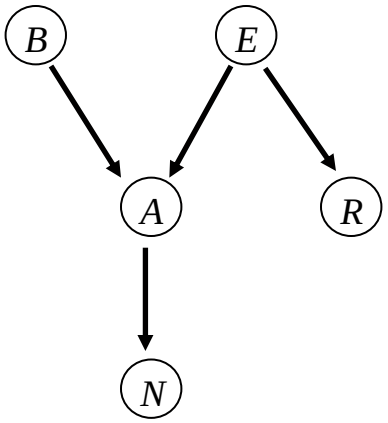
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Graphische Modelle

Niels Landwehr

Zusammenfassung Pfade



Gemeinsame Verteilung:

$$p(B,E,A,N,R) =$$
$$p(B)\,p(E)\,p(A\,|\,E,B)\,p(N\,|\,A)\,p(R\,|\,E)$$

- Zusammenfassung: ein Pfad …-*X*-*Y*-*Z*-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
    - $Descendants(Y) = \{Y' | \text{es gibt gerichteten Pfad von } Y \text{ zu } Y'\}$
  - Sonst ist der Pfad frei bei Y

D-Separation: Blockierte Pfade

- Bisher: Pfad blockiert bei einem bestimmten Knoten
- Verallgemeinerung: ein Pfad ist insgesamt blockiert, wenn er bei einem auf dem Pfad liegenden Knoten blockiert ist:
  - Seien X,X' ZV, C eine beobachtete Menge von ZV, $X,X' \notin C$
  - Ein ungerichteter Pfad $X - X_1 - \ldots - X_n - X'$ zwischen X und X' ist blockiert gegeben $\mathbf{C}$ gdw es einen Knoten X_i gibt so dass Pfad bei X_i blockiert ist gegeben C.

- D-Separation basiert auf blockierten Pfaden:
  - Seien A,B,C disjunkte Mengen von ZV.
  - Definition: A und B sind d-separiert durch C gdw jeder Pfad von einem Knoten $X \in A$ zu einem Knoten $Y \in B$ blockiert ist gegeben C.

D-Separation: Korrektheit, Vollständigkeit

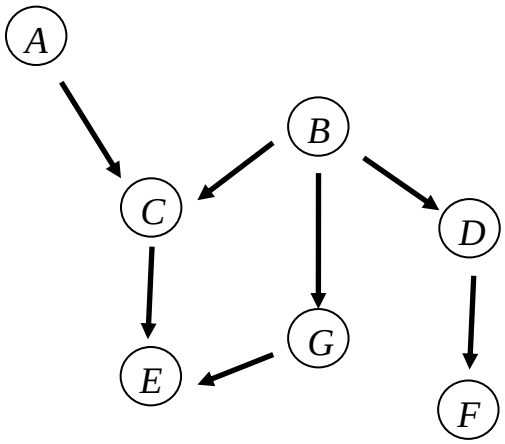
- Gegeben ein graphisches Modell über $\{X_1,...,X_N\}$ mit Graphstruktur G.
- Das graphische Modell modelliert eine Verteilung durch

$$p(X_1,...,X_N) = \prod_{i=1}^{N} p(X_i \mid pa(X_i))$$

abhängig von den bedingten Verteilungen $p(X_i \mid pa(X_i))$.

- Theorem (Korrektheit, Vollständigkeit d-separation)
  - Falls A,B d-separiert gegeben C in G, dann $p(A|B,C) = p(A|C)$
  - Es gibt keine anderen Unabhängigkeiten, die für jede Wahl der bedingten Verteilungen $p(X_i \mid pa(X_i))$ gelten.

D-Separation: Beispiel



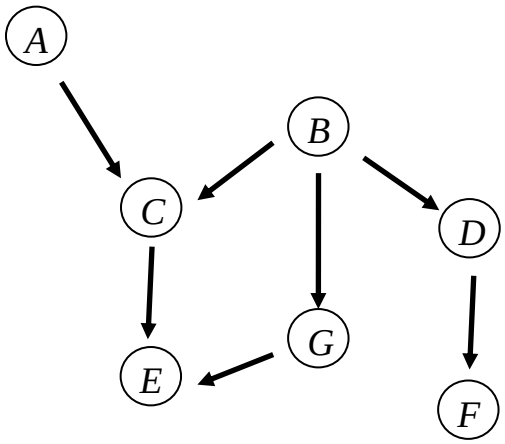
Gilt $A \perp F \mid D$?

Gilt $B \perp E \mid C$?

Gilt $A \perp E \mid C$?

- Ein Pfad …-X-Y-Z-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
  - Sonst ist der Pfad frei bei Y

D-Separation: Beispiel



Gilt $A \perp F \mid D$? Ja

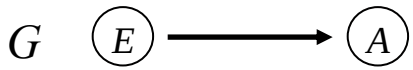
Gilt $B \perp E \mid C$? Nein: $B - G - E$

Gilt $A \perp E \mid C$? Nein: $A - C - B - G - E$

- Ein Pfad …-X-Y-Z-... ist
  - Blockiert bei Y, wenn
    - Divergierende Verbindung, und Y beobachtet, oder
    - Serielle Verbindung, und Y beobachtet, oder
    - Konvergierende Verbindung, und weder Y noch einer seiner Nachfahren $Y' \in Descendants(Y)$ beobachtet
  - Sonst ist der Pfad frei bei Y

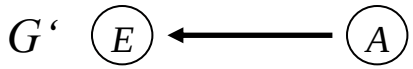
Graphische Modelle: Kausalität

- Oft werden graphische Modelle so konstruiert, dass gerichtete Kanten kausalen Einflüssen entsprechen



„Erdbeben lösen die Alarmanlage aus“

- Äquivalentes Modell



„die Alarmanlage löst Erdbeben aus“

- Definition: $I(G) = \{ (A \perp B \mid C) : A \text{ und } B \text{ sind d-separiert gegeben } C \text{ in } G \}$

„Alle Unabhängigkeitsannahmen, die durch G kodiert werden“

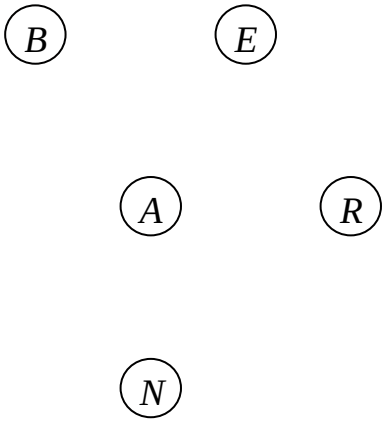
- $I(G) = I(G') = \varnothing$:
  - Keine statistischen Gründe, eines der Modelle vorzuziehen
  - Kann nicht aufgrund von Daten zwischen Modellen unterscheiden
  - Aber „kausale“ Modelle oft besser verständlich

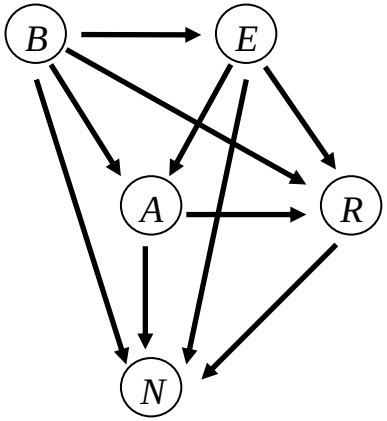
Modelle Unterschiedlicher Komplexität

- Komplexität eines Modell hängt ab von der Menge der Kanten im Graph
  - Viele Kanten: wenig Unabhängigkeitsannahmen, viele Parameter, große Klasse von Verteilungen kann repräsentiert werden
  - Wenige Kanten: viele Unabhängigkeitsannahmen, wenige Parameter, kleine Klasse von Verteilungen kann repräsentiert werden

Modelle Unterschiedlicher Komplexität

- Hinzufügen von Kanten: Familie der darstellbare Verteilungen wird grösser, $I(G)$ wird kleiner.
- Extremfälle: Graph ohne Kanten, (ungerichtet) vollständig verbundener Graph.





N Parameter (binäre Variablen)

$I(G) = \{(A \perp \mathrm{B} \mid C) : A, B, C \text{ disj. Mengen von ZV}\}$

2^N -1 Parameter (binäre Variablen)

$I(G) = \varnothing$

Überblick

- Graphische Modelle: Syntax und Semantik
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen (exakt, approximativ)
- Sequenzmodelle

Erinnerung: Lernproblem

- Erinnerung: Lernproblem
  - Trainingsdaten

$$L = \langle (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$ Merkmalsvektoren

$y_i \in \{0,1\}$ binäre Klassenlabel

$y_i \in \mathbb{R}$ reelles Label

  - Ziel: Vorhersage des Labels für Testinstanz $\mathbf{x}$

$$\mathbf{x} \mapsto y$$

  - Matrixschreibweise

Merkmalsvektoren

$$X = (\mathbf{x}_1 \quad ... \quad \mathbf{x}_N) = \begin{pmatrix} x_{11} & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1m} & x_{Nm} \end{pmatrix}$$

Zugehörige Labels

$$\mathbf{y} = \begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix}$$

Erinnerung: Bayes‘sches Lernen

- Bayescher Ansatz: Wir wenden probabilistische Überlegungen auf Daten, Modelle und Vorhersagen an
- A-priori Verteilung über Modelle $p(\theta)$ (bekannt)
- A-posteriori Verteilung $\underbrace{p(\theta \mid L)}_{posterior} \propto \underbrace{p(L \mid \theta)}_{likelihood} \underbrace{p(\theta)}_{prior}$
- Vorhersageproblem: MAP Lösung

$$\theta_* = \arg\max_\theta p(\theta \mid L) \qquad y_* = \arg\max_y p(y \mid \mathbf{x}, \theta_*)$$

- Vorhersageproblem: Bayes Lösung

$$y_* = \arg\max_y p(y \mid \mathbf{x}, L) = \arg\max_y \int p(y \mid \mathbf{x}, \theta)\, p(\theta \mid L) d\theta$$

Erinnerung: Parameterschätzung Münzwurf

- Erinnerung: Münzwurf
  - Einzelner Münzwurf Bernouilli-verteilt mit Parameter μ

    $X \in \{0,1\}$

    $X \sim \text{Bern}(X \mid \mu) = \mu^X (1-\mu)^{1-X}$

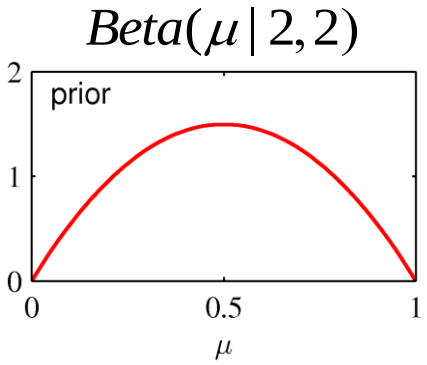
    $\mu = p(X = 1 \mid \mu)$ unbekannter Parameter
- Parameterschätzproblem:
  - Wir haben N unabhängige Münzwürfe gesehen, als Ausprägung $L=\{x_1, \ldots, x_N\}$ der ZV $X_1, \ldots, X_N$
  - Der echte Parameter μ ist unbekannt, wir wollen eine Schätzung $\hat{\mu}$ bzw. eine posterior-Verteilung $p(\mu \mid L)$
  - Bayesscher Ansatz: Posterior $\propto$ Prior x Likelihood

$$\underbrace{p(\mu \mid L)}_{\text{posterior}} \propto \underbrace{p(L \mid \mu)}_{\text{likelihood}} \underbrace{p(\mu)}_{\text{prior}}$$

Erinnerung: Parameterschätzung Münzwurf

- Prior: Beta-Verteilung über Münzparameter μ

$$p(\mu) = Beta(\mu \mid \alpha_k, \alpha_z)$$
$$= \frac{\Gamma(\alpha_k + \alpha_z)}{\Gamma(\alpha_k)\Gamma(\alpha_z)} \mu^{\alpha_k - 1}(1-\mu)^{\alpha_z - 1}$$



- Likelihood N unabhängige Münzwürfe:

$$p(X_1, ..., X_N \mid \mu) = \prod_{i=1}^{N} p(X_i \mid \mu) \qquad i.i.d.$$
$$= \prod_{i=1}^{N} \text{Bern}(X_i \mid \mu)$$
$$= \prod_{i=1}^{N} \mu^{X_i}(1-\mu)^{1-X_i}$$

Erinnerung: Parameterschätzung Münzwurf

- Münzwurfszenario als graphisches Modell?
- Zufallsvariablen in Münzwurfszenario sind $X_1, ..., X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter: Prior x Likelihood

$$p(X_1, ..., X_N, \mu) = p(\mu) \prod_{i=1}^{N} \underbrace{p(X_i \mid \mu)}_{\text{Bernoulli}}$$

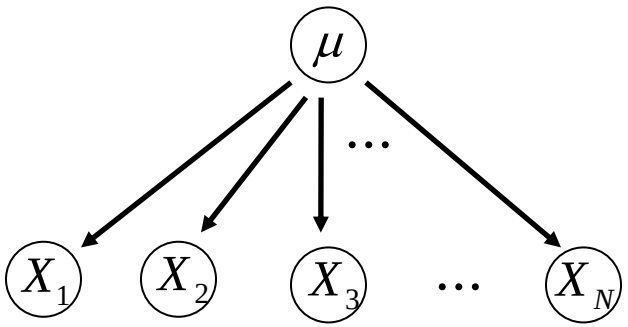
- Darstellung als graphisches Modell:

Erinnerung: Parameterschätzung Münzwurf

- Münzwurfszenario als graphisches Modell?
- Zufallsvariablen in Münzwurfszenario sind $X_1, ..., X_N, \mu$
- Gemeinsame Verteilung von Daten und Parameter: Prior x Likelihood

$$p(X_1, ..., X_N, \mu) = p(\mu) \prod_{i=1}^{N} \underbrace{p(X_i \mid \mu)}_{\text{Bernoulli}}$$

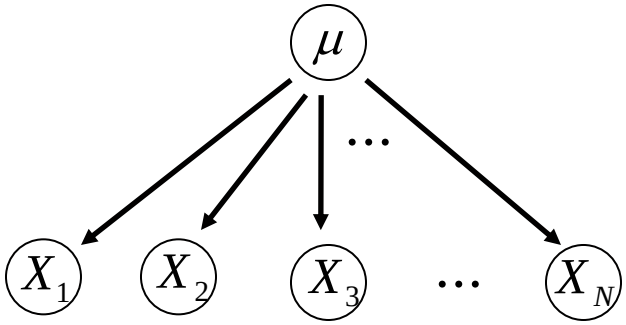
- Darstellung als graphisches Modell:



$$pa(\mu) = \varnothing$$
$$pa(X_i) = \{\mu\}$$

Schätzung eines Münzparameters als Graphisches Modell

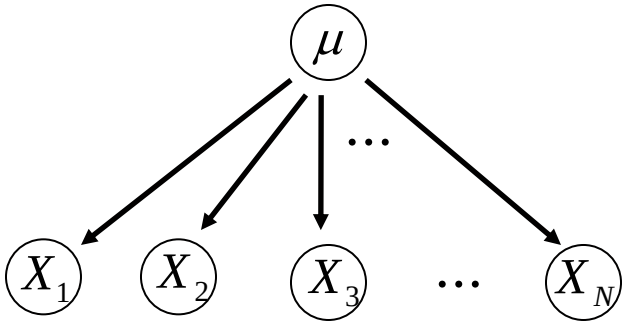
- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation
  - Gilt $X_N \perp X_1,\ldots,X_{N-1} \mid \varnothing$?

Schätzung eines Münzparameters als Graphisches Modell

- Unabhängige Münzwürfe: Darstellung als graphisches Modell



- D-separation
  - Gilt $X_N \perp X_1,...,X_{N-1} \mid \varnothing$?
  - Nein, Pfad durch μ ist nicht blockiert.
  - Intuitiv:

$$X_1 = X_2 = ... = X_{N-1} = 1 \Rightarrow \text{Wahrscheinlich } \mu > 0.5 \Rightarrow \text{Wahrscheinlich } X_N = 1$$

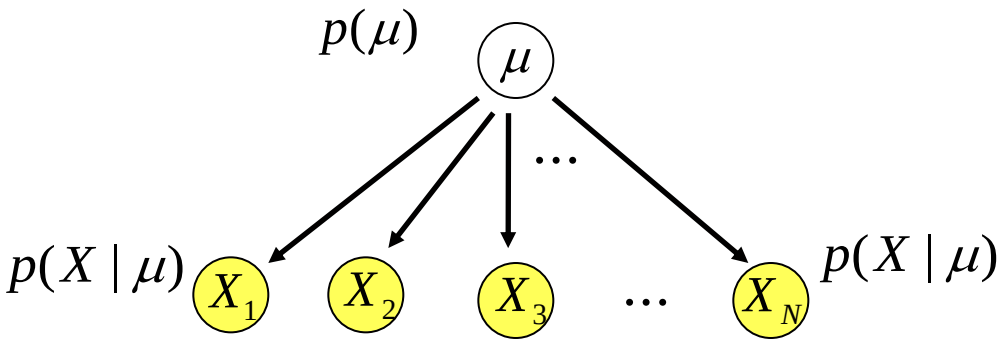
  - Der versteckte Parameter μ koppelt ZV $X_1, ..., X_N$.
  - Aber es gilt $X_N \perp X_1,...,X_{N-1} \mid \mu$

Parameterschätzung als Inferenzproblem

- MAP-Parameterschätzung Münzwurf

$$\hat{\mu} = \arg\max_{\mu} p(\mu \mid x_1, ..., x_N)$$

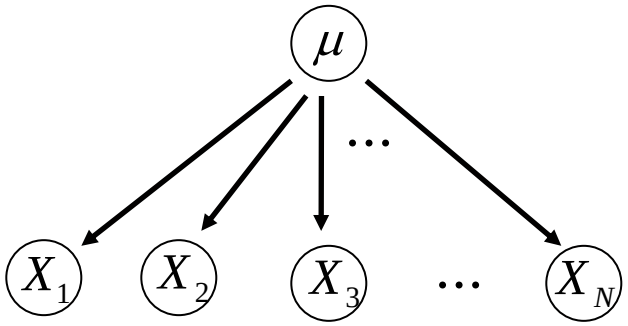
- Inferenzproblem:



  - Evidenz auf den Knoten $X_1, ..., X_N$
  - Gesucht: Verteilung $p(\mu \mid X_1, ..., X_N)$

Plate-Modelle

- Erweiterung der graphischen Modelle: Plate-Notation
- Unabhängige Münzwürfe: Darstellung als graphisches Modell

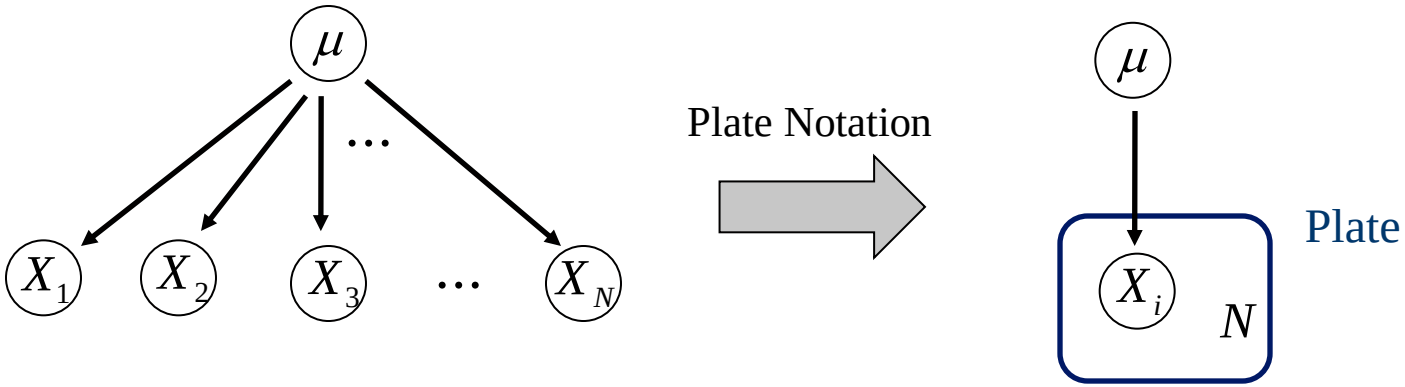


- Knoten $X_1, ..., X_N$ sind von gleicher Form
  - Gleicher Wertebereich
  - Gleiche bedingte Verteilungen $p(X_i \mid \mu) = p(X_j \mid \mu)$.

- Kurznotation in der Form einer „Schablone“: Plate Notation

Plate-Modelle

- Plate Notation für Münzwürfe



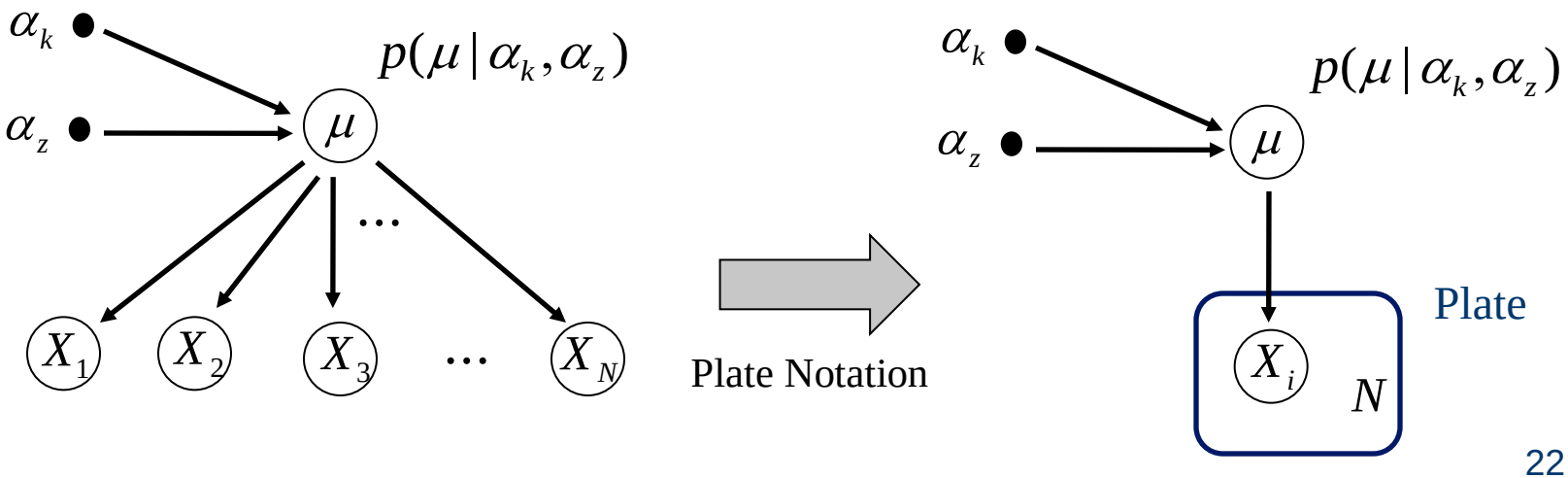
- Ein „Plate“ ist eine abkürzende Notation für N Variablen der gleichen Form
  - Bezeichnet mit Anzahl der Variablen, N
  - Variablen haben Index (z.B. X_i).
- Plate-Modelle werden im Maschinellen Lernen oft verwendet

Plate-Modelle: Hyperparameter

- Rolle der „Hyperparameter“ α_k, α_z?
  - Keine Zufallsvariablen, wir modellieren nur die gemeinsame Verteilung über $X_1, ..., X_N, \mu$ gegeben Hyperparameter

$$p(X_1, ..., X_N, \mu \mid \alpha_k, \alpha_z) = p(\mu \mid \alpha_k, \alpha_z) \prod_{i=1}^{N} p(X_i \mid \mu)$$

  - Hyperparameter keine Knoten im GM, werden aber oft zusätzlich angegeben (Notation: Punkt statt Kreis)



Erinnerung: Bayessche Lineare Regression

- Regressionslernen

$$L = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \rangle$$

$\mathbf{x}_i \in \mathbb{R}^m$ Merkmalsvektoren

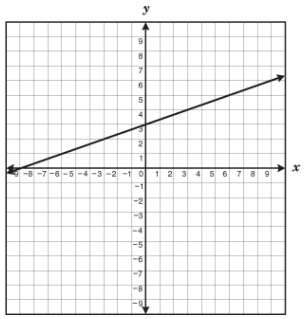
$y_i \in \mathbb{R}$ reelles Zielattribut

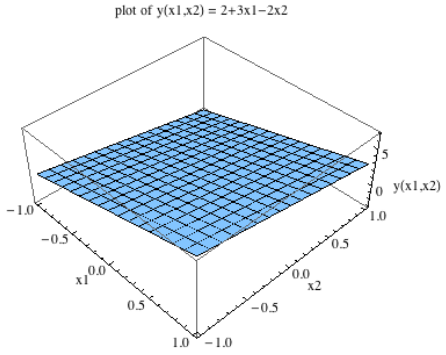
- Lineare Regression

$$f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

$$= \sum_{i=1}^{m} w_i x_i$$

$\mathbf{w}$ „Parametervektor“, „Gewichtsvektor“





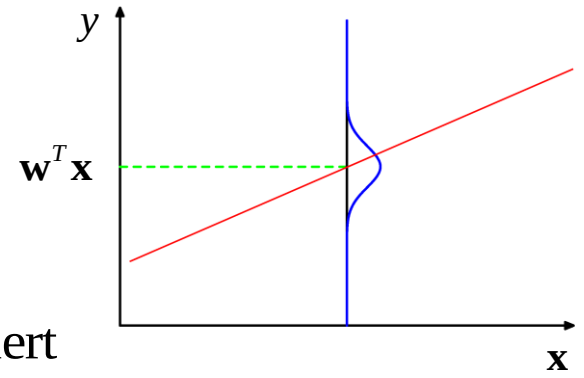
Erinnerung: Bayessche Lineare Regression

- Diskriminatives Setting: $\mathbf{x}_i$ fest, y_i generiert aus $\mathbf{x}_i$ und $\mathbf{w}$ plus Gaussschem Rauschen

$$p(y \mid \mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + N(y \mid 0, \sigma^2)$$
$$= N(y \mid \mathbf{w}^T \mathbf{x}, \sigma^2)$$

$$y_i \sim p(y \mid \mathbf{x}_i, \mathbf{w})$$

diskriminatives Modell: $p(\mathbf{x})$ nicht modelliert



- Bayessches Setting: Posterior $\propto$ Prior x Likelihood

$$\underbrace{p(\mathbf{w} \mid L)}_{posterior} \propto \underbrace{p(L \mid \mathbf{w})}_{likelihood} \underbrace{p(\mathbf{w})}_{prior}$$

Erinnerung: Bayessche Lineare Regression

- Likelihood der Daten unter einem Modell $\mathbf{w}$:

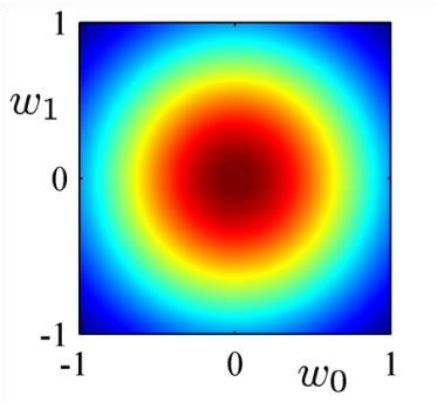
$$p(\mathbf{y} \mid X, \mathbf{w}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, \sigma^2) \qquad i.i.d.$$

$$= \prod_{i=1}^{N} N(y_i \mid \mathbf{w}^T \mathbf{x}_i, \sigma^2)$$

- Normalverteilter Prior über Modelle

$$p(\mathbf{w}) = N(\mathbf{w} \mid \mathbf{0}, \tau^2 I)$$

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$



Isotrope multivariate Normalverteilung, Mittelwert **0**, Varianz τ^2

Bayessche Lineare Regression als Graphisches Modell

- Was sind Zufallsvariablen?
  - Labels $y_1, \ldots, y_N$, Modell $\mathbf{w}$
  - Nicht: $\mathbf{x}_1, \ldots, \mathbf{x}_N$, Hyperparameter σ^2, τ^2

- Gemeinsame Verteilung über Labels und Parameter

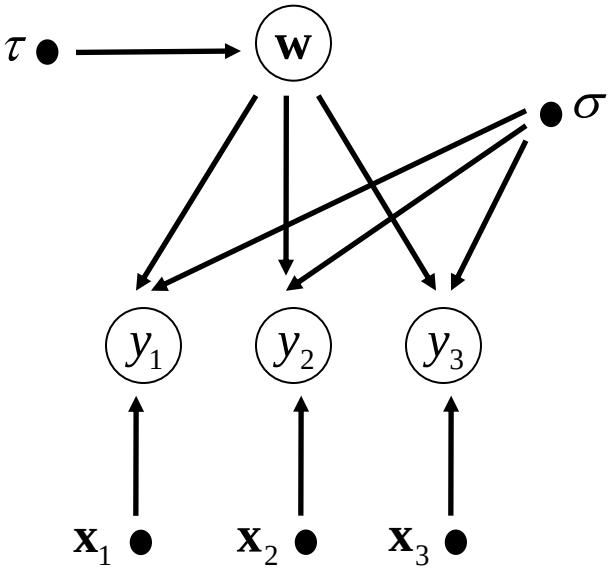
$$p(y_1,\ldots,y_N,\mathbf{w}\,|\,\mathbf{x}_1,\ldots,\mathbf{x}_N,\sigma^2,\tau^2) = \overbrace{p(\mathbf{w}\,|\,\tau^2)}^{\text{Prior}}\overbrace{p(y_1,\ldots,y_N\,|\,\mathbf{w},\mathbf{x}_1,\ldots,\mathbf{x}_N,\sigma^2)}^{\text{Likelihood}}$$
$$= p(\mathbf{w}\,|\,\tau^2)\prod_{i=1}^{N} p(y_i\,|\,\mathbf{w},\mathbf{x}_i,\sigma^2)$$

- Darstellung von Bayesscher Linearer Regression als graphisches Modell: Ablesen der Struktur aus gemeinsamer Verteilung

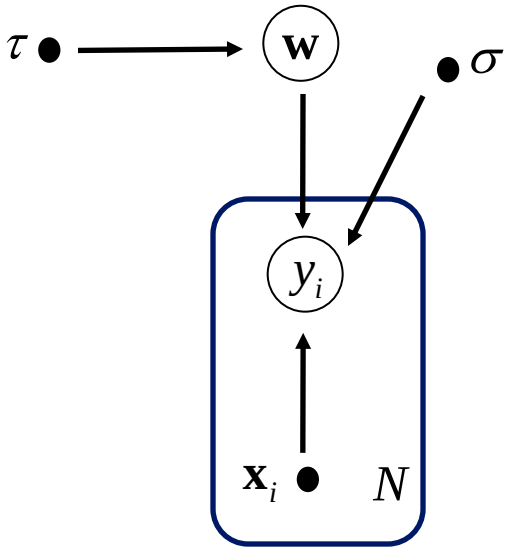
Bayessche Lineare Regression als Graphisches Modell

$$p(y_1,...,y_N,\mathbf{w}\,|\,\mathbf{x}_1,...,\mathbf{x}_N,\sigma^2,\tau^2) = p(\mathbf{w}\,|\,\tau^2)\prod_{i=1}^{N} p(y_i\,|\,\mathbf{x}_i,\mathbf{w},\sigma^2)$$

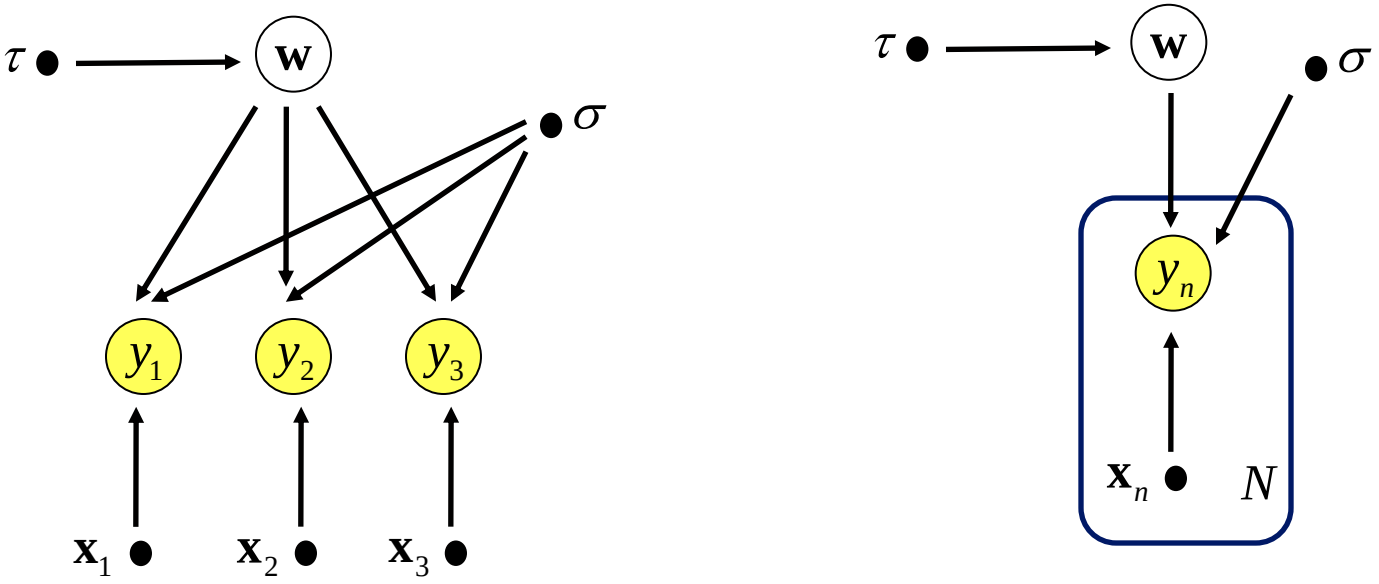
Graphisches Modell, N=3



Graphisches Modell, Plate-Notation



MAP Parameterschätzung als Inferenzproblem



- MAP Parameterschätzung: wahrscheinlichstes Modell gegeben Daten
  - $\mathbf{w}_* = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid y_1, ..., y_N, \mathbf{x}_1, ..., \mathbf{x}_N, \sigma^2, \tau^2)$
  - Inferenzproblem: was ist die Verteilung über den Knoten $\mathbf{w}$, gegeben beobachtete Knoten $y_1, ..., y_N$?

Bayes-optimale Vorhersage

- Vorhersage für neue Testinstanz $\mathbf{x}$:

$$\mathbf{x} \mapsto y$$

- Vorhersage mit MAP Modell:

$$\mathbf{w}_* = \arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{y}, X, \sigma^2, \tau^2)$$

$$y_* = \arg\max_y p(y \mid \mathbf{x}, \mathbf{w}_*, \sigma^2)$$
$$= \mathbf{w}_*^T \mathbf{x}$$

- Alternativ: Bayessche Vorhersage

$$y_* = \arg\max_y p(y \mid \mathbf{x}, \mathbf{y}, X, \sigma^2, \tau^2)$$
$$= \arg\max_y \int p(y \mid \mathbf{x}, \mathbf{w}, \sigma^2)\, p(\mathbf{w} \mid \mathbf{y}, X, \sigma^2, \tau^2) d\mathbf{w}$$

Nicht nötig, sich auf ein Modell fest zu legen

Bayessche Lineare Regression als Graphisches Modell

- Bayessche Vorhersage: Erweiterung des Modells durch neue Testinstanz (neue Zufallsvariable y)

$$p(y_1,...,y_N,y,\mathbf{w}\,|\,\mathbf{x}_1,...,\mathbf{x}_N,\mathbf{x},\sigma^2,\tau^2) = p(\mathbf{w}\,|\,\tau^2)\left(\prod_{i=1}^{N} p(y_i\,|\,\mathbf{w},\mathbf{x}_i,\sigma^2)\right) p(y\,|\,\mathbf{w},\mathbf{x},\sigma^2)$$

Graphisches Modell, N=3

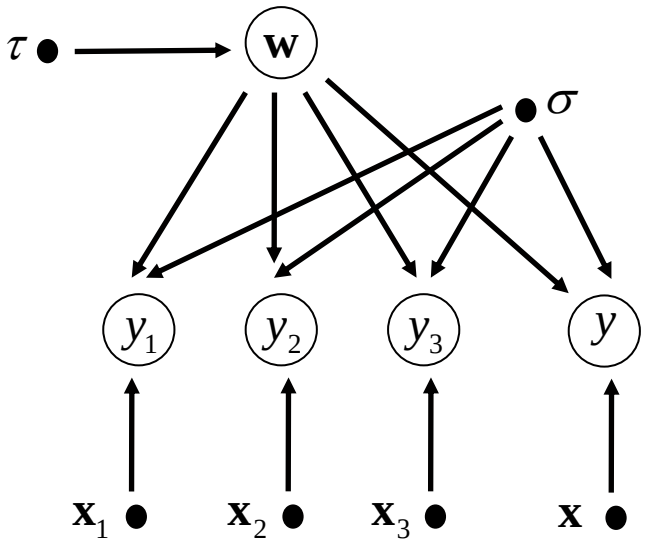
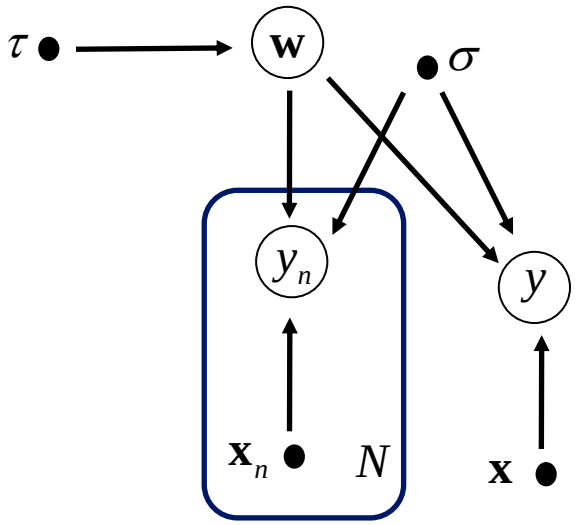
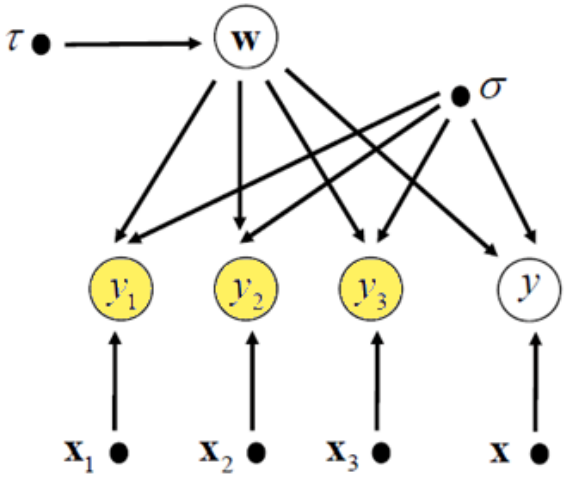
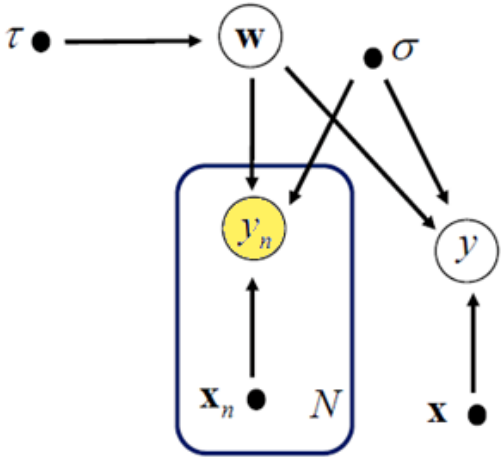


Plate Notation



Bayessche Lineare Regression als Graphisches Modell





- Bayessche Vorhersage
  - $y_* = \arg\max_y \; p(y \mid \mathbf{x}, \mathbf{y}, X, \sigma^2, \tau^2)$
  - Inferenzproblem: was ist der wahrscheinlichste Zustand für Knoten y, gegeben beobachtete Knoten $y_1, \ldots, y_N$?

Überblick

- Graphische Modelle: Syntax und Semantik
- Graphische Modelle im Maschinellen Lernen
- Inferenz in Graphischen Modellen (exakt, approximativ)
- Sequenzmodelle