Машинное обучение, ФКН ВШЭ
Семинар №15

1 EM-алгоритм

Напомним некоторые выражения, которые были рассмотрены на лекции.

Дивергенцией Кульбака-Лейблера называется функционал:

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

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:

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

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

Теперь вместо прямой максимизации логарифма правдоподобия будем максимизировать нижнюю вариационную оценку $\mathcal{L}(q, \Theta)$ поочерёдно по q и по Θ.

E-step. Максимизируем по q.

Максимизация нижней вариационной оценки эквивалентна минимизации второго слагаемого — дивергенции Кульбака-Лейблера.

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

M-step. Максимизируем по Θ.

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

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

$$p(x_i \,|\, \mu_k) = \mu_{ki}^{x_i} (1 - \mu_{ki})^{1 - x_i},$$
$$p(x \,|\, \mu_k) = \prod_{i=1}^{d} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1 - x_i}.$$

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

Введём скрытые переменные

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

На Е-шаге вычисляется апостериорное распределение на скрытых переменных:

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

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

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

Тогда параметры распределения $p(Z \mid X, \theta)$ вычисляются по следующим формулам:

$$g_{ik} \equiv p(z_{ik} = 1 \mid x_i, \theta) = \frac{\pi_k p(x_i \mid \mu_k)}{\sum_{j=1}^{K} \pi_j p(x_i \mid \mu_j)}.$$

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

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

Получаем:

$$\begin{aligned} \mathbb{E}_{Z \sim p(Z \mid X, \theta)} \log p(X, Z \mid \theta) = \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{\log \pi_k + \log p(x_i \mid \mu_k)\Big\} = \\ = \sum_{i=1}^{\ell} \sum_{k=1}^{K} g_{ik} \Big\{\log \pi_k + \sum_{m=1}^{d} (x_i \log \mu_{km} + (1 - x_i) \log(1 - \mu_{km}))\Big\} \end{aligned}$$

Дифференцируя данный функционал, нетрудно получить формулы М-шага:

$$\begin{aligned} \pi_k &= \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ik}; \\ \mu_{km} &= \frac{\sum_i g_{ik} x_i + \sum_i g_{ik}(x_i - 1)}{\sum_i g_{ik} x_i} \end{aligned}$$

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. 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.

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

Задача 3.1. *Покажите, что данное выражение действительно является функцией плотности совместного распределения X и Z. То есть покажите, что выполняется нормировка.*

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

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

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

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

Тогда параметры распределения $p(Z \mid X, \theta)$ вычисляются по следующим формулам:

$$g_{ik} \equiv p(z_{ik} = 1 \mid x_i, \theta) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)}.$$

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

$$\mathbb{E}_{Z\sim p(Z\,|\,X,\theta)}[z_{ik}] = 1 * p(z_{ik} = 1\,|\,x_i,\theta) + 0 * p(z_{ik} = 0\,|\,x_i,\theta) = g_{ik}.$$

Получаем:

$$\mathbb{E}_{Z\sim p(Z\,|\,X,\theta)} \log p(X, Z\,|\,\theta) = \sum_{i=1}^{\ell}\sum_{k=1}^{K} g_{ik}\Big\{\log \pi_k + \log \mathcal{N}(x_i\,|\,\mu_k, \Sigma_k)\Big\}.$$

Дифференцируя данный функционал, нетрудно получить формулы M-шага:

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