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 зависящую 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 когда $q = p$:

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

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.

■

Задача 1.2. *Почему ЕМ-алгоритм необходимо сходится к локальному максимуму неполного правдоподобия $p(X|\Theta)$?*

Решение. Рассмотрим очередную итерацию ЕМ-алгоритма из начального приближения Θ^{old}. Вспомним разложение логарифма неполного правдоподобия на KL-дивиргенцию и нижнюю оценку для некоторого q:

$$\log p(X|\Theta^{\text{old}}) = \mathcal{L}(q, \Theta^{\text{old}}) + \mathrm{KL}(q(Z)\|p(Z|X, \Theta^{\text{old}}))$$

На Е-шаге мы выбираем $q^*(Z)$ равное апостериорному распределению на скрытые переменные Z при условии наблюдаемых и текущих параметрах модели $p(Z|X, \Theta^{\text{old}})$, зануляя таким образом KL-дивиргенцию, то есть:

$$\log p(X|\Theta^{\text{old}}) = \mathcal{L}(q^*(Z), \Theta^{\text{old}})$$

На М-шаге мы оптимизируем левую часть равенства по параметрам Θ при фиксированом $q^*(Z)$, то есть $\mathcal{L}(q^*(Z), \Theta^{\text{new}}) > \mathcal{L}(q^*(Z), \Theta^{\text{old}})$ (при условии что Θ^{old} уже не точка оптимума), тогда справедлива следующая цепочка неравенств:

$$\log p(X|\Theta^{\text{new}}) = \mathcal{L}(q^*, \Theta^{\text{new}}) + \mathrm{KL}(q^*(Z)\|p(Z|X, \Theta^{\text{new}})) > \mathcal{L}(q^*, \Theta^{\text{old}}) = \log p(X|\Theta^{\text{old}})$$

Таким образом на каждой итерации EM-алгоритма мы увеличиваем значение неполного правдоподобия $p(X|\Theta)$.

■

2 Разделение смеси нормальных распределений

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

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

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

Введём скрытые переменные Z аналогично предыдущей задаче:

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

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

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

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

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

Введём обозначение:

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

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

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

Получаем следующую оптимизационную задачу:

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

$\pi_{\mathbf{k}}$: На параметры π_k есть ограничение $\sum_k \pi_k = 1$, поэтому воспользуемся методом множиетелей Лагранжа:

$$\mathcal{F}(\pi,\lambda)=\sum_{i=1}^{\ell}\sum_{k=1}^{K}g_{ik}\log\pi_k+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

$$\nabla_{\pi_k}\mathcal{F}=\sum_i g_{ik}\frac{1}{\pi_k}+\lambda\Rightarrow\pi_k=\frac{1}{\lambda}\sum_i g_{ik},\quad\lambda=l$$

$$\pi_k=\frac{1}{l}\sum_i g_{ik}$$

$\mu_{\mathbf{k}}$:

$$\mathcal{L}(q^*,\Theta)\propto_{\mu_k}\sum_{i=1}^{\ell}\sum_{k=1}^{K}g_{ik}\left[-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right]$$

$$\nabla_{\mu_k}\mathcal{L}=\sum_{i=1}^{\ell}g_{ik}\Sigma_k^{-1}(x_i-\mu_k)=\Sigma_k^{-1}\sum_{i=1}^{\ell}g_{ik}(x_i-\mu_k)=0\Rightarrow\sum_{i=1}^{\ell}g_{ik}(x_i-\mu_k)=\Sigma_k0=0$$

$$\mu_k=\frac{1}{l\pi_k}\sum_{i=1}^{\ell}g_{ik}x_i,\quad l\pi_k=\sum_{i=1}^{\ell}g_{ik}$$

$\boldsymbol{\Sigma}_{\mathbf{k}}$: Обозначим $\Lambda_k=\Sigma_k$, тогда:

$$\mathcal{L}(q^*,\Theta)\propto_{\Lambda_k}\sum_{i=1}^{\ell}\sum_{k=1}^{K}g_{ik}\left[-\frac{1}{2}(x_i-\mu_k)^T\Lambda_k(x_i-\mu_k)+\frac{1}{2}\log\det\Lambda_k\right]$$

$$\nabla_{\Lambda_k}\mathcal{L}=\sum_{i=1}^{\ell}g_{ik}\left[-\frac{1}{2}(x_i-\mu_k)(x_i-\mu_k)^T+\frac{1}{2}\Lambda^{-1}\right]=0$$

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

3 Разделение смеси распределений Бернулли

Рассмотрим смесь распределений Бернулли:

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

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

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

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.

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

На E-шаге вычисляется апостериорное распределение на скрытых переменных:

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

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

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

Введём обозначение:

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

Вычислим теперь матожидание полного правдоподобия:

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

Нам понадобится вспомогательная величина:

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

Получаем следующую оптимизационную задачу:

$$\mathbb{E}_{Z\sim p(Z\,|\,X,\Theta^{\text{old}})} \log p(X,Z\,|\,\Theta) = \sum_{i=1}^{\ell}\sum_{k=1}^{K} g_{ik}\Big\{\log\pi_k + \log p(x_i\,|\,\mu_k)\Big\} =$$
$$= \sum_{i=1}^{\ell}\sum_{k=1}^{K} g_{ik}\Big\{\log\pi_k + \sum_{j=1}^{d}(x_{ij}\log\mu_{kj} + (1-x_{ij})\log(1-\mu_{kj}))\Big\} \to \max_{\{\pi_k,\,\mu_k\}}$$

Дифференцируя данный функционал, можем получить формулы М-шага:

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

4 Кластеризация при помощи ЕМ-алгоритма

Несмотря на то, что выше были рассмотрены примеры применения ЕМ-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 ЕМ-алгоритм также в качестве модели *мягкой кластеризации*, предполагающей, что каждая компонента смеси является кластером со своим набором параметров (визуализация применения ЕМ-алгоритма для разделения смеси гауссиан).