

Generative Modelle

Bisher:

Lerne eine unbekannte Zielfunktion approximativ nach Beobachtung zufällig erzeugter Beispiele

Jetzt:

Finde möglichst viel über die zugrunde liegende Beispiel-Verteilung D heraus.

Die Fragestellung ist weitaus schwieriger geworden: Nimm an, dass

- D zu einer eingeschränkten Klasse von Verteilungen gehört bzw.
- bestimmte Unabhängigkeits-Eigenschaften erfüllt.

Maximum Likelihood

Maximum-Likelihood: Der Ansatz

- B ist eine Beobachtung (oder Folge von Beispielen):

Von welcher Verteilung wurde B erzeugt?

- Verteilungen werden über ihre Parameter Θ beschrieben:

$p_{\Theta}(B) :=$ Wahrscheinlichkeit von B „im Modell“ Θ .

Maximum-Likelihood: Der Ansatz

- B ist eine Beobachtung (oder Folge von Beispielen):

Von welcher Verteilung wurde B erzeugt?

- Verteilungen werden über ihre Parameter Θ beschrieben:

$p_{\Theta}(B) :=$ Wahrscheinlichkeit von B „im Modell“ Θ .

Ein Parameterwert Θ_{opt} beschreibt eine

Maximum Likelihood Hypothese (ML-Hypothese),

wenn

$$\text{prob}_{\Theta_{\text{opt}}}[B] = \max_{\Theta} \text{prob}_{\Theta}[B]$$

- 1 X ist eine binäre Zufallsvariable mit unbekannter Verteilung D .
 - ▶ Es ist $\Theta = \text{prob}_{\Theta}[X = 1]$.
 - ▶ Bestimme Θ .
- 2 Wir erhalten die Beobachtung

$$B = (x_1, \dots, x_m) \in \{0, 1\}^m.$$

Chernoff-Ungleichung: Θ wird durch $\frac{\sum_{i=1}^m x_i}{m}$ scharf approximiert.

Was liefert **Maximum-Likelihood**?

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] =$$

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] = \prod_{i=1}^m \Theta^{x_i} \cdot (1 - \Theta)^{1-x_i} =$$

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] = \prod_{i=1}^m \Theta^{x_i} \cdot (1 - \Theta)^{1-x_i} = \Theta^{\sum_{i=1}^m x_i} \cdot (1 - \Theta)^{\sum_{i=1}^m (1-x_i)}.$$

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] = \prod_{i=1}^m \Theta^{x_i} \cdot (1 - \Theta)^{1-x_i} = \Theta^{\sum_{i=1}^m x_i} \cdot (1 - \Theta)^{\sum_{i=1}^m (1-x_i)}.$$

2. Bestimme die **Log-Likelihood** durch Logarithmieren

$$L(B; \Theta) := \sum_{i=1}^m x_i \cdot \log \Theta + \sum_{i=1}^m (1 - x_i) \cdot \log(1 - \Theta).$$

3. Für welchen Wert von Θ wird B am wahrscheinlichsten?

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] = \prod_{i=1}^m \Theta^{x_i} \cdot (1 - \Theta)^{1-x_i} = \Theta^{\sum_{i=1}^m x_i} \cdot (1 - \Theta)^{\sum_{i=1}^m (1-x_i)}.$$

2. Bestimme die **Log-Likelihood** durch Logarithmieren

$$L(B; \Theta) := \sum_{i=1}^m x_i \cdot \log \Theta + \sum_{i=1}^m (1 - x_i) \cdot \log(1 - \Theta).$$

3. Für welchen Wert von Θ wird B am wahrscheinlichsten?

- ▶ Differenziere $L(B; \Theta)$ nach Θ . Nach Null-Setzung:

$$\frac{\sum_{i=1}^m x_i}{\Theta_{\text{opt}}} = \frac{\sum_{i=1}^m (1 - x_i)}{1 - \Theta_{\text{opt}}}.$$

1. Bestimme die Wahrscheinlichkeit von B im Modell Θ , also:

$$\text{prob}_{\Theta}[B] = \prod_{i=1}^m \Theta^{x_i} \cdot (1 - \Theta)^{1-x_i} = \Theta^{\sum_{i=1}^m x_i} \cdot (1 - \Theta)^{\sum_{i=1}^m (1-x_i)}.$$

2. Bestimme die **Log-Likelihood** durch Logarithmieren

$$L(B; \Theta) := \sum_{i=1}^m x_i \cdot \log \Theta + \sum_{i=1}^m (1 - x_i) \cdot \log(1 - \Theta).$$

3. Für welchen Wert von Θ wird B am wahrscheinlichsten?

- ▶ Differenziere $L(B; \Theta)$ nach Θ . Nach Null-Setzung:

$$\frac{\sum_{i=1}^m x_i}{\Theta_{\text{opt}}} = \frac{\sum_{i=1}^m (1 - x_i)}{1 - \Theta_{\text{opt}}}.$$

- ▶ Die Lösung dieser Gleichung ist $\Theta_{\text{opt}} = \frac{1}{m} \cdot \sum_{i=1}^m x_i$

Maximum-Likelihood: Der Mittelwert ist das wahrscheinlichste Modell.

Die unbekannte Normalverteilung

$$\mathcal{P}_{\Theta}[x] := \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

erzeugt $B = (x_1, \dots, x_m) \in \mathbb{R}^m$. Bestimme Erwartungswert und Varianz.

Die unbekannte Normalverteilung

$$\mathcal{P}_{\Theta}[x] := \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

erzeugt $B = (x_1, \dots, x_m) \in \mathbb{R}^m$. Bestimme Erwartungswert und Varianz.

1. Die Log-Likelihood für $\Theta = (\mu, \sigma)$ ist

$$L(x_1, \dots, x_m; \Theta) :=$$

Die unbekannte Normalverteilung

$$\mathcal{P}_{\Theta}[x] := \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

erzeugt $B = (x_1, \dots, x_m) \in \mathbb{R}^m$. Bestimme Erwartungswert und Varianz.

1. Die Log-Likelihood für $\Theta = (\mu, \sigma)$ ist

$$L(x_1, \dots, x_m; \Theta) := \sum_{i=1}^m \log \mathcal{P}_{\Theta}[x_i]$$

Die unbekannte Normalverteilung

$$\mathcal{P}_\Theta[x] := \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

erzeugt $B = (x_1, \dots, x_m) \in \mathbb{R}^m$. Bestimme Erwartungswert und Varianz.

1. Die Log-Likelihood für $\Theta = (\mu, \sigma)$ ist

$$L(x_1, \dots, x_m; \Theta) := \sum_{i=1}^m \log \mathcal{P}_\Theta[x_i] = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i - \mu)^2 - m \log(\sigma \cdot \sqrt{2\pi}).$$

Die unbekannte Normalverteilung

$$\mathcal{P}_{\Theta}[x] := \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

erzeugt $B = (x_1, \dots, x_m) \in \mathbb{R}^m$. Bestimme Erwartungswert und Varianz.

1. Die Log-Likelihood für $\Theta = (\mu, \sigma)$ ist

$$L(x_1, \dots, x_m; \Theta) := \sum_{i=1}^m \log \mathcal{P}_{\Theta}[x_i] = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i - \mu)^2 - m \log(\sigma \cdot \sqrt{2\pi}).$$

2. Differenziere nach μ :

$$\frac{\partial L}{\partial \mu}(x_1, \dots, x_m; \Theta) = \frac{1}{\sigma^2} \sum_{i=1}^m (x_i - \mu) \stackrel{!}{=} 0$$

3. Es ist

$$L(x_1, \dots, x_m; \Theta) = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i - \mu)^2 - m \log(\sigma \cdot \sqrt{2\pi}).$$

Differenziere nach σ :

$$\frac{\partial L}{\partial \sigma}(x_1, \dots, x_m; \Theta) =$$

3. Es ist

$$L(x_1, \dots, x_m; \Theta) = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i - \mu)^2 - m \log(\sigma \cdot \sqrt{2\pi}).$$

Differenziere nach σ :

$$\frac{\partial L}{\partial \sigma}(x_1, \dots, x_m; \Theta) = \frac{1}{\sigma^3} \sum_{i=1}^m (x_i - \mu)^2 - \frac{m}{\sigma} \stackrel{!}{=} 0.$$

3. Es ist

$$L(x_1, \dots, x_m; \Theta) = \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^m (x_i - \mu)^2 - m \log(\sigma \cdot \sqrt{2\pi}).$$

Differenziere nach σ :

$$\frac{\partial L}{\partial \sigma}(x_1, \dots, x_m; \Theta) = \frac{1}{\sigma^3} \sum_{i=1}^m (x_i - \mu)^2 - \frac{m}{\sigma} \stackrel{!}{=} 0.$$

Maximum-Likelihood liefert als wahrscheinlichste Parameter

$$\mu_{\text{opt}} = \frac{1}{m} \sum_{i=1}^m x_i \quad \text{und} \quad \sigma_{\text{opt}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\text{opt}})^2}.$$

Maximum-Likelihood und neuronale Netzwerke

Backpropagation versucht, den empirischen **Log-Loss**

$$\text{Loss}_D^S = -\frac{1}{s} \cdot \sum_{(x,y) \in S} \log p_w(y|x),$$

zu minimieren.

Maximum-Likelihood und neuronale Netzwerke

Backpropagation versucht, den empirischen **Log-Loss**

$$\text{Loss}_D^S = -\frac{1}{s} \cdot \sum_{(x,y) \in S} \log p_w(y|x),$$

zu minimieren. Für die tatsächliche Wahrscheinlichkeit $p(x, y)$ von (x, y) ist

$$\text{Loss}_D(w) = \mathbb{E}_{(x,y) \sim D}[\ell(w, x, y)] = - \sum_{(x,y) \in X \times Y} p(x, y) \cdot \log p_w(y|x)$$

Maximum-Likelihood und neuronale Netzwerke

Backpropagation versucht, den empirischen **Log-Loss**

$$\text{Loss}_D^S = -\frac{1}{s} \cdot \sum_{(x,y) \in S} \log p_w(y|x),$$

zu minimieren. Für die tatsächliche Wahrscheinlichkeit $p(x, y)$ von (x, y) ist

$$\begin{aligned} \text{Loss}_D(w) &= \mathbb{E}_{(x,y) \sim D}[\ell(w, x, y)] = - \sum_{(x,y) \in X \times Y} p(x, y) \cdot \log p_w(y|x) \\ &= \sum_{x \in X} p(x) \cdot \sum_{y \in Y} p(y|x) \cdot \log\left(\frac{1}{p_w(y|x)}\right) \end{aligned}$$

Maximum-Likelihood und neuronale Netzwerke

Backpropagation versucht, den empirischen **Log-Loss**

$$\text{Loss}_D^S = -\frac{1}{s} \cdot \sum_{(x,y) \in S} \log p_w(y|x),$$

zu minimieren. Für die tatsächliche Wahrscheinlichkeit $p(x, y)$ von (x, y) ist

$$\begin{aligned} \text{Loss}_D(w) &= \mathbb{E}_{(x,y) \sim D}[\ell(w, x, y)] = - \sum_{(x,y) \in X \times Y} p(x, y) \cdot \log p_w(y|x) \\ &= \sum_{x \in X} p(x) \cdot \sum_{y \in Y} p(y|x) \cdot \log \left(\frac{1}{p_w(y|x)} \right) \\ &= \underbrace{\sum_{x \in X} p(x) \cdot \sum_{y \in Y} p(y|x) \cdot \log \left(\frac{p(y|x)}{p_w(y|x)} \right)}_{\text{Kullback-Leibler Divergenz} \geq 0} + \underbrace{\sum_{x \in X} p(x) \cdot \sum_{y \in Y} p(y|x) \cdot \log \left(\frac{1}{p(y|x)} \right)}_{\text{Entropie } H(p(*|x))} \end{aligned}$$

$$\text{Loss}_D(w) \geq \sum_{x \in X} p(x) \cdot \sum_{y \in Y} p(y|x) \cdot \log \left(\frac{1}{p(y|x)} \right).$$

(mittlere Entropie)

Maximum a Posteriori Hypothese

MAP-Hypothesen

Sei \mathcal{H} eine Hypothesenklasse.

- Hypothesen entsprechen diesmal Zufallsvariablen.
- Der **Prior** bzw. die **A-priori Verteilung**

$$\text{prob}[h]$$

sei bekannt ebenso wie die Wahrscheinlichkeit

$$\text{prob}[B|h]$$

mit der B bei angenommener Hypothese h erzeugt wird.

MAP-Hypothesen

Sei \mathcal{H} eine Hypothesenklasse.

- Hypothesen entsprechen diesmal Zufallsvariablen.
- Der **Prior** bzw. die **A-priori Verteilung**

$$\text{prob}[h]$$

sei bekannt ebenso wie die Wahrscheinlichkeit

$$\text{prob}[B|h]$$

mit der B bei angenommener Hypothese h erzeugt wird.

Bestimme die **Maximum a Posteriori** (oder **MAP-**) **Hypothese**

$$\text{prob}[h^* | B] = \max_{h \in \mathcal{H}} \text{prob}[h | B]$$

Der Satz von Bayes

Bekannt: $\text{prob}[h]$ und $\text{prob}[B|h]$.

Gesucht: $\max_{h \in \mathcal{H}} \text{prob}[h|B]$.

Der Satz von Bayes

Bekannt: $\text{prob}[h]$ und $\text{prob}[B|h]$.

Gesucht: $\max_{h \in \mathcal{H}} \text{prob}[h|B]$.

$$\text{prob}[h|B] = \frac{\text{prob}[B|h] \cdot \text{prob}[h]}{\text{prob}[B]}.$$

Der Satz von Bayes

Bekannt: $\text{prob}[h]$ und $\text{prob}[B|h]$.

Gesucht: $\max_{h \in \mathcal{H}} \text{prob}[h|B]$.

$$\text{prob}[h|B] = \frac{\text{prob}[B|h] \cdot \text{prob}[h]}{\text{prob}[B]}.$$

Warum? Es ist

$$\text{prob}[h|B] \cdot \text{prob}[B] = \text{prob}[h, B]$$

Der Satz von Bayes

Bekannt: $\text{prob}[h]$ und $\text{prob}[B|h]$.

Gesucht: $\max_{h \in \mathcal{H}} \text{prob}[h|B]$.

$$\text{prob}[h|B] = \frac{\text{prob}[B|h] \cdot \text{prob}[h]}{\text{prob}[B]}.$$

Warum? Es ist

$$\text{prob}[h|B] \cdot \text{prob}[B] = \text{prob}[h, B] = \text{prob}[B|h] \cdot \text{prob}[h]$$

Wenn alle Hypothesen gleichwahrscheinlich sind:

Der Satz von Bayes

Bekannt: $\text{prob}[h]$ und $\text{prob}[B|h]$.

Gesucht: $\max_{h \in \mathcal{H}} \text{prob}[h|B]$.

$$\text{prob}[h|B] = \frac{\text{prob}[B|h] \cdot \text{prob}[h]}{\text{prob}[B]}.$$

Warum? Es ist

$$\text{prob}[h|B] \cdot \text{prob}[B] = \text{prob}[h, B] = \text{prob}[B|h] \cdot \text{prob}[h]$$

Wenn alle Hypothesen gleichwahrscheinlich sind:

ML-Hypothesen = MAP-Hypothesen

Fakten:

- 0.8% der Bevölkerung haben eine bestimmte Erkrankung,
- Wenn eine Erkrankung vorliegt: Positiver Befund in 98% aller Fälle
- Wenn nicht erkrankt: Negativer Befund in 97% aller Fälle.

Fakten:

- **0.8%** der Bevölkerung haben eine bestimmte Erkrankung,
- Wenn eine Erkrankung vorliegt: Positiver Befund in **98%** aller Fälle
- Wenn nicht erkrankt: Negativer Befund in **97%** aller Fälle.

Hypothesenklasse: $\mathcal{H} = \{ \textit{erkrankt}, \textit{nicht erkrankt} \}$.

Bestimme eine MAP-Hypothese, wobei

- **Prior**: $\text{prob}[\textit{erkrankt}] = 0,008$, $\text{prob}[\textit{nicht erkrankt}] = 0,992$.
- Die **Wahrscheinlichkeit von Befunden**:
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{erkrankt}] =$

Fakten:

- **0.8%** der Bevölkerung haben eine bestimmte Erkrankung,
- Wenn eine Erkrankung vorliegt: Positiver Befund in **98%** aller Fälle
- Wenn nicht erkrankt: Negativer Befund in **97%** aller Fälle.

Hypothesenklasse: $\mathcal{H} = \{ \textit{erkrankt}, \textit{nicht erkrankt} \}$.

Bestimme eine MAP-Hypothese, wobei

- **Prior**: $\text{prob}[\textit{erkrankt}] = 0,008$, $\text{prob}[\textit{nicht erkrankt}] = 0,992$.
- Die **Wahrscheinlichkeit von Befunden**:
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{erkrankt}] = 0.98$,
 - ▶ $\text{prob}[\text{Befund: nicht erkrankt} \mid \textit{erkrankt}] =$

Fakten:

- **0.8%** der Bevölkerung haben eine bestimmte Erkrankung,
- Wenn eine Erkrankung vorliegt: Positiver Befund in **98%** aller Fälle
- Wenn nicht erkrankt: Negativer Befund in **97%** aller Fälle.

Hypothesenklasse: $\mathcal{H} = \{ \textit{erkrankt}, \textit{nicht erkrankt} \}$.

Bestimme eine MAP-Hypothese, wobei

- **Prior**: $\text{prob}[\textit{erkrankt}] = 0,008$, $\text{prob}[\textit{nicht erkrankt}] = 0,992$.
- Die **Wahrscheinlichkeit von Befunden**:
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{erkrankt}] = 0.98$,
 - ▶ $\text{prob}[\text{Befund: nicht erkrankt} \mid \textit{erkrankt}] = 0.02$,
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{nicht-erkrankt}] =$

Fakten:

- **0.8%** der Bevölkerung haben eine bestimmte Erkrankung,
- Wenn eine Erkrankung vorliegt: Positiver Befund in **98%** aller Fälle
- Wenn nicht erkrankt: Negativer Befund in **97%** aller Fälle.

Hypothesenklasse: $\mathcal{H} = \{ \textit{erkrankt}, \textit{nicht erkrankt} \}$.

Bestimme eine MAP-Hypothese, wobei

- **Prior**: $\text{prob}[\textit{erkrankt}] = 0,008$, $\text{prob}[\textit{nicht erkrankt}] = 0,992$.
- Die **Wahrscheinlichkeit von Befunden**:
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{erkrankt}] = 0.98$,
 - ▶ $\text{prob}[\text{Befund: nicht erkrankt} \mid \textit{erkrankt}] = 0.02$,
 - ▶ $\text{prob}[\text{Befund: erkrankt} \mid \textit{nicht-erkrankt}] = 0.03$
 - ▶ $\text{prob}[\text{Befund: nicht-erkrankt} \mid \textit{nicht-erkrankt}] = 0.97$.

- $\text{prob}[\text{Befund: erkrankt} \mid \text{erkrankt}] = 0.98,$
- $\text{prob}[\text{Befund: nicht erkrankt} \mid \text{erkrankt}] = 0.02,$
- $\text{prob}[\text{Befund: erkrankt} \mid \text{nicht-erkrankt}] = 0.03$
- $\text{prob}[\text{Befund: nicht-erkrankt} \mid \text{nicht-erkrankt}] = 0.97.$

Mit dem Satz von Bayes:

$$\text{prob}[\text{erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}]$$

=

- $\text{prob}[\text{Befund: erkrankt} \mid \text{erkrankt}] = 0.98,$
- $\text{prob}[\text{Befund: nicht erkrankt} \mid \text{erkrankt}] = 0.02,$
- $\text{prob}[\text{Befund: erkrankt} \mid \text{nicht-erkrankt}] = 0.03$
- $\text{prob}[\text{Befund: nicht-erkrankt} \mid \text{nicht-erkrankt}] = 0.97.$

Mit dem Satz von Bayes:

$$\text{prob}[\text{erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}]$$

$$= 0.98 \cdot 0.008 = 0.00784$$

$$\text{prob}[\text{nicht-erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}]$$

=

- $\text{prob}[\text{Befund: erkrankt} \mid \text{erkrankt}] = 0.98,$
- $\text{prob}[\text{Befund: nicht erkrankt} \mid \text{erkrankt}] = 0.02,$
- $\text{prob}[\text{Befund: erkrankt} \mid \text{nicht-erkrankt}] = 0.03$
- $\text{prob}[\text{Befund: nicht-erkrankt} \mid \text{nicht-erkrankt}] = 0.97.$

Mit dem Satz von Bayes:

$$\text{prob}[\text{erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}]$$

$$= 0.98 \cdot 0.008 = 0.00784$$

$$\text{prob}[\text{nicht-erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}]$$

$$= 0.03 \cdot 0.992 = 0.02976.$$

Bei positivem Befund: Die MAP-Hypothese ist

- $\text{prob}[\text{Befund: erkrankt} \mid \text{erkrankt}] = 0.98$,
- $\text{prob}[\text{Befund: nicht erkrankt} \mid \text{erkrankt}] = 0.02$,
- $\text{prob}[\text{Befund: erkrankt} \mid \text{nicht-erkrankt}] = 0.03$
- $\text{prob}[\text{Befund: nicht-erkrankt} \mid \text{nicht-erkrankt}] = 0.97$.

Mit dem Satz von Bayes:

$$\begin{aligned} & \text{prob}[\text{erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}] \\ = & 0.98 \cdot 0.008 = 0.00784 \\ & \text{prob}[\text{nicht-erkrankt} \mid \text{Befund: erkrankt}] \cdot \text{prob}[\text{Befund: erkrankt}] \\ = & 0.03 \cdot 0.992 = 0.02976. \end{aligned}$$

Bei positivem Befund: Die MAP-Hypothese ist „nicht erkrankt“.
Desweiteren: Ein positiver Befund ist falsch mit Wahrscheinlichkeit

$$\frac{0.02976}{0.00784 + 0.02976} \approx 0,7914$$

Bayes-Klassifikation

Bayes-Hypothese

Sei D eine **bekannte** Verteilung über der Menge $X \times \{0, 1\}$.

Für die Beispielmenge $S = \{x_1, \dots, x_s\} \subseteq X$ ist die **Bayes-Hypothese**

$$h_D(x_i) = \begin{cases} 1 \end{cases}$$

Sei D eine **bekannte** Verteilung über der Menge $X \times \{0, 1\}$.

Für die Beispielmenge $S = \{x_1, \dots, x_s\} \subseteq X$ ist die **Bayes-Hypothese**

$$h_D(x_i) = \begin{cases} 1 & \text{prob}_D[y = 1 | x_i] \geq \frac{1}{2}, \\ 0 & \text{sonst} \end{cases}$$

Sei D eine **bekannte** Verteilung über der Menge $X \times \{0, 1\}$.

Für die Beispielmenge $S = \{x_1, \dots, x_s\} \subseteq X$ ist die **Bayes-Hypothese**

$$h_D(x_i) = \begin{cases} 1 & \text{prob}_D[y = 1 | x_i] \geq \frac{1}{2}, \\ 0 & \text{sonst} \end{cases}$$

Bei unbekannter Zielfunktion ist h_D optimal, d.h. f. a. Hypothesen h gilt

$$\text{Loss}_D(h) \geq \text{Loss}_D(h_D).$$

Aber natürlich ist D i.A. nicht bekannt: Benutze die Bayes-Hypothese als Referenz-Hypothese (wie etwa für 1-Nearest-Neighbor).

Klassifikation einzelner Beispiele

Die Hypothesenklasse \mathcal{H} bestehe aus Zufallsvariablen. Klassifiziere ein **einziges** Beispiel B mit einem Wert $c \in C$. Bekannt seien:

- 1 $\text{prob}[h | B]$ für jede Hypothese $h \in \mathcal{H}$ und
- 2 $\text{prob}[c | h]$ für jedes $c \in C$.

Gesucht ist eine **optimale Bayes-Klassifikation** $c^* \in C$, d.h. es gelte

$$\text{prob}[c^* | B] = \max_{c \in C} \text{prob}[c | B].$$

Klassifikation einzelner Beispiele

Die Hypothesenklasse \mathcal{H} bestehe aus Zufallsvariablen. Klassifiziere ein **einziges** Beispiel B mit einem Wert $c \in C$. Bekannt seien:

- 1 $\text{prob}[h | B]$ für jede Hypothese $h \in \mathcal{H}$ und
- 2 $\text{prob}[c | h]$ für jedes $c \in C$.

Gesucht ist eine **optimale Bayes-Klassifikation** $c^* \in C$, d.h. es gelte

$$\text{prob}[c^* | B] = \max_{c \in C} \text{prob}[c | B].$$

Es ist

$$\text{prob}[c | B] = \sum_{h \in \mathcal{H}} \text{prob}[c | h] \cdot \text{prob}[h | B].$$

und das vorhandene Wissen ist ausreichend.

Weighted-Majority bestimmt eine MAP-Klassifikation.

- (a) Definiere die A-priori Wahrscheinlichkeit $\text{prob}[h | B]$ der Hypothese h durch das relative *Gewicht* des „Experten“ h .
- (b) $\text{prob}[c | h]$ ist 0-1-wertig, denn Experten sind deterministisch.

Weighted Majority entscheidet sich für das Mehrheitsvotum

$$\text{prob}[c^* | B] = \max_{c \in C} \text{prob}[c | B]$$

und damit für die optimale Bayes-Klassifikation.

Naive Bayes-Klassifikation

Wie bestimmt man eine optimale Bayes-Klassifizierung **effizient**?

- 1 Es gelte $X := \Sigma_1 \times \dots \times \Sigma_n$ für Alphabete $\Sigma_1, \dots, \Sigma_n$.
- 2 Eine unbekannte Zielfunktion $f : X \rightarrow C$ ist vorgegeben.
- 3 Für ein Beispiel $B = (a_1, \dots, a_n) \in X$ bestimme

$$\max_{c \in C} \text{prob}[c \mid a_1 \dots a_n].$$

Wie bestimmt man eine optimale Bayes-Klassifizierung **effizient**?

- 1 Es gelte $X := \Sigma_1 \times \dots \times \Sigma_n$ für Alphabete $\Sigma_1, \dots, \Sigma_n$.
- 2 Eine unbekannte Zielfunktion $f : X \rightarrow C$ ist vorgegeben.
- 3 Für ein Beispiel $B = (a_1, \dots, a_n) \in X$ bestimme

$$\max_{c \in C} \text{prob}[c \mid a_1 \dots a_n].$$

Mit dem Satz von Bayes ist

$$\text{prob}[c \mid a_1 \dots a_n] = \frac{\text{prob}[a_1 \dots a_n \mid c] \cdot \text{prob}[c]}{\text{prob}[a_1 \dots a_n]}.$$

⇒ Maximiere

$$\text{prob}[a_1 \dots a_n \mid c] \cdot \text{prob}[c]$$

Maximiere $\text{prob}[a_1 \cdots a_n | c] \cdot \text{prob}[c]$.

Wir machen die **naive** Annahme der stochastischen Unabhängigkeit:

$$\text{prob}[a_1 \cdots a_n | c] = \prod_{i=1}^n \text{prob}[a_i | c].$$

Maximiere $\text{prob}[a_1 \cdots a_n | c] \cdot \text{prob}[c]$.

Wir machen die **naive** Annahme der stochastischen Unabhängigkeit:

$$\text{prob}[a_1 \cdots a_n | c] = \prod_{i=1}^n \text{prob}[a_i | c].$$

\implies Bestimme eine Klassifizierung $c^* \in C$, so dass

$$\prod_{i=1}^n \text{prob}[a_i | c^*] \cdot \text{prob}[c^*] = \max_{c \in C} \underbrace{\left(\prod_{i=1}^n \text{prob}[a_i | c] \right)} \cdot \text{prob}[c].$$

Maximiere $\text{prob}[a_1 \cdots a_n | c] \cdot \text{prob}[c]$.

Wir machen die **naive** Annahme der stochastischen Unabhängigkeit:

$$\text{prob}[a_1 \cdots a_n | c] = \prod_{i=1}^n \text{prob}[a_i | c].$$

⇒ Bestimme eine Klassifizierung $c^* \in C$, so dass

$$\prod_{i=1}^n \text{prob}[a_i | c^*] \cdot \text{prob}[c^*] = \max_{c \in C} \left(\prod_{i=1}^n \text{prob}[a_i | c] \right) \cdot \text{prob}[c].$$

Bestimme die $\sum_{i=1}^n |\Sigma_i| \cdot |C|$ Terme $\text{prob}[a_i | c]$ sowie die $|C|$ Terme $\text{prob}[c]$ durch eine Statistik auf den Trainingsdaten.

Die naive Bayes-Klassifikation ist in „vielen“ Anwendungen erfolgreich.

Naive Bayes-Klassifikation: Ein Beispiel

Wir erhalten als *interessant* bzw *uninteressant* markierte Textdateien.

Lerne den Geschmack des Nutzers!

Naive Bayes-Klassifikation: Ein Beispiel

Wir erhalten als *interessant* bzw *uninteressant* markierte Textdateien.

Lerne den Geschmack des Nutzers!

- 1 Sei V das benutzte Vokabular. Es ist $\Sigma_i = V$ für jedes $i \implies$
Repräsentiere Textdateien durch den Vektor ihrer Worte.
- 2 Mit der lächerlichen Annahme der stochastischen Unabhängigkeit folgt

$$\text{prob}[w_1 \cdots w_n | c] = \prod_{i=1}^n \text{prob}[w_i | c]$$

Naive Bayes-Klassifikation: Ein Beispiel

Wir erhalten als *interessant* bzw *uninteressant* markierte Textdateien.

Lerne den Geschmack des Nutzers!

- 1 Sei V das benutzte Vokabular. Es ist $\Sigma_i = V$ für jedes $i \implies$
Repräsentiere Textdateien durch den Vektor ihrer Worte.
- 2 Mit der lächerlichen Annahme der stochastischen Unabhängigkeit folgt

$$\text{prob}[w_1 \cdots w_n | c] = \prod_{i=1}^n \text{prob}[w_i | c]$$

- 3 Setze $\text{prob}[w | c] = \frac{\text{Häufigkeit}_c(w)}{P_c + |V|}$, wobei
 P_c die Gesamtlänge aller Textdateien mit Klassifizierung c ist.

Der Ansatz funktioniert! Und der Bag-of-Words Ansatz für SVMs?

Bayes (Belief) Netzwerke

Gemeinsame Verteilungen

Die Zufallsvariable X_1, \dots, X_n sind gegeben.

Bestimme die gemeinsame Wahrscheinlichkeitsverteilung

$$p[X_1 = x_1, \dots, X_n = x_n]$$

und werte sie aus: Z. B. bestimme die Randverteilung

$$p(X_U = x_U | X_V = x_V) = \sum_y p(X_U = x_U, X_{\{1, \dots, n\} \setminus (U \cup V)} = y | X_V = x_V)$$

für disjunkte Teilmengen $U, V \subseteq \{1, \dots, n\}$.

Gemeinsame Verteilungen

Die Zufallsvariable X_1, \dots, X_n sind gegeben.

Bestimme die gemeinsame Wahrscheinlichkeitsverteilung

$$p[X_1 = x_1, \dots, X_n = x_n]$$

und werte sie aus: Z. B. bestimme die Randverteilung

$$p(X_U = x_U | X_V = x_V) = \sum_y p(X_U = x_U, X_{\{1, \dots, n\} \setminus (U \cup V)} = y | X_V = x_V)$$

für disjunkte Teilmengen $U, V \subseteq \{1, \dots, n\}$.

Wenn die Zufallsvariablen unabhängig sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = p[X_1 = x_1] \cdots p[X_n = x_n].$$

und die Bestimmung ist trivial. Ansonsten mache Abhängigkeiten unter den Variablen explizit: Stelle ein **Bayes Netzwerk** auf.

Ein Bayes Netzwerk ($B = V, E, p$) besitzt die Komponenten:

- 1 $V = \{X_1, \dots, X_n\}$ und (V, E) ist ein **kreisfreier** gerichteter Graph.
- 2 Für den Knoten X_i seien $X_{i_1}, \dots, X_{i_{k_i}}$ alle direkten Vorgänger (bzw. die **Eltern**) von X_i . Wir fordern

$$p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k_i}} = x_{i_{k_i}}] =$$

$$p[X_i = x_i \mid X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n]$$

Ein Bayes Netzwerk ($B = V, E, p$) besitzt die Komponenten:

- 1 $V = \{X_1, \dots, X_n\}$ und (V, E) ist ein **kreisfreier** gerichteter Graph.
- 2 Für den Knoten X_i seien $X_{i_1}, \dots, X_{i_{k_i}}$ alle direkten Vorgänger (bzw. die **Eltern**) von X_i . Wir fordern

$$p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k_i}} = x_{i_{k_i}}] =$$

$$p[X_i = x_i \mid X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \dots, X_n = x_n]$$

Der Gesamteinfluss aller Variablen auf X_i entspricht genau dem Einfluss der Eltern auf X_i .

- 3 Jeder Knoten X_i erhält die Tabellen

$$p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k_i}} = x_{i_{k_i}}]$$

der bedingten Wahrscheinlichkeiten.

Die Graphstruktur eines Bayes Netzwerks zeigt alle maßgeblichen kausalen Beziehungen unter den Zufallsvariablen:

Sei B ein Bayes Netzwerk mit den Zufallsvariablen X_1, \dots, X_n .
Wenn $X_{i_1}, \dots, X_{i_{k(i)}}$ die Eltern von X_i sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k(i)}} = x_{i_{k(i)}}].$$

Wie bestimmt man die bedingten Wahrscheinlichkeiten?

Bestimmung bedingter Wahrscheinlichkeiten

- 1 Heuristische Rechnungen: Z.B. bei verrauschter ODER-Beziehung mit (fast) unabhängigen Eltern, schätze

$$\text{prob}[X = 0 | X_1 = 1, \dots, X_k = 1] \approx \prod_{i=1}^k \text{prob}[X = 0 | X_i = 1].$$

- 2 Wenn nur einige Eigenschaften (wie etwa Erwartungswerte) bekannt sind:
 - ▶ Das **Prinzip der Maximum Entropie**: Wähle eine alle Eigenschaften erfüllende Verteilung mit maximaler Entropie.
- 3 Der **EM-Algorithmus** versucht eine Schätzung mit Hilfe beobachteter Variablen.
 - ▶ Die Schätzung soll die Beobachtungen hochwahrscheinlich machen.
 - ▶ Die Werte einiger Variablen mit „Einfluss“ dürfen fehlen.

Pathfinder: Ein Diagnose-System für Lymphknotenerkrankungen

- 1 Pathfinder führt ≈ 135 Symptome, Befunde und Laborwerte und gibt 60 verschiedene Diagnosen.
- 2 Nach Setzung der Evidenz-Variablen
 - Symptome, Befunde und Laborwerte eines Patienten wird ein Überblick über die Wahrscheinlichkeiten der jeweiligen Diagnose gegeben.

Bayes Netzwerke werden auch in der

probabilistischen Inferenz

eingesetzt.

Die Komplexität der Berechnung von Randverteilungen

! Ein einfaches, schwieriges Bayes-Netzwerk, das nur aus binären Zufallsvariablen besteht:

- ▶ Die Zufallsvariablen X_1, \dots, X_n sind unabhängig mit $\text{prob}[X_i = 1] = \frac{1}{2}$.
- ▶ Jede der Zufallsvariablen K_1, \dots, K_m hängt von genau drei X -Variablen ab und es ist $\text{prob}[K_j = 1] = 7/8$.
- ▶ Für die Zufallsvariable X gilt

$$X = 1 \iff K_1 = \dots = K_m = 1.$$

Die Komplexität der Berechnung von Randverteilungen

! Ein einfaches, schwieriges Bayes-Netzwerk, das nur aus binären Zufallsvariablen besteht:

- ▶ Die Zufallsvariablen X_1, \dots, X_n sind unabhängig mit $\text{prob}[X_i = 1] = \frac{1}{2}$.
- ▶ Jede der Zufallsvariablen K_1, \dots, K_m hängt von genau drei X -Variablen ab und es ist $\text{prob}[K_j = 1] = 7/8$.
- ▶ Für die Zufallsvariable X gilt

$$X = 1 \iff K_1 = \dots = K_m = 1.$$

Schon die Frage, ob $\text{prob}[X = 1] > 0$ führt auf ein NP-vollständiges Problem.

- ✓ Eine effiziente Berechnung von Randverteilungen gelingt für **Faktorgraphen**, die einem Baum entsprechen.

Faktorgraphen, eine eingeschränkte Version

Ein **Faktorgraph** $G = (V, F, E, \text{Funktion})$ ist ein bipartiter Graph.

- (a) Die Menge $V = \{X_1, \dots, X_n\}$ der **Variablen** ist die erste, die Menge F der **Faktoren** ist die zweite Schicht von B .
- (b) Für jeden Faktor f ist **Funktion(f)** eine Funktion, die nur von den Variablen X_v für Nachbarn v von f abhängt.

Forderung:

$$\text{prob}[X_1 = x_1, \dots, X_n = x_n] = \prod_{f \in F} \text{Funktion}(f).$$

Faktorgraphen, eine eingeschränkte Version

Ein **Faktorgraph** $G = (V, F, E, \text{Funktion})$ ist ein bipartiter Graph.

- (a) Die Menge $V = \{X_1, \dots, X_n\}$ der **Variablen** ist die erste, die Menge F der **Faktoren** ist die zweite Schicht von B .
- (b) Für jeden Faktor f ist **Funktion(f)** eine Funktion, die nur von den Variablen X_v für Nachbarn v von f abhängt.

Forderung:

$$\text{prob}[X_1 = x_1, \dots, X_n = x_n] = \prod_{f \in F} \text{Funktion}(f).$$

Wenn der Faktorgraph ein **Baum mit kleinem Grad** ist, dann ist

$$\text{prob}[X_U = a | X_V = b]$$

für alle disjunkten Teilmengen $U, V \subseteq \{1, \dots, n\}$ und alle a, b effizient berechenbar.

Sei \mathcal{B} ein Bayes Netzwerk mit den Zufallsvariablen X_1, \dots, X_n .
Wenn $X_{i_1}, \dots, X_{i_{k(i)}}$ die Eltern von X_i sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k(i)}} = x_{i_{k(i)}}].$$

Baue einen Faktorgraphen $\mathcal{F} := (V, F, E, \text{Funktion})$ für \mathcal{B} :

1. $V := \{X_1, \dots, X_n\}$,

Sei \mathcal{B} ein Bayes Netzwerk mit den Zufallsvariablen X_1, \dots, X_n .
Wenn $X_{i_1}, \dots, X_{i_{k(i)}}$ die Eltern von X_i sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k(i)}} = x_{i_{k(i)}}].$$

Baue einen Faktorgraphen $\mathcal{F} := (V, F, E, \text{Funktion})$ für \mathcal{B} :

1. $V := \{X_1, \dots, X_n\}$,
2. $F := \{F_1, \dots, F_n\}$,

Bayes Netzwerke und Faktorgraphen

Sei \mathcal{B} ein Bayes Netzwerk mit den Zufallsvariablen X_1, \dots, X_n .
Wenn $X_{i_1}, \dots, X_{i_{k(i)}}$ die Eltern von X_i sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k(i)}} = x_{i_{k(i)}}].$$

Baue einen Faktorgraphen $\mathcal{F} := (V, F, E, \text{Funktion})$ für \mathcal{B} :

1. $V := \{X_1, \dots, X_n\}$,
2. $F := \{F_1, \dots, F_n\}$,
3. Kante $\{X_i, F_j\}$ gehört zu $E \iff i = j$ oder X_i ist Elternknoten von X_j ,

Bayes Netzwerke und Faktorgraphen

Sei \mathcal{B} ein Bayes Netzwerk mit den Zufallsvariablen X_1, \dots, X_n .
Wenn $X_{i_1}, \dots, X_{i_{k(i)}}$ die Eltern von X_i sind, gilt

$$p[X_1 = x_1, \dots, X_n = x_n] = \prod_{i=1}^n p[X_i = x_i \mid X_{i_1} = x_{i_1}, \dots, X_{i_{k(i)}} = x_{i_{k(i)}}].$$

Baue einen Faktorgraphen $\mathcal{F} := (V, F, E, \text{Funktion})$ für \mathcal{B} :

1. $V := \{X_1, \dots, X_n\}$,
2. $F := \{F_1, \dots, F_n\}$,
3. Kante $\{X_i, F_j\}$ gehört zu $E \iff i = j$ oder X_i ist Elternknoten von X_j ,
4. falls X_{j_1}, \dots, X_{j_k} die Elternknoten von X_j sind, ist

$$\text{Funktion}(F_j) := p[X_j = x_j \mid X_{j_1} = x_{j_1}, \dots, X_{j_k} = x_{j_k}].$$

Belief Propagation (BP) für Bäume

BP für Bäume: Das „eingeschränkte“ Ziel

T sei ein gewurzelter „Faktorbaum“.

(a) BP rechnet von den Blättern aufwärts bis zur Wurzel.

(b) Die Wurzel ist die Variable v : Bestimme die Randverteilung, d.h.

$$\text{prob}[X_v = a]$$

für alle Elemente a im Wertebereich von X_v .

BP für Bäume: Das „eingeschränkte“ Ziel

T sei ein gewurzelter „Faktorbaum“.

(a) BP rechnet von den Blättern aufwärts bis zur Wurzel.

(b) Die Wurzel ist die Variable v : Bestimme die Randverteilung, d.h.

$$\text{prob}[X_v = a]$$

für alle Elemente a im Wertebereich von X_v .

Für eine Variable oder einen Faktor t definiere

$$\text{prob}_T^{(t)}$$

als das Produkt aller Faktoren im Teilbaum von t . Des Weiteren ist

$$[\text{prob}_T^{(t)}] = \sum_y \text{prob}_T^{(t)}(y),$$

wobei über alle *Wertekombinationen* y für Variablen „unterhalb t “ summiert wird.

(F) Erreicht BP einen Faktor g , dann sendet BP die Nachricht

$$m_{g \rightarrow v} := [\text{prob}_T^{(g)}]$$

an den Elternknoten v von g .

Achtung: $[\text{prob}_T^{(g)}]$ ist eine Funktion der Werte von

- (F) Erreicht BP einen Faktor g , dann sendet BP die Nachricht

$$m_{g \rightarrow v} := [\text{prob}_T^{(g)}]$$

an den Elternknoten v von g .

Achtung: $[\text{prob}_T^{(g)}]$ ist eine Funktion der Werte von X_v .

- (V) Erreicht BP eine Variable v , dann sendet BP die Nachricht

$$m_{v \rightarrow f} := [\text{prob}_T^{(v)}]$$

an den Elternknoten f von v .

Achtung: $[\text{prob}_T^{(v)}]$ ist eine Funktion der Werte von

Belief Propagation: Die Invarianten

(F) Erreicht BP einen Faktor g , dann sendet BP die Nachricht

$$m_{g \rightarrow v} := [\text{prob}_T^{(g)}]$$

an den Elternknoten v von g .

Achtung: $[\text{prob}_T^{(g)}]$ ist eine Funktion der Werte von X_v .

(V) Erreicht BP eine Variable v , dann sendet BP die Nachricht

$$m_{v \rightarrow f} := [\text{prob}_T^{(v)}]$$

an den Elternknoten f von v .

Achtung: $[\text{prob}_T^{(v)}]$ ist eine Funktion der Werte von X_v .

BP sendet stets eine „Randverteilung“ für X_v .

BP erreicht das Blatt x .

- (a) Wenn x der Faktorknoten f ist und v sein Elternknoten ist, dann hängt f nur von x_v ab \implies BP sendet die Nachricht

BP erreicht das Blatt x .

- (a) Wenn x der Faktorknoten f ist und v sein Elternknoten ist, dann hängt f nur von x_v ab \implies BP sendet die Nachricht

$$m_{f \rightarrow v} := \text{Funktion}(f)$$

an die Variable v . Beachte, dass $[\text{prob}_T^{(f)}] = \text{Funktion}(f)$ gilt.

BP erreicht das Blatt x .

- (a) Wenn x der Faktorknoten f ist und v sein Elternknoten ist, dann hängt f nur von x_v ab \implies BP sendet die Nachricht

$$m_{f \rightarrow v} := \text{Funktion}(f)$$

an die Variable v . Beachte, dass $[\text{prob}_T^{(f)}] = \text{Funktion}(f)$ gilt.

- (b) Wenn x die Variable v ist, dann sendet BP die Nachricht

BP erreicht das Blatt x .

- (a) Wenn x der Faktorknoten f ist und v sein Elternknoten ist, dann hängt f nur von x_v ab \implies BP sendet die Nachricht

$$m_{f \rightarrow v} := \text{Funktion}(f)$$

an die Variable v . Beachte, dass $[\text{prob}_T^{(f)}] = \text{Funktion}(f)$ gilt.

- (b) Wenn x die Variable v ist, dann sendet BP die Nachricht

$$m_{v \rightarrow g} := 1$$

an den Elternknoten g . (Die **Funktion** 1 weist jedem Wert von X_v den Wert 1 zu. Beachte, dass $[\text{prob}_T^{(v)}] = 1$ gilt.)

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} =$$

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} = \text{Funktion}(f) \cdot \text{prob}_T^{(v_1)} \dots \text{prob}_T^{(v_\ell)}.$$

Invariante (V) \implies BP hat die Nachrichten

$$m_{v_i \rightarrow f} := [\text{prob}_T^{(v_i)}]$$

an f gesendet.

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} = \text{Funktion}(f) \cdot \text{prob}_T^{(v_1)} \dots \text{prob}_T^{(v_\ell)}.$$

Invariante (V) \implies BP hat die Nachrichten

$$m_{v_i \rightarrow f} := [\text{prob}_T^{(v_i)}]$$

an f gesendet. Hat f alle Nachrichten erhalten, sendet BP

$$m_{f \rightarrow v} := \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} m_{v_i \rightarrow f} \right]$$

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} = \text{Funktion}(f) \cdot \text{prob}_T^{(v_1)} \dots \text{prob}_T^{(v_\ell)}.$$

Invariante (V) \implies BP hat die Nachrichten

$$m_{v_i \rightarrow f} := [\text{prob}_T^{(v_i)}]$$

an f gesendet. Hat f alle Nachrichten erhalten, sendet BP

$$m_{f \rightarrow v} := \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} m_{v_i \rightarrow f} \right] = \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} [\text{prob}_T^{(v_i)}] \right]$$

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} = \text{Funktion}(f) \cdot \text{prob}_T^{(v_1)} \cdots \text{prob}_T^{(v_\ell)}.$$

Invariante (V) \implies BP hat die Nachrichten

$$m_{v_i \rightarrow f} := [\text{prob}_T^{(v_i)}]$$

an f gesendet. Hat f alle Nachrichten erhalten, sendet BP

$$m_{f \rightarrow v} := \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} m_{v_i \rightarrow f} \right] = \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} [\text{prob}_T^{(v_i)}] \right]$$

Distributivität
 $=$

$$\left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} \text{prob}_T^{(v_i)} \right]$$

BP erreicht den Faktor f mit Kindern v_1, \dots, v_ℓ und Elternknoten $v \implies$

$$\text{prob}_T^{(f)} = \text{Funktion}(f) \cdot \text{prob}_T^{(v_1)} \dots \text{prob}_T^{(v_\ell)}.$$

Invariante (V) \implies BP hat die Nachrichten

$$m_{v_i \rightarrow f} := [\text{prob}_T^{(v_i)}]$$

an f gesendet. Hat f alle Nachrichten erhalten, sendet BP

$$m_{f \rightarrow v} := \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} m_{v_i \rightarrow f} \right] = \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} [\text{prob}_T^{(v_i)}] \right]$$

$$\stackrel{\text{Distributivität}}{=} \left[\text{Funktion}(f) \cdot \prod_{i=1}^{\ell} \text{prob}_T^{(v_i)} \right] = [\text{prob}_T^{(f)}]$$

an den Elternknoten v von f . Invariante (F) wird bewahrt.

BP erreicht die Variable v mit Kindern f_1, \dots, f_k und Elternknoten g

\implies

$$\text{prob}_T^{(v)} = \text{prob}_T^{(f_1)} \cdots \text{prob}_T^{(f_k)}.$$

BP erreicht die Variable v mit Kindern f_1, \dots, f_k und Elternknoten g
 \implies

$$\text{prob}_T^{(v)} = \text{prob}_T^{(f_1)} \cdots \text{prob}_T^{(f_k)}.$$

Invariante (F) \implies BP hat die Nachrichten

$$m_{f_i \rightarrow v} := [\text{prob}_T^{(f_i)}]$$

von f_i an v gesendet.

BP erreicht die Variable v mit Kindern f_1, \dots, f_k und Elternknoten g
 \implies

$$\text{prob}_T^{(v)} = \text{prob}_T^{(f_1)} \cdots \text{prob}_T^{(f_k)}.$$

Invariante (F) \implies BP hat die Nachrichten

$$m_{f_i \rightarrow v} := [\text{prob}_T^{(f_i)}]$$

von f_i an v gesendet. Hat v alle Nachrichten erhalten, sendet BP

$$m_{v \rightarrow g} := \prod_{i=1}^k m_{f_i \rightarrow v}$$

BP erreicht die Variable v mit Kindern f_1, \dots, f_k und Elternknoten g
 \implies

$$\text{prob}_T^{(v)} = \text{prob}_T^{(f_1)} \cdots \text{prob}_T^{(f_k)}.$$

Invariante (F) \implies BP hat die Nachrichten

$$m_{f_i \rightarrow v} := [\text{prob}_T^{(f_i)}]$$

von f_i an v gesendet. Hat v alle Nachrichten erhalten, sendet BP

$$m_{v \rightarrow g} := \prod_{i=1}^k m_{f_i \rightarrow v} = \prod_{i=1}^k [\text{prob}_T^{(f_i)}]$$

BP erreicht die Variable v mit Kindern f_1, \dots, f_k und Elternknoten g
 \implies

$$\text{prob}_T^{(v)} = \text{prob}_T^{(f_1)} \cdots \text{prob}_T^{(f_k)}.$$

Invariante (F) \implies BP hat die Nachrichten

$$m_{f_i \rightarrow v} := [\text{prob}_T^{(f_i)}]$$

von f_i an v gesendet. Hat v alle Nachrichten erhalten, sendet BP

$$m_{v \rightarrow g} := \prod_{i=1}^k m_{f_i \rightarrow v} = \prod_{i=1}^k [\text{prob}_T^{(f_i)}] \stackrel{\text{Distributivität!}}{=} [\text{prob}_T^{(v)}]$$

von v an den Elternknoten g \implies Invariante V wird bewahrt.

BP auf Bäumen: Eine Zusammenfassung

Sei $(V, F, E, \text{Funktion})$ ein Baum der Tiefe T . Ist die Variable v Wurzel des Baums, dann bestimmt Belief Propagation

$$\text{prob}[X_v = a]$$

für alle Werte a von X_v in T Runden.

Beweis: Es ist $[\text{prob}_T^{(v)}] = \text{prob}[X_v = *]$ □

BP auf Bäumen: Eine Zusammenfassung

Sei $(V, F, E, \text{Funktion})$ ein Baum der Tiefe T . Ist die Variable v Wurzel des Baums, dann bestimmt Belief Propagation

$$\text{prob}[X_v = a]$$

für alle Werte a von X_v in T Runden.

Beweis: Es ist $[\text{prob}_T^{(v)}] = \text{prob}[X_v = *]$ □

BP ist effizient, wenn der Grad des Baums klein ist.

BP auf Bäumen: Eine Zusammenfassung

Sei $(V, F, E, \text{Funktion})$ ein Baum der Tiefe T . Ist die Variable v Wurzel des Baums, dann bestimmt Belief Propagation

$$\text{prob}[X_v = a]$$

für alle Werte a von X_v in T Runden.

Beweis: Es ist $[\text{prob}_T^{(v)}] = \text{prob}[X_v = *]$ □

BP ist effizient, wenn der Grad des Baums klein ist.

Für allgemeine Graphen wird BP zu

Loopy Belief Propagation.

Nachrichten werden nach der folgenden Vorschrift versandt:

1. Hat die Variable v die Nachrichten $m_{f \rightarrow v}$ von ihren Nachbarn f erhalten, so schickt v die Nachricht

$$m_{v \rightarrow g} := \prod_{f \text{ ist Nachbar von } v, f \neq g} m_{f \rightarrow v}$$

an ihren Nachbarn g .

Loopy Belief Propagation

Nachrichten werden nach der folgenden Vorschrift versandt:

1. Hat die Variable v die Nachrichten $m_{f \rightarrow v}$ von ihren Nachbarn f erhalten, so schickt v die Nachricht

$$m_{v \rightarrow g} := \prod_{f \text{ ist Nachbar von } v, f \neq g} m_{f \rightarrow v}$$

an ihren Nachbarn g .

2. Hat der Faktor f die Nachrichten $m_{v \rightarrow f}$ von seinem Nachbarn v erhalten, so schickt f die Nachricht

$$m_{f \rightarrow w} := [\text{Funktion}(f) \cdot \prod_{v \text{ ist Nachbar von } f, v \neq w} m_{v \rightarrow f}]$$

an seine Nachbarin w .

Expectation Maximization

Expectation Maximization: Das Modell

Beobachtete Daten x und **versteckte** Daten y liegen vor.
Bestimme ein **Modell** (mit Parametern Θ), das x „am besten erklärt“.

Die Parameter Θ definieren die gemeinsame Wahrscheinlichkeit

$$P_{\Theta}(x, y)$$

Wenn $P_{\Theta}(x)$ die Wahrscheinlichkeit von x ist, dann ist

$$P_{\Theta}(x) = \sum_y P_{\Theta}(x, y).$$

Expectation Maximization: Das Modell

Beobachtete Daten x und **versteckte** Daten y liegen vor.
Bestimme ein **Modell** (mit Parametern Θ), das x „am besten erklärt“.

Die Parameter Θ definieren die gemeinsame Wahrscheinlichkeit

$$P_{\Theta}(x, y)$$

Wenn $P_{\Theta}(x)$ die Wahrscheinlichkeit von x ist, dann ist

$$P_{\Theta}(x) = \sum_y P_{\Theta}(x, y).$$

Bestimme ein **ML-Modell** Θ^* :

$$P_{\Theta^*}(x) \stackrel{!}{=} \max_{\Theta} P_{\Theta}(x) \quad \text{bzw.} \quad \log P_{\Theta^*}(x) \stackrel{!}{=} \max_{\Theta} \log P_{\Theta}(x).$$

Hidden Markov Model (HMM)

Für M eine Markov-Kette und alle Zustände p legt Verteilung D_p die Wahrscheinlichkeit fest, mit der M in p Buchstaben $a \in \Sigma$ ausdrückt.

- 1 Die Markov-Kette M wie auch die Verteilungen D_p über einem Alphabet Σ sind unbekannt $\implies \Theta = (M, (D_p|p))$.
- 2 Wir **beobachten** die Ausgabenfolge x , wenn die Kette eine zufällige (**versteckte**) Zustandsfolge y annimmt

Hidden Markov Model (HMM)

Für M eine Markov-Kette und alle Zustände p legt Verteilung D_p die Wahrscheinlichkeit fest, mit der M in p Buchstaben $a \in \Sigma$ ausdrückt.

- 1 Die Markov-Kette M wie auch die Verteilungen D_p über einem Alphabet Σ sind unbekannt $\implies \Theta = (M, (D_p|p))$.
- 2 Wir **beobachten** die Ausgabenfolge x , wenn die Kette eine zufällige (**versteckte**) Zustandsfolge y annimmt \implies

$$P_{\Theta}(x, y)$$

ist die Wahrscheinlichkeit, dass das HMM Θ die Ausgabefolge x mit der (versteckten) Zustandsfolge y produziert

Hidden Markov Model (HMM)

Für M eine Markov-Kette und alle Zustände p legt Verteilung D_p die Wahrscheinlichkeit fest, mit der M in p Buchstaben $a \in \Sigma$ ausdrückt.

- 1 Die Markov-Kette M wie auch die Verteilungen D_p über einem Alphabet Σ sind unbekannt $\implies \Theta = (M, (D_p|p))$.
- 2 Wir **beobachten** die Ausgabenfolge x , wenn die Kette eine zufällige (**versteckte**) Zustandsfolge y annimmt \implies

$$P_{\Theta}(x, y)$$

ist die Wahrscheinlichkeit, dass das HMM Θ die Ausgabefolge x mit der (versteckten) Zustandsfolge y produziert \implies

$$\sum_y P_{\Theta}(x, y)$$

ist die Wahrscheinlichkeit, dass Θ die Folge x ausgibt.

Maximum Likelihood soll uns helfen, die HMM Θ zu bestimmen.

Es ist $P_{\Theta}(y | x) \cdot P_{\Theta}(x) = P_{\Theta}(x, y)$ und deshalb folgt

$$\log P_{\Theta}(x) = \log P_{\Theta}(x, y) - \log P_{\Theta}(y | x).$$

Es ist $P_{\Theta}(y | x) \cdot P_{\Theta}(x) = P_{\Theta}(x, y)$ und deshalb folgt

$$\log P_{\Theta}(x) = \log P_{\Theta}(x, y) - \log P_{\Theta}(y | x).$$

Multipliziere beide Seiten mit $P_{\Theta}(y | x)$ und summiere über y

$$\log P_{\Theta}(x) = \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x)$$

Es ist $P_{\Theta}(y | x) \cdot P_{\Theta}(x) = P_{\Theta}(x, y)$ und deshalb folgt

$$\log P_{\Theta}(x) = \log P_{\Theta}(x, y) - \log P_{\Theta}(y | x).$$

Multipliziere beide Seiten mit $P_{\Theta}(y | x)$ und summiere über y

$$\begin{aligned} \log P_{\Theta}(x) &= \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x) \\ &= \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x, y) - \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(y | x). \end{aligned}$$

Es ist $P_{\Theta}(y | x) \cdot P_{\Theta}(x) = P_{\Theta}(x, y)$ und deshalb folgt

$$\log P_{\Theta}(x) = \log P_{\Theta}(x, y) - \log P_{\Theta}(y | x).$$

Multipliziere beide Seiten mit $P_{\Theta}(y | x)$ und summiere über y

$$\begin{aligned} \log P_{\Theta}(x) &= \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x) \\ &= \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x, y) - \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(y | x). \end{aligned}$$

Wir möchten von Θ zu einem besseren Modell Θ' übergehen.

$$\log P_{\Theta'}(x) = \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)}_{=: Q(\Theta' | \Theta)} - \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(y | x).$$

Es ist $Q(\theta' | \theta) := \sum_y P_\theta(y | x) \cdot \log P_{\theta'}(x, y)$.

- (a) $Q(\theta' | \theta) = \mathbb{E}_{P_\theta(y|x)}[\log P_{\theta'}(x, y)]$ ist ein Erwartungswert.
- (b) Fasse $Q(\theta' | \theta)$ als Funktion von θ' auf.

Es ist $Q(\Theta' | \Theta) := \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)$.

- (a) $Q(\Theta' | \Theta) = \mathbb{E}_{P_{\Theta}(y|x)}[\log P_{\Theta'}(x, y)]$ ist ein Erwartungswert.
- (b) Fasse $Q(\Theta' | \Theta)$ als Funktion von Θ' auf.

$$0 \stackrel{!}{\leq} \log P_{\Theta'}(x) - \log P_{\Theta}(x)$$

Es ist $Q(\Theta' | \Theta) := \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)$.

(a) $Q(\Theta' | \Theta) = \mathbb{E}_{P_{\Theta}(y|x)}[\log P_{\Theta'}(x, y)]$ ist ein Erwartungswert.

(b) Fasse $Q(\Theta' | \Theta)$ als Funktion von Θ' auf.

$$\begin{aligned}
 0 & \stackrel{!}{\leq} \log P_{\Theta'}(x) - \log P_{\Theta}(x) \\
 &= \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)}_{Q(\Theta' | \Theta)} - \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(y | x) \\
 &\quad - \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x, y)}_{Q(\Theta | \Theta)} + \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(y | x)
 \end{aligned}$$

Es ist $Q(\Theta' | \Theta) := \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)$.

(a) $Q(\Theta' | \Theta) = \mathbb{E}_{P_{\Theta}(y|x)}[\log P_{\Theta'}(x, y)]$ ist ein Erwartungswert.

(b) Fasse $Q(\Theta' | \Theta)$ als Funktion von Θ' auf.

$$\begin{aligned}
 0 & \stackrel{!}{\leq} \log P_{\Theta'}(x) - \log P_{\Theta}(x) \\
 &= \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(x, y)}_{Q(\Theta' | \Theta)} - \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta'}(y | x) \\
 &\quad - \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(x, y)}_{Q(\Theta | \Theta)} + \sum_y P_{\Theta}(y | x) \cdot \log P_{\Theta}(y | x) \\
 &= Q(\Theta' | \Theta) - Q(\Theta | \Theta) + \sum_y P_{\Theta}(y | x) \cdot \log \frac{P_{\Theta}(y | x)}{P_{\Theta'}(y | x)}
 \end{aligned}$$

$$\begin{aligned}\log P_{\Theta'} - \log P_{\Theta} &= Q(\Theta' | \Theta) - Q(\Theta | \Theta) + \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log \frac{P_{\Theta}(y | x)}{P_{\Theta'}(y | x)}}_{=D(P_{\Theta}(*|x) \| P_{\Theta'}(*|x))} \\ &\geq Q(\Theta' | \Theta) - Q(\Theta | \Theta)\end{aligned}$$

$$\begin{aligned} \log P_{\Theta'} - \log P_{\Theta} &= Q(\Theta' | \Theta) - Q(\Theta | \Theta) + \underbrace{\sum_y P_{\Theta}(y | x) \cdot \log \frac{P_{\Theta}(y | x)}{P_{\Theta'}(y | x)}}_{=D(P_{\Theta}(*|x) \| P_{\Theta'}(*|x))} \\ &\geq Q(\Theta' | \Theta) - Q(\Theta | \Theta) \end{aligned}$$

Die Kullback-Leibler Divergenz $D(P_{\Theta}(*|x) \| P_{\Theta'}(*|x))$ ist stets nicht-negativ!

Falls $Q(\Theta' | \Theta) - Q(\Theta | \Theta) > 0$ ist das neue Modell Θ' besser als Θ .

Eingabe: Eine Beobachtung x .

1. Initialisiere die Parameter für ein Modell $\Theta^{(0)}$.

2. /* Expectation Schritt */

Bestimme $Q(\Theta | \Theta^{(t)})$ als Funktion von Θ .

3. /* Maximization Schritt */

Bestimme $\Theta^{(t+1)}$, so dass

$$Q(\Theta^{(t+1)} | \Theta^{(t)}) = \max_{\Theta} Q(\Theta | \Theta^{(t)})$$

und setze $t = t + 1$.

4. Wiederhole, solange $\log P_{\Theta^{(t)}}(x) - \log P_{\Theta^{(t-1)}}(x)$ zu groß ist.

Eingabe: Eine Beobachtung x .

1. Initialisiere die Parameter für ein Modell $\Theta^{(0)}$.

2. /* Expectation Schritt */

Bestimme $Q(\Theta | \Theta^{(t)})$ als Funktion von Θ .

3. /* Maximization Schritt */

Bestimme $\Theta^{(t+1)}$, so dass

$$Q(\Theta^{(t+1)} | \Theta^{(t)}) = \max_{\Theta} Q(\Theta | \Theta^{(t)})$$

und setze $t = t + 1$.

4. Wiederhole, solange $\log P_{\Theta^{(t)}}(x) - \log P_{\Theta^{(t-1)}}(x)$ zu groß ist.

Der EM-Algorithmus konvergiert i. A gegen ein lokales Maximum.
Gutes Verhalten vor Allem bei **Exponentialfamilien**^a

^az.B. mehrdimensionale Binomial-, Exponential- und Normalverteilungen.

Punkte $x_1, \dots, x_n \in \mathbb{R}$ werden durch zwei Normalverteilungen erzeugt.

- (a) Wir kennen die beiden Normalverteilungen nicht,
 - ▶ aber **beobachten** die Daten $x = (x_1, \dots, x_n)$.
- (b) Wir wissen nicht, ob x_i von der ersten ($y_i = 1$) oder der zweiten Normalverteilung ($y_i = 2$) erzeugt wurde.
 - ▶ $y = (y_1, \dots, y_n)$ ist der Vektor der **versteckten** Daten.
- (c) Bestimme die Parameter $\mu_1, \sigma_1, \mu_2, \sigma_2$ der Normalverteilungen \implies
 - ▶ $\Theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ ist das unbekannte Modell.

1. Der EM-Algorithmus beginnt mit zwei Normalverteilungen.
2. **Expectation-Schritt:**
 - ▶ Bestimme die Wahrscheinlichkeit $P_{\Theta^{(t)}}(y_i = 1 | x_i)$, dass x_i von der ersten Normalverteilung erzeugt wird, d.h
 - ▶ berechne die Dichten in x_i für die beiden durch $\Theta^{(t)}$ spezifizierten Normalverteilungen und setze sie in Relation zueinander.

1. Der EM-Algorithmus beginnt mit zwei Normalverteilungen.
2. **Expectation-Schritt:**
 - ▶ Bestimme die Wahrscheinlichkeit $P_{\Theta^{(t)}}(y_i = 1 | x_i)$, dass x_i von der ersten Normalverteilung erzeugt wird, d.h
 - ▶ berechne die Dichten in x_i für die beiden durch $\Theta^{(t)}$ spezifizierten Normalverteilungen und setze sie in Relation zueinander.
3. **Maximization-Schritt:**
 - ▶ Bestimme das optimal auf $\Theta^{(t)}$ „eingestellte“ Modell $\Theta^{(t+1)}$, d.h
 - ▶ berechne neue, optimale Mittelwerte und Varianzen für die Verteilungen $P_{\Theta^{(t)}}(*|x_i)$ für alle i .

Der k-Means Algorithmus für das Clustering Problem

1. Bestimme k Werte μ_1, \dots, μ_k .
2. „*Expectation-Schritt*“: Weise jedem Punkt x das Cluster C_i zu, für das die Distanz $\|x - \mu_i\|$ minimal ist.
3. „*Maximization Schritt*“: Bestimme die Schwerpunkte μ_i von Cluster C_i für $i = 1, \dots, k$.
4. Wiederhole gegebenenfalls.

Der k-Means Algorithmus für das Clustering Problem

1. Bestimme k Werte μ_1, \dots, μ_k .
2. „*Expectation-Schritt*“: Weise jedem Punkt x das Cluster C_j zu, für das die Distanz $\|x - \mu_j\|$ minimal ist.
3. „*Maximization Schritt*“: Bestimme die Schwerpunkte μ_j von Cluster C_j für $i = 1, \dots, k$.
4. Wiederhole gegebenenfalls.

k -Means ist eine „harte“ Variante des „weichen“ EM-Algorithmus:

- Jeder Punkt wird „hart“ einem Cluster zugewiesen,
- während EM die Zuweisung „weich“ über die Verteilung $P_{\Theta(t)}(y|x)$ ausführt.