

Байесовский выбор моделей: байесовская линейная регрессия и понятие обоснованности (evidence)

Александр Адуенко

25е сентября 2019

- 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;
- Тестирование гипотез
- Проблема множественного тестирования гипотез
- Экспоненциальное семейство распределений. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.
- Линейная регрессия: классический подход, связь МНК и принципа максимума правдоподобия, связь регуляризации и MAP-оценки для вектора параметров w .
- Апостериорное распределение на вектор параметров w в линейной регрессии и свойство сопряженности априорного распределения и правдоподобия.

Линейная регрессия: байесовский подход

Вероятностная модель линейной регрессии

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, где $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{w} \in \mathbb{R}^d$.

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2}.$$

Байесовский подход.

Пусть теперь еще $\mathbf{w} \sim p(\mathbf{w}|\alpha)$, тогда $p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha) = p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)$.

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) = \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} - \text{апостериорное распределение.}$$

$$\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) = \arg \min_{\mathbf{w}} (-\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) - \log p(\mathbf{w}|\alpha)).$$

Примеры:

- $p(\mathbf{w}|\alpha) = N(\mathbf{0}, \tau^{-1}\mathbf{I})$

$$\mathbf{w}_{MAP} = \arg \min_{\mathbf{w}} \left(\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\tau}{2} \|\mathbf{w}\|^2 \right).$$

- $p(\mathbf{w}|\alpha) = \text{Laplace}(\mathbf{0}, \tau^{-1}\mathbf{I})$

$$\mathbf{w}_{MAP} = \arg \min_{\mathbf{w}} \left(\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tau \|\mathbf{w}\|_1 \right).$$

Вопрос 1: А как получить ML оценку $\mathbf{w}_{ML} = \arg \min_{\mathbf{w}} (-\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}))$?

Вопрос 2: Получили ли мы что-то новое?

Апостериорное распределение

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) = \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)} \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\alpha).$$

Тогда $\log p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) \propto \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}|\alpha)$.

Нормальное априорное распределение.

Рассмотрим $p(\mathbf{w}|\alpha) = N(\mathbf{0}, \tau^{-1}\mathbf{I})$, тогда

$$-\log p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) \propto \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\tau}{2}\|\mathbf{w}\|^2 = \frac{1}{2\sigma^2}\mathbf{y}^\top \mathbf{y} - \frac{1}{\sigma^2}\mathbf{y}^\top \mathbf{X}\mathbf{w} +$$

$$\frac{1}{2\sigma^2}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} + \frac{\tau}{2}\mathbf{w}^\top \mathbf{w} \propto \frac{1}{2} \left(\mathbf{w}^\top (\tau\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X})\mathbf{w} - \frac{2}{\sigma^2}\mathbf{y}^\top \mathbf{X}\mathbf{w} \right) \propto$$

$$\frac{1}{2}(\mathbf{w} - \mathbf{m})^\top \Sigma^{-1}(\mathbf{w} - \mathbf{m}), \text{ где}$$

$$\mathbf{m} = \left(\mathbf{X}^\top \mathbf{X} + \tau\sigma^2\mathbf{I} \right)^{-1} \mathbf{X}^\top \mathbf{y}, \quad \Sigma = \left(\tau\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X} \right)^{-1}.$$

Таким образом, $p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) \propto e^{-\frac{1}{2}(\mathbf{w}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{w}-\mathbf{m})}$.

Вопрос 1: Что мы можем сказать про распределение $p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha)$?

Вопрос 2: Что получилось бы, если бы в качестве $p(\mathbf{w}|\alpha)$ было взято $\text{Laplace}(\mathbf{0}, \tau\mathbf{I})$?

Вопрос 3: Что получилось бы, если бы в качестве $p(\mathbf{w}|\alpha)$ была взята смесь нормальных распределений $\sum_k \pi_k N(\mathbf{m}_k, \Sigma_k)$?

Экспоненциальное семейство распределений

Распределение $p(\mathbf{x})$ в экспоненциальном семействе, если плотность вероятности (функция вероятности) представима в виде

$$p(\mathbf{x}|\Theta) = \frac{1}{Z(\Theta)} h(\mathbf{x}) \exp(\Theta^\top \mathbf{u}(\mathbf{x})).$$

Вопрос 1: как выбрать априорное распределение $p(\Theta)$, чтобы апостериорное распределение осталось в том же экспоненциальном семействе? (свойство сопряженности правдоподобия $p(\mathbf{x}|\Theta)$ и априорного распределения $p(\Theta)$)

Пусть $p(\Theta) = \frac{H(\alpha, \mathbf{v})}{Z(\Theta)^\alpha} \exp(\Theta^\top \mathbf{v})$. Тогда $p(\Theta|\mathbf{x}) = \frac{p(\mathbf{x}|\Theta)p(\Theta)}{p(\mathbf{x})} =$

$$\frac{1}{Z(\Theta)^n p(\mathbf{x})} \prod_{i=1}^n h(x_i) \exp(\Theta^\top \sum_{i=1}^n \mathbf{u}(x_i)) \cdot \frac{H(\alpha, \mathbf{v})}{Z(\Theta)^\alpha} \exp(\Theta^\top \mathbf{v}) =$$
$$\frac{1}{Z(\Theta)^{n+\alpha}} \left(H(\alpha, \mathbf{v}) \prod_{i=1}^n h(x_i)/p(\mathbf{x}) \right) \exp \left(\Theta^\top \left(\mathbf{v} + \sum_{i=1}^n \mathbf{u}(x_i) \right) \right).$$

Вопрос 2: Зачем нам свойство сопряженности?

Обоснованность (evidence)

Модель M_i : $p_i(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$

| Шаг | Наблюдаемые | Скрытые | Результат |
|----------|--|----------------------------|--|
| Обучение | $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ | \mathbf{w} | $p(\mathbf{w} \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ |
| Контроль | \mathbf{X}_{test} | \mathbf{y}_{test} | $p(\mathbf{y}_{\text{test}} \mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ |

$$p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \frac{p(\mathbf{y}_{\text{train}}, \mathbf{w}|\mathbf{X}_{\text{train}})}{\int p(\mathbf{y}_{\text{train}}, \mathbf{w}^*|\mathbf{X}_{\text{train}})d\mathbf{w}^*}$$

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

Модель M_i : $p_i(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})p_i(\mathbf{w})$

Пусть имеется $K > 1$ моделей.

Процесс порождения выборки:

- Природа выбирает модель из K доступных моделей с априорными вероятностями $p(M_i)$, $i = 1, \dots, K$.
- Для выбранной модели i^* природа сэмплирует вектор параметров \mathbf{w}^* из априорного распределения $p_{i^*}(\mathbf{w})$
- Имея i^* , \mathbf{w}^* природа выбирает $\mathbf{X}_{\text{train}}$ и сэмплирует $\mathbf{y}_{\text{train}}$ из $p_{i^*}(\mathbf{y}|\mathbf{X}_{\text{train}}, \mathbf{w}^*)$
- $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ даны наблюдателю.
- Природа выбирает \mathbf{X}_{test} и сэмплирует \mathbf{y}_{test} из $p_{i^*}(\mathbf{y}|\mathbf{X}_{\text{test}}, \mathbf{w}^*)$

Обоснованность (evidence)

Модель M_i : $p_i(\mathbf{y}, \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})p_i(\mathbf{w})$

Общая модель M : $p(\mathbf{y}, \mathbf{w}, M_i|\mathbf{X}) = p(M_i)p_i(\mathbf{w})p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})$

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

$$\sum_{i=1}^K p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}, M_i)p(M_i|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) =$$

$$\sum_{i=1}^K p_i(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})p(M_i|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$$

$$p(M_i|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \frac{p(\mathbf{y}_{\text{train}}, M_i|\mathbf{X}_{\text{train}})}{P(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})} \propto p(\mathbf{y}_{\text{train}}, M_i|\mathbf{X}_{\text{train}}) =$$

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

Пример выбора модели

a – applicant, r – reviewer

$$a, r = \begin{cases} 0, \text{ нет PhD,} \\ 1, \text{ PhD.} \end{cases}$$

d – decision

$$d = \begin{cases} 1, \text{ принять,} \\ 0, \text{ отвергнуть.} \end{cases}$$

| $r = 0$ | $d = 0$ | $d = 1$ |
|---------|---------|---------|
| $a = 0$ | 9 | 0 |
| $a = 1$ | 132 | 19 |

| $r = 1$ | $d = 0$ | $d = 1$ |
|---------|---------|---------|
| $a = 0$ | 97 | 6 |
| $a = 1$ | 52 | 11 |

Случаи:

- 1 $p(d|a, r) = p(d)$
- 2 $p(d|a, r) = p(d|a)$
- 3 $p(d|a, r) = p(d|r)$
- 4 $p(d|a, r) = p(d|a, r)$

$$1) p(d|a, r) = p(d)$$

Поэтому $p(d|w) = \text{Be}(w)$. **Prior** : $p(w) = U[0, 1]$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, w)p(w)dw = \int_0^1 C_9^0(1-w)^9 C_{103}^{97}w^6(1-w)^{97} C_{151}^{132}w^{19}(1-w)^{132} C_{63}^{52}w^{11}(1-w)^{52}dw = 2.8 \cdot 10^{-51} CCCC$$

$$2) p(d|a, r) = p(d|a)$$

Поэтому $p(d|a=0) = \text{Be}(w_1)$, $p(d|a=1) = \text{Be}(w_2)$.

Prior : $p(w_1) = U[0, 1]$, $p(w_2) = U[0, 1]$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, w_1, w_2)p(w_1)p(w_2)dw_1dw_2 = \int_0^1 \int_0^1 C_9^0(1-w_1)^9 C_{103}^{97}w_1^6(1-w_1)^{97} C_{151}^{132}w_2^{19}(1-w_2)^{132} C_{63}^{52}w_2^{11}(1-w_2)^{52}dw_1dw_2 = 4.7 \cdot 10^{-51} CCCC$$

$$3) p(d|a, r) = p(d|r)$$

Поэтому $p(d|r = 0) = \text{Be}(w_1)$, $p(d|r = 1) = \text{Be}(w_2)$.

Prior : $p(w_1) = U[0, 1]$, $p(w_2) = U[0, 1]$

$$p(\mathbf{y}|\mathbf{X}) = 0.27 \cdot 10^{-51} CCCCC$$

$$4) p(d|a, r) = p(d|a, r)$$

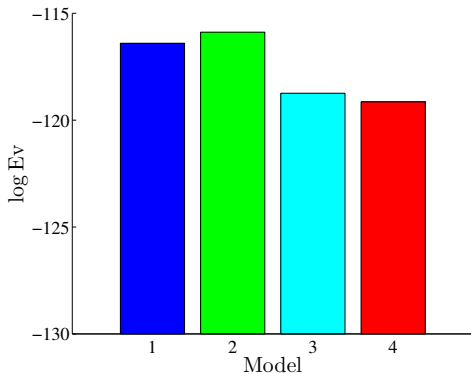
Поэтому $p(d|a = 0, r = 0) = \text{Be}(w_1)$, $p(d|a = 0, r = 1) = \text{Be}(w_2)$,

$p(d|a = 1, r = 0) = \text{Be}(w_3)$, $p(d|a = 1, r = 1) = \text{Be}(w_4)$.

Prior : $p(w_1) = U[0, 1]$, $p(w_2) = U[0, 1]$,

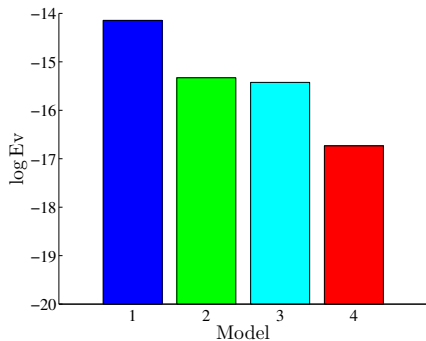
$p(w_3) = U[0, 1]$, $p(w_4) = U[0, 1]$

$$p(\mathbf{y}|\mathbf{X}) = 0.18 \cdot 10^{-51} CCCCC$$

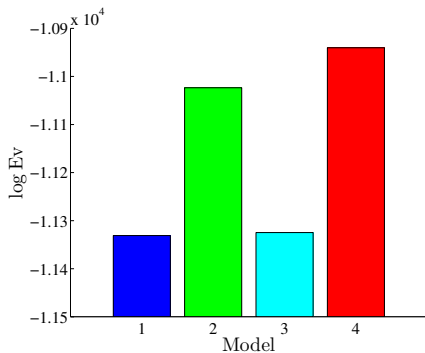


Сравнение обоснованностей, 326 объектов в выборке

Выбор модели: зависимость от размера выборки



Сравнение обоснованностей, 33
объекта в выборке



Сравнение обоснованностей, 32600
объектов в выборке

$$\text{Evidence : } p_i(\mathbf{y}|\mathbf{X}) = \int p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})p_i(\mathbf{w})d\mathbf{w}$$

$$p_i(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p_i(\mathbf{y}|\mathbf{X}, \mathbf{w})p_i(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}.$$

Предположения:

- w одномерный
- Априорное распределение $p_i(w)$ плоское с шириной Δw_{prior}
- Апостериорное распределение $p_i(w|\mathbf{X}, \mathbf{y})$ сконцентрировано вокруг w_{MP} с шириной Δw_{post}

Тогда: $\log p_i(\mathbf{y}|\mathbf{X}) \approx \log p_i(\mathbf{y}|\mathbf{X}, w_{MP}) + \log \left(\frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}} \right)$.

Для M -мерного \mathbf{w} : $\log p_i(\mathbf{y}|\mathbf{X}) \approx \log p_i(\mathbf{y}|\mathbf{X}, \mathbf{w}_{MP}) + M \log \left(\frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}} \right)$.

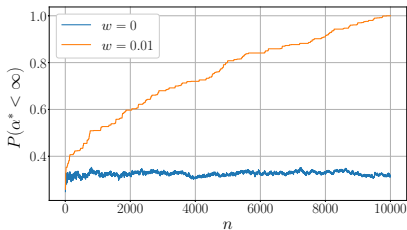
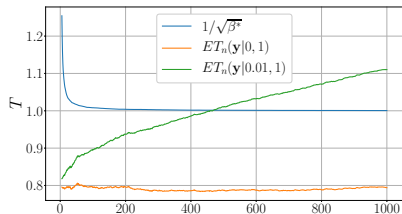
Пример оптимизации evidence

$$y_i = w + \varepsilon_i, \varepsilon_i \sim N(\varepsilon|0, \beta^{-1})$$

$$y_1|w, \dots, y_n|w \sim N(y_i|w, \beta^{-1}), w \sim N(w|0, \alpha^{-1}).$$

$$p(\mathbf{y}|\alpha, \beta) = \frac{\beta^{n/2} \alpha^{1/2}}{(2\pi)^{n/2} \sqrt{n\beta + \alpha}} \exp \left(-\frac{1}{2} \beta \sum_{i=1}^n y_i^2 + \frac{\beta^2 (\sum_{i=1}^n y_i)^2}{2(n\beta + \alpha)} \right).$$

$$(\alpha^*, \beta^*) = \arg \max_{\alpha, \beta} p(\mathbf{y}|\alpha, \beta).$$



$$\alpha^* = \begin{cases} \frac{n^2 \beta^*}{\beta^* (\sum_{i=1}^n y_i)^2 - n}, & \underbrace{\frac{|\sum_{i=1}^n y_i|}{\sqrt{n}}}_{T_n(\mathbf{y}|w, \beta)} > \frac{1}{\sqrt{\beta^*}}, \\ +\infty, & \text{иначе.} \end{cases} \quad \frac{1}{\beta^*} = \frac{\sum_{i=1}^n (y_i - \bar{\mathbf{y}})^2}{n - 1}.$$

- 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 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 Agresti, Alan. Analysis of ordinal categorical data. Vol. 656. John Wiley & Sons, 2010.
- 6 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
- 7 Conjugate priors: <https://people.eecs.berkeley.edu/jordan/courses/260-spring10/other-readings/chapter9.pdf>