Лекция 15
ЕМ-алгоритм

Е. А. Соколов
ФКН ВШЭ

1 июня 2022 г.

Изучив ЕМ-алгоритм, возникает вопрос, зачем, было придумывать что-то новое, если уже есть хорошие методы оптимизации? Оказывается, что между данным методом и градиентным подъемом есть довольно сильная связь.

1 Связь ЕМ-алгоритма и градиентного подъёма

Теорема 1.1. *Для смеси гауссиан шаг, сделанный в ЕМ-алгоритме — это шаг градиентного подъёма, масштабированный на матрицу P, то есть:*

$$\theta^{i+1} - \theta^i = P(\theta^i) \cdot \nabla_\theta log(p(X|\theta^i)) \tag{1.1}$$

Доказательство.

Рассмотрим шаг ЕМ-алгоритма для смеси гауссиан.

Сначала вычисляются апостериорные вероятности $p(Z|X, \theta)$:

$$g_{ik} = \frac{\pi_k^{old} \mathcal{N}(x_i | \mu_k^{old}, \Sigma_k^{old})}{\sum\limits_s \pi_s^{old} \mathcal{N}(x_i | \mu_s^{old} \Sigma_k^{old})}$$

Затем пересчитываются параметры распредления:

$$\pi_k = \frac{1}{l} \sum_{i=1}^{l} g_{ik}, \qquad \mu_k = \frac{1}{l\pi_k} \sum_{i=1}^{l} g_{ik} x_i, \qquad \Sigma_k = \frac{1}{l\pi_k} \sum_{i=1}^{l} g_{ik} (x_i - \mu_k)(x_i - \mu_k)^T$$

Теперь, убедимся, что утверждение теоремы выполнено для параметра π_k. Для этого продифференцируем логарифм неполного правдоподобия по этому параметру:

$$\frac{\partial}{\partial \pi_k} \log p(x_i|\theta) = \frac{\partial}{\partial \pi_k} \sum_{i=1}^{l} log \left(\sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i | \mu_k, \Sigma_k) \right) = \sum_{i=1}^{l} \frac{\mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum\limits_s \pi_s \mathcal{N}(x_i | \mu_s \Sigma_s)}$$

Рассмотрим разницу старого и нового значения параметра π_k:

$$\pi_k^{new} - \pi_k^{old} = \frac{1}{l}\sum_{i=1}^{l}\frac{\pi_k^{old}\mathcal{N}(x_i|\mu_k^{old},\Sigma_k^{old})}{\sum\limits_s \pi_s^{old}\mathcal{N}(x_i|\mu_s^{old}\Sigma_s^{old})} - \pi_k^{old} =$$

$$= \frac{1}{l}\left(\begin{pmatrix} \pi_1^{old} & 0 & \cdots & 0 \\ 0 & \pi_2^{old} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_K^{old} \end{pmatrix} - \begin{pmatrix} \pi_1^{old} \\ \vdots \\ \pi_K^{old} \end{pmatrix} \cdot \begin{pmatrix} \pi_1^{old} & \dots & \pi_K^{old} \end{pmatrix}\right) \cdot \underbrace{\begin{pmatrix} \sum\limits_{i=1}^{l} \frac{\mathcal{N}(x_i|\mu_1,\Sigma_1)}{\sum\limits_s \pi_s \mathcal{N}(x_i|\mu_s\Sigma_s)} \\ \vdots \\ \sum\limits_{i=1}^{l} \frac{\mathcal{N}(x_i|\mu_K,\Sigma_K)}{\sum\limits_s \pi_s \mathcal{N}(x_i|\mu_s\Sigma_s)} \end{pmatrix}}_{\nabla_\pi \log p(x_i|\theta)} =$$

$$= \underbrace{\frac{1}{l}\left(\begin{pmatrix} \pi_1^{old} & 0 & \cdots & 0 \\ 0 & \pi_2^{old} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_K^{old} \end{pmatrix} - \begin{pmatrix} \pi_1^{old} \\ \vdots \\ \pi_K^{old} \end{pmatrix} \cdot \begin{pmatrix} \pi_1^{old} & \dots & \pi_K^{old} \end{pmatrix}\right)}_{P_\pi^i} \cdot \nabla_\pi \log p(x_i|\theta)$$

То есть мы явно предъявили матрицу P_π^i, такую, что утверждение теоремы верно. Аналогично можно показать для μ и Σ. ■

Данный вид EM-алгоритма похож на один из методов второго порядка, а именно метод Ньютона, который является улучшений версией градиентного спуска, в силу более быстрой сходимости. Посмотрим на сколько EM-алгоритм ускоряет градиентный спуск.

Можно заметить, что в EM-алгоритме параметры пересчитываются, как некоторая функция от прошлых значений, то есть

$$\theta^{i+1} = M(\theta^i) \tag{1.2}$$

Для всех случаев, где EM-алгоритм применяется в теореме 1.1 было показано, что разница старого и нового значения параметра вычисляется как градиент, умноженный на какую-то матрицу P, перепишем выражение учитывая (1.2) и продифференцируем по θ^i.

$$\theta^{i+1} - \theta^i = P(\theta^i)\cdot\nabla_\theta log(p(X|\theta^i)) = M(\theta^i) - \theta^i = P_\theta^i \cdot \nabla_\theta log(p(X|\theta^i)$$
$$\frac{d}{d\theta^i} = M'(\theta^i) - I = P'(\theta^i)\cdot\nabla_\theta log(p(X|\theta^i)) + P(\theta^i)\cdot\underbrace{\nabla_\theta^2 log(p(X|\theta^i)}_{S(\theta^i)}$$

Заметим, что слагаемое $\nabla_\theta log(p(X|\theta^i)) \approx 0$ когда мы находимся около θ^* или в плоском регионе. Положим, что мы там, тогда

$$M'(\theta^i) - I \approx P(\theta^i)\cdot S(\theta^i) \implies P(\theta^i) \approx \big(I - M'(\theta^i)\big)\cdot\big(-S(\theta^i)\big)^{-1}$$

Отсюда можно заметить, что если собственные значения $M(\theta^i) \approx 0$, то $P(\theta^i) \approx (-S(\theta^i))^{-1}$, тогда формула (1.1) — в точности шаг метода Ньютона. Получается, что если мы находимся около θ^* или в плоском регионе и собственные значения $M(\theta^i) \approx 0$, тогда EM-алгоритм преобретает суперлинейную скорость сходимости.

Теперь осталось понять, где выполнено условие, что собственные значения $M(\theta^i) \approx 0$. Утверждается, что это выполнено, если известной информации больше, чем неизвестной. Другими словами, это зависит от того, насколько хорошо мы можем описать данные распределениями.

Список литературы

[1] *Ruslan Salakhutdinov, Sam Roweis, Zoubin Ghahramani*. Optimization with EM and Expectation-Conjugate-Gradient. // Published in ICML 21 August 2003.