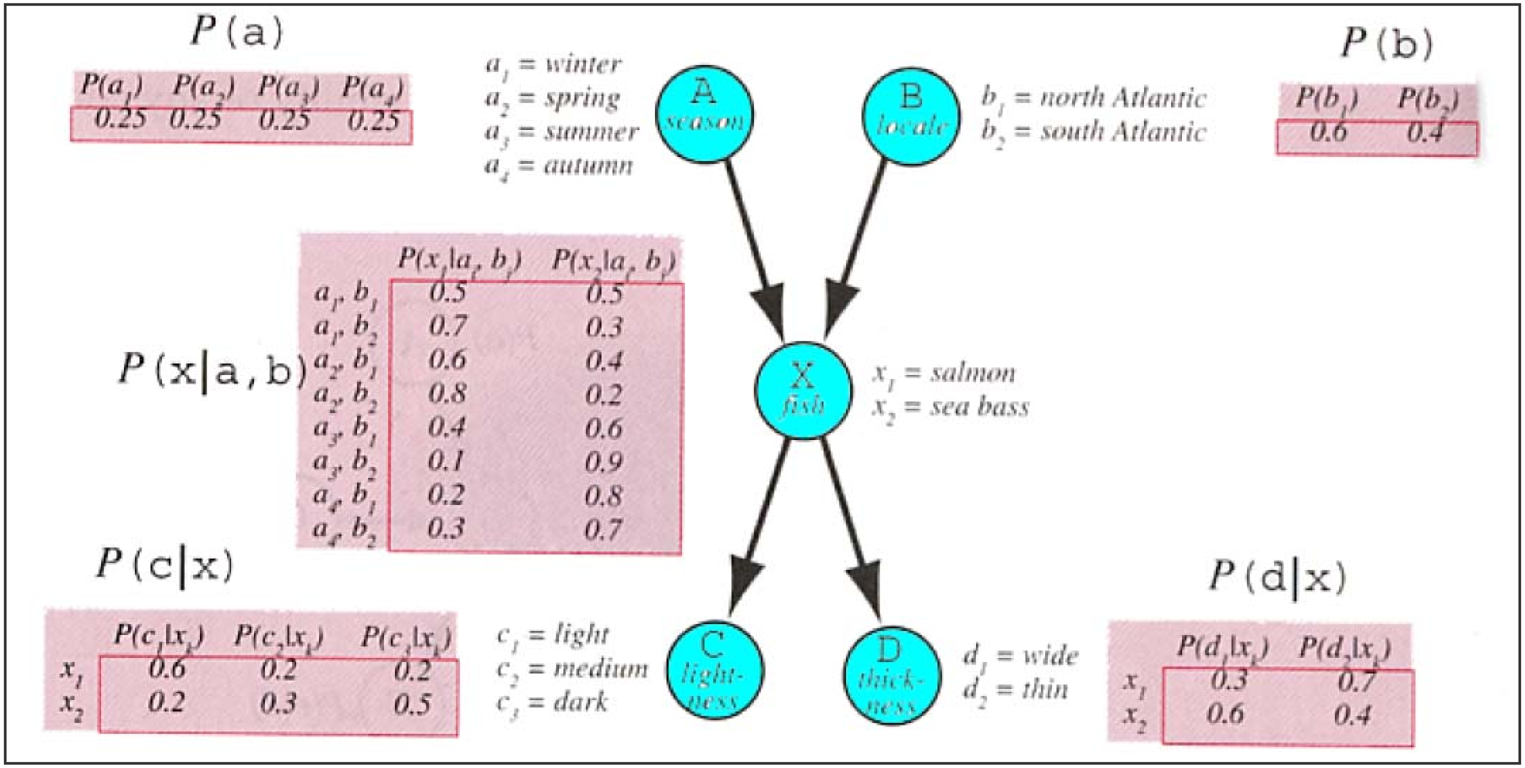
$$\boxed{\operatorname{prob}(A \mid c_r) = \frac{\operatorname{prob}(A, c_r)}{\operatorname{prob}(c_r)}}$$

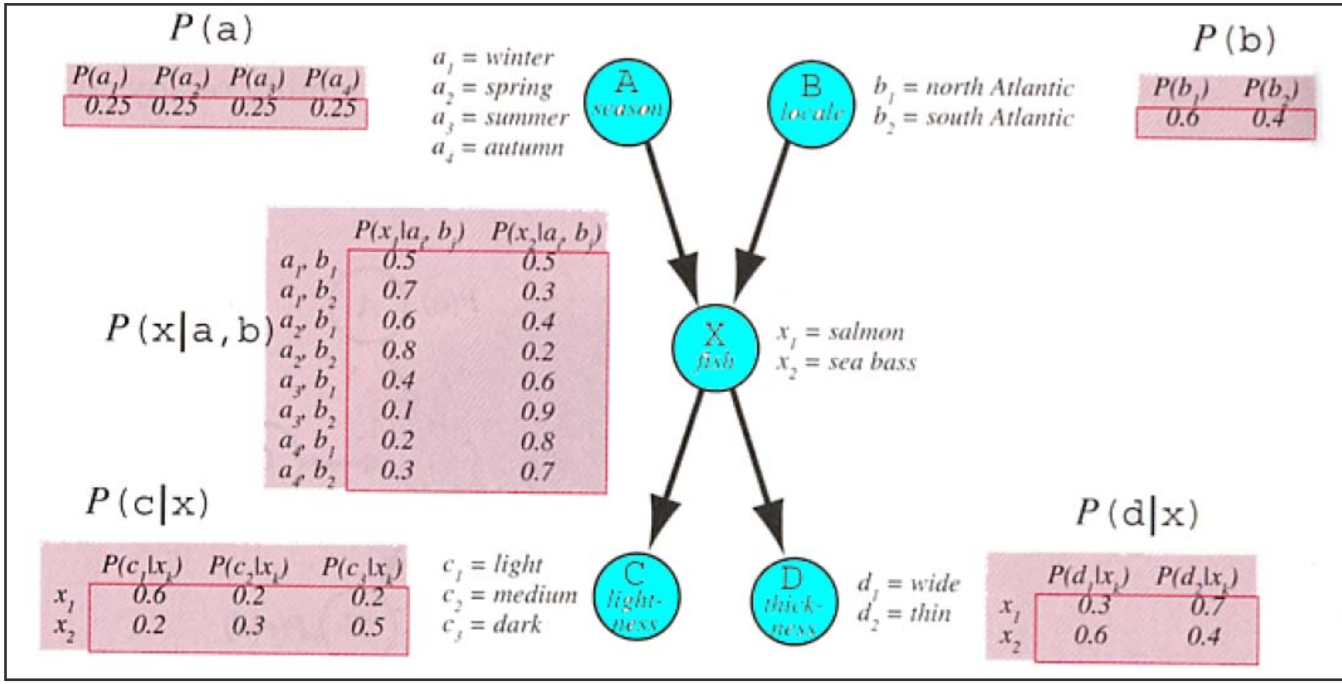
Dabei ist

$$\boxed{\operatorname{prob}(c_r) = \sum_{\substack{a_j:\ \text{alle möglichen Werte von A} \\ b_k:\ \text{alle möglichen Werte von B}}} \sum \operatorname{prob}(A = a_j, B = b_k, C = c_r)}$$

Kausalnetze

Beispiel: Fisch ist hell und im Südatlantik gefangen, mehr wissen wir nicht.





$$
\begin{aligned}
P(x_1|c_1,b_2) &= \frac{P(x_1,c_1,b_2)}{P(c_1,b_2)} \qquad (100)\\
&= \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1,\mathbf{a},b_2,c_1,\mathbf{d})\\
&= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a})P(b_2)P(x_1|\mathbf{a},b_2)P(c_1|x_1)P(\mathbf{d}|x_1)\\
&= \alpha P(b_2)P(c_1|x_1)\\
&\quad \times \left[\sum_{\mathbf{a}} P(\mathbf{a})P(x_1|\mathbf{a},b_2)\right]\left[\sum_{\mathbf{d}} P(\mathbf{d}|x_1)\right]\\
&= \alpha P(b_2)P(c_1|x_1)\\
&\quad \times [P(a_1)P(x_1|a_1,b_2)+P(a_2)P(x_1|a_2,b_2)+P(a_3)P(x_1|a_3,b_2)+P(a_4)P(x_1|a_4,b_2)]\\
&\quad \times \underbrace{[P(d_1|x_1)+P(d_2|x_1)]}_{=1}\\
&= \alpha(0.4)(0.6)\,[(0.25)(0.7)+(0.25)(0.8)+(0.25)(0.1)+(0.25)(0.3)]\,1.0\\
&= \alpha\ 0.114.
\end{aligned}
$$

Kausalnetze

Fortsetzung Beispiel

Entsprechende Rechung ergibt:

$$P(x_2|c_1, b_2) = \alpha \; 0.066$$

Normierung auf 1:

$$P(x_1|c_1, b_2) = 0.63 \text{ and } P(x_2|c_1, b_2) = 0.37.$$

Also ist Fisch wahrscheinlich Lachs.