Wissensentdeckung in Datenbanken

Probabilistische Graphische Modelle II

Nico Piatkowski und Uwe Ligges

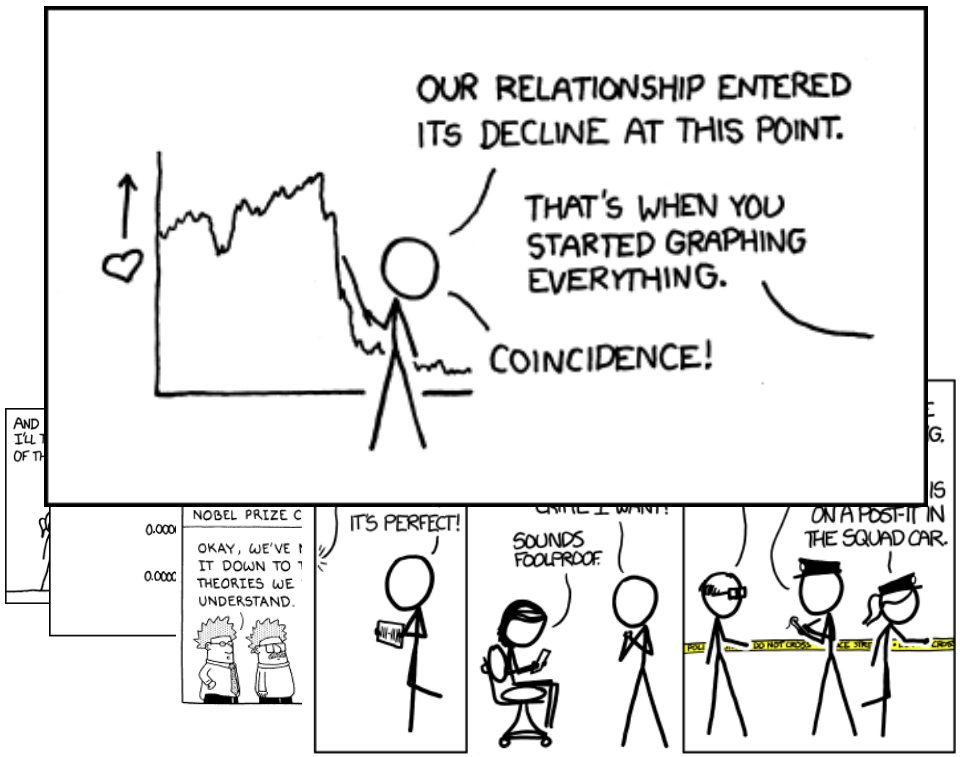
Informatik—Künstliche Intelligenz
Computergestützte Statistik
Technische Universität Dortmund

27.06.2017

Überblick

Was bisher geschah...

- Modellklassen
- Verlustfunktionen
- Numerische Optimierung
- Regularisierung
- Überanpassung
- SQL, Häufige Mengen
- SVM, xDA, Bäume, . . .
- Graphische Modelle



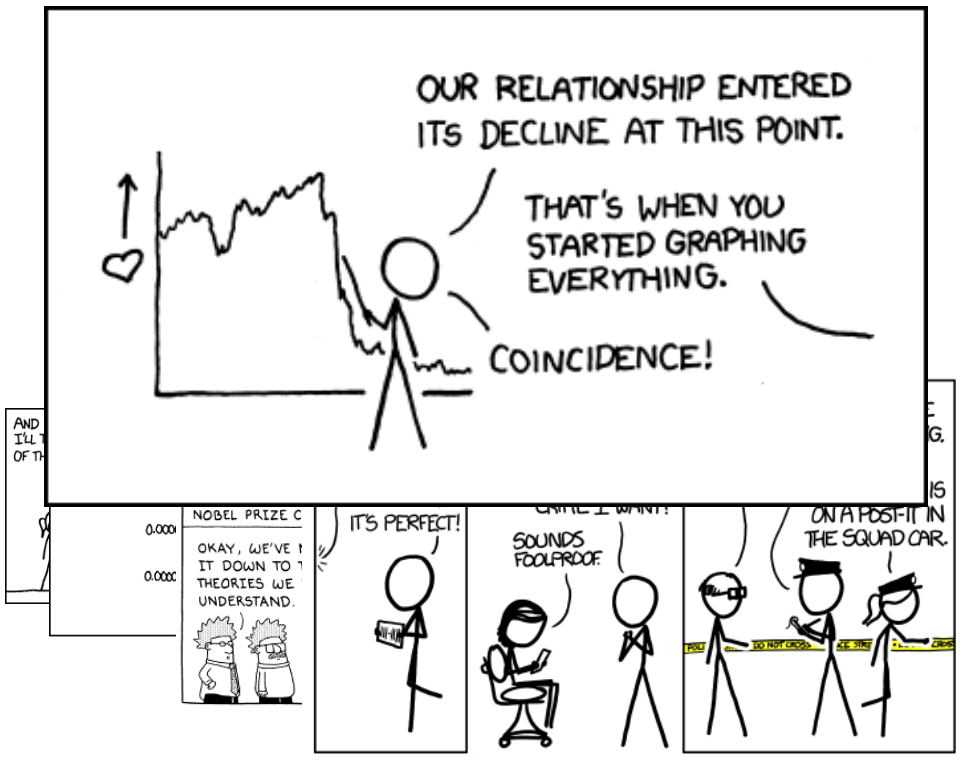
Heute

- Graphische Modelle—Theorie und Algorithmen

Überblick

Was bisher geschah...

- Modellklassen
- Verlustfunktionen
- Numerische Optimierung
- Regularisierung
- Überanpassung
- SQL, Häufige Mengen
- SVM, xDA, Bäume, ...
- Graphische Modelle

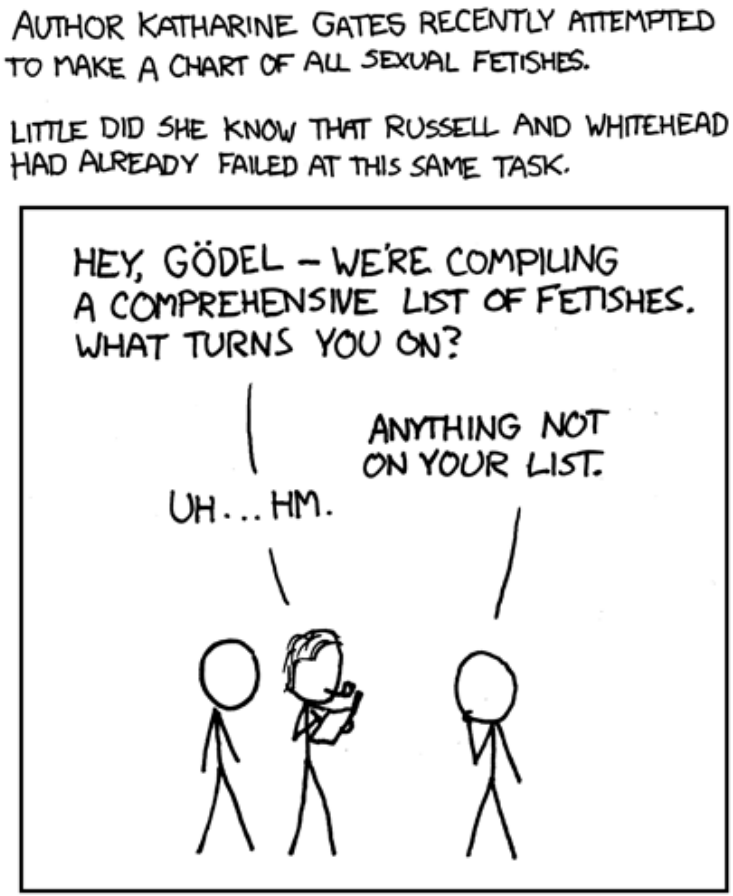


Heute

- Graphische Modelle—Theorie und Algorithmen

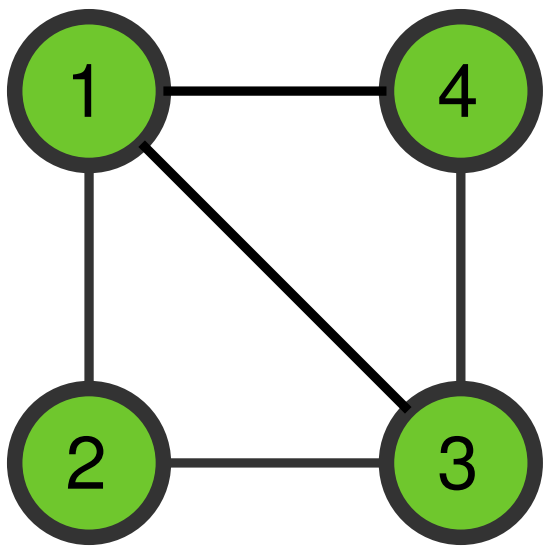
Überblick

- Suffiziente Statistiken
- Maximum-Entropie
- Gradient
- Randverteilung
- Belief Propagation
- Gibbs Sampling



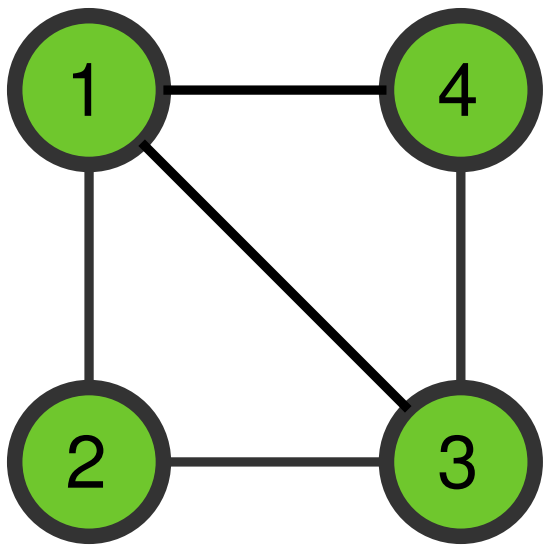
Graph

- $G = (V, E)$ mit Knotenmenge V und Kantenmenge E
- Hier: $V = \{1, 2, 3, 4\}$, $E = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{3, 4\}\}$
- Cliquen: $\mathcal{C}(G) = V \cup E \cup \{\{1, 2, 3\}, \{1, 3, 4\}\}$



Graph

- $G = (V, E)$ mit Knotenmenge V und Kantenmenge E
- Hier: $V = \{1, 2, 3, 4\}$, $E = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{3, 4\}\}$
- Cliquen: $\mathcal{C}(G) = V \cup E \cup \{\{1, 2, 3\}, \{1, 3, 4\}\}$



Suffiziente Statistik

Daten $\mathcal{D}$, Modell mit Parameter $\boldsymbol{\beta}$

Funktion ϕ ist eine suffiziente Statistik $\Leftrightarrow \beta \perp\!\!\!\perp \mathcal{D} \mid \phi(\mathcal{D})$

mit $\phi(\mathcal{D}) = \sum_{\boldsymbol{x} \in \mathcal{D}} \phi(\boldsymbol{x})$

Für diskrete $\mathcal{X}$:

ϕ ist immer gegeben durch $\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C) = \prod_{v \in C} \mathbb{1}_{\boldsymbol{x}_v = \boldsymbol{y}_v}$

mit $C \in \mathcal{C}(G)$, $\boldsymbol{x}_C \in \mathcal{X}_C$, $\boldsymbol{x} \in \mathcal{X}$

$\phi_C(\boldsymbol{x}_C) = (\phi_{C=\boldsymbol{y}_C^1}(\boldsymbol{x}_C), \phi_{C=\boldsymbol{y}_C^2}(\boldsymbol{x}_C), \dots) = (\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C))_{\boldsymbol{y}_C \in \mathcal{X}_C}$

Suffiziente Statistik

Daten $\mathcal{D}$, Modell mit Parameter $\boldsymbol{\beta}$

Funktion ϕ ist eine suffiziente Statistik $\Leftrightarrow \beta \perp\!\!\!\perp \mathcal{D} \mid \phi(\mathcal{D})$

mit $\phi(\mathcal{D}) = \sum_{\boldsymbol{x} \in \mathcal{D}} \phi(\boldsymbol{x})$

Für diskrete $\mathcal{X}$:

ϕ ist immer gegeben durch $\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C) = \prod_{v \in C} \mathbb{1}_{\boldsymbol{x}_v = \boldsymbol{y}_v}$

mit $C \in \mathcal{C}(G)$, $\boldsymbol{x}_C \in \mathcal{X}_C$, $\boldsymbol{x} \in \mathcal{X}$
$\phi_C(\boldsymbol{x}_C) = (\phi_{C=\boldsymbol{y}_C^1}(\boldsymbol{x}_C), \phi_{C=\boldsymbol{y}_C^2}(\boldsymbol{x}_C), \dots) = (\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C))_{\boldsymbol{y}_C \in \mathcal{X}_C}$

Beispiel: Suffiziente Statistik

Beispiel: Suffiziente Statistik

Beispiel: Suffiziente Statistik

Beispiel: Suffiziente Statistik

Beispiel: Suffiziente Statistik

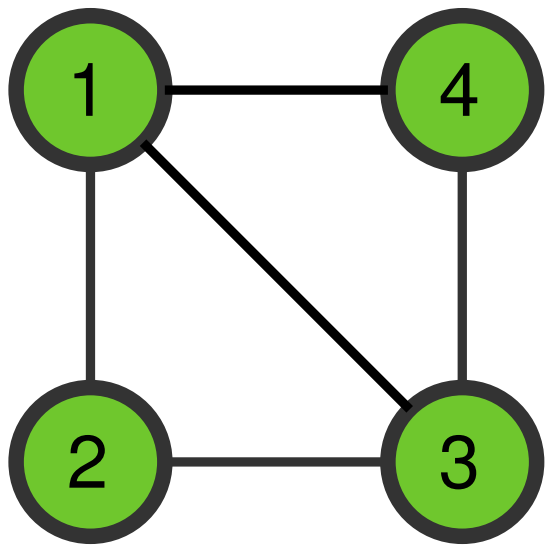
Beispiel: Suffiziente Statistik

$V = \{1,2,3,4\}$, $E = \{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{3,4\}\}$,
$\mathcal{C}(G) = V \cup E \cup \{\{1,2,3\},\{1,3,4\}\}$

$$\phi_C(\boldsymbol{x}_C) = (\phi_{C=\boldsymbol{y}_C^1}(\boldsymbol{x}_C), \phi_{C=\boldsymbol{y}_C^2}(\boldsymbol{x}_C), \dots) = (\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C))_{\boldsymbol{y}_C \in \mathcal{X}_C}$$

$\mathcal{X}_1 = \{1,2,3,4,5\}$, $\mathcal{X}_2 = \{-, ♫\}$,
$\mathcal{X}_3 = \{-, \text{Punk}, \text{Pop}, \dots\}$, $\mathcal{X}_4 = \{■, ■, ■\}$

$\boldsymbol{x} = (2, -, -, ■)$
$\boldsymbol{x}' = (1, -, \text{Pop}, ■)$
$\boldsymbol{x}'' = (1, ♫, \text{Punk}, ■)$
$\dots$



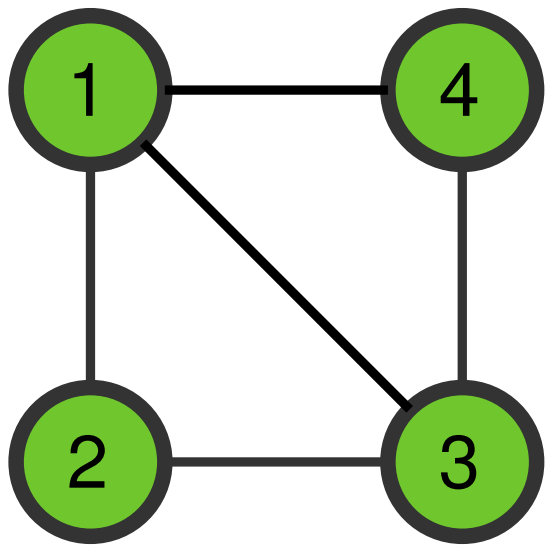
Beispiel: Suffiziente Statistik

$V = \{1,2,3,4\}$, $E = \{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{3,4\}\}$,
$\mathcal{C}(G) = V \cup E \cup \{\{1,2,3\},\{1,3,4\}\}$

$$\phi_C(\boldsymbol{x}_C) = (\phi_{C=\boldsymbol{y}_C^1}(\boldsymbol{x}_C), \phi_{C=\boldsymbol{y}_C^2}(\boldsymbol{x}_C), \dots) = (\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C))_{\boldsymbol{y}_C \in \mathcal{X}_C}$$

$\mathcal{X}_1 = \{1,2,3,4,5\}$, $\mathcal{X}_2 = \{-, ♫\}$,
$\mathcal{X}_3 = \{-, \text{Punk}, \text{Pop}, \dots\}$, $\mathcal{X}_4 = \{■, ■, ■\}$

$x = (2, -, -, ■)$
$x' = (1, -, \text{Pop}, ■)$
$x'' = (1, ♫, \text{Punk}, ■)$
$\dots$



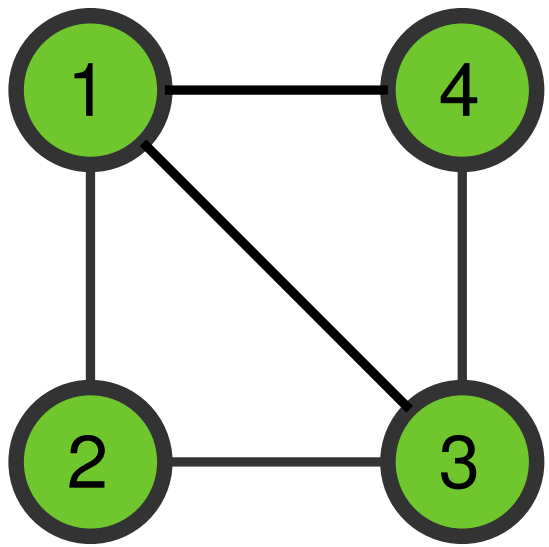
Beispiel: Suffiziente Statistik

$V = \{1,2,3,4\}$, $E = \{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{3,4\}\}$,
$\mathcal{C}(G) = V \cup E \cup \{\{1,2,3\},\{1,3,4\}\}$

$$\phi_C(\boldsymbol{x}_C) = (\phi_{C=\boldsymbol{y}_C^1}(\boldsymbol{x}_C), \phi_{C=\boldsymbol{y}_C^2}(\boldsymbol{x}_C), \dots) = (\phi_{C=\boldsymbol{y}_C}(\boldsymbol{x}_C))_{\boldsymbol{y}_C \in \mathcal{X}_C}$$

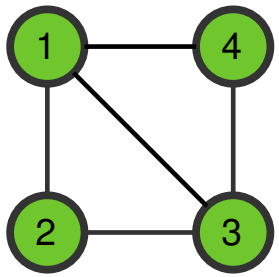
$\mathcal{X}_1 = \{1,2,3,4,5\}$, $\mathcal{X}_2 = \{-, ♫\}$,
$\mathcal{X}_3 = \{-, \text{Punk}, \text{Pop}, \dots\}$, $\mathcal{X}_4 = \{■, ■, ■\}$

$\boldsymbol{x} = (2, -, -, ■)$
$\boldsymbol{x}' = (1, -, \text{Pop}, ■)$
$\boldsymbol{x}'' = (1, ♫, \text{Punk}, ■)$
$\dots$



Beispiel: Suffiziente Statistik II

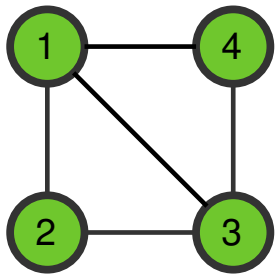
$C = \{1, 2, 3\}$, $\mathcal{X}_C = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$
$\mathcal{X}_1 = \{1, 2, 3, 4, 5\}$, $\mathcal{X}_2 = \{-, \text{♫}\}$,
$\mathcal{X}_3 = \{-, \text{Punk}, \text{Pop}, \text{Rock}, \text{Schlager}\}$
$\mathcal{X}_C = \{(1, -, -), (2, -, -), \ldots, (5, -, -), (1, \text{♫}, -), \ldots,$
$(5, \text{♫}, -), (1, \text{♫}, \text{Punk}), \ldots, (5, \text{♫}, \text{Schlager})\}$



$$\phi_{\{1,2,3\}}(5, \text{♫}, \text{Schlager}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \qquad \phi_{\{1,2,3\}}(2, \text{♫}, -) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Beispiel: Suffiziente Statistik II

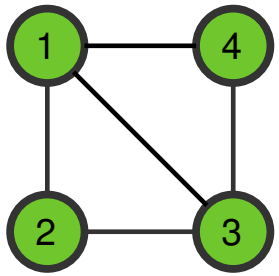
$C = \{1, 2, 3\}$, $\mathcal{X}_C = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3$
$\mathcal{X}_1 = \{1, 2, 3, 4, 5\}$, $\mathcal{X}_2 = \{-, \text{♫}\}$,
$\mathcal{X}_3 = \{-, \text{Punk}, \text{Pop}, \text{Rock}, \text{Schlager}\}$
$\mathcal{X}_C = \{(1, -, -), (2, -, -), \ldots, (5, -, -), (1, \text{♫}, -), \ldots,$
$(5, \text{♫}, -), (1, \text{♫}, \text{Punk}), \ldots, (5, \text{♫}, \text{Schlager})\}$



$$\phi_{\{1,2,3\}}(5, \text{♫}, \text{Schlager}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \qquad \phi_{\{1,2,3\}}(2, \text{♫}, -) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$$

Exponentialfamilie

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) &= \frac{1}{Z(\boldsymbol{\beta})} \exp\Big(\big\langle \boldsymbol{\beta}_{\{1,2,3\}}, \phi_{\{1,2,3\}}(\boldsymbol{x}_{\{1,2,3\}}) \big\rangle \Big) \\ &\quad \exp\Big(\big\langle \boldsymbol{\beta}_{\{1,3,4\}}, \phi_{\{1,3,4\}}(\boldsymbol{x}_{\{1,3,4\}}) \big\rangle \Big) \\ &= \exp\big(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta}) \big)\end{aligned}$$



$$A(\boldsymbol{\beta}) = \log Z(\boldsymbol{\beta}) = \log \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle)$$

Maximum Entropie Prinzip: Aufgabe

Gegeben: Daten $\mathcal{X}$ (Realisierungen einer ZV $\boldsymbol{X}$)
und beliebige Funktion $f : \mathcal{X} \to \mathbb{R}^d$

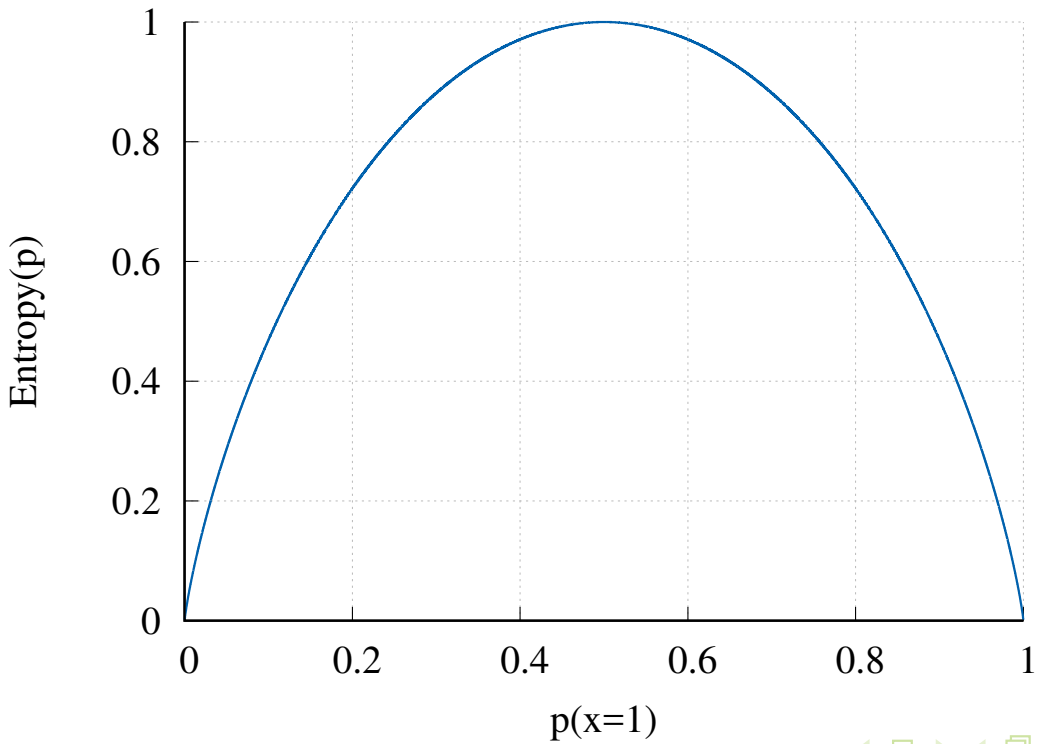
Gesucht: $\mathbb{P}$ mit $\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})] = \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})] = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$

Problem: Viele $\mathbb{P}$ haben die gesuchte Eigenschaft!!

Maximum Entropie Prinzip: Intuition

Entropie $\mathcal{H}$ einer Zufallsvariable $\boldsymbol{X}$ mit $\mathbb{P}$

$$\mathcal{H}(\boldsymbol{X}) = -\sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}(\boldsymbol{x}) \log \mathbb{P}(\boldsymbol{x})$$

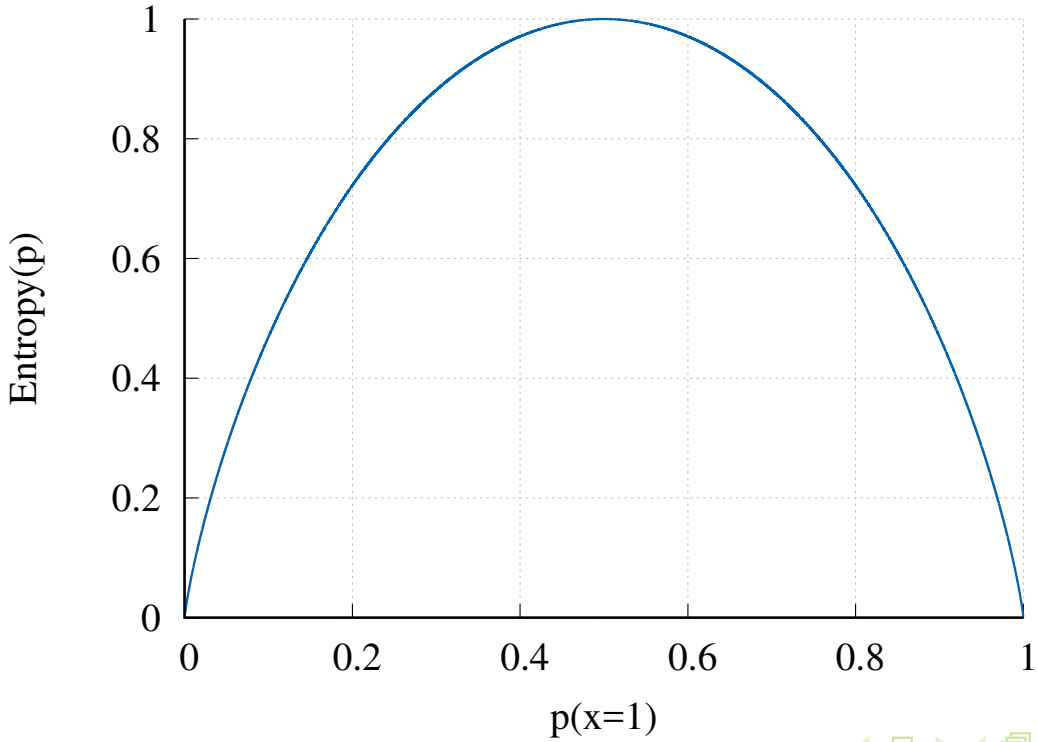


Maximum Entropie Prinzip: Intuition

Sei $\mathcal{P}$ der Raum aller Wahrscheinlichkeitsfunktionen

$$\max_{\mathbb{P} \in \mathcal{P}} \mathcal{H}(\mathbb{P})$$

s.t. $\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})] = \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})] \leftarrow d$ Nebenbedingungen



Maximum Entropie Prinzip: Lösung

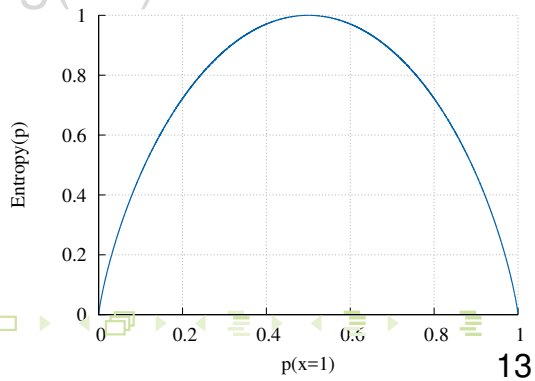
Umformulieren in Lagrange Funktion

$$\max_{\mathbb{P}\in\mathcal{P}} \mathcal{H}(\mathbb{P}) + \sum_{i=1}^{d} \boldsymbol{\lambda}_i \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})]_i - \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})]_i$$

ableiten (nach $\mathbb{P}$!) und $= 0$ setzen liefert

$$\mathbb{P}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\lambda}, f(\boldsymbol{x}) \rangle - A(\boldsymbol{\lambda}))$$

Also: Parameter von Exponentialfamilien sind Lagrange-Multiplikatoren der Nebenbedingung(en)
$\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})] = \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})]$



Maximum Entropie Prinzip: Lösung

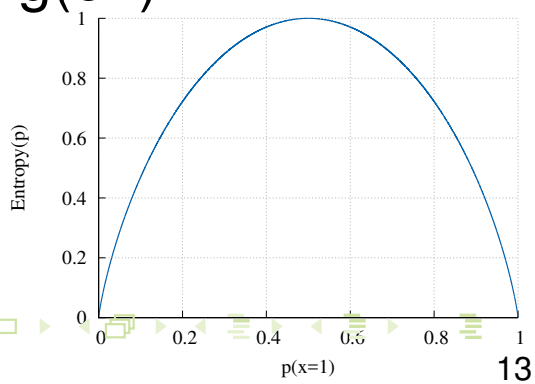
Umformulieren in Lagrange Funktion

$$\max_{\mathbb{P}\in\mathcal{P}} \mathcal{H}(\mathbb{P}) + \sum_{i=1}^{d} \boldsymbol{\lambda}_i \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})]_i - \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})]_i$$

ableiten (nach $\mathbb{P}$!) und $= 0$ setzen liefert

$$\mathbb{P}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\lambda}, f(\boldsymbol{x}) \rangle - A(\boldsymbol{\lambda}))$$

Also: Parameter von Exponentialfamilien sind Lagrange-Multiplikatoren der Nebenbedingung(en)
$\mathbb{E}_{\mathbb{P}}[f(\boldsymbol{X})] = \tilde{\mathbb{E}}_{\mathcal{D}}[f(\boldsymbol{X})]$



Parameterlernen durch Gradientenabstieg

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär) Verlustfunktion: Negative mittlere log-Likelihood

$$\ell(\boldsymbol{\beta}, \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \log \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x})$$

$$= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle + A(\boldsymbol{\beta})$$

$$= -\langle \boldsymbol{\beta}, \tilde{\boldsymbol{\mu}} \rangle + A(\boldsymbol{\beta})$$

Partielle Ableitung:

$$\frac{\partial \ell(\boldsymbol{\beta}, \mathcal{D})}{\partial \boldsymbol{\beta}_i} = -\tilde{\boldsymbol{\mu}}_i + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) = \hat{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_i$$

Parameterlernen durch Gradientenabstieg

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär) Verlustfunktion: Negative mittlere log-Likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}, \mathcal{D}) &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \log \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) \\ &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle + A(\boldsymbol{\beta}) \\ &= -\langle \boldsymbol{\beta}, \tilde{\boldsymbol{\mu}} \rangle + A(\boldsymbol{\beta})\end{aligned}$$

Partielle Ableitung:

$$\frac{\partial \ell(\boldsymbol{\beta}, \mathcal{D})}{\partial \boldsymbol{\beta}_i} = -\tilde{\boldsymbol{\mu}}_i + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) = \hat{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_i$$

Parameterlernen durch Gradientenabstieg

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär) Verlustfunktion: Negative mittlere log-Likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}, \mathcal{D}) &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \log \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) \\ &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle + A(\boldsymbol{\beta}) \\ &= -\langle \boldsymbol{\beta}, \tilde{\boldsymbol{\mu}} \rangle + A(\boldsymbol{\beta})\end{aligned}$$

Partielle Ableitung:

$$\frac{\partial \ell(\boldsymbol{\beta}, \mathcal{D})}{\partial \boldsymbol{\beta}_i} = -\tilde{\boldsymbol{\mu}}_i + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) = \hat{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_i$$

Parameterlernen durch Gradientenabstieg

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär) Verlustfunktion: Negative mittlere log-Likelihood

$$\begin{aligned}\ell(\boldsymbol{\beta}, \mathcal{D}) &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \log \mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) \\ &= -\frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{x} \in \mathcal{D}} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle + A(\boldsymbol{\beta}) \\ &= -\langle \boldsymbol{\beta}, \tilde{\boldsymbol{\mu}} \rangle + A(\boldsymbol{\beta})\end{aligned}$$

Partielle Ableitung:

$$\frac{\partial \ell(\boldsymbol{\beta}, \mathcal{D})}{\partial \boldsymbol{\beta}_i} = -\tilde{\boldsymbol{\mu}}_i + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) = \hat{\boldsymbol{\mu}}_i - \tilde{\boldsymbol{\mu}}_i$$

Marginalisierung

Wenn ϕ binär, dann ist $\boldsymbol{\mu}_i = \mathbb{E}[\phi_i(\boldsymbol{X})]$ die Wahrscheinlichkeit für $\phi_i(\boldsymbol{X}) = 1$

Annahme: Paarweises Modell ≡ Nur die Kantengewichte sind relevant

$$\mathbb{P}(\boldsymbol{X}_v = x) = \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \mathbb{P}(\boldsymbol{y}, x)$$

Ausnutzen der Faktorisierung sowie der Distributivität:

$$\mathbb{P}(\boldsymbol{X}_v = x) = \frac{1}{Z} \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \prod_{C \in \mathcal{C}(G)} \exp(\langle \boldsymbol{\beta}_C, \phi_C(\boldsymbol{x}_C) \rangle)$$

wobei $\boldsymbol{x} = (\boldsymbol{y}, x)$

Marginalisierung

Wenn ϕ binär, dann ist $\boldsymbol{\mu}_i = \mathbb{E}[\phi_i(\boldsymbol{X})]$ die Wahrscheinlichkeit für $\phi_i(\boldsymbol{X}) = 1$

Annahme: Paarweises Modell ≡ Nur die Kantengewichte sind relevant

$$\mathbb{P}(\boldsymbol{X}_v = x) = \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \mathbb{P}(\boldsymbol{y}, x)$$

Ausnutzen der Faktorisierung sowie der Distributivität:

$$\mathbb{P}(\boldsymbol{X}_v = x) = \frac{1}{Z} \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \prod_{C \in \mathcal{C}(G)} \exp(\langle \boldsymbol{\beta}_C, \phi_C(\boldsymbol{x}_C) \rangle)$$

wobei $\boldsymbol{x} = (\boldsymbol{y}, x)$

Marginalisierung

Wenn ϕ binär, dann ist $\boldsymbol{\mu}_i = \mathbb{E}[\phi_i(\boldsymbol{X})]$ die Wahrscheinlichkeit für $\phi_i(\boldsymbol{X}) = 1$

Annahme: Paarweises Modell $\equiv$ Nur die Kantengewichte sind relevant

$$\mathbb{P}(\boldsymbol{X}_v = x) = \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \mathbb{P}(\boldsymbol{y}, x)$$

Ausnutzen der Faktorisierung sowie der Distributivität:

$$\mathbb{P}(\boldsymbol{X}_v = x) = \frac{1}{Z} \sum_{\boldsymbol{y} \in \mathcal{X}_{V \setminus \{v\}}} \prod_{C \in \mathcal{C}(G)} \exp(\langle \boldsymbol{\beta}_C, \phi_C(\boldsymbol{x}_C) \rangle)$$

wobei $\boldsymbol{x} = (\boldsymbol{y}, x)$

Gibbs Sampling

- Idee: Erzeuge neue Stichprobe gemäß $\mathbb{P}_{\boldsymbol{\beta}}$ und berechne $\hat{\boldsymbol{\mu}}_i$ durch "abzählen"
- Aber: Wie erzeugt man neue Samples aus $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{X})$?

$\Rightarrow$ Ausnutzung bedingter Unabhängigkeiten!

Gibbs Sampling

- Idee: Erzeuge neue Stichprobe gemäß $\mathbb{P}_{\boldsymbol{\beta}}$ und berechne $\hat{\boldsymbol{\mu}}_i$ durch "abzählen"
- Aber: Wie erzeugt man neue Samples aus $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{X})$?

$\Rightarrow$ Ausnutzung bedingter Unabhängigkeiten!