Wissensentdeckung in Datenbanken

Belief Propagation, Strukturlernen

Nico Piatkowski und Uwe Ligges

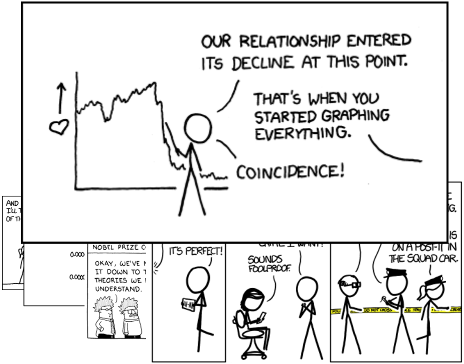
Informatik—Künstliche Intelligenz
Computergestützte Statistik
Technische Universität Dortmund

29.06.2017

Überblick

Was bisher geschah...

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



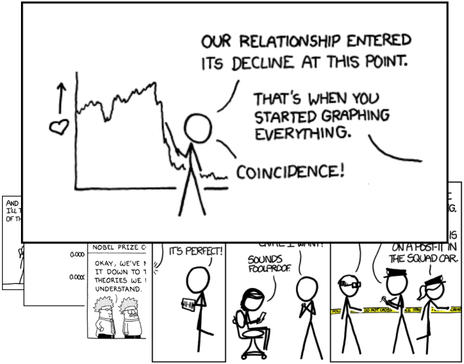
Heute

- Graphische Modelle—Inferenz und Strukturlernen

Überblick

Was bisher geschah...

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

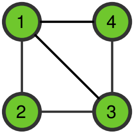


Heute

- Graphische Modelle—Inferenz und Strukturlernen

Mehr Eigenschaften von Graphen

Nachbarschaft: Für Knoten $v \in V$ in Graph $G = (V, E)$



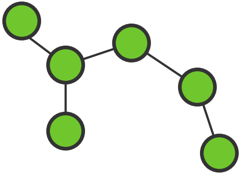
$$\mathcal{N}(v) = \{u \in V \mid \{v, u\} \in E\}$$

Pfad: Folge $(v_1, v_2, \dots, v_m)$ von Knoten in der sich kein Knoten wiederholt

Kreis: Pfad $(v_1, v_2, \dots, v_m)$ mit $\{v_1, v_m\} \in E$

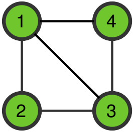
Baum:

- Graph ohne Kreise ("kreisfrei")
- Maximale Cliquengröße $= 2$



Mehr Eigenschaften von Graphen

Nachbarschaft: Für Knoten $v \in V$ in Graph $G = (V, E)$



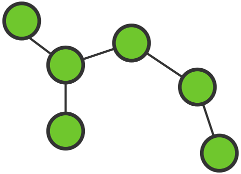
$$\mathcal{N}(v) = \{u \in V \mid \{v, u\} \in E\}$$

Pfad: Folge $(v_1, v_2, \ldots, v_m)$ von Knoten in der sich kein Knoten wiederholt

Kreis: Pfad $(v_1, v_2, \ldots, v_m)$ mit $\{v_1, v_m\} \in E$

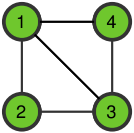
Baum:

- Graph ohne Kreise ("kreisfrei")
- Maximale Cliquengröße $= 2$



Mehr Eigenschaften von Graphen

Nachbarschaft: Für Knoten $v \in V$ in Graph $G = (V, E)$



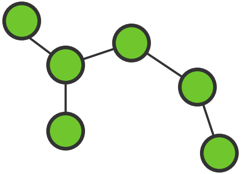
$$\mathcal{N}(v) = \{u \in V \mid \{v, u\} \in E\}$$

Pfad: Folge $(v_1, v_2, \ldots, v_m)$ von Knoten in der sich kein Knoten wiederholt

Kreis: Pfad $(v_1, v_2, \ldots, v_m)$ mit $\{v_1, v_m\} \in E$

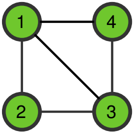
Baum:

- Graph ohne Kreise ("kreisfrei")
- Maximale Cliquengröße $= 2$



Mehr Eigenschaften von Graphen

Nachbarschaft: Für Knoten $v \in V$ in Graph $G = (V, E)$



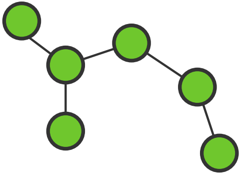
$$\mathcal{N}(v) = \{u \in V \mid \{v, u\} \in E\}$$

Pfad: Folge $(v_1, v_2, \ldots, v_m)$ von Knoten in der sich kein Knoten wiederholt

Kreis: Pfad $(v_1, v_2, \ldots, v_m)$ mit $\{v_1, v_m\} \in E$

Baum:

- Graph ohne Kreise (“kreisfrei”)
- Maximale Cliquengröße $= 2$



Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|) \sum_{\boldsymbol{x} \in \mathcal{D}} \phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta}))$

Negative mittlere log-Likelihood:

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

Partielle Ableitung nach β_i:

$$\frac{\partial \ell(\boldsymbol{\beta}; \mathcal{D})}{\partial \beta_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial \beta_i} A(\boldsymbol{\beta})$$

Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|)\sum_{\boldsymbol{x}\in\mathcal{D}}\phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle\boldsymbol{\beta},\phi(\boldsymbol{x})\rangle - A(\boldsymbol{\beta}))$

Negative mittlere log-Likelihood:

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

Partielle Ableitung nach $\boldsymbol{\beta}_i$:

$$\frac{\partial\ell(\boldsymbol{\beta};\mathcal{D})}{\partial\boldsymbol{\beta}_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial\boldsymbol{\beta}_i}A(\boldsymbol{\beta})$$

Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|)\sum_{\boldsymbol{x}\in\mathcal{D}} \phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle\boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\beta}))$
Negative mittlere log-Likelihood:

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

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

$$= -\left\langle\boldsymbol{\beta}, \tilde{\mathbb{E}}[\phi(\boldsymbol{X})]\right\rangle + A(\boldsymbol{\beta})$$

Partielle Ableitung nach $\boldsymbol{\beta}_i$:

$$\frac{\partial\ell(\boldsymbol{\beta};\mathcal{D})}{\partial\boldsymbol{\beta}_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial\boldsymbol{\beta}_i}A(\boldsymbol{\beta})$$

Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|)\sum_{\boldsymbol{x}\in\mathcal{D}}\phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle\boldsymbol{\beta},\phi(\boldsymbol{x})\rangle - A(\boldsymbol{\beta}))$
Negative mittlere log-Likelihood:

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

Partielle Ableitung nach $\boldsymbol{\beta}_i$:

$$\frac{\partial\ell(\boldsymbol{\beta};\mathcal{D})}{\partial\boldsymbol{\beta}_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial\boldsymbol{\beta}_i}A(\boldsymbol{\beta})$$

Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|)\sum_{\boldsymbol{x}\in\mathcal{D}} \phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle\boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\beta}))$
Negative mittlere log-Likelihood:

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

Partielle Ableitung nach β_i:

$$\frac{\partial\ell(\boldsymbol{\beta};\mathcal{D})}{\partial\boldsymbol{\beta}_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial\boldsymbol{\beta}_i}A(\boldsymbol{\beta})$$

Gradient der Log-Likelihood von Exponentialfamilien

Gegeben Datensatz $\mathcal{D}$, Funktion ϕ (binär),
$\tilde{\mathbb{E}}[\phi(\boldsymbol{X})] = (1/|\mathcal{D}|) \sum_{\boldsymbol{x} \in \mathcal{D}} \phi(\boldsymbol{x})$, $\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta}))$
Negative mittlere log-Likelihood:

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

Partielle Ableitung nach $\boldsymbol{\beta}_i$:

$$\frac{\partial \ell(\boldsymbol{\beta}; \mathcal{D})}{\partial \boldsymbol{\beta}_i} = -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta})$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ (= $\log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}_i} \log \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta})) \phi(\boldsymbol{x})_i \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]
\end{aligned}$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ (= $\log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

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

$$= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle)$$

$$= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle$$

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

$$= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ ($= \log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}_i} \log \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta})) \phi(\boldsymbol{x})_i \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]
\end{aligned}
$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ (= $\log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}_i} \log \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta})) \phi(\boldsymbol{x})_i \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]
\end{aligned}$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ ($= \log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}_i} \log \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x}) \rangle - A(\boldsymbol{\beta})) \phi(\boldsymbol{x})_i \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]
\end{aligned}
$$

Ableitung der Normalisierung

Erinnerung: Form von $A(\boldsymbol{\beta})$ (= $\log Z(\boldsymbol{\beta})$) ist abhängig von $\mathcal{X}$ (diskret?,reel?,endlich?).Hier:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}_i} \log \sum_{\boldsymbol{x}\in\mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\beta}_i} \sum_{\boldsymbol{x}\in\mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle) \\
&= \frac{1}{Z(\boldsymbol{\beta})} \sum_{\boldsymbol{x}\in\mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle) \frac{\partial}{\partial \boldsymbol{\beta}_i} \langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} \exp(\langle \boldsymbol{\beta}, \phi(\boldsymbol{x})\rangle - A(\boldsymbol{\beta}))\phi(\boldsymbol{x})_i \\
&= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]
\end{aligned}
$$

Gradient der Log-Likelihood von Exponentialfamilien

Also gilt für die **partielle Ableitung** nach $\boldsymbol{\beta}_i$:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta}; \mathcal{D})}{\partial \boldsymbol{\beta}_i} &= -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) \\ &= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i] - \tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i]\end{aligned}$$

Ableitung ist beschränkt $\frac{\partial \ell(\boldsymbol{\beta}; \mathcal{D})}{\partial \boldsymbol{\beta}_i} \in [-1; 1] \Rightarrow \ell$ ist Lipschitz stetig
Man kann zeigen: $\nabla \ell(\boldsymbol{\beta}; \mathcal{D})$ auch Lipschitz stetig ($L = 2|\mathcal{C}(G)|$)

Jetzt: Wie berechnet man $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]$?

Jedes i entspricht einem Paar von Clique und Zustand, d.h. $\phi(\boldsymbol{X})_i = \phi(\boldsymbol{X})_{C=\boldsymbol{y}}$ für ein $C \in \mathcal{C}(g)$ und $\boldsymbol{y} \in \mathcal{X}_C$

Gradient der Log-Likelihood von Exponentialfamilien

Also gilt für die **partielle Ableitung** nach $\boldsymbol{\beta}_i$:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta};\mathcal{D})}{\partial \boldsymbol{\beta}_i} &= -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) \\ &= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i] - \tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i]\end{aligned}$$

Ableitung ist beschränkt $\frac{\partial \ell(\boldsymbol{\beta};\mathcal{D})}{\partial \boldsymbol{\beta}_i} \in [-1;1] \Rightarrow \ell$ ist Lipschitz stetig
Man kann zeigen: $\nabla \ell(\boldsymbol{\beta};\mathcal{D})$ auch Lipschitz stetig ($L = 2|\mathcal{C}(G)|$)

Jetzt: Wie berechnet man $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]$?

Jedes i entspricht einem Paar von Clique und Zustand, d.h.
$\phi(\boldsymbol{X})_i = \phi(\boldsymbol{X})_{C=\boldsymbol{y}}$ für ein $C \in \mathcal{C}(g)$ und $\boldsymbol{y} \in \mathcal{X}_C$

Gradient der Log-Likelihood von Exponentialfamilien

Also gilt für die **partielle Ableitung** nach $\boldsymbol{\beta}_i$:

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta};\mathcal{D})}{\partial \boldsymbol{\beta}_i} &= -\tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i] + \frac{\partial}{\partial \boldsymbol{\beta}_i} A(\boldsymbol{\beta}) \\ &= \mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i] - \tilde{\mathbb{E}}[\phi(\boldsymbol{X})_i]\end{aligned}$$

Ableitung ist beschränkt $\frac{\partial \ell(\boldsymbol{\beta};\mathcal{D})}{\partial \boldsymbol{\beta}_i} \in [-1;1] \Rightarrow \ell$ ist Lipschitz stetig
Man kann zeigen: $\nabla \ell(\boldsymbol{\beta};\mathcal{D})$ auch Lipschitz stetig ($L = 2|\mathcal{C}(G)|$)

Jetzt: Wie berechnet man $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_i]$?

Jedes i entspricht einem Paar von Clique und Zustand, d.h.
$\phi(\boldsymbol{X})_i = \phi(\boldsymbol{X})_{C=\boldsymbol{y}}$ für ein $C \in \mathcal{C}(g)$ und $\boldsymbol{y} \in \mathcal{X}_C$

Marginalisierung

Wenn Z binäre Zufallsvariable ($\mathcal{Z} = \{0, 1\}$), dann
$\mathbb{E}[Z] = \mathbb{P}(Z = 1)$

Ziel: Berechnung von $\mathbb{E}_{\beta}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

Allgemein:

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

Ausnutzen der Faktorisierung sowie der Distributivität:

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

wobei $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x})$ und $\boldsymbol{z}_U$ sind die Werte der Knoten in $U \subseteq V$

Marginalisierung

Wenn Z binäre Zufallsvariable ($\mathcal{Z} = \{0, 1\}$), dann $\mathbb{E}[Z] = \mathbb{P}(Z = 1)$

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

Allgemein:

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

Ausnutzen der Faktorisierung sowie der Distributivität:

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

wobei $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x})$ und $\boldsymbol{z}_U$ sind die Werte der Knoten in $U \subseteq V$

Marginalisierung

Wenn Z binäre Zufallsvariable ($\mathcal{Z} = \{0, 1\}$), dann
$\mathbb{E}[Z] = \mathbb{P}(Z = 1)$

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

Allgemein:

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

Ausnutzen der Faktorisierung sowie der Distributivität:

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

wobei $\boldsymbol{z} = (\boldsymbol{y}, \boldsymbol{x})$ und $\boldsymbol{z}_U$ sind die Werte der Knoten in $U \subseteq V$

Marginalisierung in Bäumen

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

Wenn G ein Baum ist mit $V = \{1, 2, \ldots, n\}$, dann ist

$$\mathbb{P}_{\boldsymbol{\beta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\beta})} \prod_{uv \in E} \exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle)$$

und

$$\begin{aligned} Z(\boldsymbol{\beta}) &= \sum_{\boldsymbol{x} \in \mathcal{X}} \prod_{uv \in E} \exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle) \\ &= \sum_{\boldsymbol{x}_1 \in \mathcal{X}_1} \sum_{\boldsymbol{x}_2 \in \mathcal{X}_2} \cdots \sum_{\boldsymbol{x}_n \in \mathcal{X}_n} \prod_{uv \in E} \exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle) \end{aligned}$$

$\rightarrow$ Distributivität ausnutzen!

Belief Propagation

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

$$Z(\boldsymbol{\beta}) = \sum_{\boldsymbol{x}_1 \in \mathcal{X}_1} \sum_{\boldsymbol{x}_2 \in \mathcal{X}_2} \cdots \sum_{\boldsymbol{x}_n \in \mathcal{X}_n} \prod_{uv \in E} \underbrace{\exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle)}_{\psi_{uv}(\boldsymbol{x}_{uv})}$$

Distributivität ausnutzen. . .

$$m_{v \to u}(x) = \sum_{y \in \mathcal{X}_v} \psi_{uv}(yx) \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y)$$

$$Z(\boldsymbol{\beta}) = \sum_{y \in \mathcal{X}_v} \prod_{w \in \mathcal{N}(v)} m_{w \to v}(y)$$

$$\mathbb{P}(\boldsymbol{X}_{uv} = xy) = \frac{\psi_{uv}(yx)}{Z(\boldsymbol{\beta})} \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y) \prod_{w \in \mathcal{N}(u) \setminus \{v\}} m_{w \to u}(y)$$

Belief Propagation

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

$$Z(\boldsymbol{\beta}) = \sum_{\boldsymbol{x}_1 \in \mathcal{X}_1} \sum_{\boldsymbol{x}_2 \in \mathcal{X}_2} \cdots \sum_{\boldsymbol{x}_n \in \mathcal{X}_n} \prod_{uv \in E} \underbrace{\exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle)}_{\psi_{uv}(\boldsymbol{x}_{uv})}$$

Distributivität ausnutzen...

$$m_{v \to u}(x) = \sum_{y \in \mathcal{X}_v} \psi_{uv}(yx) \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y)$$

$$Z(\boldsymbol{\beta}) = \sum_{y \in \mathcal{X}_v} \prod_{w \in \mathcal{N}(v)} m_{w \to v}(y)$$

$$\mathbb{P}(\boldsymbol{X}_{uv} = xy) = \frac{\psi_{uv}(yx)}{Z(\boldsymbol{\beta})} \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y) \prod_{w \in \mathcal{N}(u) \setminus \{v\}} m_{w \to u}(y)$$

Belief Propagation

Ziel: Berechnung von $\mathbb{E}_{\boldsymbol{\beta}}[\phi(\boldsymbol{X})_{C=\boldsymbol{y}}] = \mathbb{P}(\boldsymbol{X}_C = \boldsymbol{y})$

$$Z(\boldsymbol{\beta}) = \sum_{\boldsymbol{x}_1 \in \mathcal{X}_1} \sum_{\boldsymbol{x}_2 \in \mathcal{X}_2} \cdots \sum_{\boldsymbol{x}_n \in \mathcal{X}_n} \prod_{uv \in E} \underbrace{\exp(\langle \boldsymbol{\beta}_{uv}, \phi_{uv}(\boldsymbol{x}_{uv}) \rangle)}_{\psi_{uv}(\boldsymbol{x}_{uv})}$$

Distributivität ausnutzen...

$$m_{v \to u}(x) = \sum_{y \in \mathcal{X}_v} \psi_{uv}(yx) \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y)$$

$$Z(\boldsymbol{\beta}) = \sum_{y \in \mathcal{X}_v} \prod_{w \in \mathcal{N}(v)} m_{w \to v}(y)$$

$$\mathbb{P}(\boldsymbol{X}_{uv} = xy) = \frac{\psi_{uv}(yx)}{Z(\boldsymbol{\beta})} \prod_{w \in \mathcal{N}(v) \setminus \{u\}} m_{w \to v}(y) \prod_{w \in \mathcal{N}(u) \setminus \{v\}} m_{w \to u}(y)$$

Gibbs Sampling

- Problem: Belief Propagation nur exakt wenn G ein Baum ist!
- 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!

Beobachtung: Wenn ganze Nachbarschaft eines Knotens v beobachtet ist, kann $\mathbb{P}_v(x \mid x_{\mathcal{N}(v)})$ einfach berechnet werden!

Gibbs Sampling

- Problem: Belief Propagation nur exakt wenn G ein Baum ist!
- 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!

Beobachtung: Wenn ganze Nachbarschaft eines Knotens v beobachtet ist, kann $\mathbb{P}_v(x \mid \boldsymbol{x}_{\mathcal{N}(v)})$ einfach berechnet werden!

Gibbs Sampling: Algorithmus

1. Erzeuge $\boldsymbol{x}$ zufällig Gleichverteilt (das entspricht **NICHT** $\mathbb{P}_{\boldsymbol{\beta}}$!)
2. Besuche jeden Knoten $v \in V$ und weise gemäß $\mathbb{P}_v(x \mid \boldsymbol{x}_{\mathcal{N}(v)})$ neuen Wert zu
3. Wiederhole Schritt 2 so oft wie möglich

Man kann zeigen: Nach endlicher Anzahl von Schritten ist $\boldsymbol{x}$ ein echtes Sample aus $\mathbb{P}_{\boldsymbol{\beta}}$!

Dann: Nutze den Algorithmus um “viele” (so viele wie möglich) Samples zu erzeugen und berechne $\mathbb{P}(\boldsymbol{X}_{uv} = xy)$ (für alle Kanten $\{v, u\} \in E$) durch “abzählen”

Gibbs Sampling: Algorithmus

1. Erzeuge $\boldsymbol{x}$ zufällig Gleichverteilt (das entspricht **NICHT** $\mathbb{P}_{\boldsymbol{\beta}}$!)
2. Besuche jeden Knoten $v \in V$ und weise gemäß $\mathbb{P}_v(x \mid \boldsymbol{x}_{\mathcal{N}(v)})$ neuen Wert zu
3. Wiederhole Schritt 2 so oft wie möglich

Man kann zeigen: Nach endlicher Anzahl von Schritten ist $\boldsymbol{x}$ ein echtes Sample aus $\mathbb{P}_{\boldsymbol{\beta}}$!

Dann: Nutze den Algorithmus um "viele" (so viele wie möglich) Samples zu erzeugen und berechne $\mathbb{P}(\boldsymbol{X}_{uv} = xy)$ (für alle Kanten $\{v, u\} \in E$) durch "abzählen"

Chow-Liu Bäume

Minimierung der Distanz zwischen optimalem Graph und "bestem" Baum T
Hier: Distanz gemessen durch Kullback-Leiber Divergenz

$$\text{KL}(\mathbb{P}^*, \mathbb{P}_T) = \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}^*(\boldsymbol{x}) \log \frac{\mathbb{P}^*(\boldsymbol{x})}{\mathbb{P}_T(\boldsymbol{x})}$$

Kann umgeform werden zu

$$\text{KL}(\mathbb{P}^*, \mathbb{P}_T) = -\mathcal{H}(\mathbb{P}^*) + \sum_{v \in V} \mathcal{H}(\mathbb{P}^*_v) - \underbrace{\sum_{vu \in E(T)} I(\boldsymbol{X}_v, \boldsymbol{X}_u)}_{\text{Maximaler Spannbaum!}}$$

mit $I(\boldsymbol{X}_v, \boldsymbol{X}_u) = \text{KL}(\mathbb{P}_{vu}, \mathbb{P}_v \mathbb{P}_u)$ (Allgemeines Maß für Unabhängigkeit!)

Chow-Liu Bäume

Minimierung der Distanz zwischen optimalem Graph und "bestem" Baum T
Hier: Distanz gemessen durch Kullback-Leiber Divergenz

$$\mathrm{KL}(\mathbb{P}^*, \mathbb{P}_T) = \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{P}^*(\boldsymbol{x}) \log \frac{\mathbb{P}^*(\boldsymbol{x})}{\mathbb{P}_T(\boldsymbol{x})}$$

Kann umgeform werden zu

$$\mathrm{KL}(\mathbb{P}^*, \mathbb{P}_T) = -\mathcal{H}(\mathbb{P}^*) + \sum_{v \in V} \mathcal{H}(\mathbb{P}_v^*) - \underbrace{\sum_{vu \in E(T)} I(\boldsymbol{X}_v, \boldsymbol{X}_u)}_{\text{Maximaler Spannbaum!}}$$

mit $I(\boldsymbol{X}_v, \boldsymbol{X}_u) = \mathrm{KL}(\mathbb{P}_{vu}, \mathbb{P}_v \mathbb{P}_u)$ (Allgemeines Maß für Unabhängigkeit!)

Graphen mittels Regularisierung

Baobachtung: Sind ist kompletter Parametervektor $\boldsymbol{\beta}_{vu}$ einer Kante $= 0$, so hat diese Kante keinen Einfluss auf $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$!

Idee: Minimiere $\ell(\boldsymbol{\beta}; \mathcal{D}) + \lambda \|\boldsymbol{\beta}\|_1$

$\|\cdot\|_1$ nicht differenzierbar!! → Proximaler Gradientenabstieg (nächste Woche)

Spoiler:

$$\mathrm{prox}_{\lambda\|\cdot\|_1}(\boldsymbol{\beta}_i) = \begin{cases} \boldsymbol{\beta}_i - \lambda & , \boldsymbol{\beta}_i > \lambda \\ \boldsymbol{\beta}_i + \lambda & , \boldsymbol{\beta}_i < -\lambda \\ 0 & , \text{sonst} \end{cases}$$