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

Александр Адуенко

3е октября 2023

Содержание предыдущих лекций

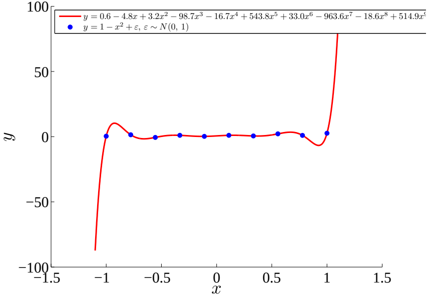
- Формула Байеса: $\mathsf{P}(A|B) = \dfrac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B)}$;
- Формула полной вероятности: $\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|\overline{A})\mathsf{P}(\overline{A})$;
- Определение априорных вероятностей и selection bias;
- Тестирование гипотез
- 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 $\mathbf{w}$.

Линейная регрессия: классический подход

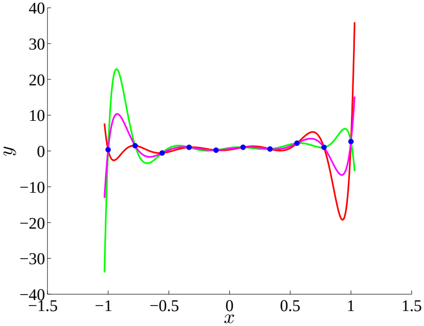
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, где $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{w} \in \mathbb{R}^d$.

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

Оптимизационный задача: $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}$.



$n = d$



$n < d$

Вопросы:

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

Линейная регрессия: классический подход

Оптимизационный задача: $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \to \min\limits_{\mathbf{w}}.$

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

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

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

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

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

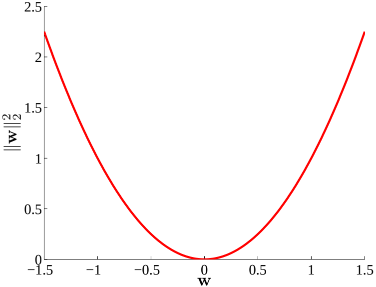
Принцип максимума правдоподобия: $\hat{\mathbf{w}}_{ML} = \arg\max\limits_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\ \mathbf{w})$

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

Регуляризация: классический подход

Квадратическая регуляризация

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tau\|w\|_2^2 \to \min_{\mathbf{w}}$$



Свойства:

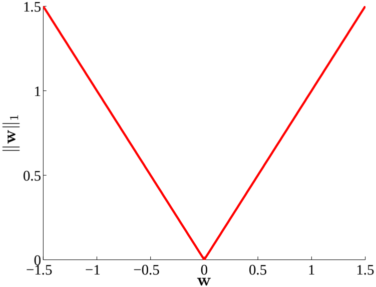
\+ Разрешимость

\+ Есть аналитическое решение

— Слабо поощряет разреженность

l_1–regularization

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tau\|w\|_1 \to \min_{\mathbf{w}}$$



Свойства:

\+ Разрешимость

— Нет аналитического решения

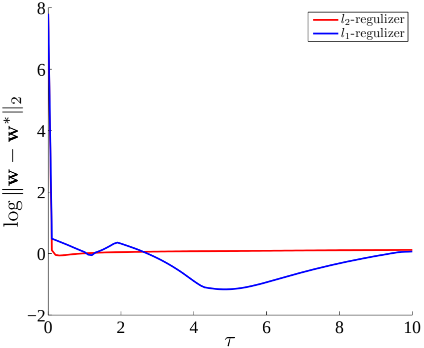
— Недифференцируемая целевая функция

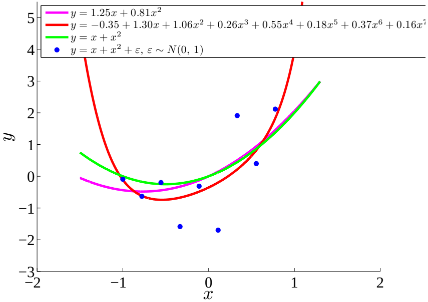
\+ Поощряет разреженность

Пример с регрессией на полиномы

Данные

$y = x + x^2 + \varepsilon,\ \varepsilon \sim \mathcal{N}(0,\ 1),$
$y_i \sim p(y|x_i),\ i = 1,\ \dots,\ 10,$ где $x_1,\ \dots,\ x_{10}$
выбраны равномерно на $[-1,\ 1]$.





Зависимость точности от параметра регуляризации τ

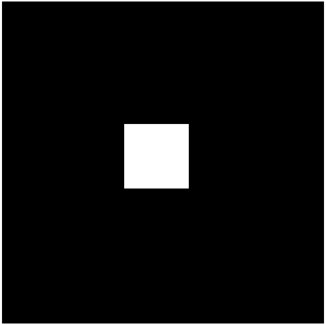
Наилучшие полиномы

Пример "томография"

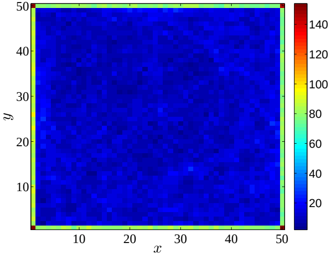
Постановка задачи

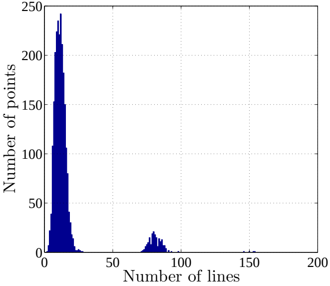
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \ \beta^{-1}\mathbf{I}),$
$\mathbf{y} \in \mathbb{R}^m, \ \mathbf{X} \in \mathbb{R}^{m \times n^2}, \ m < n^2.$
$\mathbf{w} \in [0, \ 1]^{n^2}.$

Параметры: $m = 1000, \ n = 50.$



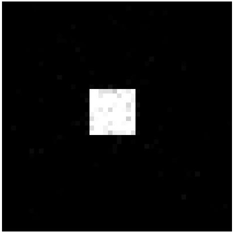
Настоящий $\mathbf{w}$



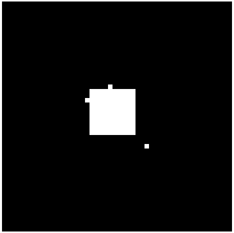


Распределение точек по числу линий

l_1–регуляризация

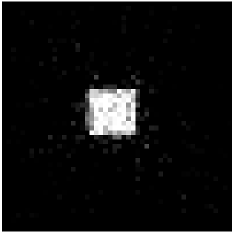


$\hat{\mathbf{w}}$

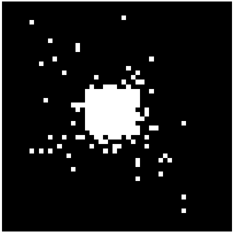


$[\hat{\mathbf{w}} > 0.05]$

Квадратическая регуляризация



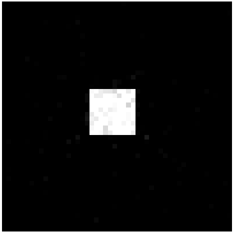
$\hat{\mathbf{w}}$



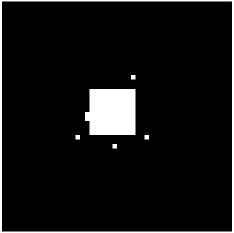
$[\hat{\mathbf{w}} > 0.05]$

Пример “томография”, $\beta = 4$

l_1–регуляризация

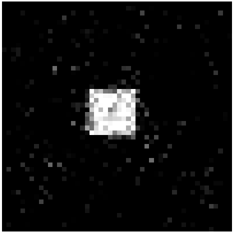


$\hat{\mathbf{w}}$

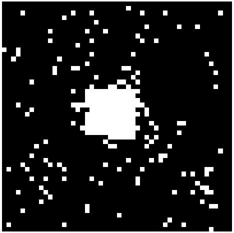


$[\hat{\mathbf{w}} > 0.05]$

Квадратическая регуляризация



$\hat{\mathbf{w}}$



$[\hat{\mathbf{w}} > 0.05]$

Линейная регрессия: байесовский подход

Вероятностная модель линейной регрессии

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

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

Байесовский подход.

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

$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \alpha) = \dfrac{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \alpha)}{p(\mathbf{y}|\mathbf{X}, \alpha)}$ – 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:

- $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I})$
  $\mathbf{w}_{MAP} = \arg\min_{\mathbf{w}} \left(\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\tau}{2}\|\mathbf{w}\|^2\right)$.
- $p(\mathbf{w}|\alpha) = \mathrm{Laplace}(\mathbf{0}, \tau^{-1}\mathbf{I})$
  $\mathbf{w}_{MAP} = \arg\min_{\mathbf{w}} \left(\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tau\|\mathbf{w}\|_1\right)$.

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

$-\log p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \alpha) \propto \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{\tau}{2}\|\mathbf{w}\|^2 = \frac{1}{2\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{y} - \frac{1}{\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \frac{1}{2\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w} + \frac{\tau}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} \propto \frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}(\tau\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{w} - \frac{2}{\sigma^2}\mathbf{y}^{\mathsf{T}}\mathbf{X}\mathbf{w}\right) \propto \frac{1}{2}(\mathbf{w} - \mathbf{m})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{w} - \mathbf{m})$, где

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

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

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 $\mathrm{Laplace}(\mathbf{0},\ \tau\mathbf{I})$?

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

Распределение $p(\mathbf{x})$ в экспоненциальном семействе, если плотность вероятности (функция вероятности) представима в виде
$p(\mathbf{x}|\boldsymbol{\Theta}) = \dfrac{1}{Z(\boldsymbol{\Theta})} h(\mathbf{x}) \exp(\boldsymbol{\Theta}^{\mathsf{T}} \mathbf{u}(\mathbf{x}))$.

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

Пусть $p(\boldsymbol{\Theta}) = \dfrac{H(\alpha,\, \mathbf{v})}{Z(\boldsymbol{\Theta})^{\alpha}} \exp(\boldsymbol{\Theta}^{\mathsf{T}} \mathbf{v})$. Тогда $p(\boldsymbol{\Theta}|\mathbf{x}) = \dfrac{p(\mathbf{x}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\mathbf{x})} =$

$$\frac{1}{Z(\boldsymbol{\Theta})^n p(\mathbf{x})} \prod_{i=1}^{n} h(x_i) \exp(\boldsymbol{\Theta}^{\mathsf{T}} \sum_{i=1}^{n} \mathbf{u}(x_i)) \cdot \frac{H(\alpha,\, \mathbf{v})}{Z(\boldsymbol{\Theta})^{\alpha}} \exp(\boldsymbol{\Theta}^{\mathsf{T}} \mathbf{v}) =$$

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

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

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

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

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

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

Обоснованность (evidence)

Модель M_i: $p_i(\mathbf{y},\ \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w})p_i(\mathbf{w})$
Пусть имеется $K > 1$ моделей.
Процесс порождения выборки:

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

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

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X},\ w)p(w)dw = \int_0^1 C_9^0(1-w)^9$$

$C_{103}^{97}w^6(1-w)^{97}C_{151}^{132}w^{19}(1-w)^{132}C_{63}^{52}w^{11}(1-w)^{52}dw = 2.8\cdot 10^{-51}CCCC$

2) $p(d|a,\ r) = p(d|a)$
Поэтому $p(d|a=0) = \mathrm{Be}(w_1)$, $p(d|a=1) = \mathrm{Be}(w_2)$.
$\mathbf{Prior}:\ p(w_1) = U[0,\ 1]$, $p(w_2) = U[0,\ 1]$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, w_1, w_2)p(w_1)p(w_2)dw_1dw_2 = \int_0^1\int_0^1 C_9^0(1-w_1)^9C_{103}^{97}$$

$w_1^6(1-w_1)^{97}C_{151}^{132}w_2^{19}(1-w_2)^{132}C_{63}^{52}w_2^{11}(1-w_2)^{52}dw_1dw_2 = 4.7\cdot 10^{-51}CCCC$

Пример выбора модели

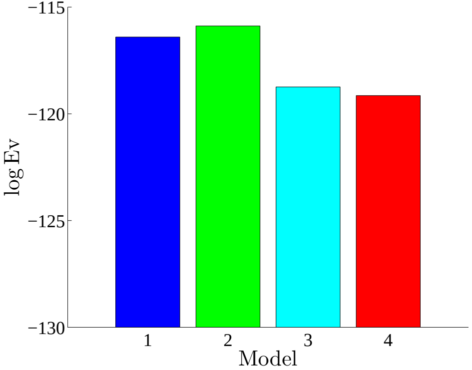
3)$p(d|a,\, r) = p(d|r)$
Поэтому $p(d|r=0) = \mathrm{Be}(w_1)$, $p(d|r=1) = \mathrm{Be}(w_2)$.
$\textbf{Prior}:\ p(w_1) = U[0,\, 1],\, p(w_2) = U[0,\, 1]$

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

4) $p(d|a,\, r) = p(d|a,\, r)$
Поэтому $p(d|a=0,\, r=0) = \mathrm{Be}(w_1)$, $p(d|a=0,\, r=1) = \mathrm{Be}(w_2)$,
$p(d|a=1,\, r=0) = \mathrm{Be}(w_3)$, $p(d|a=1,\, r=1) = \mathrm{Be}(w_4)$.
$\textbf{Prior}:\ p(w_1) = U[0,\, 1],\, p(w_2) = U[0,\, 1],$
$p(w_3) = U[0,\, 1],\, p(w_4) = U[0,\, 1]$

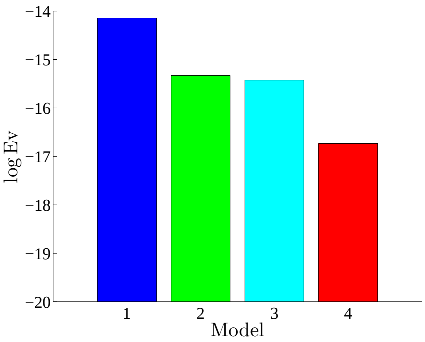
$$p(\mathbf{y}|\mathbf{X}) = 0.18 \cdot 10^{-51} CCCC$$

Пример выбора модели

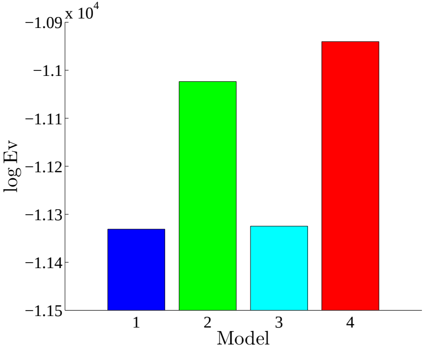


Сравнение обоснованностей, 326 объектов в выборке

Выбор модели: зависимость от размера выборки



Сравнение обоснованностей, 33 объекта в выборке



Сравнение обоснованностей, 32600 объектов в выборке

Простое понимание обоснованности

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

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

Предположения:

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

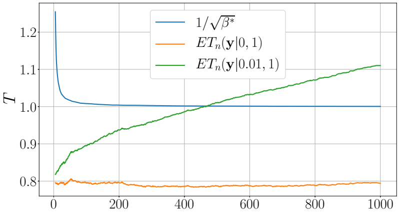
Пример оптимизации evidence

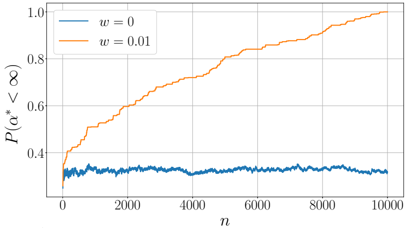
$$y_i = w + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\varepsilon|0,\ \beta^{-1})$$

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

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

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





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

Литература

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