

Байесовский выбор моделей: Сходимость EM-алгоритма.

Константин Яковлев

1е мая 2024

EM-алгоритм

Пусть \mathbf{D} – наблюдаемые переменные, \mathbf{Z} – скрытые переменные.

$$p(\mathbf{D}, \mathbf{Z}|\Theta) = p(\mathbf{D}|\mathbf{Z}, \Theta)p(\mathbf{Z}|\Theta), \quad \Theta \in \Omega.$$

Вопрос 1: как решить задачу $p(\mathbf{D}|\Theta) = \int p(\mathbf{D}, \mathbf{Z}|\Theta)d\mathbf{Z} \rightarrow \max_{\Theta \in \Omega}$?

EM-алгоритм

$$\log p(\mathbf{D}|\Theta) = \underbrace{E_{q(\mathbf{Z})} \log \frac{p(\mathbf{D}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})}}_{F(q, \Theta)} + \underbrace{D_{\text{KL}}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{D}, \Theta))}_{\geq 0}.$$

Идея 1: $p(\mathbf{D}|\Theta) \rightarrow \max_{\Theta}$ заменим на $F(q, \Theta) \rightarrow \max_{q, \Theta}$.

Идея 2: Пошагово оптимизируем по Θ и q , то есть

1 E-шаг: $q^s = F(q, \Theta^{s-1}) \rightarrow \max_q$;

2 M-шаг: $\Theta^s = F(q^s, \Theta) \rightarrow \max_{\Theta \in \Omega}$.

$$Q_n(\Theta'|\Theta) := \frac{1}{n} \sum_{i=1}^n E_{p(z_i|D_i, \Theta)} \log p(D_i, z_i|\Theta'), \quad Q(\Theta'|\Theta) := EQ_n(\Theta'|\Theta),$$

$$M_n(\Theta) := \arg \max_{\Theta' \in \Omega} Q_n(\Theta'|\Theta), \quad M(\Theta) := \arg \max_{\Theta' \in \Omega} Q(\Theta'|\Theta).$$

Вопрос 2: чему соответствуют $M(\cdot)$ и $M_n(\cdot)$?

Пример: разделение смеси гауссиан

Пусть задано $p(\mathbf{x}|\Theta) = \frac{1}{2}\mathcal{N}(\mathbf{x}|\Theta^*, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\mathbf{x}|-\Theta^*, \sigma^2\mathbf{I})$, $\mathbf{x} \in \mathbb{R}^d$.

Скрытая переменная $z \in \{0, 1\}$, $p(z) = \text{Ber}(0.5)$.

$p(\mathbf{x}|z=0, \Theta) = \mathcal{N}(\mathbf{x}|-\Theta, \sigma^2\mathbf{I})$, $p(\mathbf{x}|z=1, \Theta) = \mathcal{N}(\mathbf{x}|\Theta, \sigma^2\mathbf{I})$

Выполняем E-шаг (считаем Θ фиксированным)

$q(z_i = 1) = p(z_i = 1|\mathbf{x}_i, \Theta)$,

$p(z_i = 1|\mathbf{x}_i, \Theta) = e^{-\frac{\|\mathbf{x}_i - \Theta\|^2}{2\sigma^2}} [e^{-\frac{\|\mathbf{x}_i - \Theta\|^2}{2\sigma^2}} + e^{-\frac{\|\mathbf{x}_i + \Theta\|^2}{2\sigma^2}}]^{-1} := w_{\Theta}(\mathbf{x}_i)$.

Выполняем M-шаг (считаем $q(z_i)$ фиксированным)

$Q_n(\Theta'|\Theta) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^n [w_{\Theta}(\mathbf{x}_i)\|\mathbf{x}_i - \Theta'\|^2 + (1 - w_{\Theta}(\mathbf{x}_i))\|\mathbf{x}_i + \Theta'\|^2]$,

$M_n(\Theta) = \frac{2}{n} \sum_{i=1}^n w_{\Theta}(\mathbf{x}_i)\mathbf{x}_i - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, $M(\Theta) = 2\mathbb{E}[w_{\Theta}(\mathbf{x})\mathbf{x}]$.

Анализ сходимости EM-алгоритма I

Предложение 1 (self-consistency)

Пусть $\Theta^* \in \arg \max_{\Theta \in \Omega} \mathbb{E} \log p(D_i | \Theta)$. Тогда $\Theta^* \in \arg \max_{\Theta \in \Omega} Q(\Theta | \Theta^*)$.

Предположение 1

Пусть $q(\cdot) := Q(\cdot | \Theta^*)$ является λ -сильно вогнутой в шаре $\mathbb{B}_r(\Theta^*)$:

$$q(\Theta_1) - q(\Theta_2) - \langle \nabla q(\Theta_2), \Theta_1 - \Theta_2 \rangle \leq -\frac{\lambda}{2} \|\Theta_1 - \Theta_2\|^2$$

Вопрос 3: выполнено ли условие выше для смеси гауссиан?

Предположение 2 (First-order Stability (FOS))

Пусть для любого $\Theta \in \mathbb{B}_r(\Theta^*)$ и некоторого $\gamma \geq 0$

$$\|\nabla Q(M(\Theta) | \Theta^*) - \nabla Q(M(\Theta) | \Theta)\|_2 \leq \gamma \|\Theta - \Theta^*\|_2$$

Замечание: для смеси гауссиан выполнено

$$\frac{2}{\sigma^2} \mathbb{E}[\|(w_{\Theta}(\mathbf{x}) - w_{\Theta^*}(\mathbf{x}))\mathbf{x}\|_2] \leq \gamma \|\Theta - \Theta^*\|_2.$$

Теорема 1

Для некоторого $r > 0$ и чисел $0 \leq \gamma < \lambda$ предположим, что $Q(\cdot | \Theta^*)$ является λ -сильно вогнутой в $\mathbb{B}_r(\Theta^*)$, а также выполнено условие FOS(γ) в $\mathbb{B}_r(\Theta^*)$. Тогда для любого $\Theta \in \mathbb{B}_r(\Theta^*)$

$$\|M(\Theta) - \Theta^*\|_2 \leq \frac{\gamma}{\lambda} \|\Theta - \Theta^*\|_2.$$

Доказательство: запишем условие оптимальности первого порядка:

$$\begin{aligned} \langle \nabla Q(\Theta^* | \Theta^*), M(\Theta) - \Theta^* \rangle &\leq 0, \quad \langle \nabla Q(M(\Theta) | \Theta), \Theta^* - M(\Theta) \rangle \leq 0. \\ \Rightarrow \langle \nabla Q(M(\Theta) | \Theta^*) - \nabla Q(\Theta^* | \Theta^*), \Theta^* - M(\Theta) \rangle &\leq \\ \langle \nabla Q(M(\Theta) | \Theta^*) - \nabla Q(M(\Theta) | \Theta), \Theta^* - M(\Theta) \rangle. \end{aligned}$$

Воспользуемся λ -сильной вогнутостью:

$$\langle \nabla Q(M(\Theta) | \Theta^*) - \nabla Q(\Theta^* | \Theta^*), \Theta^* - M(\Theta) \rangle \geq \lambda \|\Theta^* - M(\Theta)\|_2^2$$

Воспользуемся FOS(γ), а также неравенством КБШ:

$$\lambda \|\Theta^* - M(\Theta)\|_2^2 \leq \gamma \|\Theta^* - M(\Theta)\|_2 \cdot \|\Theta^* - \Theta\|_2$$

Замечание: рассуждения верны для $M(\Theta) \leftarrow \text{proj}_{\mathbb{B}_r(\Theta^*)}(M(\Theta))$

Предположение 3

Пусть для заданного $\delta \in (0, 1)$ и объема выборки n с вероятностью хотя бы $1 - \delta$ выполнено $\sup_{\Theta \in \mathbb{B}_r(\Theta^*)} \|M_n(\Theta) - M(\Theta)\|_2 \leq \varepsilon(n, \delta)$.

Теорема 2

Пусть оператор M является сжимающим в $\mathbb{B}_r(\Theta^*)$ с параметром $\kappa \in (0, 1)$. Пусть $\Theta^0 \in \mathbb{B}_r(\Theta^*)$, а также $\varepsilon(n, \delta) \leq (1 - \kappa)\|\Theta^* - \Theta^0\|$. Тогда с вероятностью хотя бы $1 - \delta$:

$$\|\Theta^t - \Theta^*\|_2 \leq \kappa^t \|\Theta^0 - \Theta^*\|_2 + (1 - \kappa)^{-1} \varepsilon(n, \delta).$$

Доказательство: Докажем по индукции, что с вероятностью хотя бы $1 - \delta$ выполнено $\|\Theta^{t+1} - \Theta^*\|_2 \leq \kappa \|\Theta^t - \Theta^*\|_2 + \varepsilon(n, \delta) \leq r$. База очевидна. Докажем переход:

$$\|\Theta^{t+1} - \Theta^*\|_2 = \|M_n(\Theta^t) - \Theta^*\|_2 \leq \|M_n(\Theta^t) - M(\Theta^t)\|_2 + \|M(\Theta^t) - \Theta^*\|_2 \leq \varepsilon(n, \delta) + \kappa \|\Theta^t - \Theta^*\|_2 \leq r(1 - \kappa) + \kappa r \leq r.$$

$$\Rightarrow \|\Theta^t - \Theta^*\|_2 \leq \kappa^t \|\Theta^0 - \Theta^*\|_2 + \left(\sum_{s=0}^{t-1} \kappa^s \right) \varepsilon(n, \delta) \leq \kappa^t \|\Theta^0 - \Theta^*\|_2 + \frac{\varepsilon(n, \delta)}{1 - \kappa}.$$

Сходимость EM-алгоритма для модели разделения смеси гауссиан

Теорема 3

Пусть для достаточно большого η выполнено $\frac{\|\Theta^*\|_2}{\sigma} > \eta$. Тогда найдется универсальная константа $c > 0$ такая, что оператор M является сжимающим в шаре $\mathbb{B}_r(\Theta^*)$, где $\kappa(\eta) \leq e^{-c\eta^2}$, $r = \frac{\|\Theta^*\|_2}{4}$.

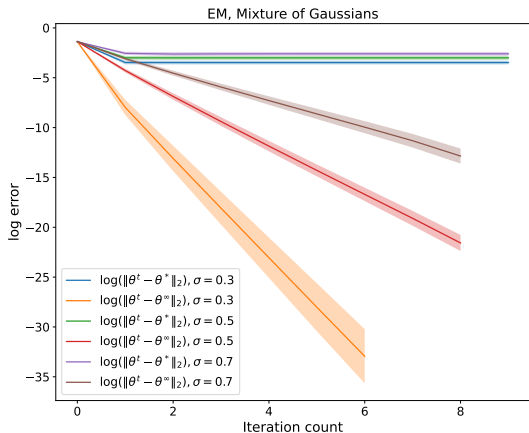
Теорема 4

Пусть выполнены условия теоремы 3. Пусть также $n \geq c_1 d \log(1/\delta)$. Тогда для любого $\Theta^0 \in \mathbb{B}_{\|\Theta^*\|/4}(\Theta^*)$ с вероятностью хотя бы $1 - \delta$ выполнено

$$\|\Theta^t - \Theta^*\|_2 \leq \kappa^t \|\Theta^0 - \Theta^*\|_2 + \frac{c_2 \|\Theta^*\|_2 \sqrt{\|\Theta^*\|_2^2 + \sigma^2}}{1 - \kappa} \sqrt{\frac{d}{n} \log(1/\delta)},$$

где c_1, c_2 – универсальные константы.

Сходимость EM-алгоритма на примере модели разделения смеси гауссиан



Параметры эксперимента

- $n = 10^4$
- $d = 10$
- $\|\Theta^*\|_2 = 1$
- $\|\Theta^0 - \Theta^*\|_2 = \frac{\|\Theta^*\|_2}{4}$

Замечание: для $\|\Theta^t - \Theta^\infty\|_2$ также можно показать линейную сходимость.

- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171, 498-505.
- 2 Balakrishnan S., Wainwright M. J., Yu B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. – 2017.