Семинары по ЕМ-алгоритму

Евгений Соколов
sokolov.evg@gmail.com

16 марта 2017 г.

1 ЕМ-алгоритм

§1.1 Смеси распределений

Говорят, что распределение $p(x)$ является *смесью распределений*, если его плотность имеет вид

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x), \qquad \sum_{k=1}^{K} \pi_k = 1, \qquad \pi_k \geqslant 0, \tag{1.1}$$

где $p_k(x)$ — распределения компонент смеси, π_k — априорные вероятности компонент, K — число компонент. Будем считать, что распределения компонент смеси принадлежат некоторому параметрическому семейству: $p_k(x) = \varphi(x \,|\, \theta_k)$.

Рассмотрим следующий эксперимент: сначала из дискретного распределения $\{\pi_1, \dots, \pi_K\}$ выбирается номер k, а затем из распределения $\varphi(x \,|\, \theta_k)$ выбирается значение x. Покажем, что распределение переменной x будет представлять собой смесь вида (1.1).

Введем *скрытую переменную* z, отвечающую за выбор компоненты смеси. Пусть она представляет собой K-мерный бинарный случайный вектор, ровно одна компонента которого равна единице:

$$z \in \{0, 1\}^K, \qquad \sum_{k=1}^{K} z_k = 1.$$

Вероятность того, что единице будет равна k-я компонента, равна π_k:

$$p(z_k = 1) = \pi_k.$$

Запишем распределение сразу всего вектора:

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

Если номер компоненты смеси известен, то случайная величина x имеет распределение $\varphi(x \,|\, \theta_k)$:

$$p(x \,|\, z_k = 1) = \varphi(x \,|\, \theta_k),$$

или, что то же самое,

$$p(x \,|\, z) = \prod_{k=1}^{K} \Bigl[\varphi(x \,|\, \theta_k)\Bigr]^{z_k}.$$

Запишем совместное распределение переменных x и z:

$$p(x, z) = p(z)p(x \,|\, z) = \prod_{k=1}^{K} \Bigl[\pi_k \varphi(x \,|\, \theta_k)\Bigr]^{z_k}.$$

Чтобы найти распределение переменной x, нужно избавиться от скрытой переменной:

$$p(x) = \sum_{z} p(x, z).$$

Суммирование здесь ведется по всем возможным значениям z, то есть по всем K-мерным бинарным векторам с одной единицей:

$$p(x) = \sum_{z} p(x, z) = \sum_{k=1}^{K} \pi_k \varphi(x \,|\, \theta_k).$$

Мы получили, что распределение переменной x в описанном эксперименте представляет собой смесь K компонент.

§1.2 Модели со скрытыми переменными

Рассмотрим вероятностную модель с наблюдаемыми переменными X и параметрами Θ, для которой задано правдоподобие $\log p(X \,|\, \Theta)$. Предположим, что в модели также существуют *скрытые переменные* Z, описывающие ее внутреннее состояние. Тогда правдоподобие $\log p(X \,|\, \Theta)$ называется *неполным*, а правдоподобие $\log p(X, Z \,|\, \Theta)$ — *полным*. Они связаны соотношением

$$\log p(X \,|\, \Theta) = \log \left\{ \sum_{Z} p(X, Z \,|\, \Theta) \right\}.$$

Как правило, знание скрытых переменных существенно упрощает правдоподобие и позволяет достаточно просто оценить параметры Θ.

Рассмотрим пример со смесями распределений. В качестве наблюдаемых переменных здесь выступает выборка $X^\ell = \{x_1, \dots, x_\ell\}$, в качестве скрытых переменных — номера компонент, из которых сгенерированы объекты $Z = \{z_1, \dots, z_\ell\}$ (здесь каждый из z_i является K-мерным вектором), в качестве параметров — априорные вероятности и параметры компонент $\Theta = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$. Неполное правдоподобие имеет вид

$$\log p(X \,|\, \Theta) = \sum_{i=1}^{\ell} \log \left\{ \sum_{k=1}^{K} \pi_k p(x_i \,|\, \theta_k) \right\}.$$

Правдоподобие здесь имеет вид «логарифм суммы». Если приравнять нулю его градиент, то получатся сложные уравнения, не имеющие аналитического решения. Данное правдоподобие сложно вычислять, оно не является вогнутым и имеет много локальных экстремумов, поэтому применение итерационных методов для его непосредственной максимизации приводит к медленной сходимости.

Рассмотрим теперь полное правдоподобие:

$$\log p(X, Z \,|\, \Theta) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} z_{ik} \Bigl\{ \log \pi_k + \log \varphi(x_i \,|\, \theta_k) \Bigr\}.$$

Оно имеет вид «сумма логарифмов», и позволяет аналитически найти оценки максимального правдоподобия на параметры Θ при известных переменных X и Z. Проблема же заключается в том, что нам не известны скрытые переменные Z, поэтому их необходимо оценивать одновременно с параметрами, что никак не легче максимизации неполного правдоподобия. Решение данной проблемы предлагается в *EM-алгоритме.*

§1.3 EM-алгоритм

EM-алгоритм решает задачу максимизации полного правоподобия путем попеременной оптимизации по параметрам и по скрытым переменным.

Опишем сначала «наивный» способ оптимизации. Зафиксируем некоторое начальное приближение для параметров Θ^{old}. При известных наблюдаемых переменных X и параметрах Θ^{old} мы можем оценить скрытые переменные, найдя их наиболее правдоподобные значения:

$$Z^* = \arg\max_Z p(Z \,|\, X, \Theta^{\text{old}}) = \arg\max_Z p(X, Z \,|\, \Theta^{\text{old}}).$$

Зная скрытые переменные, мы можем теперь найти следующее приближение для параметров:

$$\Theta^{\text{new}} = \arg\max_{\Theta} p(X, Z^* \,|\, \Theta).$$

Повторяя итерации до сходимости, мы получим некоторый итоговый вектор параметров Θ^*. Данная процедура, однако, не гарантирует сходимости и может выдать неоптимальное решение.

Гораздо лучшие результаты можно получить, воспользовавшись байесовским подходом. Как и прежде, зафиксируем вектор параметров Θ^{old}, но вместо точечной оценки вычислим апостериорное распределение на скрытых переменных $p(Z \,|\, X, \Theta^{\text{old}})$. В этом заключается *E-шаг* EM-алгоритма.

Усредним логарифм полного правдоподобия по всем возможным значениям скрытых переменных Z с весами, равными апостериорным вероятностям этих переменных $p(Z \,|\, X, \Theta^{\text{old}})$:

$$Q(\Theta, \Theta^{\text{old}}) = \mathbb{E}_{Z \sim p(Z \,|\, X, \Theta^{\text{old}})} \log p(X, Z \,|\, \Theta) = \sum_Z p(Z \,|\, X, \Theta^{\text{old}}) \log p(X, Z \,|\, \Theta).$$

Формально говоря, мы нашли матожидание логарифма полного правдоподобия по апостериорному распределению на скрытых переменных. На *M-шаге* новый вектор параметров находится как максимизатор данного матожидания:

$$\Theta^{\text{new}} = \arg\max_{\Theta} Q(\Theta, \Theta^{\text{old}}) = \arg\max_{\Theta} \sum_{Z} p(Z \mid X, \Theta^{\text{old}}) \log p(X, Z \mid \Theta).$$

Далее мы рассмотрим условия сходимости описанного итерационного процесса и увидим, что при достаточно общих условиях любая из его сходящихся подпоследовательностей сойдется к стационарной точке неполного правдоподобия.

§1.4 Вывод формул для задачи разделения смеси

На лекциях EM-алгоритм для разделения смеси гауссиан был выведен следующим образом: неполное правдоподобие продифференцировано по параметрам и приравнено к нулю, в полученных уравнениях сгруппированы и зафиксированы выражения, имеющие смысл вероятности получения объектов из определенных компонент смеси; после этого из уравнений стало возможно аналитически выразить параметры. В данном разделе мы выведем формулы, исходя из описанной общей схемы EM-алгоритма, и увидим, что они совпадают с формулами из лекций.

1.4.1 Разделение смеси нормальных распределения

Запишем полное правдоподобие для смеси нормальных распределений:

$$\log p(X, Z \mid \Theta) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} z_{ik} \Big\{ \log \pi_k + \log \varphi(x_i \mid \theta_k) \Big\}.$$

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

$$p(Z \mid X, \Theta^{\text{old}}) = \frac{p(X, Z \mid \Theta^{\text{old}})}{p(X \mid \Theta^{\text{old}})} \propto \prod_{i=1}^{\ell} \prod_{k=1}^{K} \Big[\pi_k^{\text{old}} \mathcal{N}(x_i \mid \mu_k^{\text{old}}, \Sigma_k^{\text{old}}) \Big]^{z_{ik}}.$$

Заметим, что данное распределение распадается в произведение распределений, соответствующих отдельным объектам $p(z_i \mid x_i, \Theta)$:

$$p(Z \mid X, \Theta^{\text{old}}) = \prod_{i=1}^{\ell} p(z_i \mid x_i, \Theta^{\text{old}}).$$

Иными словами, величины $\{z_i\}$ независимы при известных объектах $\{x_i\}$. Вектор z_i имеет K возможных значений. Запишем их вероятности, воспользовавшись формулой Байеса:

$$\begin{aligned} g_{ik} \equiv p(z_{ik} = 1 \mid x_i, \Theta^{\text{old}}) &= \frac{p(z_{ik} = 1) p(x_i \mid z_{ik} = 1, \Theta^{\text{old}})}{\sum_{j=1}^{K} p(z_{ij} = 1) p(x_i \mid z_{ij} = 1, \Theta^{\text{old}})} = \\ &= \frac{\pi_k^{\text{old}} \mathcal{N}(x_i \mid \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^{K} \pi_j^{\text{old}} \mathcal{N}(x_i \mid \mu_j^{\text{old}}, \Sigma_j^{\text{old}})}. \end{aligned}$$

Вычислим теперь матожидание полного правдоподобия:

$$
\begin{aligned}
Q(\Theta, \Theta^{\text{old}}) &= \mathbb{E}_{Z \sim p(Z \,|\, X, \Theta^{\text{old}})} \log p(X, Z \,|\, \Theta) = \\
&= \mathbb{E}_{Z \sim p(Z \,|\, X, \Theta^{\text{old}})} \sum_{i=1}^{\ell} \sum_{k=1}^{K} z_{ik} \Big\{ \log \pi_k + \log \mathcal{N}(x_i \,|\, \mu_k, \Sigma_k) \Big\} = \\
&= \sum_{i=1}^{\ell} \sum_{k=1}^{K} \mathbb{E}_{Z \sim p(Z \,|\, X, \Theta^{\text{old}})}[z_{ik}] \Big\{ \log \pi_k + \log \mathcal{N}(x_i \,|\, \mu_k, \Sigma_k) \Big\}.
\end{aligned}
$$

Нам понадобится вспомогательная величина:

$$
\begin{aligned}
\mathbb{E}_{Z \sim p(Z \,|\, X, \Theta^{\text{old}})}[z_{ik}] &= 1 * p(z_{ik} = 1 \,|\, x_i, \Theta) + 0 * p(z_{ik} = 0 \,|\, x_i, \Theta) = \\
&= \frac{\pi_k^{\text{old}} \mathcal{N}(x_i \,|\, \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{j=1}^{K} \pi_j^{\text{old}} \mathcal{N}(x_i \,|\, \mu_j^{\text{old}}, \Sigma_j^{\text{old}})} = \\
&= g_{ik}.
\end{aligned}
$$

Получаем:

$$
Q(\Theta, \Theta^{\text{old}}) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{ \log \pi_k + \log \mathcal{N}(x_i \,|\, \mu_k, \Sigma_k) \Big\}.
$$

Дифференцируя данный функционал, нетрудно получить формулы М-шага:

$$
\begin{aligned}
\pi_k &= \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ik}; \\
\mu_k &= \frac{1}{\ell \pi_k} \sum_{i=1}^{\ell} g_{ik} x_i; \\
\Sigma_k &= \frac{1}{\ell \pi_k} \sum_{i=1}^{\ell} g_{ik} (x_i - \mu_k)(x_i - \mu_k)^T.
\end{aligned}
$$

§1.5 Дивергенция Кульбака-Лейблера

Дивергенция Кульбака-Лейблера — это мера расстояния между двумя вероятностными распределениями, которая определяется как

$$
\text{KL}\,(q \,\|\, p) = \int q(x) \log \frac{q(x)}{p(x)} dx.
$$

В случае с дискретными распределениями она принимает вид

$$
\text{KL}\,(q \,\|\, p) = \sum_{x} q(x) \log \frac{q(x)}{p(x)}.
$$

KL-дивергенция определена только в том случае, когда из $p(x) = 0$ следует $q(x) = 0$. При вычислении мы считаем, что $0 \log 0 = 0$ и $0 \log \frac{0}{0} = 0$.

Задача 1.1. *Покажите, что KL-дивергенция неотрицательна.*

Решение. Нам понадобится неравенство Йенсена в интегральной форме: для любой вогнутой функции $f(x)$ выполнено

$$f\left(\int \alpha(x)y(x)dx\right) \geqslant \int \alpha(x)f(y(x))dx, \quad \int \alpha(x)dx = 1, \quad \alpha(x) \geqslant 0.$$

Пользуясь данным неравенством и вогнутостью логарифма, получаем:

$$\mathrm{KL}\,(q \,\|\, p) = -\int q(x)\log\frac{p(x)}{q(x)}dx \geqslant -\log\left(\int q(x)\frac{p(x)}{q(x)}dx\right) = -\log\left(\int p(x)dx\right) = 0.$$

Можно доказать данное утверждение и без использования неравенства Йенсена, если в определении используется натуральный логарифм. Заметим, что при $y > 0$ имеет место неравенство $\ln y \leqslant y - 1$, которое обращается в равенство только при $y = 1$. Тогда:

$$\mathrm{KL}\,(q \,\|\, p) = -\int q(x)\log\frac{p(x)}{q(x)}dx \geqslant -\int q(x)\left(\frac{p(x)}{q(x)} - 1\right)dx = \int q(x)dx - \int p(x)dx = 0.$$

■

Неравенство Йенсена обращается в равенство тогда и только тогда, когда $y(x) = \mathrm{const}$. В нашем случае это означает, что $\frac{p(x)}{q(x)} = \mathrm{const}$. Поскольку q и p — вероятностные распределения, это возможно только при их равенстве. Мы получили важное свойство KL-дивергенции: она обращается в нуль тогда и только тогда, когда ее аргументы равны.

Задача 1.2. *Пусть заданы выборка X^ℓ и распределение на объектах $p(x \mid \theta)$, параметр которого мы хотим настроить под данную выборку. Эмпирическим распределением называется дискретное распределение на объектах, присваивающее каждому объекту из обучающей выборки вероятность $1/\ell$:*

$$\hat{p}(x \mid X^\ell) = \sum_{i=1}^{\ell} \frac{1}{\ell}[x = x_i].$$

Покажите, что максимизация правдоподобия эквивалентна минимизации дивергенции Кульбака-Лейблера между эмпирическим распределением и модельным распределением: KL $\left(\hat{p}(x \mid X^\ell) \,\middle\|\, p(x \mid \theta)\right)$.

Решение. Распишем указанную дивергенцию:

$$\mathrm{KL}\left(\hat{p}(x \mid X^\ell) \,\middle\|\, p(x \mid \theta)\right) = \sum_{i=1}^{\ell} \frac{1}{\ell}\log\frac{1/\ell}{p(x_i \mid \theta)} = \\ = \sum_{i=1}^{\ell} \frac{1}{\ell}\log\frac{1}{\ell} - \frac{1}{\ell}\sum_{i=1}^{\ell}\log p(x_i \mid \theta) \to \min_\theta.$$

Отбросим константные члены:

$$\sum_{i=1}^{\ell}\log p(x_i \mid \theta) \to \max_\theta.$$

Мы получили задачу максимизации логарифма правдоподобия.

■

Таким образом, метод максимума правдоподобия старается подобрать такие параметры модели, чтобы она давала равномерное распределение на объектах выборки и присваивала нулевую вероятность всем остальным объектам.

§1.6 Обоснование EM-алгоритма

Представим неполное правдоподобие в виде суммы двух функций:

$$\log p(X \mid \Theta) = \mathcal{L}(q, \Theta) + \mathrm{KL}\left(q \parallel p\right), \tag{1.2}$$

где

$$\mathcal{L}(q, \Theta) = \sum_Z q(Z) \log \frac{p(X, Z \mid \Theta)}{q(Z)},$$
$$\mathrm{KL}\left(q \parallel p\right) = -\sum_Z q(Z) \log \frac{p(Z \mid X, \Theta)}{q(Z)}.$$

Здесь $q(Z)$ — это произвольное распределение на скрытых переменных.

Задача 1.3. *Докажите, что это представление корректно.*

Решение.

$$\begin{aligned}
&\sum_Z q(Z) \log \frac{p(X, Z \mid \Theta)}{q(Z)} - \sum_Z q(Z) \log \frac{p(Z \mid X, \Theta)}{q(Z)} = \\
&\quad = \sum_Z q(Z) \log \frac{p(X, Z \mid \Theta)}{p(Z \mid X, \Theta)} = \\
&\quad = \sum_Z q(Z) \log p(X \mid \Theta) = \\
&\quad = \log p(X \mid \Theta) \sum_Z q(Z) = \\
&\quad = \log p(X \mid \Theta).
\end{aligned}$$

■

Заметим, что $\mathcal{L}(q, \Theta)$ — это нижняя оценка на логарифм правдоподобия:

$$\log p(X \mid \Theta) = \mathcal{L}(q, \Theta) + \underbrace{\mathrm{KL}\left(q \parallel p\right)}_{\geqslant 0} \geqslant \mathcal{L}(q, \Theta).$$

Чем «правильнее» выбрано распределение $q(Z)$, тем точнее эта оценка. Будем по очереди максимизировать нижнюю оценку $\mathcal{L}(q, \Theta)$ по q и по Θ. Зафиксируем сначала вектор параметров Θ^{old} и найдем максимум по q. Заметим, что в разложении (1.2) левая часть не зависит от q, поэтому нижняя оценка будет максимальна тогда, когда KL-дивергенция будет минимальна. Мы знаем, что минимум дивергенции равен нулю и достигается на равных распределениях. Таким образом, нижняя оценка достигнет своего максимума на $q = p(Z \mid X, \Theta^{\text{old}})$. Мы получили E-шаг EM-алгоритма — вычисление апостериорного распределения на скрытых переменных.

Зафиксируем теперь q и найдем максимум нижней оценки по Θ. Преобразуем задачу:

$$\begin{aligned}
\mathcal{L}(q, \Theta) &= \sum_Z q(Z) \log \frac{p(X, Z \mid \Theta)}{q(Z)} = \\
&= \sum_Z q(Z) \log p(X, Z \mid \Theta) - \sum_Z q(Z) \log q(Z) = \\
&= \sum_Z p(Z \mid X, \Theta^{\text{old}}) \log p(X, Z \mid \Theta) + \text{const}(\Theta) = \\
&= Q(\Theta, \Theta^{\text{old}}) + \text{const}(\Theta) \to \max_{\Theta}.
\end{aligned}$$

Мы получили оптимизационную задачу с M-шага EM-алгоритма.

Описанный способ вывода E- и M-шагов позволяет получить важное свойство EM-алгоритма — *на каждой его итерации значение правдоподобия не уменьшается.* Действительно, после E-шага значение нижней оценки совпадает со значением правдоподобия, а значит, максимизация оценки на M-шаге приведет и к максимизации правдоподобия:

$$\log p(X \mid \Theta^{\text{new}}) = \mathcal{L}(q, \Theta^{\text{new}}) + \text{KL}\,(\, q \,\|\, p) \geqslant \mathcal{L}(q, \Theta^{\text{new}}) \geqslant \mathcal{L}(q, \Theta^{\text{old}}) = \log p(X \mid \Theta^{\text{old}}).$$

Если правдоподобие ограничено сверху, то последовательность значений правдоподобия $\{p(X \mid \Theta^i)\}_i$ обязательно сойдется. Здесь мы обозначили последовательность параметров, генерируемую EM-алгоритмом, через $\{\Theta^i\}_i$.

Существуют и более сильные утверждения о сходимости.

Теорема 1.1 ([1]). *Пусть $Q(\Theta, \Theta^{old})$ непрерывна по Θ и Θ^{old}. Тогда все предельные точки последовательности $\{\Theta^i\}_i$ являются стационарными точками неполного правдоподобия $p(X \mid \Theta)$, а последовательность $\{p(X \mid \Theta^i)\}_i$ монотонно сходится к значению правдоподобия $L^* = p(X \mid \Theta^*)$ в одной из стационарных точек Θ^*.*

Обратим внимание на тот факт, что сходимость последовательности $\{\Theta^i\}_i$ не гарантируется — у нее может быть несколько подпоследовательностей, каждая из которых будет сходиться к своей стационарной точке. Также отметим, что речь идет только о сходимости к *стационарной* точке; сходимость к локальному максимуму гарантируется лишь для некоторых семейств распределений (например, для экспоненциальных [1]).

Покажем одно из свойств EM-алгоритма.

Задача 1.4. *Докажите, что если Θ^i не является стационарной точкой логарифма правдоподобия, то следующее приближение Θ^{i+1}, выданное EM-алгоритмом, будет отличаться от Θ^i.*

Решение. Пусть Θ^i не является стационарной точкой, то есть

$$\nabla_\Theta \log p(X \mid \Theta)|_{\Theta^i} \neq 0.$$

Выполним E-шаг, найдем апостериорное распределение $q(\Theta^i)$, и запишем разложение правдоподобия:

$$\log p(X \mid \Theta^i) = \mathcal{L}(q, \Theta^i) + \underbrace{\mathrm{KL}\left(q(\Theta^i) \,\|\, p\right)}_{=0}.$$

KL-дивергенция здесь равна нулю в силу выбора распределения $q(\Theta^i)$. Поскольку на данном распределении достигается минимум дивергенции, ее градиент равен нулю:

$$\nabla_\Theta \mathrm{KL}\left(q(\Theta) \,\|\, p\right)\big|_{\Theta^i} = 0.$$

Получаем:

$$\nabla_\Theta \mathcal{L}(q, \Theta)\big|_{\Theta^i} = \nabla_\Theta \log p(X \mid \Theta)\big|_{\Theta^i} - \underbrace{\nabla_\Theta \mathrm{KL}\left(q(\Theta) \,\|\, p\right)\big|_{\Theta^i}}_{=0} = \nabla_\Theta \log p(X \mid \Theta)\big|_{\Theta^i} \neq 0.$$

Таким образом, точка Θ^i не является максимумом нижней оценки, и поэтому на M-шаге будет сделан переход к новой точке $\Theta^{i+1} \neq \Theta^i$.

∎

Список литературы

[1] *Wu, C. F. Jeff* (1983). On the Convergence Properties of the EM Algorithm. // Annals of Statistics, 11(1), p. 95-103.