Байесовский выбор моделей: Оценки скорости сходимости EM-алгоритма.

Константин Яковлев

30е апреля 2024

Пусть $\mathbf{D}$ – наблюдаемые переменные, $\mathbf{Z}$ – скрытые переменные.
$p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}) = p(\mathbf{D}|\mathbf{Z},\ \boldsymbol{\Theta})p(\mathbf{Z}|\boldsymbol{\Theta}),\ \boldsymbol{\Theta} \in \boldsymbol{\Omega}$ – выпуклое множество.

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

ЕМ-алгоритм

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

Идея 1: $p(\mathbf{D}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$ заменим на $F(q,\ \boldsymbol{\Theta}) \to \max\limits_{q,\ \boldsymbol{\Theta}}$.

Идея 2: Пошагово оптимизируем по $\boldsymbol{\Theta}$ и q, то есть

1. Е-шаг: $q^s = F(q,\ \boldsymbol{\Theta}^{s-1}) \to \max\limits_{q}$;
2. М-шаг: $\boldsymbol{\Theta}^s = F(q^s,\ \boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}\in\boldsymbol{\Omega}}$.

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

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

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

Пример: разделение смеси гауссиан

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

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

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

Выполняем E-шаг (считаем $\boldsymbol{\Theta}$ фиксированным)

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

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

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

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

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

Анализ сходимости EM-алгоритма I

Предложение 1 (self-consistency)

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

Предположение 1

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

$$q(\mathbf{\Theta}_1) - q(\mathbf{\Theta}_2) - \langle \nabla q(\mathbf{\Theta}_2), \mathbf{\Theta}_1 - \mathbf{\Theta}_2 \rangle \leq -\frac{\lambda}{2} \|\mathbf{\Theta}_1 - \mathbf{\Theta}_2\|^2$$

Вопрос 3: выполнено ли условие выше для смеси гауссиан?

Предположение 2 (First-order Stability (FOS))

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

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

Замечание: для смеси гауссиан выполнено

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

Теорема 1

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

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

Доказательство: запишем условие оптимальности первого порядка:

$$\langle \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta}^*), M(\mathbf{\Theta}) - \mathbf{\Theta}^* \rangle \le 0, \quad \langle \nabla Q(M(\mathbf{\Theta})|\mathbf{\Theta}), \mathbf{\Theta}^* - M(\mathbf{\Theta}) \rangle \le 0.$$

$$\Rightarrow \langle \nabla Q(M(\mathbf{\Theta})|\mathbf{\Theta}^*) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta}^*), \mathbf{\Theta}^* - M(\mathbf{\Theta}) \rangle \le$$
$$\langle \nabla Q(M(\mathbf{\Theta})|\mathbf{\Theta}^*) - \nabla Q(M(\mathbf{\Theta})|\mathbf{\Theta}), \mathbf{\Theta}^* - M(\mathbf{\Theta}) \rangle.$$

Воспользуемся λ-сильной вогнутостью:

$$\langle \nabla Q(M(\mathbf{\Theta})|\mathbf{\Theta}^*) - \nabla Q(\mathbf{\Theta}^*|\mathbf{\Theta}^*), \mathbf{\Theta}^* - M(\mathbf{\Theta}) \rangle \ge \lambda \|\mathbf{\Theta}^* - M(\mathbf{\Theta})\|_2^2$$

Воспользуемся $\mathrm{FOS}(\gamma)$, а также неравенством КБШ:

$$\lambda \|\mathbf{\Theta}^* - M(\mathbf{\Theta})\|_2^2 \le \gamma \|\mathbf{\Theta}^* - M(\mathbf{\Theta})\|_2 \cdot \|\mathbf{\Theta}^* - \mathbf{\Theta}\|_2$$

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

Анализ сходимости EM-алгоритма III

Предположение 3

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

Теорема 2

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

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

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

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

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

Сходимость EM-алгоритма для модели разделения смеси гауссиан

Теорема 3

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

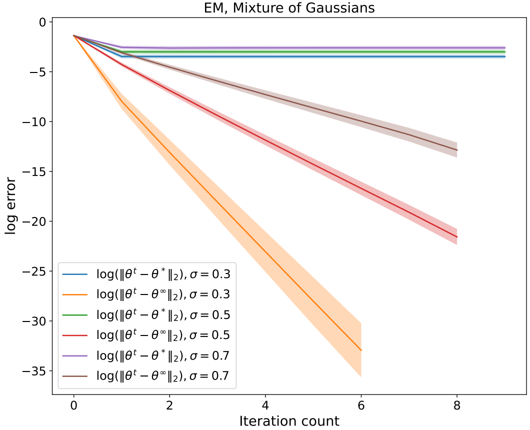
Теорема 4

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

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

где c_1, c_2 – универсальные константы.

Сходимость EM-алгоритма на примере модели разделения смеси гауссиан



Параметры эксперимента

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

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

Литература

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.