Байесовский выбор моделей: Вариационный EM-алгоритм (воспоминание)

Александр Адуенко

27е февраля 2024

- ЕМ-алгоритм. Использование ЕМ-алгоритма для отбора признаков в байесовской линейной регрессии.

EM-алгоритм: воспоминание

Пусть $\mathbf{D} = (\mathbf{X}, \mathbf{y})$ – наблюдаемые переменные, $\mathbf{Z}$ – скрытые переменные.
$p(\mathbf{D}, \mathbf{Z}|\boldsymbol{\Theta}) = p(\mathbf{D}|\mathbf{Z}, \boldsymbol{\Theta})p(\mathbf{Z}|\boldsymbol{\Theta})$.

Вопрос 1: как решить задачу $p(\mathbf{D}|\boldsymbol{\Theta}) = \int p(\mathbf{D}, \mathbf{Z}|\boldsymbol{\Theta})d\mathbf{Z} \to \max\limits_{\boldsymbol{\Theta}}$?

EM-алгоритм

Введем $F(q, \boldsymbol{\Theta}) = -\int q(\mathbf{Z}) \log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{D}, \mathbf{Z}|\boldsymbol{\Theta})d\mathbf{Z} =$

$$-\int q(\mathbf{Z}) \log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{D}, \boldsymbol{\Theta})d\mathbf{Z} + \int \log p(\mathbf{D}|\boldsymbol{\Theta})q(\mathbf{Z})d\mathbf{Z} =$$

$$\log p(\mathbf{D}|\boldsymbol{\Theta}) - \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{D},\boldsymbol{\Theta})}d\mathbf{Z} = \log p(\mathbf{D}|\boldsymbol{\Theta}) - D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D}, \boldsymbol{\Theta})).$$

Идея 1: $p(\mathbf{D}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$ заменим на $F(q, \boldsymbol{\Theta}) \to \max\limits_{q, \boldsymbol{\Theta}}$.

Идея 2: Пошагово оптимизируем по $\boldsymbol{\Theta}$ и q, то есть

1. Е-шаг: $q^s = F(q, \boldsymbol{\Theta}^{s-1}) \to \max\limits_{q \in Q}$;
2. М-шаг: $\boldsymbol{\Theta}^s = F(q^s, \boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

Вопрос: Зачем $q \in Q$? Как Е-шаг был выполнен при максимизации обоснованности для модели линейной регрессии?

Вариационный ЕМ-алгоритм. Е-шаг

$$F(q,\ \boldsymbol{\Theta}^{s-1}) \to \max_{q\in Q} \iff D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D},\ \boldsymbol{\Theta})) \to \min_{q\in Q}.$$

$$D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D},\ \boldsymbol{\Theta})) = \log p(\mathbf{D}|\boldsymbol{\Theta}) + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta})} d\mathbf{Z}.$$

Пусть $Q = \left\{ q:\ q(\mathbf{Z}) = \prod_{k=1}^{K} q(\mathbf{Z}_k) \right\}$, тогда

$$D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D},\ \boldsymbol{\Theta})) \propto \int \prod_{k=1}^{K} q(\mathbf{Z}_k) \log \frac{\prod_{j=1}^{K} q(\mathbf{Z}_j)}{p(\mathbf{D},\ \mathbf{Z}_1,\ \ldots,\ \mathbf{Z}_K|\boldsymbol{\Theta})} d\mathbf{Z}_1 \ldots d\mathbf{Z}_K =$$

$$\int q(\mathbf{Z}_k) \log q(\mathbf{Z}_k) \underbrace{\left[\prod_{j\neq k} \int q(\mathbf{Z}_j) d\mathbf{Z}_j \right]}_{=1} d\mathbf{Z}_k - \sum_{j\neq k} C_j \underbrace{\int q(\mathbf{Z}_k) d\mathbf{Z}_k}_{=1} -$$

$$\int q(\mathbf{Z}_k) \underbrace{\left[\int \prod_{j\neq k} q(\mathbf{Z}_j) \log p(\mathbf{D},\ \mathbf{Z}_1,\ \ldots,\ \mathbf{Z}_K|\boldsymbol{\Theta}) d\mathbf{Z}_{j\neq k} \right]}_{\mathsf{E}_{q\setminus k} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})} d\mathbf{Z}_k \propto$$

$$\int q(\mathbf{Z}_k) \log \frac{q(\mathbf{Z}_k)}{\frac{1}{C} e^{\mathsf{E}_{q\setminus k} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})}} d\mathbf{Z}_k \to \min_{q(\mathbf{Z}_k)}.$$

Вариационный ЕМ-алгоритм

$$F(q,\ \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} = \log p(\mathbf{D}|\boldsymbol{\Theta}) - D_{\mathrm{KL}}(q \| p(\mathbf{Z}|\mathbf{D},\ \boldsymbol{\Theta})).$$

Е-шаг. $\displaystyle\int q(\mathbf{Z}_k) \log \frac{q(\mathbf{Z}_k)}{\frac{1}{C} e^{\mathsf{E}_{q\setminus k} \log p(\mathbf{D},\mathbf{Z}|\boldsymbol{\Theta})}} d\mathbf{Z}_k \to \min_{q(\mathbf{Z}_k)}.$

Полный алгоритм

Пошагово оптимизируем по $\boldsymbol{\Theta}$ и $q(\mathbf{Z}_k),\ k = 1,\ \dots,\ K$, то есть

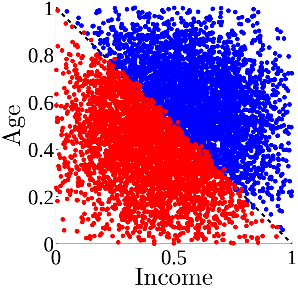
1. Е-шаг: $\log q(\mathbf{Z}_k^s) \propto \mathsf{E}_{q\setminus k} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}^{s-1})$;
2. М-шаг: $\mathsf{E}_{q^s} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

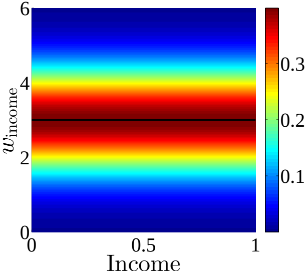
Вопрос 1: зачем нужна факторизация? Чем полученные итеративные формулы лучше формул полного ЕМ-алгоритма?

Вопрос 2: как понять, что в конкретной задаче формулы Е и М-шагов выписаны верно?

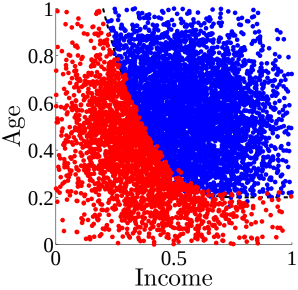
Нарушение свойства $p(\mathbf{w}|\mathbf{x}_i) = p(\mathbf{w})$

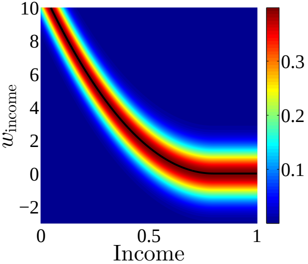
Предполагаемый результат





Реальные данные





Вопрос: как можно учесть указанную нелинейность в модели?

Смесь моделей линейной регрессии

Вероятностная модель генерации данных

- Веса моделей в смеси $\boldsymbol{\pi}$ получены из априорного распределения $p(\boldsymbol{\pi}|\mu)$;
- Векторы параметров моделей $\mathbf{w}_k$ получены из нормального распределения $p(\mathbf{w}_k|\mathbf{A}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}),\ k = 1, \dots, K$;
- Для каждого объекта $\mathbf{x}_i$ выбрана модель f_{k_i}, которой он описывается, причем $p(k_i = k) = \pi_k$;
- Для каждого объекта $\mathbf{x}_i$ целевая переменная y_i определена в соответствии с моделью f_{k_i}: $y_i \sim \mathcal{N}(y_i|\mathbf{w}_{k_i}^{\mathsf{T}}\mathbf{x}_i,\ \sigma_{k_i}^2)$.

Совместное правдоподобие модели

$$p(\mathbf{y},\ \mathbf{w}_1,\ \dots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \dots,\ \mathbf{A}_K,\ \sigma_1^2,\ \dots,\ \sigma_K^2,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu})\prod_{k=1}^{K}\mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})\prod_{i=1}^{m}\left(\sum_{l=1}^{K}\pi_l\mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i,\ \sigma_l^2)\right).$$

Введем матрицу скрытых переменных $\mathbf{Z} = \|z_{ik}\|$, где $z_{ik} = 1 \iff k_i = k$.

$$p(\mathbf{y},\ \mathbf{w}_1,\ \dots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \dots,\ \mathbf{A}_K,\ \sigma_1^2,\ \dots,\ \sigma_K^2,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu})\prod_{k=1}^{K}\mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})\prod_{i=1}^{m}\prod_{l=1}^{K}\left(\pi_l\mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i,\ \sigma_l^2)\right)^{z_{il}}.$$

Получение MAP-оценки

Пусть $p(\boldsymbol{\pi}|\boldsymbol{\mu}) = \mathrm{Dir}(\boldsymbol{\mu}) = \frac{\Gamma(\sum_k \mu_k)}{\prod_l \Gamma(\mu_l)} \prod_k \pi_k^{\mu_k - 1}$.

$$p(\mathbf{y},\, \mathbf{W},\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\mu}) \propto$$

$$\prod_{k=1}^{K} \pi_k^{\mu_k - 1} \prod_{k=1}^{K} \sqrt{\det \mathbf{A}_k} \exp(-\tfrac{1}{2}\mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k) \prod_{i=1}^{m} \prod_{l=1}^{K} \left(\pi_l \mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i,\, \sigma_l^2)\right)^{z_{il}}.$$

$$(\boldsymbol{\pi}^*,\, \mathbf{w}_1^*,\, \ldots,\, \mathbf{w}_K^*) = \arg\max_{\boldsymbol{\pi},\, \mathbf{W}} p(\mathbf{y},\, \mathbf{W},\, \boldsymbol{\pi}|\mathbf{X},\, \mathbf{A},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\mu}).$$

E-шаг. $\log q(\mathbf{Z}) \propto \sum_{i=1}^{m} \sum_{l=1}^{K} z_{il} (\log \pi_l + \log \mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i,\, \sigma_l^2))$, откуда

$$\gamma_{il} = p(z_{il} = 1) \propto \pi_l \mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i,\, \sigma_l^2).$$

M-шаг. $\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{W},\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\mu}) \to \max\limits_{\boldsymbol{\pi},\, \mathbf{W}}$.

$$\mathbf{w}_k^* = \left(\sigma_k^2 \mathbf{A}_k + \sum_{i=1}^{m} \gamma_{ik} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}\right)^{-1} \sum_{i=1}^{m} \gamma_{ik} y_i \mathbf{x}_i.$$

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \sum_{k=1}^{K} \log \pi_k \Big(\underbrace{\sum_{i=1}^{m} \gamma_{ik}}_{\gamma_k} + \mu_k - 1\Big) \Longrightarrow \pi_k \propto \max(0,\, \gamma_k + \mu_k - 1).$$

Получение апостериорного распределения

Вопрос: как получить $p(\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}|\mathbf{X}, \mathbf{y}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\mu})$?

$$p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}|\mathbf{X}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\mu}) \propto$$
$$\prod_{k=1}^{K} \pi_k^{\mu_k - 1} \prod_{k=1}^{K} \sqrt{\det \mathbf{A}_k} \exp(-\tfrac{1}{2}\mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l \mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i, \sigma_l^2) \right).$$

Идея: найдем $q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\pi}) = q(\mathbf{Z})q(\mathbf{W})q(\boldsymbol{\pi})$, наиболее близкое к $p(\mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y})$.

$$\log q(\mathbf{Z}) \propto \mathsf{E}_{q \backslash \mathbf{Z}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\mu}) \propto$$
$$\sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \left(\mathsf{E} \log \pi_k + \mathsf{E} \log \mathcal{N}(y_i|\mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i, \sigma_k^2)) \right)$$
$$\Longrightarrow p(z_{ik} = 1) \propto$$
$$\exp \left(\mathsf{E} \log \pi_k - \frac{1}{2\sigma_k^2}(y_i^2 - 2\mathsf{E}\mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_k \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) - \frac{1}{2} \log \sigma_k^2 \right).$$

$$\log q(\boldsymbol{\pi}) \propto \mathsf{E}_{q \backslash \boldsymbol{\pi}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\mu}) \propto$$
$$\sum_{k=1}^{K} \log \pi_k \left(\mu_k - 1 + \sum_{i=1}^{m} \mathsf{E} z_{ik} \right) \Longrightarrow \boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\,\boldsymbol{\mu} + \boldsymbol{\gamma}).$$

Получение апостериорного распределения (продолжение)

$$\log q(\mathbf{w}_1, \dots, \mathbf{w}_K) \propto \mathsf{E}_{q_{\setminus \mathbf{w}}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}, \boldsymbol{\sigma}^2, \boldsymbol{\mu}) \propto$$
$$\sum_{k=1}^{K} \left(-\frac{1}{2}\mathbf{w}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{w}_k + \frac{1}{2\sigma_k^2}\sum_{i=1}^{m} \mathsf{E} z_{ik}(y_i^2 - 2\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i y_i + \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_k) \right).$$

Вопрос 1: Какую структуру имеет $q(\mathbf{w}_1, \dots, \mathbf{w}_K)$?

Вопрос 2: Какой вид имеет распределение $q(\mathbf{w}_k)$?

$$q(\mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_k q_k(\mathbf{w}_k),\; q_k(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{m}_k, \boldsymbol{\Sigma}_k^{-1}),$$ где

$$\boldsymbol{\Sigma}_k = \mathbf{A}_k + \frac{1}{\sigma_k^2}\sum_{i=1}^{m}\gamma_{ik}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}},$$

$$\mathbf{m}_k = \left(\sigma_k^2\mathbf{A}_k + \sum_{i=1}^{m}\gamma_{ik}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)^{-1}\sum_{i=1}^{m}\gamma_{ik}y_i\mathbf{x}_i.$$

Вопрос 3: Как распределение $q_k(\mathbf{w}_k)$ соотносится с MAP-оценкой для $\mathbf{w}_k$, полученной ранее?

Вопрос 4: Как определить $\mathbf{A}$, $\boldsymbol{\sigma}^2$?

Максимизация обоснованности

$$p(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\sigma}^2,\ \boldsymbol{\mu}) \to \max_{\mathbf{A},\ \boldsymbol{\sigma}^2,\ \{\boldsymbol{\mu}\}}$$

Идея: воспользуемся вариационным ЕМ-алгоритмом.

Е-шаг: найдем $q(\mathbf{Z},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}) = q(\mathbf{Z})q(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K)q(\boldsymbol{\pi})$,
наиболее близкое к $p(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{y})$.
М-шаг: $\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{W},\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\sigma}^2,\ \boldsymbol{\mu}) \to \max\limits_{\mathbf{A},\ \boldsymbol{\sigma}^2}$.

$$2\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{W},\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\sigma}^2,\ \boldsymbol{\mu}) \propto \sum_{k=1}^{K}\left[\log\det\mathbf{A}_k - \mathsf{E}_q\mathbf{w}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{w}_k\right] +$$

$$\sum_{i=1}^{m}\sum_{l=1}^{K}\mathsf{E}z_{il}\left(\log\frac{1}{\sigma_l^2} - \frac{1}{\sigma_l^2}(y_i^2 - 2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_l + \mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_l\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i)\right).$$

$$\sum_{k=1}^{K}\left[\sum_{j=1}^{n}\log\alpha_{kj} - \alpha_{kj}\mathsf{E}w_{kj}^2\right] \to \max_{\alpha_{11},\ \ldots,\ \alpha_{Kn}},$$ где $\mathbf{A}_k = \operatorname{diag}(\alpha_{k1},\ \ldots,\ \alpha_{kn})$.

$\alpha_{kj} = \dfrac{1}{\mathsf{E}w_{kj}^2}$. **Hint:** $\alpha_{kj} = \dfrac{1 - \alpha_{kj}^{\text{old}}\mathsf{D}w_{kj}}{\mathsf{E}^2 w_{kj}}$.

Определение гиперпараметров смеси моделей

М-шаг: $\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{W},\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\mu}) \to \max\limits_{\mathbf{A},\, \boldsymbol{\sigma}^2}.$

$$2\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{W},\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A},\, \boldsymbol{\sigma}^2,\, \boldsymbol{\mu}) \propto \sum_{k=1}^{K} \left[\log\det \mathbf{A}_k - \mathsf{E}_q \mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k\right] +$$

$$\sum_{i=1}^{m} \sum_{l=1}^{K} \mathsf{E} z_{il} \left(\log \frac{1}{\sigma_l^2} - \frac{1}{\sigma_l^2}(y_i^2 - 2y_i \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_l + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_l \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i)\right).$$

$$\sum_{k=1}^{K} \left[\sum_{j=1}^{n} \log \alpha_{kj} - \alpha_{kj} \mathsf{E} w_{kj}^2\right] \to \max_{\alpha_{11},\, \dots,\, \alpha_{Kn}},$$ где $\mathbf{A}_k = \mathrm{diag}(\alpha_{k1},\, \dots,\, \alpha_{kn}).$

$\alpha_{kj} = \dfrac{1}{\mathsf{E} w_{kj}^2}$. **Hint:** $\alpha_{kj} = \dfrac{1 - \alpha_{kj}^{\text{old}} \mathsf{D} w_{kj}}{\mathsf{E}^2 w_{kj}}.$

$$\sigma_l^2 = \frac{\sum_{i=1}^{m} \mathsf{E} z_{il} (y_i^2 - 2y_i \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_l + \mathbf{x}_i^{\mathsf{T}} \mathsf{E}\mathbf{w}_l \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i)}{\sum_{i=1}^{m} \mathsf{E} z_{il}}.$$

Вопрос: Как получить $\mathsf{E} w_{kj}$, $\mathsf{D} w_{kj}$, $\mathsf{E}\mathbf{w}_l \mathbf{w}_l^{\mathsf{T}}$, $\mathsf{E} z_{il}$?

Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171, 481-489, 498-505.
2. MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
3. MacKay, David JC. "The evidence framework applied to classification networks."Neural computation 4.5 (1992): 720-736.
4. Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
5. Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
6. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.