Машинное обучение, ФКН ВШЭ
Семинар №17

§1.1 EM-алгоритм для PCA

Рассмотрим следующую вероятностную модель выделения главных компонент

$$p(x \mid t, \theta) = \mathcal{N}(x \mid Wt + \mu, \sigma^2 I), \quad p(t) = \mathcal{N}(t \mid 0, I), \tag{1.1}$$

где x – вектор признаков одного объекта, t – его представление в пространстве меньшей размерности, W и μ – параметры линейного преобразования из пространства меньшей размерности.

$$\begin{aligned}
&p(x \mid \theta) = \int p(x \mid t, \theta) p(t) dt \\
&= \int \frac{1}{(2\pi)^{d/2}(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{1}{2\sigma^2}(x - Wt - \mu)^T(x - Wt - \mu) - \frac{1}{2}t^T t\right) dt \\
&= \int \frac{1}{(2\pi)^{d/2}(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{1}{2\sigma^2}\left[t^T(W^TW + \sigma^2 I)t - 2(x-\mu)^T Wt + (x-\mu)^T(x-\mu)\right]\right) dt \\
&= \left\{ \Sigma_t^{-1} = \frac{1}{\sigma^2}(W^TW + \sigma^2 I), \quad \mu_t^T \Sigma_t^{-1} = \frac{1}{\sigma^2}(x-\mu)^T W \right\} = \\
&= \int \frac{1}{(2\pi)^{d/2}(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{1}{2}\left[t^T\Sigma_t^{-1}t - 2\mu_t^T\Sigma_t^{-1}t + \mu_t^T\Sigma_t^{-1}\mu_t\right] - \right. \\
&\left. -\frac{1}{2}\left[\frac{1}{\sigma^2}(x-\mu)^T(x-\mu) - \frac{1}{\sigma^2}(x-\mu)^T W(W^TW + \sigma^2 I)^{-1}W^T(x-\mu)\right]\right) dt =
\end{aligned}$$

используя тождество Вудбери

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}, \tag{1.2}$$

получаем

$$= \exp\left(-\frac{1}{2}(x-\mu)^T(WW^T + \sigma^2 I)^{-1}(x-\mu)\right) \int \frac{1}{(2\pi)^{d/2}(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{1}{2}(t-\mu_t)^T\Sigma_t^{-1}(t-\mu_t)\right) dt.$$

Таким образом

$$p(x \mid \theta) = \mathcal{N}(x \mid \mu, (WW^T + \sigma^2 I)).$$

Заметим, что вообще-то мы можем решать задачу максимизации неполного правдоподобия и без EM-алгоритма, решая следующую оптимизационную задачу:

$$p(X \mid \theta) = \prod_{i=1}^{\ell} p(x_i \mid \theta) = \mathcal{N}(x_i \mid \mu, (WW^T + \sigma^2 I)) \to \max_{W,\mu,\sigma} \tag{1.3}$$

Задача 1.1. *Задача для самостоятельного решения. Решите задачу (1.3) – найдите оптимальные значения W, μ, σ, выведите оценки сложности вычисления этих оценок. Предложите способ вычисления соответствующих t_i.*

Вместо прямого решения задачи (1.3) мы будем решать её с помощью EM-алгоритма. На **E-шаге** мы вычисляем распределение на скрытые переменные $T = \{t_i\}_{i=1}^{\ell}$:

$$q(T) = p(T \mid X, \theta) = \prod_{i=1}^{\ell} p(t_i \mid x_i, \theta). \tag{1.4}$$

На самом деле, распределение $p(t_i \mid x_i, \theta)$ мы уже нашли неявным образом, когда искали правдоподобие $p(x \mid \theta)$.

$$p(t_i \mid x_i, \theta) = \frac{p(x_i \mid t_i, \theta) p(t_i)}{p(x_i \mid \theta)} = \mathcal{N}(t_i \mid \mu_t^i, \Sigma_t^i), \tag{1.5}$$

$$\mu_t^i = (W^T W + \sigma^2 I)^{-1} W^T (x_i - \mu), \tag{1.6}$$

$$\Sigma_t^i = \sigma^2 (W^T W + \sigma^2 I)^{-1}. \tag{1.7}$$

На **M-шаге** мы находим новые оценки на переменные, решая оптимизационную задачу:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{T\sim q(T)} \log p(X, T \mid \theta) = \arg\max_{\theta} \mathbb{E}_{T\sim q(T)} \sum_{i=1}^{\ell} \log p(x_i \mid t_i, \theta) p(t_i) \tag{1.8}$$

$$= \arg\max_{\theta} \mathbb{E}_{T\sim q(T)} \sum_{i=1}^{\ell} \Big(\log p(x_i \mid t_i, \theta) + \log p(t_i) \Big) \tag{1.9}$$

$$= \arg\max_{\theta} \sum_{i=1}^{\ell} \Big(\int q(t_i) \log p(x_i \mid t_i, \theta) dt_i \Big) \tag{1.10}$$

Продифференцируем функционал по параметрам θ и запишем необходимое условие максимума:

$$\frac{\partial}{\partial \mu} \int q(t_i) \Big(-\frac{1}{2\sigma^2} (x_i - Wt_i - \mu)^T (x_i - Wt_i - \mu) - \frac{D}{2} \log(2\pi\sigma^2) \Big) dt_i =$$
$$\frac{\partial}{\partial \mu} \int q(t_i) \Big(-\frac{1}{2\sigma^2} \mu^T \mu + \frac{1}{\sigma^2} (x_i - Wt_i)^T \mu \Big) dt_i = \Big(-\frac{1}{\sigma^2} \mu + \frac{1}{\sigma^2} (x_i - W\mu_t^i) \Big)$$

$$\frac{\partial}{\partial \mu}\sum_{i=1}^{\ell}\left(\int q(t_i)\log p(x_i\,|\,t_i,\theta)dt_i\right)=0, \tag{1.11}$$

$$\sum_{i=1}^{\ell}(x_i - W\mu_t^i) = \ell\mu, \tag{1.12}$$

$$\mu = \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - W\mu_t^i). \tag{1.13}$$

$$\begin{aligned}
&\frac{\partial}{\partial W}\int q(t_i)\Big(-\frac{1}{2\sigma^2}(x_i - Wt_i - \mu)^T(x_i - Wt_i - \mu) - \frac{D}{2}\log(2\pi\sigma^2)\Big)dt_i = \\
&\frac{\partial}{\partial W}\int q(t_i)\Big(-\frac{1}{2\sigma^2}t_i^T W^T W t_i + \frac{1}{\sigma^2}(x_i-\mu)^T W t_i\Big)dt_i = \\
&\frac{-1}{2\sigma^2}\frac{\partial}{\partial W}\Big(\mathrm{tr}(\Sigma_t^i W^T W) + (\mu_t^i)^T W^T W \mu_t^i - 2(x_i-\mu)^T W\mu_t^i\Big) = \\
&\frac{-1}{2\sigma^2}\Big(2W\Sigma_t^i + 2W\mu_t^i(\mu_t^i)^T - 2(x_i-\mu)(\mu_t^i)^T\Big)
\end{aligned}$$

$$\frac{\partial}{\partial W}\sum_{i=1}^{\ell}\left(\int q(t_i)\log p(x_i\,|\,t_i,\theta)dt_i\right)=0, \tag{1.14}$$

$$2W\sum_{i=1}^{\ell}(\Sigma_t^i + \mu_t^i(\mu_t^i)^T) = 2\sum_{i=1}^{\ell}(x_i-\mu)(\mu_t^i)^T, \tag{1.15}$$

$$W = \left(\sum_{i=1}^{\ell}(x_i-\mu)(\mu_t^i)^T\right)\left(\sum_{i=1}^{\ell}(\Sigma_t^i + \mu_t^i(\mu_t^i)^T)\right)^{-1}. \tag{1.16}$$

$$\begin{aligned}
&\frac{\partial}{\partial \sigma}\int q(t_i)\Big(-\frac{1}{2\sigma^2}(x_i - Wt_i - \mu)^T(x_i - Wt_i - \mu) - \frac{D}{2}\log(2\pi\sigma^2)\Big)dt_i = \\
&-\frac{D}{\sigma} + \frac{1}{\sigma^3}\int q(t_i)\Big((x_i - Wt_i - \mu)^T(x_i - Wt_i - \mu)\Big)dt_i = \\
&-\frac{D}{\sigma} + \frac{1}{\sigma^3}\Big((x_i-\mu)^T(x_i-\mu) - 2(x_i-\mu)^T W\mu_t^i + \mathrm{tr}(\Sigma_t^i W^T W) + (\mu_t^i)^T W^T W\mu_t^i\Big)
\end{aligned}$$

$$\frac{\partial}{\partial \sigma}\sum_{i=1}^{\ell}\left(\int q(t_i)\log p(x_i\,|\,t_i,\theta)dt_i\right)=0, \tag{1.17}$$

$$\sigma^2 = \frac{1}{\ell D}\sum_{i=1}^{\ell}\Big((x_i-\mu)^T(x_i-\mu) - 2(x_i-\mu)^T W\mu_t^i + \mathrm{tr}(\Sigma_t^i W^T W) + (\mu_t^i)^T W^T W\mu_t^i\Big) \tag{1.18}$$

§1.2 EM PCA с пропусками

Полученный EM-алгоритм легко обобщается на случай, когда в данных есть пропуски. Обозначим через K_i множество индексов известных значений признаков для объекта x_i, а через U_i для неизвестных, соответственно. В таком случае вероятностная модель запишется следующим образом:

$$p((x_{i,K_i}, x_{i,U_i}) \,|\, t, \theta) = \mathcal{N}((x_{i,K_i}, x_{i,U_i}) \,|\, (W_{K_i} t_i + \mu_{K_i}, W_{U_i} t_i + \mu_{U_i}), \sigma^2 I), \tag{1.19}$$

$$p(t_i) = \mathcal{N}(t_i \,|\, 0, I) \tag{1.20}$$

Легко показать, что для E-шага формулы для скрытых переменных будут следующими:

$$q((x_{i,U_i}, t_i)) = \mathcal{N}((x_{i,U_i}, t_i) \,|\, m_i, S_i), \tag{1.21}$$

$$m_i = (W_{U_i} M W_{K_i}^T x_{i,K_i}, M W_{K_i}^T x_{i,K_i}) \tag{1.22}$$

$$S_i = \sigma^2 \begin{bmatrix} I + W_{U_i} M W_{U_i}^T & -W_{U_i} M \\ -M W_{U_i}^T & M \end{bmatrix}, \tag{1.23}$$

$$M = (W_{K_i}^T W_{K_i} + \sigma^2 I)^{-1} \tag{1.24}$$

Для M-шага:

$$W = \left(\sum_{i:j \in K_i} x_{ij} \mathbb{E} t_i^T + \sum_{i:j \in U_i} \mathbb{E} x_{ij} t_i^T \right) \left(\sum_i \mathbb{E} t_i t_i^T \right)^{-1}, \tag{1.25}$$

$$\sigma^2 = \frac{1}{\ell D} \sum_i \Bigg(x_{i,K_i}^T x_{i,K_i} + \mathrm{tr}(\mathbb{E} x_{i,U_i}^T x_{i,U_i}) - 2\mathbb{E} t_i^T W_{K_i}^T x_{i,K_i} - 2\mathrm{tr}(W_{U_i}^T \mathbb{E} x_{i,U_i} t_i^T), \tag{1.26}$$

$$+\mathrm{tr}(W_{K_i}^T W_{K_i} \mathbb{E} t_i t_i^T) + \mathrm{tr}(U_{K_i}^T W_{U_i} \mathbb{E} t_i t_i^T) \Bigg) \tag{1.27}$$