Машинное обучение, ФКН ВШЭ
Семинар №16

1 EM-алгоритм

Напомним некоторые выражения, которые были рассмотрены на лекции.

Дивергенцией Кульбака-Лейблера называется функционал:

$$\mathrm{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{1.1}$$

Данный функционал имеет смысл «расстояния» между распределениями и обладает следующими свойствами:

- $\mathrm{KL}(p\|q) \geqslant 0, \ \forall p, \, q;$
- $\mathrm{KL}(p\|q) = 0 \iff p = q.$

EM-алгоритм — итерационный метод максимизации правдоподобия выборки. Пусть есть следующая задача:

$$\log p(X|\Theta) \to \max_{\Theta} \tag{1.2}$$

Пусть в модели существуют скрытые переменные Z, описывающее её внутреннее состояние. Для некоторого распределения $q(Z)$ на скрытых переменных верно:

$$\begin{aligned}\log p(X|\Theta) = &\int q(Z) \log p(X|\Theta) dZ = \\ &\int q(Z) \log \frac{p(X, Z|\Theta)}{p(Z|X, \Theta)} dZ = \int q(Z) \log \frac{p(X, Z|\Theta) q(Z)}{p(Z|X, \Theta) q(Z)} dZ = \\ &\int q(Z) \log \frac{p(X, Z|\Theta)}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z|X, \Theta)} dZ = \\ &\mathcal{L}(q, \Theta) + \mathrm{KL}(q\|p).\end{aligned}$$

Так как $\mathrm{KL}(q\|p) \geqslant 0$, то $\log p(X|\Theta) \geqslant \mathcal{L}(q, \Theta)$.

Напомним, что мы хотели бы максимизировать левую часть получившегося неравенства, не зависящее от распределения q, которое, в свою очередь, может быть выбрано произвольно, поэтому чем «правильнее» будет выбрано q, тем точнее будет полученная нижняя оценка в правой части неравенства. Вместо решения исходной задачи 1.2 будем максимизировать нижнюю оценку $\mathcal{L}(q, \Theta)$ поочерёдно по q и по Θ.

E-step. Максимизируем по q.

Из полученного ранее следует, что максимум $\mathcal{L}(q, \Theta)$ по q достигается в том случае, когда достигается минимум $\mathrm{KL}(q\|p)$, то есть:

$$q^*(Z) = \arg\max_q \mathcal{L}(q, \Theta^{\text{old}}) = \arg\min_q \int q(Z) \log \frac{q(Z)}{p(Z|X, \Theta^{\text{old}})} dZ = p(Z|X, \Theta^{\text{old}})$$

M-step. Максимизируем по Θ.

$$\begin{aligned}\Theta^{\text{new}} &= \arg\max_\Theta \int q^*(Z) \log \frac{p(X, Z|\Theta)}{q^*(Z)} dZ = \arg\max_\Theta \int q^*(Z) \log p(X, Z|\Theta) dZ \\ &= \arg\max_\Theta \mathbb{E}_{Z\sim q^*(Z)} \log p(X, Z|\Theta)\end{aligned}$$

Задача 1.1. *Зачем необходимо приводить исходную оптимизационную задачу 1.2 к оптимизационной задаче на М-шаге?*

Решение. Оптимизируемая функция в задаче

$$\log p(X|\Theta) \to \max_\Theta \tag{1.3}$$

часто оказывается невыпуклой. За счёт того, что скрытые переменные Z мы можем ввести произвольным образом, мы можем подобрать их так, чтобы задача

$$\Theta^* = \arg\max_\Theta \mathbb{E}_Z \log p(X, Z|\Theta)$$

имела удобный для оптимизации вид, например, чтобы распределение $p(X, Z|\Theta)$ находилось в классе экспоненциальных распределений.

■

2 Разделение смеси распределений Бернулли

Рассмотрим смесь распределений Бернулли:

$$p(x \,|\, \mu, \pi) = \sum_{k=1}^{K} \pi_k p(x \,|\, \mu_k),$$

где $x \in \mathbb{R}^d$, $\mu = \{\mu_1, \ldots, \mu_K\}$, $\mu_k \in [0, 1]^d$, $\pi = \{\pi_1, \ldots, \pi_K\}$, $\sum_{i=1}^{K} \pi_k = 1$, и

$$\begin{aligned}p(x_j \,|\, \mu_k) &= \mu_{kj}^{x_j}(1 - \mu_{kj})^{1-x_j}, \\ p(x \,|\, \mu_k) &= \prod_{j=1}^{d} \mu_{kj}^{x_j}(1 - \mu_{kj})^{1-x_j}.\end{aligned}$$

Иными словами, k-я компонента смеси — это такое распределение на d-мерных бинарных векторах, что j-я координата вектора имеет распределение Бернулли с параметром μ_{kj}.

Введём скрытые переменные Z. Переменная z_{ik} имеет смысл принадлежности объекта компоненте смеси: принимает значение 1, если i-ый объект обучающей выборки принадлежит k-ой компоненте смеси, и 0 иначе.

$$p(X, Z \mid \Theta) = \prod_{i=1}^{\ell} \prod_{k=1}^{K} \Big[\pi_k p(x_i \mid \mu_k)\Big]^{z_{ik}}$$

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

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

Заметим, что данное распределение факторизуется в произведение распределений, соответствующих отдельным объектам $p(z_i \mid x_i, \Theta^{\text{old}})$:

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

Введём обозначение:

$$g_{ik} \equiv p(z_{ik} = 1 \mid x_i, \Theta^{\text{old}}) = \frac{\pi_k^{\text{old}} p(x_i \mid \mu_k^{\text{old}})}{\sum_{s=1}^{K} \pi_s^{\text{old}} p(x_i \mid \mu_s^{\text{old}})}.$$

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

$$\mathbb{E}_{Z \sim p(Z \mid X, \Theta^{\text{old}})}[z_{ik}] = 1 * p(z_{ik} = 1 \mid x_i, \Theta^{\text{old}}) + 0 * p(z_{ik} = 0 \mid x_i, \Theta^{\text{old}}) = g_{ik}.$$

Получаем следующую оптимизационную задачу:

$$\mathbb{E}_{Z \sim p(Z \mid X, \Theta^{\text{old}})} \log p(X, Z \mid \Theta) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{\log \pi_k + \log p(x_i \mid \mu_k)\Big\} =$$

$$= \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{\log \pi_k + \sum_{j=1}^{d} (x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj}))\Big\} \to \max_{\{\pi_k, \mu_k\}}$$

Дифференцируя данный функционал, можем получить формулы M-шага:

$$\pi_k^{\text{new}} = \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ik};$$

$$\mu_{kj}^{\text{new}} = \frac{\sum_i g_{ik} x_{ij} + \sum_i g_{ik}(x_{ij} - 1)}{\sum_i g_{ik} x_{ij}}.$$

3 Разделение смеси нормальных распределений

Рассмотрим смесь нормальных распределений. В таком случае плотность вероятности нашей выборки описывается следующим образом:

$$p(X \mid \Theta) = \prod_{i=1}^{\ell} p(x_i \mid \Theta) = \prod_{i=1}^{\ell} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k),$$

где i – индекс объекта выборки, k – индекс компоненты смеси, $\pi_1, \dots \pi_K$ – априорные вероятности компонент.

Введём скрытые переменные Z аналогично предыдущей задаче:

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

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

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

Заметим, что данное распределение факторизуется в произведение распределений, соответствующих отдельным объектам $p(z_i \mid x_i, \Theta^{\text{old}})$:

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

Введём обозначение:

$$g_{ik} \equiv p(z_{ik} = 1 \mid x_i, \Theta^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(x_i \mid \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_{s=1}^{K} \pi_s^{\text{old}} \mathcal{N}(x_i \mid \mu_s^{\text{old}}, \Sigma_s^{\text{old}})}.$$

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

$$\mathbb{E}_{Z \sim p(Z \mid X, \Theta^{\text{old}})}[z_{ik}] = 1 * p(z_{ik} = 1 \mid x_i, \Theta^{\text{old}}) + 0 * p(z_{ik} = 0 \mid x_i, \Theta^{\text{old}}) = g_{ik}.$$

Получаем следующую оптимизационную задачу:

$$\mathbb{E}_{Z \sim p(Z \mid X, \Theta^{\text{old}})} \log p(X, Z \mid \Theta) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{\log \pi_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)\Big\} \to \max_{\{\pi_k, \mu_k, \Sigma_k\}}$$

Дифференцируя данный функционал, нетрудно получить формулы М-шага:

$$\pi_k^{\text{new}} = \frac{1}{\ell}\sum_{i=1}^{\ell} g_{ik};$$

$$\mu_k^{\text{new}} = \frac{1}{\ell\pi_k}\sum_{i=1}^{\ell} g_{ik}x_i;$$

$$\Sigma_k^{\text{new}} = \frac{1}{\ell\pi_k}\sum_{i=1}^{\ell} g_{ik}(x_i - \mu_k)(x_i - \mu_k)^T.$$

4 Кластеризация при помощи EM-алгоритма

Несмотря на то, что выше были рассмотрены примеры применения EM-алгоритма лишь в задаче разделения смесей распределений, он может быть применен к любой модели, в которой можно ввести скрытые переменные таким образом, что решения задач обоих шагов алгоритма могут быть получены в явном виде. Тем не менее, как говорилось на лекции, модель смеси распределений позволяет рассматривать EM-алгоритм также в качестве модели *мягкой кластеризации*, предполагающей, что каждая компонента смеси является кластером со своим набором параметров (визуализация применения EM-алгоритма для разделения смеси гауссиан).