

Байесовский выбор моделей: EM-алгоритм и вариационный EM-алгоритм.

Александр Адуенко

5е ноября 2024

Содержание предыдущих лекций

- Формула Байеса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;
- Формула полной вероятности: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$;
- 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
$$p(M_i | \mathbf{X}_{train}, \mathbf{y}_{train}) \propto p(M_i) p_i(\mathbf{y}_{train} | \mathbf{X}_{train}).$$
- Обоснованность: понимание и связь со статистической значимостью.
- БЛогР: обоснованность и отбор признаков, апостериорное распределение. Нелинейная разделяющая; выбросы и пропуски.

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

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

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

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

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

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

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

$$F(q, \mathbf{A}, \beta) = - \int q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w} + \int q(\mathbf{w}) \log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \beta) d\mathbf{w} =$$

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

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

$$F(q, \mathbf{A}, \beta) \rightarrow \max_q \iff q(\mathbf{w}) = p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \mathbf{A}, \beta) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \Sigma_0^{-1}), \text{ где}$$

$$\Sigma_0 = \mathbf{A} + \beta \mathbf{X}^T \mathbf{X}, \mathbf{w}_0 = \beta \Sigma_0^{-1} \mathbf{X}^T \mathbf{y}.$$

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

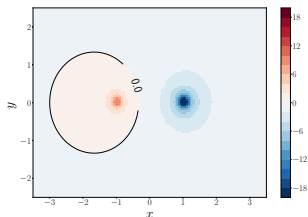
Потенциал поля точечного заряда: $\varphi = k \frac{q}{r}$.

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

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

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



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

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

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

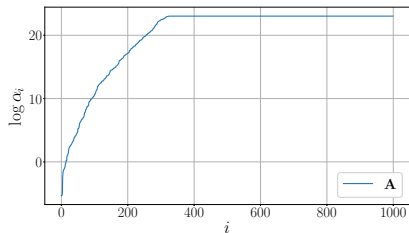
Шаг 1: $p(\mathbf{y}_{\text{train}} | \Phi_{\text{train}}, \mathbf{A}, \beta) \rightarrow \max_{\mathbf{A}, \beta}$ позволит отобрать признаки.

Шаг 2: Прогноз для тестовой выборки:

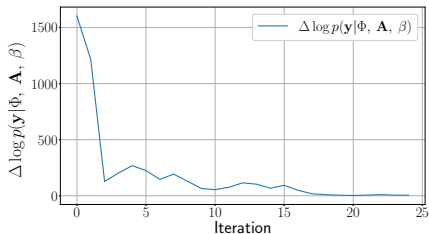
$$p(\mathbf{y}_{\text{test}} | \Phi_{\text{test}}, \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}} | \mathbf{w}, \Phi_{\text{test}}) p(\mathbf{w} | \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) d\mathbf{w}$$

Результаты для задачи восстановления потенциала

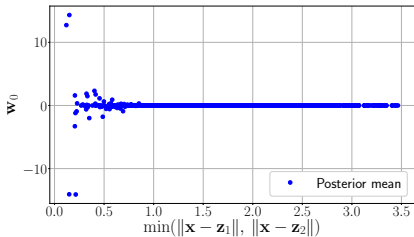
Оптимальный α



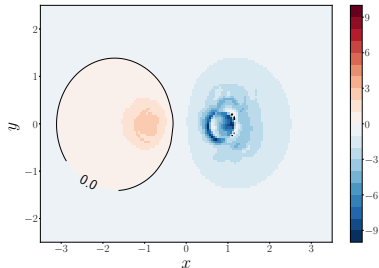
Обоснованность по итерациям



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

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

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

$$\int q(\mathbf{Z}_k) \underbrace{\left[\int \prod_{j \neq k} q(\mathbf{Z}_j) \log p(\mathbf{D}, \mathbf{Z}_1, \dots, \mathbf{Z}_K | \Theta) d\mathbf{Z}_{j \neq k} \right]}_{=1} d\mathbf{Z}_k \propto$$

$$\mathbb{E}_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z} | \Theta)$$

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

Вариационный EM-алгоритм

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

$$\text{E-шаг. } \int q(\mathbf{Z}_k) \log \frac{q(\mathbf{Z}_k)}{\frac{1}{C} e^{E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z} | \Theta)}} d\mathbf{Z}_k \rightarrow \min_{q(\mathbf{Z}_k)}.$$

Полный алгоритм

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

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

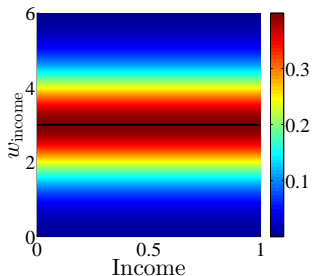
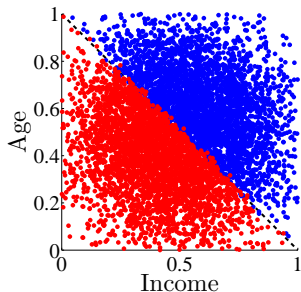
2 M-шаг: $E_{q^s} \log p(\mathbf{D}, \mathbf{Z} | \Theta) \rightarrow \max_{\Theta}$.

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

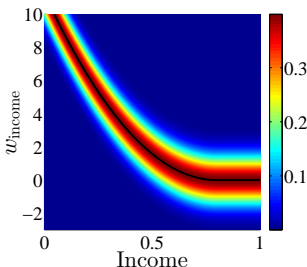
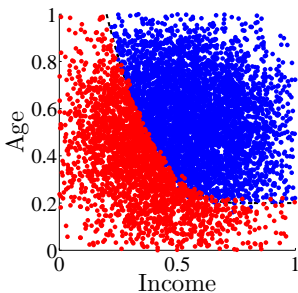
Вопрос 2: как понять, что в конкретной задаче формулы E и M-шагов выписаны верно?

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

Предполагаемый результат



Реальные данные



Вопрос: Как можно учесть указанную нелинейность в модели?

- 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 Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
- 4 Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models." *Statistica Sinica* (2003): 461-476.
- 5 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.
- 6 Baghishani, Hossein, and Mohsen Mohammadzadeh. "Asymptotic normality of posterior distributions for generalized linear mixed models." *Journal of Multivariate Analysis* 111 (2012): 66-77.
- 7 Jaakkola, Tommi, and Michael Jordan. "A variational approach to Bayesian logistic regression models and their extensions." *Sixth International Workshop on Artificial Intelligence and Statistics*. Vol. 82. No. 4. 1997