

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

Александр Адуенко

26е сентября 2018

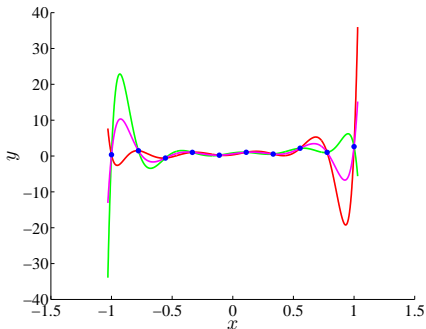
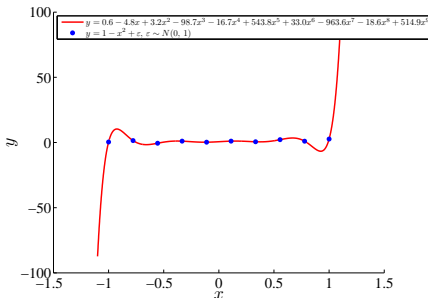
- Формула Байеса: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$;
- Формула полной вероятности: $P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$;
- Определение априорных вероятностей и selection bias;
- Тестирование гипотез
 - Ошибка первого рода и мощность критерия;
 - Критическая область и как ее определить;
- Проблема множественного тестирования гипотез
 - Проблема ложных открытий при независимом одновременном тестировании множества гипотез;
 - FWER и FDR как обобщения вероятности ошибки первого рода;
 - Поправка Бонферрони как консервативное средство контроля FWER;
 - Поправка Бенджамини-Хохберга для контроля FDR для положительно регрессионно зависимых гипотез.
- Экспоненциальное семейство распределений. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.

Линейная регрессия: классический подход

$y = \mathbf{X}\mathbf{w} + \epsilon$, где $y \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{w} \in \mathbb{R}^d$.

МНК (формула Гаусса): $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$.

Оптимизационная задача: $\|y - \mathbf{X}\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$.



$n = d$

$n < d$

Вопросы:

- Что делать, если $n < d$ ($\mathbf{X}^T \mathbf{X}$ вырождена)?
- Почему именно такая оптимизационная задача? Как связана с вероятностной моделью генерации данных?

Линейная регрессия: классический подход

Оптимизационная задача: $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \rightarrow \min_{\mathbf{w}}$.

Пример. Пусть измеряется температура y_i в серверной комнате в момент времени x_i после включения отопления и считается, что нагрев происходит линейно, то есть $\mathbf{X} = [\mathbf{1}, \mathbf{x}]$.

Предположим, что $\varepsilon_i = N(0, 1)^\circ\text{C} / -500 + N(0, 1)^\circ\text{C}$ с $p = 1/2$.

Замечание. Пусть $w = 1^\circ\text{C}/\text{час}$, а $w_0 = 20^\circ\text{C}$.

Выборка: $(0, 20.3)$, $(1, -480.5)$, $(2, 20.8)$, $(3, -476.3)$.

МНК-оценка: $w_0 = -80.44$; $w_1 = -98.85$.

Вопрос: почему МНК не сработал?

Вероятностная модель линейной регрессии

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

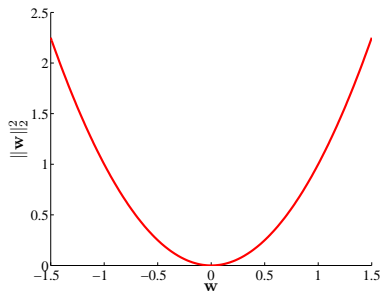
$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2}.$$

Принцип максимума правдоподобия: $\hat{\mathbf{w}}_{ML} = \arg \max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w})$

$$\hat{\mathbf{w}}_{ML} = \arg \min_{\mathbf{w}} -\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \arg \min_{\mathbf{w}} \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2.$$

Квадратическая регуляризация

$$\|y - Xw\|^2 + \tau \|w\|_2^2 \rightarrow \min_w$$

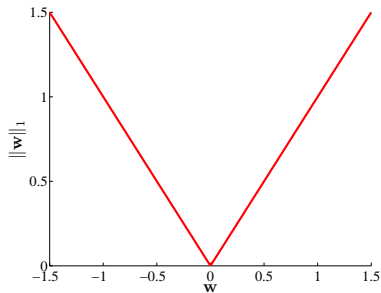


Свойства:

- + Разрешимость
- + Есть аналитическое решение
- Слабо поощряет разреженность

l_1 -regularization

$$\|y - Xw\|^2 + \tau \|w\|_1 \rightarrow \min_w$$



Свойства:

- + Разрешимость
- Нет аналитического решения
- Недифференцируемая целевая функция
- + Поощряет разреженность

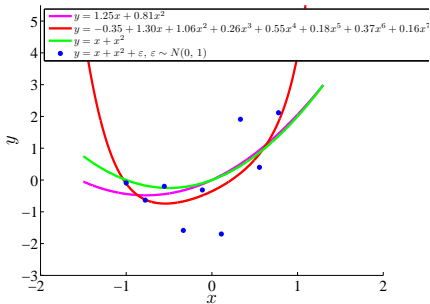
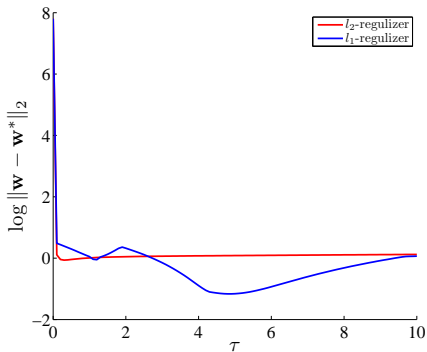
Пример с регрессией на полиномы

Данные

$$y = x + x^2 + \varepsilon, \varepsilon \sim N(0, 1),$$

$y_i \sim p(y|x_i)$, $i = 1, \dots, 10$, где x_1, \dots, x_{10}

выбраны равномерно на $[-1, 1]$.



Зависимость точности от параметра регуляризации τ Наилучшие полиномы

Пример “томография”

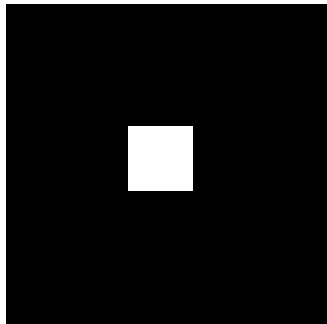
Постановка задачи

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

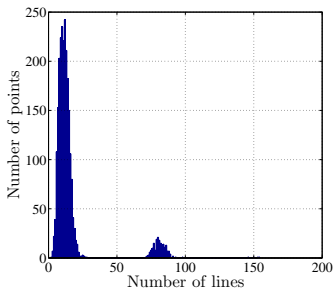
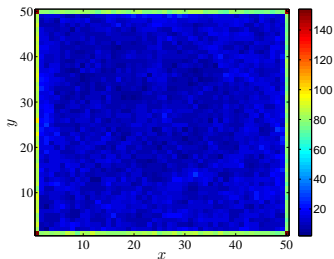
$$\mathbf{y} \in \mathbb{R}^m, \mathbf{X} \in \mathbb{R}^{m \times n^2}, m < n^2.$$

$$\mathbf{w} \in [0, 1]^{n^2}.$$

Параметры: $m = 1000$, $n = 50$.

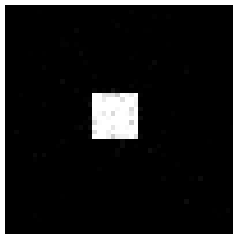


Настоящий \mathbf{w}

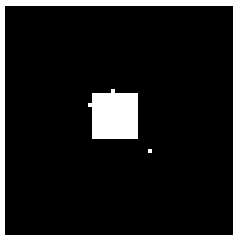


Распределение точек по числу линий

l_1 -регуляризация

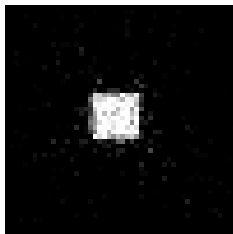


\hat{w}

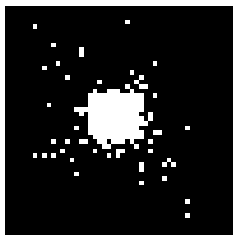


$[\hat{w} > 0.05]$

Квадратическая регуляризация

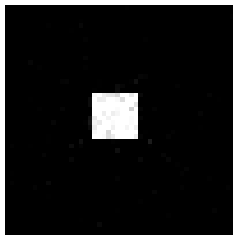


\hat{w}

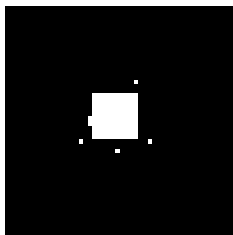


$[\hat{w} > 0.05]$

l_1 -регуляризация

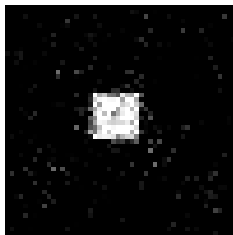


\hat{w}

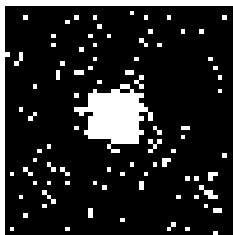


$[\hat{w} > 0.05]$

Квадратическая регуляризация



\hat{w}



$[\hat{w} > 0.05]$

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

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) = \frac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)}$$
 – 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 : $p_i(T, \theta|X) = p_i(T|X, \theta)p(\theta)$

Шаг	Наблюдаемые	Скрытые	Результат
Обучение	$(X_{\text{train}}, T_{\text{train}})$	θ	$p(\theta X_{\text{train}}, T_{\text{train}})$
Контроль	X_{test}	T_{test}	$p(T_{\text{test}} X_{\text{test}}, X_{\text{train}}, T_{\text{train}})$

$$p(\theta|X_{\text{train}}, T_{\text{train}}) = \frac{p(T_{\text{train}}, \theta|X_{\text{train}})}{\int p(T_{\text{train}}, \theta^*|X_{\text{train}})d\theta^*}$$

$$\begin{aligned} p(T_{\text{test}}|X_{\text{test}}, X_{\text{train}}, T_{\text{train}}) &= \int p(T_{\text{test}}, \theta|X_{\text{test}}, X_{\text{train}}, T_{\text{train}})d\theta = \\ &= \int p(T_{\text{test}}|\theta, X_{\text{test}}, X_{\text{train}}, T_{\text{train}})p(\theta|X_{\text{test}}, X_{\text{train}}, T_{\text{train}})d\theta = \\ &= \int p(T_{\text{test}}|\theta, X_{\text{test}})p(\theta|X_{\text{train}}, T_{\text{train}})d\theta \end{aligned}$$

Модель M_i : $p_i(T, \theta|X) = p_i(T|X, \theta)p_i(\theta)$

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 θ^* из априорного распределения $p_{i^*}(\theta)$
- Имея i^* , θ^* природа выбирает X_{train} и сэмплирует T_{train} из $p_{i^*}(T|X_{\text{train}}, \theta^*)$
- $(X_{\text{train}}, T_{\text{train}})$ даны наблюдателю.
- Природа выбирает X_{test} и сэмплирует T_{test} из $p_{i^*}(T|X_{\text{test}}, \theta^*)$

Обоснованность (evidence)

Модель M_i : $p_i(T, \theta|X) = p_i(T|X, \theta)p_i(\theta)$

Общая модель M : $p(T, \theta, M_i|X) = p(M_i)p_i(\theta)p_i(T|X, \theta)$

$$p(T_{\text{test}}|X_{\text{test}}, X_{\text{train}}, T_{\text{train}}) =$$

$$\sum_{i=1}^K p_i(T_{\text{test}}|X_{\text{test}}, X_{\text{train}}, T_{\text{train}})p(M_i|X_{\text{test}}, X_{\text{train}}, T_{\text{train}}) =$$

$$\sum_{i=1}^K p_i(T_{\text{test}}|X_{\text{test}}, X_{\text{train}}, T_{\text{train}})p(M_i|X_{\text{train}}, T_{\text{train}})$$

$$p(M_i|X_{\text{train}}, T_{\text{train}}) = \frac{p(T_{\text{train}}, M_i|X_{\text{train}})}{P(T_{\text{train}}|X_{\text{train}})} \propto p(T_{\text{train}}, M_i|X_{\text{train}}) =$$

$$\int p(T_{\text{train}}, \theta, M_i|X_{\text{train}})d\theta = p(M_i)p_i(T_{\text{train}}|X_{\text{train}})$$

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

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

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

Поэтому $p(d|a=0) = \text{Be}(\theta_1)$, $p(d|a=1) = \text{Be}(\theta_2)$.

Prior : $p(\theta_1) = U[0, 1]$, $p(\theta_2) = U[0, 1]$

$$p(T|X) = \int p(T|X, \theta_1, \theta_2)p(\theta_1)p(\theta_2)d\theta_1d\theta_2 = \int_0^1 \int_0^1 C_9^0(1-\theta_1)^9 C_{103}^{97}\theta_1^6(1-\theta_1)^{97} C_{151}^{132}\theta_2^{19}(1-\theta_2)^{132} C_{63}^{52}\theta_2^{11}(1-\theta_2)^{52}d\theta_1d\theta_2 = 4.7 \cdot 10^{-51} CCCC$$

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

Поэтому $p(d|r = 0) = \text{Be}(\theta_1)$, $p(d|r = 1) = \text{Be}(\theta_2)$.

Prior : $p(\theta_1) = U[0, 1]$, $p(\theta_2) = U[0, 1]$

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

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

Поэтому $p(d|a = 0, r = 0) = \text{Be}(\theta_1)$, $p(d|a = 0, r = 1) = \text{Be}(\theta_2)$,

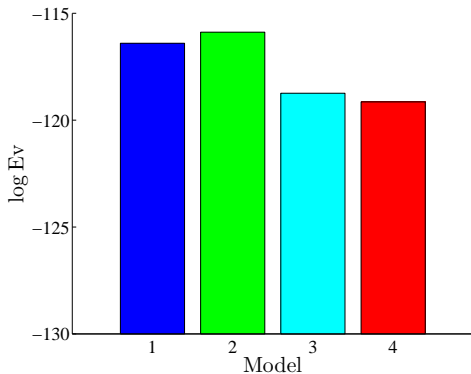
$p(d|a = 1, r = 0) = \text{Be}(\theta_3)$, $p(d|a = 1, r = 1) = \text{Be}(\theta_4)$.

Prior : $p(\theta_1) = U[0, 1]$, $p(\theta_2) = U[0, 1]$,

$p(\theta_3) = U[0, 1]$, $p(\theta_4) = U[0, 1]$

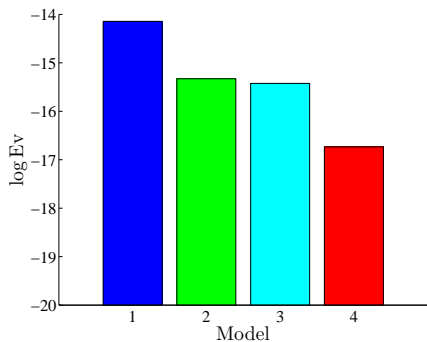
$$p(T|X) = 0.18 \cdot 10^{-51} CCCCC$$

Пример выбора модели

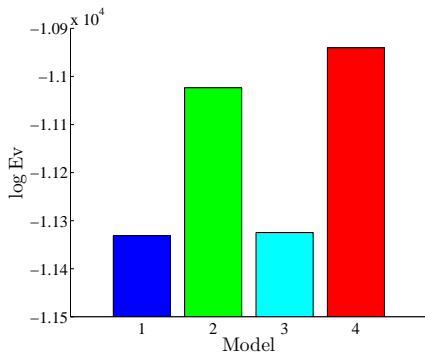


Сравнение обоснованностей, 326 объектов в выборке

Выбор модели: зависимость от размера выборки



Сравнение обоснованностей, 33
объекта в выборке



Сравнение обоснованностей, 32600
объектов в выборке

$$\text{Evidence} : p_i(T|X) = \int p_i(T|X, \theta)p_i(\theta)d\theta$$

$$p_i(\theta|X, T) = \frac{p_i(T|X, \theta)p_i(\theta)}{p(T|X)}.$$

Предположения:

- θ одномерный
- Априорное распределение $p_i(\theta)$ плоское с шириной $\Delta\theta_{\text{prior}}$
- Апостериорное распределение $p_i(\theta|X, T)$ сконцентрировано вокруг θ_{MP} с шириной $\Delta\theta_{\text{post}}$

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

$$\text{Для } M\text{-мерного } \theta: \log p_i(T|X) \approx \log p_i(T|X, \theta_{MP}) + M \log \left(\frac{\Delta\theta_{\text{post}}}{\Delta\theta_{\text{prior}}} \right).$$

Пример оптимизации evidence

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

$$t_1, \dots, t_n \sim N(t|\theta, \beta^{-1}), \theta \sim N(\theta|0, \alpha^{-1}).$$

Evidence: $p(t|\alpha, \beta)$

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

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

$$\alpha^* = \begin{cases} \frac{n^2 \beta}{\beta (\sum_{i=1}^n t_i)^2 - n}, & \beta \left(\sum_{i=1}^n t_i \right)^2 > n, \\ +\infty, & \text{иначе.} \end{cases} \quad \beta^* = \frac{n-1}{\sum_{i=1}^n (t_i - \bar{t})^2}.$$

- 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>