

Байесовский выбор моделей: вариационный EM-алгоритм

Александр Адуенко

16е октября 2019

Содержание предыдущих лекций

- Формула Байеса и формула полной вероятности;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейства. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.
- Линейная регрессия: связь МНК и w_{ML} , регуляризации и w_{MAP} .
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:

$$p(\mathbf{y}_{test} | \mathbf{X}_{test}, \mathbf{X}_{train}, \mathbf{y}_{train}) = \int p(\mathbf{y}_{test} | \mathbf{w}, \mathbf{X}_{test}) p(\mathbf{w} | \mathbf{X}_{train}, \mathbf{y}_{train}) d\mathbf{w}.$$

- Связь апостериорной вероятности модели и обоснованности
- Обоснованность: понимание и связь со статистической значимостью.
- Логистическая регрессия: проблемы ML-оценки w ; связь априорного распределения с отбором признаков.
- Связь апостериорного распределения и неопределенности в прогнозе; обработка нелинейных зависимостей и учёт пропусков в данных и выбросов.

EM-алгоритм

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

Вопрос 1: как решить задачу $p(\mathbf{D}|\Theta) = \int p(\mathbf{D}, \mathbf{Z}|\Theta)d\mathbf{Z} \rightarrow \max_{\Theta}$?

Пример 1. $\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$, $\mathbf{w} \sim N(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$, $\varepsilon \sim N(\mathbf{0}, \beta^{-1}\mathbf{I})$

$p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{A})$.

$\log p(\mathbf{y}|\mathbf{X}, \underbrace{\mathbf{A}, \beta^{-1}}) \propto -\frac{1}{2} \log \det(\beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T) - \frac{1}{2}\mathbf{y}^T (\beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T)^{-1}\mathbf{y}$.

EM-алгоритм[Ⓟ]

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

Идея 1: $p(\mathbf{D}|\Theta) \rightarrow \max_{\Theta}$ заменим на $F(q, \Theta) \rightarrow \max_{q, \Theta}$.

Идея 2: Пошагово оптимизируем по Θ и q , то есть

1 E-шаг: $q^s = F(q, \Theta^{s-1}) \rightarrow \max_q$;

2 M-шаг: $\Theta^s = F(q^s, \Theta) \rightarrow \max_{\Theta}$.

EM-алгоритм для максимизации обоснованности

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \mathbf{w} \sim N(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}), \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \beta^{-1}\mathbf{I})$$

$$p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{A}) = .$$

$$\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) \propto \frac{m}{2} \log \beta - \frac{\beta}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2} \log \det \mathbf{A} - \frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w}.$$

$$F(q, \mathbf{A}, \beta) = - \int q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w} + \int q(\mathbf{w}) \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) d\mathbf{w} = \\ \log p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) - D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \beta)) \rightarrow \max_{q, \mathbf{A}, \beta}.$$

E-шаг (считаем \mathbf{A}, β фиксированными)

$$F(q, \mathbf{A}, \beta) \rightarrow \max_q \iff q(\mathbf{w}) = p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \beta) = N(\mathbf{w}|\mathbf{w}_0, \boldsymbol{\Sigma}_0^{-1}), \text{ где}$$

$$\boldsymbol{\Sigma}_0 = \mathbf{A} + \beta \mathbf{X}^T \mathbf{X}, \mathbf{w}_0 = \beta \boldsymbol{\Sigma}_0^{-1} \mathbf{X}^T \mathbf{y}.$$

M-шаг (считаем $q(\mathbf{w})$ фиксированным)

$$E_{q(\mathbf{w})} \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) = \int q(\mathbf{w}) \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) d\mathbf{w} \rightarrow \max_{\mathbf{A}, \beta}.$$

$$\tilde{F}(\mathbf{A}, \beta) = \frac{m}{2} \log \beta - \frac{\beta}{2} \mathbb{E} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2} \sum_{j=1}^n \log \alpha_j - \frac{1}{2} \sum_{j=1}^n \alpha_j \mathbb{E} w_j^2 \rightarrow \max_{\mathbf{A}, \beta}.$$

$$\frac{\partial F}{\partial \alpha_j} = \frac{1}{2\alpha_j} - \frac{1}{2} \mathbb{E} w_j^2 = 0 \iff \alpha_j = \frac{1}{\mathbb{E} w_j^2}.$$

$$\text{Hint: } 1 = \alpha_j (\mathbb{E}^2 w_j + D w_j) \implies \alpha_j^{\text{new}} = \frac{1 - \alpha_j^{\text{old}} D w_j}{\mathbb{E}^2 w_j}.$$

$$\frac{\partial F}{\partial \beta} = \frac{m}{2\beta} - \frac{1}{2} \mathbb{E} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = 0 \iff \beta = \frac{m}{\mathbb{E} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2}.$$

EM-алгоритм для максимизации обоснованности

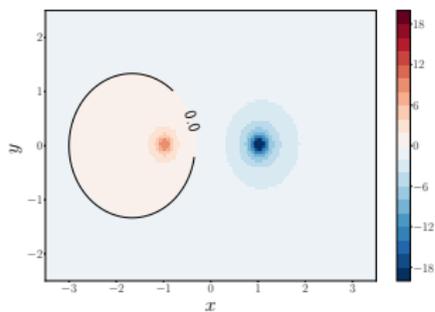
Потенциал поля точечного заряда: $\varphi = k \frac{q}{r}$.

Пусть имеется несколько зарядов q_1, \dots, q_l в точках $\mathbf{z}_1, \dots, \mathbf{z}_l$.

Тогда $\varphi(\mathbf{x}) = k \sum_{i=1}^l \frac{q_i}{\|\mathbf{x} - \mathbf{z}_i\|}$. По набору точек $\mathbf{x}_1, \dots, \mathbf{x}_m$ и измеренным

$$y_i = \varphi(\mathbf{x}_i) - \underbrace{\varphi(\infty)}_{=0} + \varepsilon_i, \quad \varepsilon_i \sim N(\varepsilon_i | 0, \beta^{-1})$$

требуется оценить $\varphi(\mathbf{x})$ для \mathbf{x} из тестовой выборки.



$$\mathbf{y} = \Phi \mathbf{w} + \varepsilon, \quad \varepsilon \sim N(\varepsilon | \mathbf{0}, \beta^{-1} \mathbf{I}), \quad \text{где}$$

$$\Phi = \left\| \frac{1}{\delta + \|\mathbf{x}_i - \mathbf{x}_j\|} \right\|, \quad i, j = \overline{1, m};$$

$$\mathbf{w} \sim p(\mathbf{w} | \mathbf{A}) = N(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1}).$$

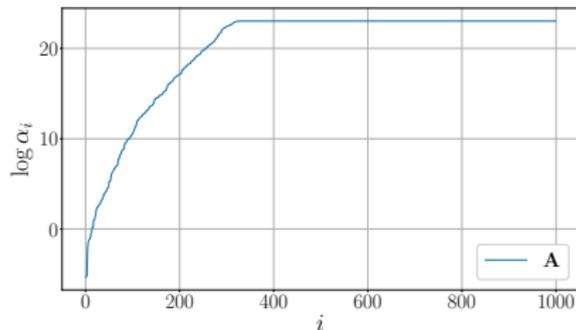
Шаг 1: $p(\mathbf{y}_{\text{train}} | \Phi_{\text{train}}, \mathbf{A}, \beta) \rightarrow \max_{\mathbf{A}, \beta}$ позволит отобрать признаки.

Шаг 2: Прогноз для тестовой выборки:

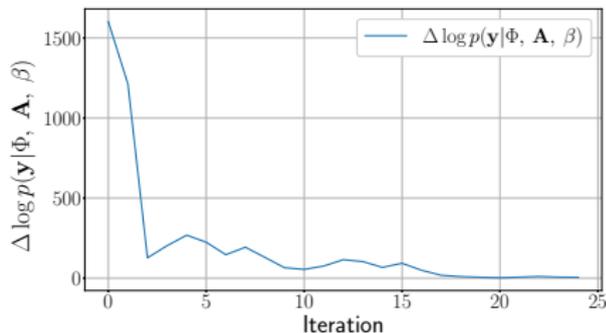
$$p(\mathbf{y}_{\text{test}} | \Phi_{\text{test}}, \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}} | \mathbf{w}, \Phi_{\text{test}}) p(\mathbf{w} | \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) d\mathbf{w}$$

Результаты для задачи восстановления потенциала

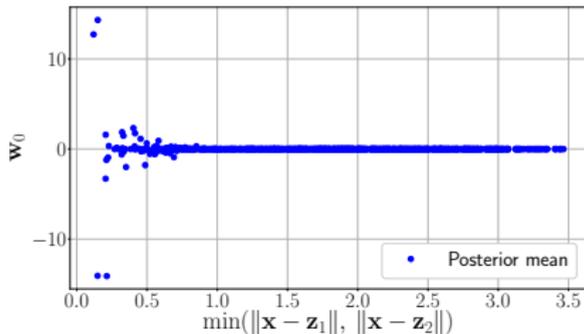
Оптимальный α



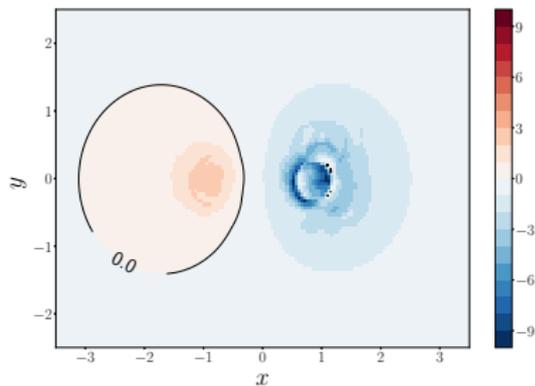
Обоснованность по итерациям



Среднее апостериорного распределения w_0



Восстановленный потенциал



EM-алгоритм: воспоминание

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

Вопрос 1: как решить задачу $p(\mathbf{D}|\Theta) = \int p(\mathbf{D}, \mathbf{Z}|\Theta)d\mathbf{Z} \rightarrow \max_{\Theta}$?

EM-алгоритм

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

Идея 1: $p(\mathbf{D}|\Theta) \rightarrow \max_{\Theta}$ заменим на $F(q, \Theta) \rightarrow \max_{q, \Theta}$.

Идея 2: Пошагово оптимизируем по Θ и q , то есть

1 E-шаг: $q^s = F(q, \Theta^{s-1}) \rightarrow \max_{q \in Q}$;

2 M-шаг: $\Theta^s = F(q^s, \Theta) \rightarrow \max_{\Theta}$.

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

Вариационный EM-алгоритм. E-шаг

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

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

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

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

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

$$\int q(\mathbf{Z}_k) \underbrace{\left[\int \prod_{j \neq k} q(\mathbf{Z}_j) \log p(\mathbf{D}, \mathbf{Z}_1, \dots, \mathbf{Z}_K | \Theta) d\mathbf{Z}_{j \neq k} \right]}_C d\mathbf{Z}_k =$$

$$C \int q(\mathbf{Z}_k) \log \frac{C q(\mathbf{Z}_k)}{e^{E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z} | \Theta)}} d\mathbf{Z}_k \rightarrow \min_{q(\mathbf{Z}_k)}$$

Вариационный EM-алгоритм

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

$$\text{E-шаг. } C \int q(\mathbf{Z}_k) \log \frac{Cq(\mathbf{Z}_k)}{e^{E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z}|\Theta)}} d\mathbf{Z}_k \rightarrow \min_{q(\mathbf{Z}_k)}.$$

Полный алгоритм

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

1 E-шаг: $\log q(\mathbf{Z}_k^s) \propto E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z}|\Theta^{s-1})$;

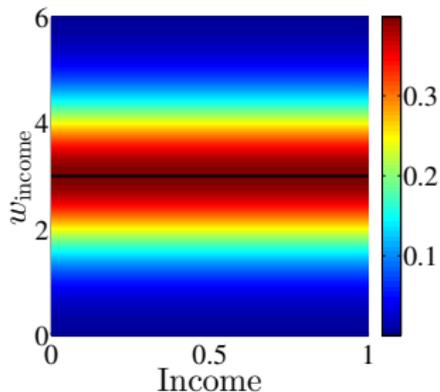
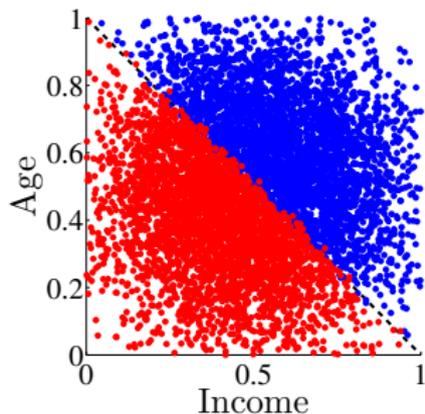
2 M-шаг: $E_{q^s} \log p(\mathbf{D}, \mathbf{Z}|\Theta) \rightarrow \max_{\Theta}$.

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

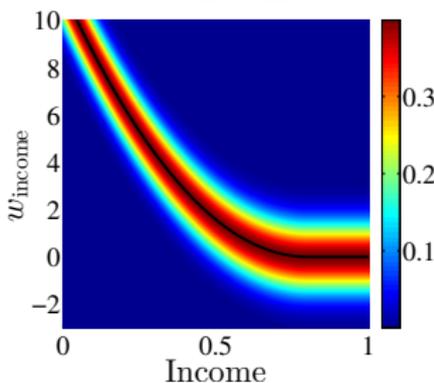
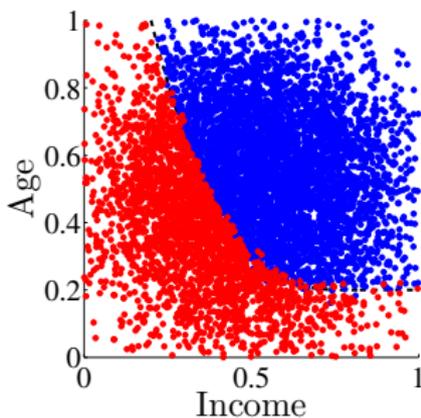
Вопрос 2: как понять, что в конкретной задаче формулы E и M-шагов выписаны верно?

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 логистической регрессии

Вероятностная модель генерации данных

- Веса моделей в смеси π получены из априорного распределения $p(\pi|\mu)$;
- 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 класс y_i определен в соответствии с моделью f_{k_i} : $y_i \sim \text{Be}(\sigma(\mathbf{w}_{k_i}^\top \mathbf{x}_i))$.

Совместное правдоподобие модели

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

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

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

Получение MAP-оценки

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

$$p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \propto \prod_{k=1}^K \pi_k^{\mu_k - 1} \prod_{k=1}^K \sqrt{\det \mathbf{A}_k} \exp\left(-\frac{1}{2} \mathbf{w}_k^\top \mathbf{A}_k \mathbf{w}_k\right) \prod_{i=1}^m \prod_{l=1}^K \left(\pi_l \sigma(y_i \mathbf{w}_l^\top \mathbf{x}_i)\right)^{z_{il}}.$$

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

$$\mathbf{E}\text{-шаг. } \log q(\mathbf{Z}) \propto \sum_{i=1}^m \sum_{l=1}^K z_{il} (\log \pi_k + \log \sigma(y_i \mathbf{w}_l^\top \mathbf{x}_i)), \text{ откуда}$$

$$\gamma_{ik} = p(z_{ik} = 1) \propto \pi_k \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i).$$

$$\mathbf{M}\text{-шаг. } \mathbf{E}_q \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \rightarrow \max_{\boldsymbol{\pi}, \mathbf{w}_1, \dots, \mathbf{w}_K}.$$

$$\mathbf{w}_k^* = \arg \max_{\mathbf{w}_k} \left[-\frac{1}{2} \mathbf{w}_k^\top \mathbf{A}_k \mathbf{w}_k + \sum_{i=1}^m \gamma_{ik} \log \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i) \right].$$

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

Получение апостериорного распределения

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

$$p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi} | \mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \propto \prod_{k=1}^K \pi_k^{\mu_k - 1} \prod_{k=1}^K \sqrt{\det \mathbf{A}_k} \exp\left(-\frac{1}{2} \mathbf{w}_k^\top \mathbf{A}_k \mathbf{w}_k\right) \prod_{i=1}^m \left(\sum_{l=1}^K \pi_l \sigma(y_i \mathbf{w}_l^\top \mathbf{x}_i) \right).$$

Идея: найдем $q(\mathbf{Z}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}) = q(\mathbf{Z})q(\mathbf{w}_1, \dots, \mathbf{w}_K)q(\boldsymbol{\pi})$, наиболее близкое к $p(\mathbf{w}_1, \dots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{y})$.

$$\log q(\mathbf{Z}) \propto \mathbb{E}_{q_{\mathbf{Z}}} p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \propto$$

$$\sum_{i=1}^m \sum_{k=1}^K z_{ik} \left(\mathbb{E} \log \pi_k + \mathbb{E} \log \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i) \right)$$

$$\implies p(z_{ik} = 1) \propto \exp \left(\mathbb{E} \log \pi_k + \mathbb{E} \log \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i) \right).$$

$$\log q(\boldsymbol{\pi}) \propto \mathbb{E}_{q_{\boldsymbol{\pi}}} p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \propto$$

$$\sum_{k=1}^K \log \pi_k \left(\mu_k - 1 + \sum_{i=1}^m \mathbb{E} z_{ik} \right) \implies \boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\pi}, \boldsymbol{\mu} + \boldsymbol{\gamma}).$$

Получение апостериорного распределения (продолжение)

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

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

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

Варианты аппроксимации $q(\mathbf{w}_k)$:

- Аппроксимация Лапласа: $q(\mathbf{w}_k) \approx N(\mathbf{w}_k | \mathbf{w}_k^*, \boldsymbol{\Sigma}_k^{-1})$;
- Ищем $q(\mathbf{w}_k) = N(\mathbf{w}_k | \mathbf{m}_k, \boldsymbol{\Sigma}_k^{-1})$ такое, что $D_{KL}(q \| C_k f_k(\mathbf{w}_k)) \rightarrow \min_q$
 - Численно (если число признаков n невелико);
 - С помощью VLB для сигмоиды и соответствующей верхней оценки для $D_{KL}(q \| C_k f_k(\mathbf{w}_k))$.

Вопрос 3: Как определить $\mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}$?

Максимизация обоснованности

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \rightarrow \max_{\mathbf{A}_1, \dots, \mathbf{A}_K, \{\boldsymbol{\mu}\}}$$

Идея: воспользуемся вариационным EM-алгоритмом.

Е-шаг: найдем $q(\mathbf{Z}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}) = q(\mathbf{Z})q(\mathbf{w}_1, \dots, \mathbf{w}_K)q(\boldsymbol{\pi})$, наиболее близкое к $p(\mathbf{w}_1, \dots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y})$.

М-шаг: $E_q \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \rightarrow \max_{\mathbf{A}_1, \dots, \mathbf{A}_K}$.

Замечание: можно сразу сделать использовать VLB для сигмоидной функции, чтобы получить нижнюю границу обоснованности

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \geq \tilde{p}(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu})$$

и тогда на Е-шаге автоматически $q(\mathbf{w}_k)$ будет нормальным.

Литература

- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171, 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.
- 7 Fahrmeir, Ludwig, and Heinz Kaufmann. "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models." *The Annals of Statistics* (1985): 342-368.
- 8 Baghishani, Hossein, and Mohsen Mohammadzadeh. "Asymptotic normality of posterior distributions for generalized linear mixed models." *Journal of Multivariate Analysis* 111 (2012): 66-77.