Оценивание гиперпараметров графических моделей

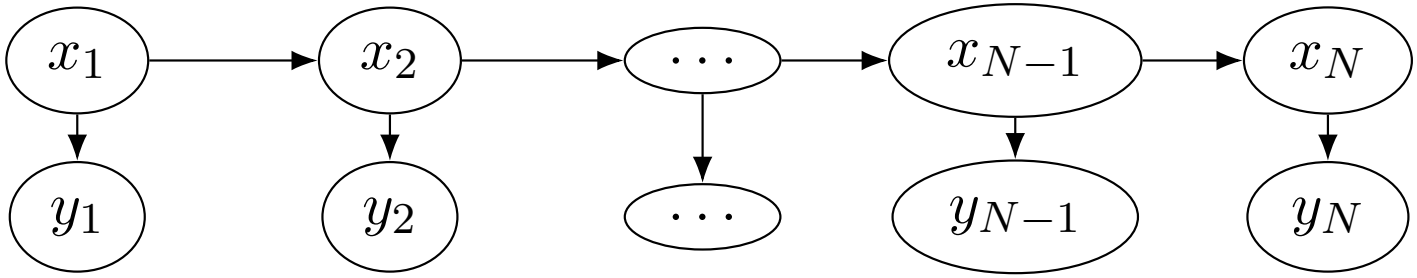
Александр Адуенко

25е мая 2022

Содержание предыдущих лекций

- Выбор априорного распределения. Распределение Джефриса.
- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Гамильтоновы методы Монте-Карло.
- 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 $x_i \in [K]$, $\mathbf{A} = \|a_{ij}\| = \|\mathsf{P}(x_l = j|x_{l-1} = i)\|$, $\pi_k = \mathsf{P}(x_1 = k)$.

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

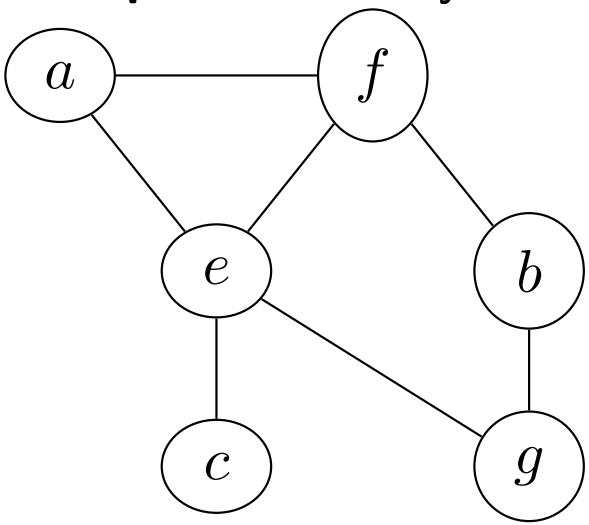
Задачи:

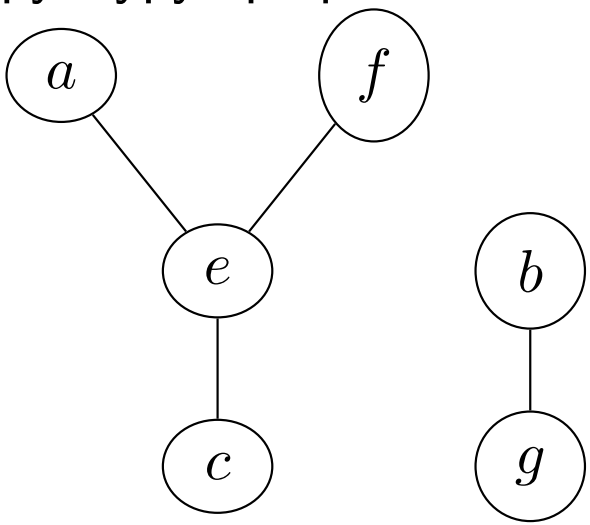
- $p(x_i|\mathbf{y},\,\Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}_C|\mathbf{y},\,\Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}|\mathbf{y},\,\Theta) \to \max\limits_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum / Graph-Cut / α – расширение / TRW;
- $p(\mathbf{x}|\mathbf{y},\,\Theta)$ – сэмплирование;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

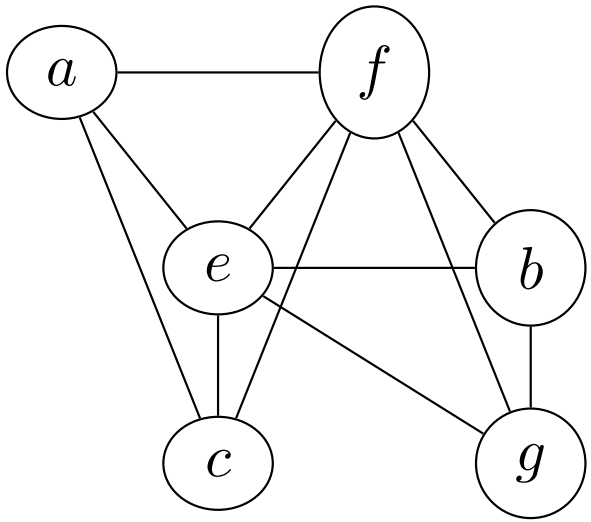
Обучение параметров графических моделей

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

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 $\mathbf{y} = [\mathbf{y}_1, \, , \ldots, \, y_K]^\mathsf{T}$, $\mathbf{y}_i \in \mathbb{R}^D$.

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

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

Идея: Ввести априорное распределение на $\mathbf{\Omega}$,
$p(\mathbf{\Omega}) \propto I[\mathbf{\Omega} > 0]\exp(-\lambda\|\mathbf{\Omega}\|_1)$.

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

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

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?

Оценка гиперпараметров ориентированной ГМ

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

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

Вопрос 1: Что изменилось по отношению к общему случаю?

$$\log p(\mathbf{y}|\Theta) = \sum_{i} \log p(\mathbf{y}_i|Pa_i,\ \Theta_i) \to \max_{\Theta}.$$

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 $\Theta_i^{kl} = \mathsf{P}(\mathbf{y}_i = k|Pa_i = l)$. Что получим для Θ_i^{kl}?

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

Оценка гиперпараметров ориентированной ГМ 2

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

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

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ.

Введем $F(q,\ \Theta) = -\int q(\mathbf{x})\log q(\mathbf{x})d\mathbf{x} + \int q(\mathbf{x})\log p(\mathbf{y},\ \mathbf{x}|\Theta)d\mathbf{x} =$

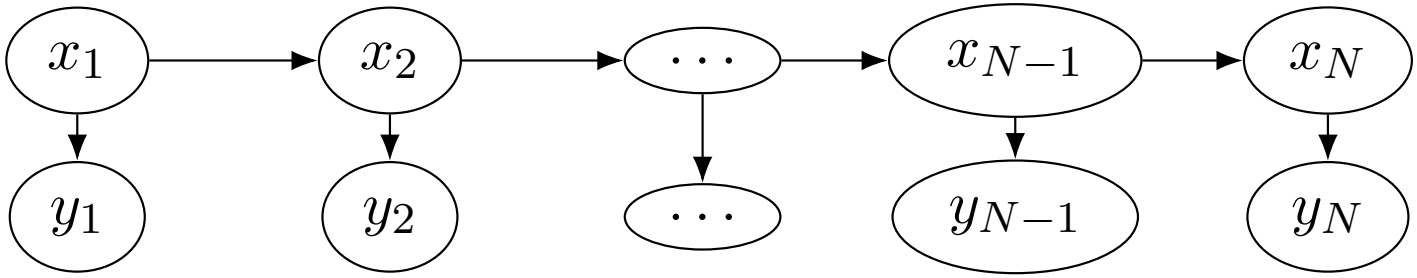
$\log p(\mathbf{y}|\Theta) - D_{\mathrm{KL}}(q\|p(\mathbf{x}|\mathbf{y},\ \Theta)) \to \max\limits_{q,\ \Theta}.$

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

M-шаг. $\sum\limits_{j=1}^{m}\sum\limits_{i=1}^{d} \mathrm{E}_{q(\mathbf{x})} \log p(\mathbf{x}_i^j\ /\ \mathbf{y}_i^j|Pa_i^j,\ \Theta_i) \to \max\limits_{\Theta}.$

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

Пример: Оценка параметров CMM



Задача: $p(\mathbf{y}|\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$, 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: $\mathsf{E}_q \log p(\mathbf{y},\,\mathbf{z}|\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$.

Пример: Оценка параметров СММ 2 (М-шаг)

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

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

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

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

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

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

Оценка параметров неориентированной ГМ

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

Вопрос: Пусть все переменные наблюдаемые $\mathbf{x} = \emptyset$;
пусть дополнительно все параметры Θ_i в разных факторах независимы.
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 $\mathbf{x} = \emptyset$.

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

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

Идея: Оценить $\nabla_\Theta \log Z(\Theta)$ и построить градиентный алгоритм максимизации $\log p(\mathbf{y}_1,\ \ldots,\ y_m|\Theta)$ по Θ, например:
$\Theta^{n+1} = \Theta^n + \lambda\nabla_\Theta \log p(\mathbf{y}_1,\ \ldots,\ \mathbf{y}_m|\Theta^n)$.

Оценка $Z(\Theta)$: Importance Sampling

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

Пусть $p_0(\mathbf{y})$ – некоторое предложное 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 $(p_0(\mathbf{y}),\ Z_0)$?

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

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

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

Оценка $Z(\Theta)$: Bridge Sampling

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

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

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

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

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_*?

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

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

Оценка параметров неориентированной ГМ 2

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

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

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

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ.

Введем $F(q,\ \Theta) = -\int q(\mathbf{x})\log q(\mathbf{x})d\mathbf{x} + \int q(\mathbf{x})\log p(\mathbf{y},\ \mathbf{x}|\Theta)d\mathbf{x} =$

$\log p(\mathbf{y}|\Theta) - D_{\mathrm{KL}}(q\|p(\mathbf{x}|\mathbf{y},\ \Theta)) \to \max\limits_{q,\ \Theta}.$

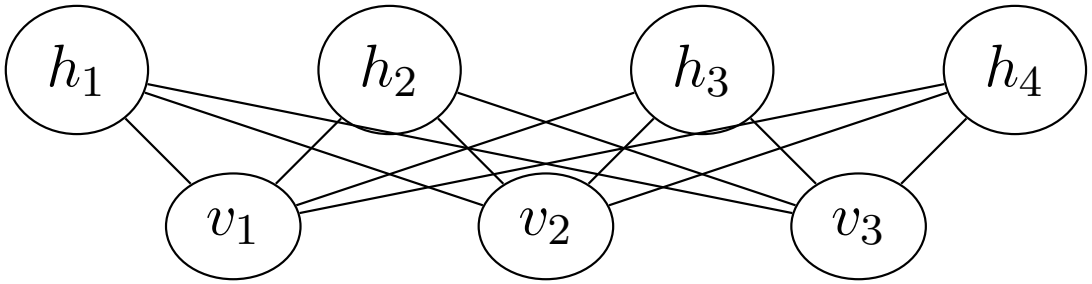
E-шаг. $q(\mathbf{x}) = \arg\min\limits_{q\in Q} D_{KL}(q||p(\mathbf{x}|\mathbf{y},\ \Theta))$.

M-шаг. $-m\log Z(\Theta) + \sum\limits_{j=1}^{m}\sum\limits_{i=1}^{d} \mathsf{E}_{q(\mathbf{x})}\log\psi_i(\mathbf{x}_i^j,\ \mathbf{y}_i^j|\Theta_i) \to \max\limits_{\Theta}.$

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



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

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

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

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

M-шаг:

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

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

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

Литература

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.