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

Александр Адуенко

31е октября 2023

Содержание предыдущих лекций

- Формула Байеса: $\mathsf{P}(A|B) = \dfrac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B)}$;
- Формула полной вероятности: $\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|\overline{A})\mathsf{P}(\overline{A})$;
- 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 $\mathbf{w}_{\mathrm{ML}}$, регуляризации и $\mathbf{w}_{\mathrm{MAP}}$.
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:
  $$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})d\mathbf{w}.$$
- Связь апостериорной вероятности модели и обоснованности $p(M_i|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) \propto p(M_i)p_i(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})$.
- Обоснованность: понимание и связь со статистической значимостью.
- БЛогР: обоснованность и отбор признаков, апостериорное распределение. Нелинейная разделяющая; выбросы и пропуски.

ЕМ-алгоритм

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

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

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

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

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

ЕМ-алгоритм

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

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

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

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

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

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

EM-алгоритм для максимизации обоснованности

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

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

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

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

$\log p(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \beta) - D_{\mathrm{KL}}(q(\mathbf{w})\|p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta)) \to \max\limits_{q,\ \mathbf{A},\ \beta}.$

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

$F(q,\ \mathbf{A},\ \beta) \to \max\limits_{q} \Longleftrightarrow q(\mathbf{w}) = p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\ \boldsymbol{\Sigma}_0^{-1})$, где

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

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

$\mathsf{E}_{q(\mathbf{w})}\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) = \int q(\mathbf{w})\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta)d\mathbf{w} \to \max\limits_{\mathbf{A},\ \beta}.$

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

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

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

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

EM-алгоритм для максимизации обоснованности

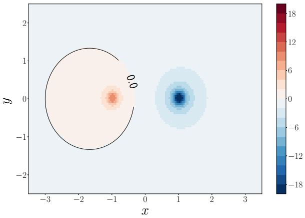
Потенциал поля точечного заряда: $\varphi = k\frac{q}{r}$.

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

Тогда $\varphi(\mathbf{x}) = k\sum\limits_{i=1}^{l} \frac{q_l}{\|\mathbf{x}-\mathbf{z}_l\|}$. По набору точек $\mathbf{x}_1, \ldots, \mathbf{x}_m$ и измеренным

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

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



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

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

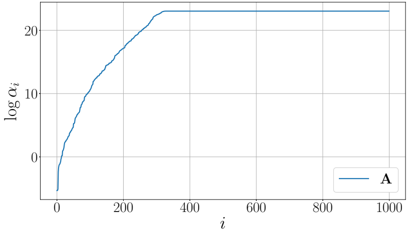
$$\mathbf{w} \sim p(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \ \mathbf{A}^{-1}).$$

Шаг 1: $p(\mathbf{y}_{\text{train}}|\mathbf{\Phi}_{\text{train}}, \ \mathbf{A}, \ \beta) \to \max\limits_{\mathbf{A}, \beta}$ позволит отобрать признаки.

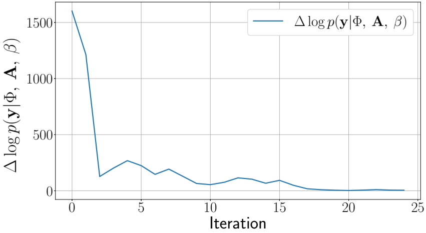
Шаг 2: Прогноз для тестовой выборки:

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

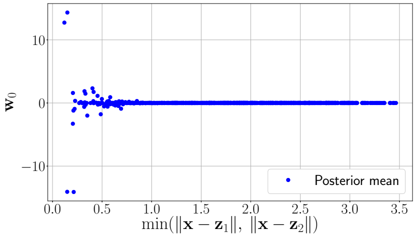
Оптимальный $\boldsymbol{\alpha}$



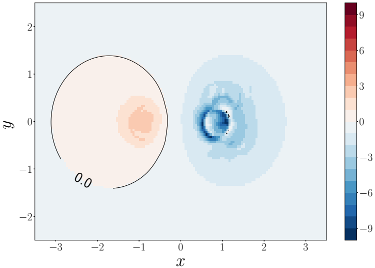
Обоснованность по итерациям



Среднее апостериорного распределения $\mathbf{w}_0$



Восстановленный потенциал



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

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

Вариационный EM-алгоритм. E-шаг

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

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

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

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

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

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

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

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

E-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

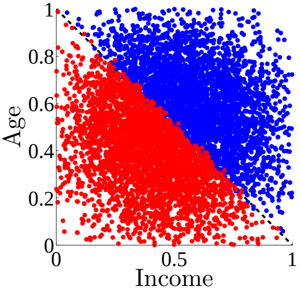
1. E-шаг: $\log q(\mathbf{Z}_k^s) \propto \mathsf{E}_{q\setminus k} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}^{s-1})$;
2. M-шаг: $\mathsf{E}_{q^s} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

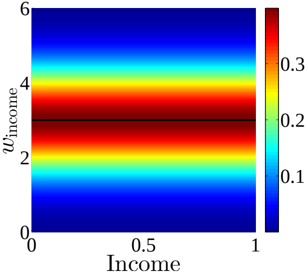
Вопрос 1: зачем нужна факторизация? Чем полученные итеративные формулы лучше формул полного EM-алгоритма?

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

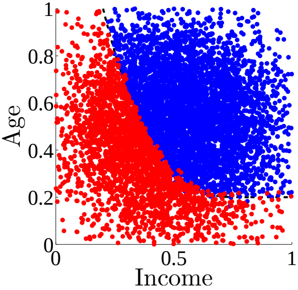
Нарушение свойства $p(\mathbf{w}|\mathbf{x}_i) = p(\mathbf{w})$

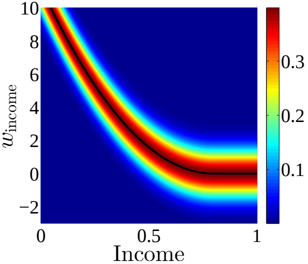
Предполагаемый результат





Реальные данные





Вопрос: как можно учесть указанную нелинейность в модели?

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. 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