

Оценивание гиперпараметров графических моделей

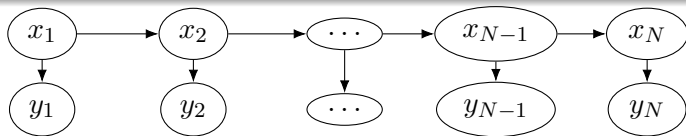
Александр Адуенко

6е мая 2024

Содержание предыдущих лекций

- EM-алгоритм и его сходимость. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.
- Скрытые марковские модели (СММ) и алгоритм Витерби. Алгоритм Max-Sum как обобщение алгоритма Витерби.
- Алгоритм Баума-Велча для определения параметров СММ.
- Алгоритмы на основе разрезов графов. Алгоритм α – расширение.
- Алгоритм TRW для приближенного вывода в циклических графических моделях с общей энергией.

Вывод в графических моделях



$$p(\mathbf{x}, \mathbf{y}) = p(x_1) \prod_{i=2}^N p(x_i | x_{i-1}) \prod_{i=1}^N p(y_i | x_i).$$

Пусть $x_i \in [K]$, $\mathbf{A} = \|a_{ij}\| = \|P(x_l = j | x_{l-1} = i)\|$, $\pi_k = P(x_1 = k)$.

$$p(\mathbf{x}, \mathbf{y} | \mathbf{A}, \boldsymbol{\pi}, \mathbf{B}) = p(x_1 | \boldsymbol{\pi}) \prod_{i=2}^N p(x_i | x_{i-1}, \mathbf{A}) \prod_{i=1}^N p(y_i | x_i, \mathbf{B}).$$

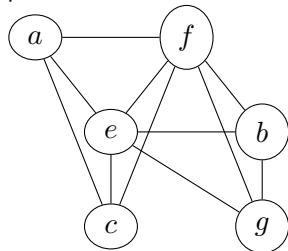
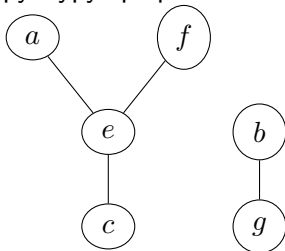
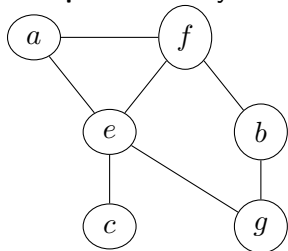
Задачи:

- $p(x_i | \mathbf{y}, \Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}_C | \mathbf{y}, \Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x} | \mathbf{y}, \Theta) \rightarrow \max_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum / Graph-Cut / α – расширение / TRW;
- $p(\mathbf{x} | \mathbf{y}, \Theta)$ – сэмплирование;
- $p(\mathbf{y} | \Theta) \rightarrow \max_{\Theta}$ – алгоритм Баума-Велча.

Обучение параметров графических моделей

$$p(\mathbf{y}|\Theta) \rightarrow \max_{\Theta}, p(\mathbf{x}, \mathbf{y}|\Theta) = \frac{1}{Z} \prod_i \psi_i(\mathbf{x}_i, \mathbf{y}_i|\Theta_i).$$

Вопрос: Как обучить структуру графической модели?



$$p(a, b, c, e, f, g) = \frac{1}{Z_1} \psi_{afe}(a, f, e) \psi_{ec}(e, c) \psi_{eg}(e, g) \psi_{bg}(b, g) \psi_{bf}(b, f);$$

$$p(a, b, c, e, f, g) = \frac{1}{Z_2} \psi_{ae}(a, e) \psi_{fe}(f, e) \psi_{ce}(c, e) \psi_{bg}(b, g);$$

$$p(a, b, c, e, f, g) = \frac{1}{Z_3} \psi_{afec}(a, f, e, c) \psi_{efbg}(e, f, b, g);$$

Пример: Обучение структуры ГМ

Пусть $\mathbf{y} = [\mathbf{y}_1, \dots, \mathbf{y}_K]^\top$, $\mathbf{y}_i \in \mathbb{R}^D$.

$$p(\mathbf{y}) = \frac{1}{Z} \exp(-E(\mathbf{y})), \quad E(\mathbf{y}) = -\frac{1}{2} \mathbf{y}^\top \mathbf{\Omega} \mathbf{y} = -\frac{1}{2} \sum_{k,l=1}^K \mathbf{y}_k^\top \mathbf{\Omega}_{kl} \mathbf{y}_l.$$

$\mathbf{\Omega}_{kl} = \mathbf{0} \iff \mathbf{y}_k, \mathbf{y}_l$ – условно независимы при условии остальных переменных.

Идея: Ввести априорное распределение на $\mathbf{\Omega}$,

$$p(\mathbf{\Omega}) \propto I[\mathbf{\Omega} > 0] \exp(-\lambda \|\mathbf{\Omega}\|_1).$$

$$\log p(\mathbf{y}, \mathbf{\Omega}) \propto -\log I[\mathbf{\Omega} > 0] - \lambda \|\mathbf{\Omega}\|_1 + \frac{m}{2} \log \det \mathbf{\Omega} - \frac{1}{2} \text{tr} \left(\mathbf{\Omega} \sum_{j=1}^m \mathbf{y}^j \mathbf{y}^{j\top} \right).$$

$$\log p(\mathbf{\Omega} | \mathbf{y}, \lambda) \propto \log p(\mathbf{y}, \mathbf{\Omega}) \rightarrow \max_{\mathbf{\Omega}}$$

Вопрос 1: Как изменить $p(\mathbf{\Omega})$, чтобы убрать разреживание структуры внутри компонент одной переменной \mathbf{y}_k ?

Вопрос 2: Как обобщить обучение структуры на случай с ненаблюдаемыми переменными?

$$p(\mathbf{y}|\Theta) \rightarrow \max_{\Theta}, p(\mathbf{x}, \mathbf{y}|\Theta) = \prod_{i=1}^d p(\mathbf{x}_i / \mathbf{y}_i | Pa_i, \Theta_i).$$

Пусть все переменные наблюдаемые, то есть $\mathbf{x} = \emptyset$.

Вопрос 1: Что изменилось по отношению к общему случаю?

$$\log p(\mathbf{y}|\Theta) = \sum_i \log p(\mathbf{y}_i | Pa_i, \Theta_i) \rightarrow \max_{\Theta}.$$

Вопрос 2: Что можно сказать про задачу, если Θ_i – непересекающиеся во всех факторах?

Вопрос 3: Пусть $\mathbf{y}_i \in [K]$, $Pa_i \in [L]$. Тогда $\Theta_i^{kl} = P(\mathbf{y}_i = k | Pa_i = l)$. Что получим для Θ_i^{kl} ?

Вопрос 4: Что делать, если $\mathbf{x} \neq \emptyset$?

$$p(\mathbf{y}|\Theta) \rightarrow \max_{\Theta}, p(\mathbf{x}, \mathbf{y}|\Theta) = \prod_{i=1}^d p(\mathbf{x}_i / \mathbf{y}_i | Pa_i, \Theta_i).$$

$$p(\mathbf{y}|\Theta) = \int \prod_{i=1}^d p(\mathbf{x}_i / \mathbf{y}_i | Pa_i, \Theta_i) d\mathbf{x} \rightarrow \max_{\Theta}.$$

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ .

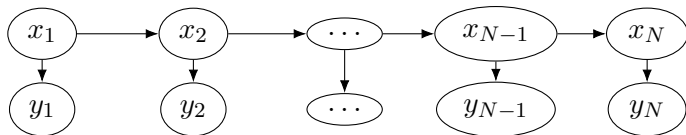
$$\text{Введем } F(q, \Theta) = - \int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} + \int q(\mathbf{x}) \log p(\mathbf{y}, \mathbf{x}|\Theta) d\mathbf{x} = \log p(\mathbf{y}|\Theta) - D_{KL}(q||p(\mathbf{x}|\mathbf{y}, \Theta)) \rightarrow \max_{q, \Theta}.$$

$$\text{E-шаг. } q(\mathbf{x}) = \arg \min_{q \in Q} D_{KL}(q||p(\mathbf{x}|\mathbf{y}, \Theta)).$$

$$\text{M-шаг. } \sum_{j=1}^m \sum_{i=1}^d E_{q(\mathbf{x})} \log p(\mathbf{x}_i^j / \mathbf{y}_i^j | Pa_i^j, \Theta_i) \rightarrow \max_{\Theta}.$$

Вопрос: Что достаточно знать о $q(\mathbf{x})$ для проведения M-шага?

Пример: Оценка параметров СММ



Задача: $p(\mathbf{y}|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \rightarrow \max_{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}}$, где $\mathbf{B} = (\mathbf{m}, \boldsymbol{\sigma}^2)$.

$$p(\mathbf{y}, \mathbf{z}|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{z_{1k}} \prod_{i=2}^N \prod_{k=1}^K \prod_{l=1}^K a_{kl}^{z_{i-1,k} z_{il}} \prod_{i=1}^N \prod_{k=1}^K \mathcal{N}(y_i | m_k, \sigma_k^2)^{z_{ik}}.$$

$$\log p(\mathbf{y}, \mathbf{z}|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) = \sum_{k=1}^K z_{1k} \log \pi_k + \sum_{i=2}^N \sum_{k=1}^K \sum_{l=1}^K z_{i-1,k} z_{il} \log a_{kl} +$$

$$\sum_{i=1}^N \sum_{k=1}^K z_{ik} \left(-\frac{1}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} (y_i - m_k)^2 \right).$$

E-шаг: $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{y}, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

M-шаг: $E_q \log p(\mathbf{y}, \mathbf{z}|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \rightarrow \max_{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}}$.

Пример: Оценка параметров СММ 2 (M-шаг)

$$\begin{aligned} \mathbb{E}_q \log p(\mathbf{y}, \mathbf{z} | \mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) &= \sum_{k=1}^K \mathbb{E} z_{1k} \log \pi_k + \sum_{i=2}^N \sum_{k=1}^K \sum_{l=1}^K \mathbb{E} z_{i-1,k} z_{il} \log a_{kl} + \\ &\sum_{i=1}^N \sum_{k=1}^K \mathbb{E} z_{ik} \left(-\frac{1}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} (y_i - m_k)^2 \right). \end{aligned}$$

$$\mathbb{E}_q \log p(\mathbf{y}, \mathbf{z} | \mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \rightarrow \max_{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}}.$$

$$\pi_k = \mathbb{E} z_{1k}, \quad a_{kl} \propto \sum_{i=2}^N \mathbb{E} z_{i-1,k} z_{il};$$

$$m_k = \frac{\sum_{i=1}^N \mathbb{E} z_{ik} y_i}{\sum_{i=1}^N \mathbb{E} z_{ik}}, \quad \sigma_k^2 = \frac{\sum_{i=1}^N \mathbb{E} z_{ik} (y_i - m_k)^2}{\sum_{i=1}^N \mathbb{E} z_{ik}}.$$

Вопрос: Что требуется знать про $q(\mathbf{Z})$, чтобы осуществить M-шаг?

Оценка параметров неориентированной ГМ

$$p(\mathbf{y}|\Theta) \rightarrow \max_{\Theta}, p(\mathbf{x}, \mathbf{y}|\Theta) = \frac{1}{Z} \prod_i \psi_i(\mathbf{x}_i, \mathbf{y}_i|\Theta_i).$$

Вопрос: Пусть все переменные наблюдаемые $\mathbf{x} = \emptyset$;

пусть дополнительно все параметры Θ_i в разных факторах независимы.

$$\text{Верно ли } \Theta_i^* = \arg \max_{\Theta_i} \sum_{j=1}^m \log \psi_i(\mathbf{y}_i^j|\Theta_i)?$$

Пусть все переменные наблюдаемые $\mathbf{x} = \emptyset$.

$$\log p(\mathbf{y}_1, \dots, \mathbf{y}_m|\Theta) = \sum_{j=1}^m \sum_i \log \psi_i(\mathbf{y}_i^j|\Theta_i) - m \log Z(\Theta) \rightarrow \max_{\Theta}.$$

$$\nabla_{\Theta} \log p(\mathbf{y}_1, \dots, \mathbf{y}_m|\Theta) = \sum_{j=1}^m \sum_i \nabla_{\Theta} \log \psi_i(\mathbf{y}_i^j|\Theta_i) - m \nabla_{\Theta} \log Z(\Theta).$$

Идея: Оценить $\nabla_{\Theta} \log Z(\Theta)$ и построить градиентный алгоритм максимизации $\log p(\mathbf{y}_1, \dots, \mathbf{y}_m|\Theta)$ по Θ , например:

$$\Theta^{n+1} = \Theta^n + \lambda \nabla_{\Theta} \log p(\mathbf{y}_1, \dots, \mathbf{y}_m|\Theta^n).$$

Оценка $Z(\Theta)$: Importance Sampling

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)}\tilde{p}(\mathbf{y}|\Theta), \quad Z(\Theta) = \int \tilde{p}(\mathbf{y}|\Theta)d\mathbf{y}.$$

Пусть $p_0(\mathbf{y})$ – некоторое предположное распределение.

$$Z = \int \frac{p_0(\mathbf{y})}{p_0(\mathbf{y})}\tilde{p}(\mathbf{y})d\mathbf{y} = \int p_0(\mathbf{y})\frac{\tilde{p}(\mathbf{y})}{\frac{1}{Z_0}\tilde{p}_0(\mathbf{y})}d\mathbf{y} = Z_0 \int p_0(\mathbf{y})\frac{\tilde{p}(\mathbf{y})}{\tilde{p}_0(\mathbf{y})}d\mathbf{y}.$$

Выборочная оценка: $\hat{Z} = \frac{Z_0}{K} \sum_{k=1}^K \frac{\tilde{p}(\mathbf{y}_k)}{\tilde{p}_0(\mathbf{y}_k)}$, $\mathbf{y}_k \sim p_0$.

Вопрос 1: Чем отличаются выборочные оценки \hat{Z} , построенные для разных $(p_0(\mathbf{y}), Z_0)$?

$$D\hat{Z} = \frac{Z_0}{K^2} \sum_{k=1}^K \left(\frac{\tilde{p}(\mathbf{y}_k)}{\tilde{p}_0(\mathbf{y}_k)} - \hat{Z} \right)^2.$$

Вопрос 2: Как зависит дисперсия оценки $D\hat{Z}$ от количества сэмплов K ?

Замечание: Схема эффективна, если $p_0(\mathbf{y}) \approx p(\mathbf{y})$.

Оценка $Z(\Theta)$: Bridge Sampling

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)} \tilde{p}(\mathbf{y}|\Theta), \quad Z(\Theta) = \int \tilde{p}(\mathbf{y}|\Theta) d\mathbf{y}.$$

Пусть $p_0(\mathbf{y})$ – некоторое предположеное распределение, а $p_*(\mathbf{y})$ – интерполирующее распределение между p_0 и p .

$$\hat{Z}_* = \frac{Z_0}{K} \sum_{k=1}^K \frac{\tilde{p}_*(\mathbf{y}_k^0)}{\tilde{p}_0(\mathbf{y}_k^0)}, \quad \mathbf{y}_k^0 \sim p_0; \quad \hat{Z}_* = \frac{Z}{K} \sum_{k=1}^K \frac{\tilde{p}_*(\mathbf{y}_k)}{\tilde{p}(\mathbf{y}_k)}, \quad \mathbf{y}_k \sim p.$$

$$\frac{Z}{Z_0} \approx \frac{\sum_{k=1}^K \frac{\tilde{p}_*(\mathbf{y}_k^0)}{\tilde{p}_0(\mathbf{y}_k^0)}}{\sum_{k=1}^K \frac{\tilde{p}_*(\mathbf{y}_k)}{\tilde{p}(\mathbf{y}_k)}}.$$

Вопрос 1: Пусть p_0 и p_* заданы. Что дополнительно требуется в Bridge Sampling против Importance Sampling с p_0 ?

Вопрос 2: Как выбрать p_* ?

$$p_*^{\text{opt}} \propto \frac{\tilde{p}_0(\mathbf{y})\tilde{p}(\mathbf{y})}{\frac{Z}{Z_0}\tilde{p}_0(\mathbf{y}) + \tilde{p}(\mathbf{y})} \text{ – зависит от } Z!$$

Идея: Итеративно обновлять $\frac{Z}{Z_0}$ и p_* .

Оценка параметров неориентированной ГМ 2

$$p(\mathbf{y}|\Theta) \rightarrow \max_{\Theta}, p(\mathbf{x}, \mathbf{y}|\Theta) = \frac{1}{Z} \prod_i \psi_i(\mathbf{x}_i, \mathbf{y}_i|\Theta_i).$$

Пусть теперь есть ненаблюдаемые переменные, то есть $\mathbf{x} \neq \emptyset$.

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)} \int \prod_{i=1}^d \psi_i(\mathbf{x}_i, \mathbf{y}_i|\Theta_i) d\mathbf{x} \rightarrow \max_{\Theta}.$$

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ .

$$\text{Введем } F(q, \Theta) = - \int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} + \int q(\mathbf{x}) \log p(\mathbf{y}, \mathbf{x}|\Theta) d\mathbf{x} = \log p(\mathbf{y}|\Theta) - D_{KL}(q||p(\mathbf{x}|\mathbf{y}, \Theta)) \rightarrow \max_{q, \Theta}.$$

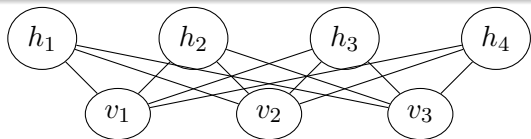
$$\text{E-шаг. } q(\mathbf{x}) = \arg \min_{q \in Q} D_{KL}(q||p(\mathbf{x}|\mathbf{y}, \Theta)).$$

$$\text{M-шаг. } -m \log Z(\Theta) + \sum_{j=1}^m \sum_{i=1}^d E_{q(\mathbf{x})} \log \psi_i(\mathbf{x}_i^j, \mathbf{y}_i^j|\Theta_i) \rightarrow \max_{\Theta}.$$

Идея: На E-шаге, сэмплировать $\mathbf{x} \sim p(\mathbf{x}|\mathbf{y}, \Theta^n)$.

На M-шаге - градиентный шаг в направлении увеличения $F(q, \Theta)$.

Пример: Restricted Boltzmann Machine



$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top \mathbf{h} - \mathbf{v}^\top \mathbf{W} \mathbf{h}, \quad v_i, h_j \in \{0, 1\}.$$

$$p(\mathbf{v}|\Theta) = p(\mathbf{v}|\mathbf{b}, \mathbf{c}, \mathbf{W}) = \frac{1}{Z(\mathbf{b}, \mathbf{c}, \mathbf{W})} \int \exp(-E(\mathbf{v}, \mathbf{h})) d\mathbf{h} \rightarrow \max_{\mathbf{b}, \mathbf{c}, \mathbf{W}}.$$

E-шаг: $\mathbf{h}_1, \dots, \mathbf{h}_K \sim p(\mathbf{h}|\mathbf{v}, \mathbf{b}^n, \mathbf{c}^n, \mathbf{W}^n)$;

$$p(\mathbf{h}|\mathbf{v}) = \prod_j p(h_j|\mathbf{v}), \quad P(h_j = 1|\mathbf{v}) = \sigma(c_j + \mathbf{v}^\top \mathbf{w}_j), \quad \mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_H].$$

M-шаг: $g(\mathbf{b}, \mathbf{c}, \mathbf{W}) =$

$$-K \log Z(\mathbf{b}, \mathbf{c}, \mathbf{W}) + \mathbf{b}^\top \sum_{l=1}^K \mathbf{v}_l + \mathbf{c}^\top \sum_{l=1}^K \mathbf{h}_l + \mathbf{v}^\top \mathbf{W} \sum_{l=1}^K \mathbf{h}_l \rightarrow \max_{\mathbf{b}, \mathbf{c}, \mathbf{W}}.$$

$$\frac{\partial g(\mathbf{b}, \mathbf{c}, \mathbf{W})}{\partial w_{ij}} = -K \frac{\partial \log Z(\mathbf{b}^n, \mathbf{c}^n, \mathbf{W}^n)}{\partial w_{ij}} + v_i h_j.$$

Свойство: $\nabla_{\Theta} \log Z(\Theta) = \mathbb{E}_{(\mathbf{v}, \mathbf{h}) \sim p(\mathbf{v}, \mathbf{h})} \nabla_{\Theta} \log \tilde{p}(\mathbf{v}, \mathbf{h}|\Theta).$

- 1 Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016: 598-621.
- 2 Ghahramani Z. Graphical models: parameter learning. URL: <https://mlg.eng.cam.ac.uk/zoubin/papers/graphical-models02.pdf>
- 3 Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
- 4 Mestres, Adria Caballe, Natalia Bochkina, and Claus Mayer. "Selection of the regularization parameter in graphical models using network characteristics." Journal of Computational and Graphical Statistics 27.2 (2018): 323-333.
- 5 Gronau, Quentin F., et al. "A tutorial on bridge sampling." Journal of mathematical psychology 81 (2017): 80-97.
- 6 Gelman, Andrew, and Xiao-Li Meng. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling." Statistical science (1998): 163-185.