

Байесовский выбор моделей: обоснованность и отбор признаков в логистической регрессии

Александр Адуенко

27е октября 2021

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.
- Байесовская логистическая регрессия (БЛР): обоснованность и отбор признаков.

Отбор признаков в БЛР

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

$$p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = \prod_{j=1}^m \sigma(y_j \mathbf{w}^\top \mathbf{x}_j), \quad p(\mathbf{w} | \mathbf{A}) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1}).$$

Полный прогноз:

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

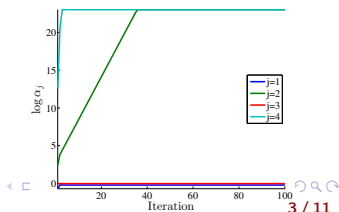
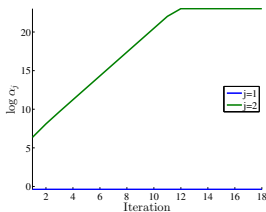
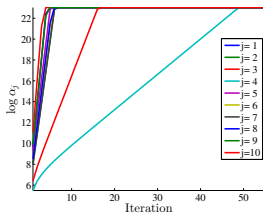
где $p(\mathbf{A} | \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) \propto p(\mathbf{A}) p(\mathbf{y}_{\text{train}} | \mathbf{X}_{\text{train}}, \mathbf{A})$.

Приближенный прогноз:

$$p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) \approx p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{A}^*), \quad \text{где}$$

$$\mathbf{A}^* = \arg \max_{\mathbf{A}} p(\mathbf{y} | \mathbf{X}, \mathbf{A}) = \arg \max_{\mathbf{A}} \int \underbrace{p(\mathbf{y} | \mathbf{X}, \mathbf{w}) p(\mathbf{w} | \mathbf{A})}_{Q(\mathbf{w})} d\mathbf{w}.$$

Иллюстрация отбора признаков в логистической регрессии



Апостериорное распределение в логистической регрессии

$$p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{A}) = p(\mathbf{y} | \mathbf{X}, \mathbf{w}) p(\mathbf{w} | \mathbf{A}), \text{ где } p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = \prod_{j=1}^m \sigma(y_j \mathbf{w}^\top \mathbf{x}_j).$$

$$p(\mathbf{w} | \mathbf{X}, \mathbf{y}, \mathbf{A}) = \frac{p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{A})}{p(\mathbf{y} | \mathbf{X}, \mathbf{A})} = \frac{\prod_{j=1}^m \sigma(y_j \mathbf{w}^\top \mathbf{x}_j) \mathcal{N}(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1})}{p(\mathbf{y} | \mathbf{X}, \mathbf{A})}.$$

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

Вопрос 1: Как определить \mathbf{w}_{MAP} ? Единственное ли решение?

$$q(\mathbf{w}) = -\log p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{A}) = -\log p(\mathbf{w} | \mathbf{A}) - \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) = q(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\top \mathbf{H}(\mathbf{w} - \mathbf{w}_{\text{MAP}}) + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3), \text{ где } \mathbf{H} = \mathbf{A} + \mathbf{X}^\top \mathbf{R} \mathbf{X}, \text{ где } \mathbf{R} = \text{diag}(\sigma(\mathbf{w}_{\text{MAP}}^\top \mathbf{x}_j) \sigma(-\mathbf{w}_{\text{MAP}}^\top \mathbf{x}_j)).$$

Нормальная аппроксимация: $p(\mathbf{w} | \mathbf{X}, \mathbf{y}, \mathbf{A}) \approx \mathcal{N}(\mathbf{w} | \mathbf{w}_{\text{MAP}}, \mathbf{H}^{-1})$.

Пример. Пусть $n = 1$, $\mathbf{w}_{\text{MAP}} = 1$.

Вопрос 2: Что можно сказать про принадлежность объектов с $x = 0$; 1; -1; 5; -5 к классу 1?

Вопрос 3: Как результат зависит от неопределенности h^{-1} ? Что происходит при $h \rightarrow 0$ и при $h \rightarrow \infty$?

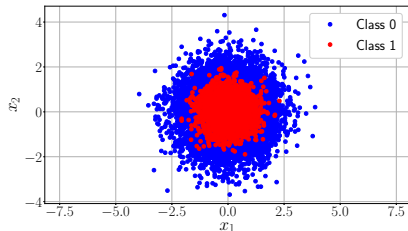
Нелинейная разделяющая поверхность

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

Прогноз вероятности класса 1 в зависимости от неопределенности h^{-1}

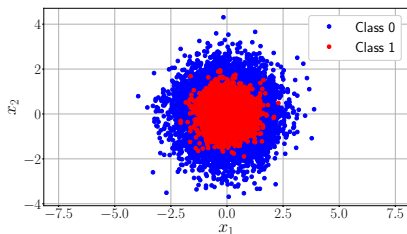
	$x = 5$	$x = 1$	$x = 0$	$x = -1$	$x = -5$
$h = \infty$	0.0067	0.269	0.5	0.731	0.9933
$h = 1$	0.169	0.301	0.5	0.699	0.831
$h = 0$	0.5	0.5	0.5	0.5	0.5

Вопрос 1: как учесть в модели, что классы не сбалансированы?



Вопрос 2: что делать, если разделяющая поверхность нелинейна?

Выбросы и пропуски в данных



Вопрос 1: что делать, если разделяющая поверхность нелинейна?

Идея:

$$\mathbf{x} \mapsto \varphi(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_i), i = 1, \dots, m].$$

Вопрос 2: Чему соответствует отбор признаков при замене $\mathbf{x} \mapsto \varphi(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_i), i = 1, \dots, m]$?

Вопрос 3: Что если значения части признаков не заданы или некорректны? Что происходит при замене на среднее / медиану?

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

Пусть $\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{Z}$, $\tilde{\mathbf{X}} \cdot \mathbf{Z} = \mathbf{0}$, где \mathbf{Z} – матрица значений пропусков.

Новая модель: $p(y, \mathbf{w}, \mathbf{Z} | \tilde{\mathbf{X}}, \mathbf{A}) = p(y | \tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{w})p(\mathbf{w} | \mathbf{A})p(\mathbf{Z} | \tilde{\mathbf{X}})$.

$$p(\mathbf{w} | y, \tilde{\mathbf{X}}, \mathbf{A}) \propto p(y, \mathbf{w} | \tilde{\mathbf{X}}, \mathbf{A}) = \int p(y, \mathbf{w}, \mathbf{Z} | \tilde{\mathbf{X}}, \mathbf{A}) d\mathbf{Z} =$$

$$\int p(y | \tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{w}) p(\mathbf{w} | \mathbf{A}) \underbrace{p(\mathbf{Z} | \tilde{\mathbf{X}})}_{\text{вес}} d\mathbf{Z}.$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

EM-алгоритм для максимизации обоснованности

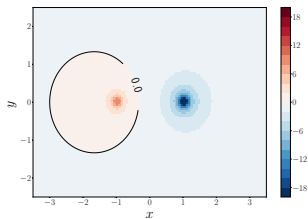
Потенциал поля точечного заряда: $\varphi = k \frac{q}{r}$.

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

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

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

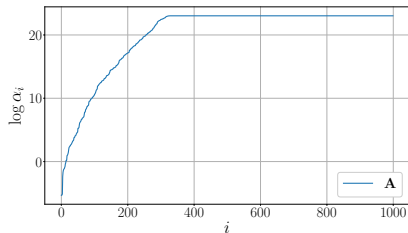
Шаг 1: $p(\mathbf{y}_{\text{train}} | \Phi_{\text{train}}, \mathbf{A}, \beta) \rightarrow \max_{\mathbf{A}, \beta}$ позволит отобрать признаки.

Шаг 2: Прогноз для тестовой выборки:

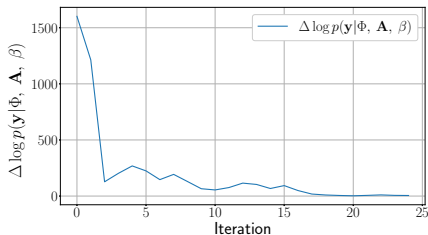
$$p(\mathbf{y}_{\text{test}} | \Phi_{\text{test}}, \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}} | \mathbf{w}, \Phi_{\text{test}}) p(\mathbf{w} | \Phi_{\text{train}}, \mathbf{y}_{\text{train}}) d\mathbf{w}$$

Результаты для задачи восстановления потенциала

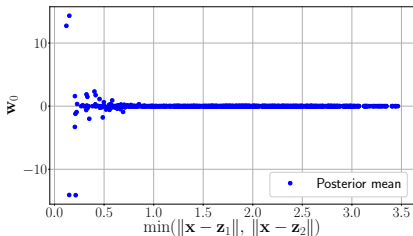
Оптимальный α



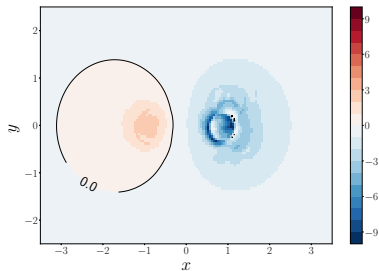
Обоснованность по итерациям



Среднее апостериорного распределения w_0



Восстановленный потенциал



- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171.
- 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