Задание 4 по курсу «Байесовский выбор моделей»

Общая информация

- Время сдачи задания: 10е декабря, 21:00 по Москве;
- Максимальная базовая оценка за задание 100 баллов, так что при желании можно выполнять не всё;
- Оценка автора наилучшей работы удваивается (с учетом баллов сверх 100), но не более, чем до 250 баллов;
- Вопросы и само задание принимаются по почте: aduenko1@gmail.com & iakovlev.kd@phystech.edu (отправлять на обе сразу);
- Тема письма: вопрос по заданию #4 или решение задания #4;
- Опоздание на неделю снижает оценку в 2 раза, опоздание на час на $0.5^{1/(7\cdot 24)} = 0.41\%$;
- Работы опоздавших не участвуют в конкурсе на лучшую работу;
- Задание не принимается после его разбора и / или после объявления об этом.

Задача (байесовский метод главных компонент). Рассмотрим вероятностную модель метода главных компонент, считая, что для каждого объекта $\mathbf{x}_i \in \mathbb{R}^n$ существует описание $\mathbf{z}_i \in \mathbb{R}^d$ в признаковом пространстве меньшей размерности, причем $\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i$, где $\boldsymbol{\mu} \in \mathbb{R}^n$ – есть некоторое смещение (на случай нецентрированности признаков), а $\boldsymbol{\varepsilon}_i \in \mathbb{R}^n$ – шумовой вектор.
Пусть имеется выборка $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ независимых объектов. Пусть

$$p(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|\mathbf{0}, \mathbf{I}), p(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{\varepsilon}_i|\mathbf{0}, \sigma^2\mathbf{I}).$$

Считая $\mathbf{W}$, σ^2, $\boldsymbol{\mu}$ – неизвестными параметрами задачи, а d фиксированным
а) выписать $p(\mathbf{X}, \mathbf{Z}|\mathbf{W}, \boldsymbol{\mu}, \sigma)$ (3 балла);
б) найти $p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma)$ (3 балла);
в) с помощью EM-алгоритма решить задачу нахождения наиболее правдоподобных оценок $\mathbf{W}$, $\boldsymbol{\mu}$, σ, то есть решить задачу

$$p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma) \to \max_{\mathbf{W}, \boldsymbol{\mu}, \sigma},$$

получив итеративные формулы пересчета для E и M шагов (25 баллов). Каково апостериорное распределение $p(\mathbf{z}_i|\mathbf{x}_i, \mathbf{W}, \sigma, \boldsymbol{\mu})$? (10 баллов) Как изменить вероятностную модель, чтобы учесть, что в данных есть пропуски? (10 баллов)
г) сгенерировать признаковую матрицу $\mathbf{X} \in \mathbb{R}^{m\times n}$, $m = 1000$, $n = 10$ для $d = 2$ путем генерации $\mathbf{Z} \in \mathbb{R}^{m\times d}$ поэлементно независимо из $\mathcal{N}(0, 1)$ и выполнения преобразования $\mathbf{X} = \mathbf{Z}\mathbf{W}^\mathsf{T} + \boldsymbol{\varepsilon}$ для $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0}, \mathbf{I})$, где $\mathbf{W} \in \mathbb{R}^{n\times d}$ матрица преобразования, выбранная Вами. Сравнить результат работы алгоритма из п. в) с обычным методом главных компонент для $d = 2$ (10 баллов);
д) (автоматическое определение числа компонент) Считаем, что $d = n$. Введем априорное распределение на $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ вида

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{j=1}^{n} \left(\sqrt{\frac{\alpha_j}{2\pi}}\right)^n \exp\left(-\frac{\alpha_j}{2}\mathbf{w}_j^\mathsf{T}\mathbf{w}_j\right),$$

где $\mathbf{w}_j$ – столбцы матрицы $\mathbf{W}$.
Если $\alpha_j \to \infty$, то $\mathbf{w}_j^\mathsf{T}\mathbf{w}_j \to 0$, то есть происходит исключение соответствующей компоненты из разложения $\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, что соответствует сокращению числа главных компонент.
С помощью вариационного EM-алгоритма решить задачу (50 баллов)

$$p(\mathbf{X}|\boldsymbol{\mu},\, \sigma,\, \boldsymbol{\alpha}) \to \max_{\boldsymbol{\alpha},\, \boldsymbol{\mu},\, \sigma}.$$

Подсказка: в качестве скрытых переменных рассмотреть $(\mathbf{Z},\, \mathbf{W})$ и на E-шаге использовать вариационное приближение $p(\mathbf{Z},\, \mathbf{W}|\mathbf{X},\, \boldsymbol{\mu},\, \sigma,\, \boldsymbol{\alpha}) \approx q(\mathbf{Z})q(\mathbf{W})$.
е) Для матрицы признаков из п. г) воспользоваться результатом пункта д) и проверить, происходит ли исключение восьми лишних главных компонент в ходе максимизации обоснованности (20 баллов).

$x_i \in \mathbb{R}^n \qquad z_i \in \mathbb{R}^d \qquad p(z_i) = \mathcal{N}(z_i | 0, I)$

$p(\varepsilon_i) = \mathcal{N}(\varepsilon_i | 0, \sigma^2 I) \in \mathbb{R}^n$

$x_i = \underline{W} z_i + \underline{\mu} + \varepsilon_i$

а) $p(X, Z | W, \mu, \sigma^2) = \prod_{i=1}^{m} \mathcal{N}(z_i | 0, I)\, \mathcal{N}(x_i - Wz_i - \mu | 0, \sigma^2 I) =$

$= \prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2} z_i^T z_i} \frac{1}{(2\pi)^{\frac{n}{2}} \sigma} \exp\left[-\frac{1}{2\sigma^2} (x_i - Wz_i - \mu)^T (x_i - Wz_i - \mu)\right] =$

$= \prod_{i=1}^{m} \mathcal{N}(z_i | 0, I) \cdot \mathcal{N}(\varepsilon_i | 0, \sigma^2 I)$

б)

$p(x_i | W, \mu, \sigma) = \mu + \mathcal{N}(0, \sigma^2 I) + \mathcal{N}(0, WW^T) \sim \mathcal{N}(\mu, WW^T + \sigma^2 I)$

$p(X | W, \mu, \sigma) = \prod^{m} \mathcal{N}(x_i | \mu, WW^T + \sigma^2 I)$

в) EM:

E шаг

$\theta = (W, \mu, \sigma)$

$Z = Z$

$q(z) = p(z | \mathcal{D}, \theta) = p(z | X, W, \mu, \sigma) \propto p(z, x | W, \mu, \sigma) =$

$\prod_{i=1}^{m} \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2} z_i^T z_i} \frac{1}{(2\pi)^{\frac{n}{2}} \sigma} \exp\left[-\frac{1}{2\sigma^2} (x_i - Wz_i - \mu)^T (x_i - Wz_i - \mu)\right] =$

рассм

$z_i^T z_i + \frac{1}{\sigma^2} (x_i - Wz_i - \mu)^T (x_i - Wz_i - \mu)$

$x_i - \mu = y_i$

$$z^T z + \frac{1}{\sigma^2}(y - Wz)^T(y - Wz) = z^2 + \frac{1}{\sigma^2}(y^2 - 2y^T W z + z^T W^T W z) =$$

$$= z^T\left(I + \frac{1}{\sigma^2} W^T W\right) z - \frac{2}{\sigma^2} y^T W z$$

$$(x-\mu)^T B (x-\mu) = (x^T B - \mu^T B)(x-\mu) = x^T B x - 2\mu^T B x + \mu^T B \mu$$

$$B = \left(I + \frac{1}{\sigma^2} W^T W\right) \qquad \mu^T B = \frac{1}{\sigma^2} y^T W$$

$$B^T \mu = \frac{1}{\sigma^2} W^T y$$

$$\mu_0 = \frac{1}{\sigma^2} B^{-1} W^T y$$

$$= \prod^m (\sim) \exp\left(-\frac{1}{2}(z_i - \mu_0) B (z_i - \mu_0)\right)$$

$$\Rightarrow \boxed{\begin{gathered} p(z \mid X, \theta) = \prod^m \mathcal{N}\left(z_i \,\middle|\, \frac{1}{\sigma^2} B^{-1} W^T (x_i - \mu),\ B^{-1}\right) \\ B = \left(I + \frac{1}{\sigma^2} W^T W\right) \end{gathered}}$$

обозначим $\mu_0^i = \frac{1}{\sigma^2} B^{-1} W^T (x_i - \mu) \qquad \Sigma_0 = B^{-1}$

$$\Rightarrow p(z \mid X, \theta) = \prod \mathcal{N}(z_i \mid \mu_0^i, B^{-1})$$

Тогда

E шаг:

$$q^s(z) = p(z \mid X, \theta^{s-1})$$

M шаг:

$$\theta^s = F(q^s, \theta) \propto \int q(z) \log p(D, z|\theta)\, dz \to \max_\theta$$

$$\log p(X, Z|\theta) \propto -m\log\sigma - \frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\sum_i^m (x_i - Wz_i - \mu)^T (x_i - Wz_i - \mu)$$

$\Rightarrow$ (Выкидываем $\sum z^2$, т.к. не зав от параметров)

$$F(q^s, \theta) \propto -m\log\sigma - \frac{1}{2\sigma^2}\sum \Big[(x_i - \mu)^2 - 2(x_i-\mu)^T W \overbrace{\mathbb{E}\, z_i}^{\mu_0^i} + \mathbb{E}_z\, z_i^T W^T W z_i\Big] \ominus$$

$$\boxed{\begin{array}{l}\text{Лемма если } A = A^T \quad y \sim \mathcal{N}(\mu, \Sigma) \\ \text{то } \mathbb{E}\, y^T A y = Tr(A\Sigma) + \mu^T A \mu\end{array}}$$

$$\ominus -m\log\sigma - \frac{1}{2\sigma^2} S$$

$$S = \sum (x_i - \mu)^2 - 2(x_i - \mu)^T W \mu_0^i + Tr(W^T W \Sigma_0) + \mu_0^{i\,T} W^T W \mu_0^i$$

найдем оптимальные параметры

- $\dfrac{\partial F}{\partial \sigma} = \dfrac{-m}{\sigma} + \dfrac{1}{\sigma^3} S = 0 \Rightarrow \sigma = \left[\dfrac{S}{m}\right]^{\frac{1}{2}}$

- $\dfrac{\partial F}{\partial \mu} \propto \sum -(x_i - \mu) + W\mu_0^i = -\sum x_i + W(\sum \mu_0^i) + m\mu = 0$

$$\Rightarrow \mu = \frac{1}{m}\Big[\sum x_i - W(\sum \mu_0^i)\Big]$$

$$\underset{dW}{dF} \propto \sum -2(x_i - \mu)^T dW \mu_0^i + Tr(dW^T W \Sigma_0) + Tr(W^T dW \Sigma_0) + 2\mu_0^{i\,T} dW W W \mu_0^i\Big] = 0$$

$$dF \propto \sum -2\,Tr\left(dW \mu_0^i (x_i-\mu)^T\right) + 2\,Tr\left(dW^T W \Sigma_0\right) + 2\,Tr\left(dW\left(W \mu_0^i \mu_0^{i^T}\right)\right) =$$

$$= 2\,Tr\left(dW \cdot \left[\sum_i^m -(x_i-\mu)\mu_0^{i^T} + W\Sigma_0 + W \mu_0^i \mu_0^{i^T}\right]\right) = 0$$

$$\Rightarrow \quad W\left(\Sigma_0 + \sum_i^m \mu_0^i \mu_0^{i^T}\right) = \sum_i^m (x_i-\mu)\mu_0^{i^T}$$

$$W = \left[\sum_i^m (x_i-\mu)\mu_0^{i^T}\right]\left[\Sigma_0 + \sum \mu_0^i \mu_0^{i^T}\right]^{-1}$$

Итого EM алгоритм во всей красе:

E:

$$q^s(z) = p(z \mid X, \theta^{s-1}) = \prod \mathcal{N}(z_i \mid \mu_{0s-1}^i, \Sigma_{0s-1})$$

$$\mu_{0s}^i = \frac{1}{\sigma_s^2} B^{-1} W_s^T (x_i - \mu_s) \qquad \Sigma_{0s} = B^{-1} = \left(I + \frac{1}{\sigma^2} W_s^T W_s\right)^{-1}$$

M:

$$T_s = \sum (x_i-\mu_s)^2 - 2(x_i-\mu_s)^T W \mu_{0s}^i + Tr\left(W_s^T W_s \Sigma_{0s}\right) + \mu_0^{i^T} W_s^T W_s \mu_{0s}^i$$

$$\sigma_{s+1} = \left[\frac{T_s}{m}\right]^{\frac{1}{2}}$$

$$\mu_{s+1} = \frac{1}{m}\left[\sum_i^m x_i - W_s\left(\sum_i^m \mu_{0s}^i\right)\right]$$

$$W_{s+1} = \left[\sum_i^m (x_i-\mu_s)\mu_{0s}^{i^T}\right]\left[\Sigma_{0s} + \sum \mu_{0s}^i \mu_{0s}^{i^T}\right]^{-1}$$

апостериорное распр

$$p(z_i \mid x_i, W, \sigma, \mu) = \mathcal{N}\left(z_i \,\Big|\, \frac{1}{\sigma^2} B^{-1} W^T (x_i - \mu),\ B^{-1}\right)$$

$$B^{-1} = \left(I + \frac{1}{\sigma^2} W_s^T W_s\right)^{-1}$$

т.е. нормальное

д) $d = n$

априорное распр на W

$$p(W|\alpha) = \prod^{n} \left(\sqrt{\frac{\alpha_i}{2\pi}}\right)^{n} \exp\left(-\frac{\alpha_i}{2} W_i^T W_i\right)$$

если $\alpha_i \to \infty$, то W_i исключ-ся

$$p(X | \mu, \sigma, \alpha) \xrightarrow[\alpha, \mu, \sigma]{} \max$$

скрытые переменные $\{W, Z\}$

$$\Theta = \{\alpha, \mu, \sigma\}$$

вероятностная модель:

$$p(Z, W, X | \mu, \sigma, \alpha) = p(Z, X | W, \mu, \sigma) \cdot p(W|\alpha)$$

Var EM

$$p(Z, W | X, \mu, \sigma, \alpha) \approx q(Z)\, q(W)$$

$$\log q(Z) = \mathbb{E}_W \log p(Z, W, X | \mu, \sigma, \alpha) =$$

$$= \mathbb{E}_W \log(Z, X | W, \mu, \sigma) + \mathbb{E}_W \log p(W|\alpha) \; \propto$$

$$\log p(X, Z|\theta) \propto -m \log \sigma - \frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\sum_i^m (x_i - W z_i - \mu)^T (x_i - W z_i - \mu)$$

$$\propto -\frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\underbrace{\sum_i^m (x_i-\mu)^2 - 2(x_i-\mu)^T \underbrace{E W}_{\mathbb{E}_W} z_i + z_i^T \mathbb{E}_W W^T W z_i}_{} =$$

$$= -\frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\sum_i^m \left[(x_i-\mu)^2 - 2(x_i-\mu)^T \mathbb{E}_W W z_i + z_i^T \mathbb{E}_W W^T W z_i\right] =$$

$$-\frac{1}{2}\sum_i^m \left[z_i^T \left(I + \frac{1}{\sigma^2}\mathbb{E}_W W^T W\right) z_i - \frac{2}{\sigma^2}\left(\mathbb{E}_W W (x_i-\mu)\right)^T z_i + (x_i-\mu)^2\right]$$

$$B = I + \frac{1}{\sigma^2}\mathbb{E}_W W^T W$$

$$\Rightarrow \quad q(z_i) \sim \mathcal{N}\left(z_i \,\middle|\, \frac{1}{\sigma^2} B^{-1} \mathbb{E}_W W (x_i - \mu),\ B^{-1}\right)$$

$$\log q(W) = \mathbb{E}_Z \log(Z, X | W, \mu, \sigma) + \mathbb{E}_Z \log p(W|\alpha) \propto$$

$$-\frac{1}{2\sigma^2}\sum^m \Big[-2\underbrace{(x_i-\mu)^T W \mathbb{E}_Z z_i}_{Tr(\mathbb{E}_Z z_i (x_i-\mu)^T W)} + \underbrace{\mathbb{E}_Z z_i^T W^T W z_i}_{Tr(W^T W \mathbb{E}_Z z_i z_i^T)}\Big] + \underbrace{\sum^n -\frac{1}{2}\alpha_i w_i^T w_i}_{-\frac{1}{2}Tr(W^T W \operatorname{diag}\alpha)} =$$

$$-\frac{1}{2} Tr\left[-2\left[\frac{1}{\sigma^2}\sum^m \mathbb{E}_Z z_i (x_i-\mu)^T\right] W + W^T W \left[\frac{1}{\sigma^2}\sum^m \mathbb{E}_Z z_i z_i^T + \operatorname{diag}\alpha\right]\right] =$$

$$B = \left[\frac{1}{\sigma^2}\sum^m \mathbb{E}_Z z_i z_i^T + \operatorname{diag}\alpha\right]$$

$$\mu_0^j = \left(\frac{1}{\sigma^2} B^{-1}\left[\sum^m \mathbb{E}_Z z_i (x_i-\mu)^T\right]\right)_j \leftarrow j \text{ строка}$$

$$q(w_j) = \mathcal{N}\left(w_j \,\middle|\, \frac{1}{\sigma^2} B^{-1}\left[\sum^m \mathbb{E}_Z z_i (x_i-\mu)\right]_j,\ B^{-1}\right)$$

это был E шаг

M шаг

$$\theta^s = F(q^s, \theta) \propto \int q(z) \log p(D, z|\theta)\, dz \to \max_\theta$$

$$\log p(X, Z, W \mid \mu, \sigma, \alpha) \propto$$

$$-m\log\sigma - \frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\sum_i^m (x_i - Wz_i - \mu)^T (x_i - Wz_i - \mu) +$$

$$+ \frac{n}{2}\sum^n \alpha_i - \frac{1}{2}\sum_i^n \alpha_i w_i^T w_i =$$

$$-m\log\sigma - \frac{1}{2}\sum_i^m z_i^2 - \frac{1}{2\sigma^2}\sum_i^m \left[(x_i - \mu)^2 - 2(x_i - \mu)^T W z_i + z_i^T W^T W z_i\right]$$

$$+ \frac{1}{2}\sum_i^n \alpha_i - \frac{1}{2}\sum^n \alpha_i w_i^T w_i$$

$$Tr(W^T W z_i z_i^T)$$

$\Rightarrow$

$$\mathbb{E}_{p,w} \log(X, Z, W \mid \sigma \mu, \alpha) \propto$$

$$-m\log\sigma - \frac{1}{2\sigma^2}\sum_i^m \left[(x_i - \mu)^2 - 2(x_i - \mu)^T \mathbb{E} W z_i + Tr(\mathbb{E}_w W^T W \cdot \mathbb{E}_z z_i z_i^T)\right]$$

$$+ \frac{1}{2}\sum_i^n \log\alpha_i - \frac{1}{2}\sum_i^n \alpha_i \mathbb{E}_w w_i^T w_i$$

производные:

- $$\frac{\partial}{\partial\sigma} F = -\frac{m}{\sigma} + \frac{1}{\sigma^3}\sum^m \left[(x_i - \mu)^2 + 2(x_i - \mu^T \mathbb{E} W z_i + Tr(\mathbb{E}_w W^T W \cdot \mathbb{E}_z z_i z_i^T)\right] = 0$$

$$\sigma = \left[\frac{1}{m}\sum^m \left[(x_i - \mu)^2 + 2(x_i - \mu^T \mathbb{E} W z_i + Tr(\mathbb{E}_w W^T W \cdot \mathbb{E}_z z_i z_i^T)\right]\right]^{\frac{1}{2}}$$

- $$\frac{\partial}{\partial\alpha_i} F = \frac{1}{2\alpha_i} - \frac{1}{2}\mathbb{E}_w w_i^T w_i^T = 0 \quad \Rightarrow \alpha_i = \frac{1}{\mathbb{E}_w w_i^T w_i}$$

$$\cdot \frac{\partial}{\partial \mu} \propto \sum_i^m -(x_i - \mu) + \mathbb{E} W z_i = 0$$

$$\mu = \frac{1}{m}\left[\sum^m x_i - \mathbb{E} W \mathbb{E} z_i\right]$$

E шаг: $B_s = I + \frac{1}{\sigma_s^2} \mathbb{E}_{w_s} W^T W$

$$q^{s+1}(z_i) \sim \mathcal{N}\left(z_i \,\middle|\, \frac{1}{\sigma_s^2} B_s^{-1} \mathbb{E}_w W (x_i - \mu_s),\ B_s^{-1}\right)$$

$$P = \left[\frac{1}{\sigma^2} \sum^m \mathbb{E}_z z_i z_i^T + \operatorname{diag} \alpha\right]$$

$$\mu_0^j = \left(\frac{1}{\sigma^2} P^{-1} \left[\sum^m \mathbb{E}_z z_i (x_i - \mu)^T\right]\right)_j$$ ← j строка

$$q(w_j) = \mathcal{N}\left(w_i \,\middle|\, \mu_0^{j\,T},\ P^{-1}\right)$$

M шаг:

$$\sigma^{s+1} = \left[\frac{1}{m} \sum^m \left[(x_i - \mu_s)^2 + 2 (x_i - \mu)^T \mathbb{E}_s W z_i + \operatorname{Tr}\left(\mathbb{E}_{w_s} W^T W \cdot \mathbb{E}_{z_s} z_i z_i^T\right)\right]\right]^{\frac{1}{2}}$$

$$\alpha_i = \frac{1}{\mathbb{E}_{w_s} w_i^T w_i}$$

$$\mu = \frac{1}{m}\left[\sum^m x_i - \mathbb{E}_{w_s} W \mathbb{E}_{z_s} z_i\right]$$