

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

Александр Адуенко

24е сентября 2024

- 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.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.

Наивный байесовский классификатор

Пусть имеется K классов $C = \{C_1, \dots, C_K\}$ и $\mathbf{x} \in \mathbb{R}^n$.

Требуется построить классификатор $f(\cdot) : \mathbb{R}^n \rightarrow C$.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \propto p(C_k)p(\mathbf{x}|C_k).$$

$$p(C_k)p(\mathbf{x}|C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_1, \dots, x_{n-1}, C_k).$$

«Наивность»: $p(x_i|x_1, \dots, x_{i-1}, C_k) = p(x_i|C_k)$.

$$p(C_k|\mathbf{x}) = \frac{p(C_k) \prod_{i=1}^n p(x_i|C_k)}{p(\mathbf{x})}.$$

Классификатор: $f(\mathbf{x}) = \arg \max_k \left(p(C_k) \prod_{i=1}^n p(x_i|C_k) \right)$.

Вопросы:

- Как определить $p(C_k)$ и $p(x_i|C_k)$?
- Насколько плоха «наивность», и зачем она вводится?
- Почему классификатор такого вида?

Вопрос: как определить $p(C_k)$ и $p(x_i|C_k)$?

- 1** Определяем $p(C_k)$ частотно по выборке, а для $p(x_i|C_k)$ строим параметрическую модель и используем ML-оценки ее параметров по выборке;
- 2** Аналогично п.1, но используем непараметрическое оценивание плотностей;
- 3** Вводим априорное распределение на вектор вероятностей $[p(C_1), \dots, p(C_K)]^\top$, параметрическую модель на $p(x_i|C_k)$ с неизвестными параметрами, и априорное распределение на параметры моделей.

Вопрос: насколько плоха «наивность», и зачем она вводится?

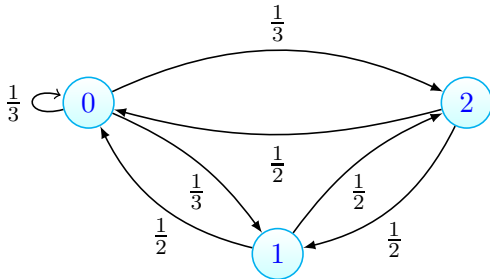
Пример: $K = 2$,

$$p(\mathbf{x}|C_1) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right), \quad p(\mathbf{x}|C_2) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right).$$

Наивный байесовский классификатор: продолжение

Пример. Классификация пользователей по интересующему атрибуту (например, полу, возрасту, достатку, интересу к некоторому товару) по истории x переходов между веб-страницами.

Предположение: переходы между страницами для каждого класса C_k описываются марковской цепью с некоторыми вероятностями перехода (разными для разных классов) между состояниями (веб-страницами).



$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \propto p(C_k)p(\mathbf{x}|C_k).$$

$$p(C_k)p(\mathbf{x}|C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_1, \dots, x_{n-1}, C_k) = p(C_k)p(x_1|C_k)p(x_2|x_1, C_k) \cdot \dots \cdot p(x_n|x_{n-1}, C_k).$$

Вопрос: как оценить $p(x_1|C_k)$, $p(C_k)$ и $p(x_i|x_{i-1}, C_k)$?

Классификатор:

$$f(\mathbf{x}) = \arg \max_k p(C_k | \mathbf{x}) = \arg \max_k \left(p(C_k) \prod_{i=1}^n p(x_i | C_k) \right).$$

Вопрос. Пусть $p(C_k | \mathbf{x})$ известна точно. Какой классификатор оптимален?

Пусть $K = 2$ и $P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$ есть матрица штрафа.

Пример 1. $p_{11} = p_{22} = 0$, $p_{12} = 0$, $p_{21} = 1$;

Пример 2. $p_{11} = p_{22} = 0$, $p_{12} = 1$, $p_{21} = 1$;

Пример 3. $p_{11} = p_{22} = 0$, $p_{12} = 1$, $p_{21} = 10$;

Пример 4. $p_{11} = -1$, $p_{22} = -100$, $p_{12} = 1$, $p_{21} = 1$.

Экспоненциальное семейство распределений

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

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

Распределение	Плотность	$\mathbf{u}(\mathbf{x})$	Θ	$Z(\Theta)$
Be(p)	$p^x (1-p)^{1-x}$	x	$\log \frac{p}{1-p}$	$\frac{1}{1-p}$
Poison(λ)	$\frac{\lambda^x}{x!} e^{-\lambda}$	x	$\log \lambda$	e^λ
$\Gamma(\alpha, \beta)$	$\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$	$[\log x, x]$	$[\alpha, -\beta]$	$\frac{\Gamma(\alpha)}{\beta^\alpha}$
$B(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$[\log x, \log(1-x)]$	$[\alpha, \beta]$	$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
Dir(α)	$\frac{\Gamma(\sum \alpha_i)}{\prod_j \Gamma(\alpha_j)} \prod_i p_i^{\alpha_i-1}$	$[\log p_i]$	α	$\frac{\prod_j \Gamma(\alpha_j)}{\Gamma(\sum \alpha_i)}$
$N(\mathbf{m}, \Sigma^{-1})$	$\frac{\sqrt{\det \Sigma}}{(2\pi)^{n/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma (\mathbf{x}-\mathbf{m})}$	$[\mathbf{x}, \mathbf{xx}^\top]$	$[\Sigma \mathbf{m}, -\frac{1}{2}\Sigma]$	$\frac{(2\pi)^{n/2} e^{-\frac{1}{2}\mathbf{m}^\top \Sigma \mathbf{m}}}{\sqrt{\det \Sigma}}$

Пример: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2} = \frac{1}{\underbrace{\sqrt{2\pi\sigma e \frac{m^2}{2\sigma^2}}}_{Z(\Theta)}} e^{\underbrace{\frac{u_1(x)}{x}}_{\frac{\theta_1}{\sigma^2}} + \underbrace{\frac{u_2(x)}{x^2}}_{\frac{\theta_2}{2\sigma^2}}}$

$$Z(\Theta) = \sqrt{-\pi/\theta_2} e^{-\frac{\theta_1^2}{4\theta_2}}.$$

Достаточные статистики.

Статистика $T(\mathbf{x})$ называется **достаточной** относительно параметра Θ , если $p(\mathbf{x}|T(\mathbf{x}) = t, \Theta) = p(\mathbf{x}|T(\mathbf{x}) = t)$.

Пример: $p(\mathbf{x}|\Theta) = \frac{1}{Z^n(\Theta)} \exp(\theta_1 \sum_{i=1}^n x_i + \theta_2 \sum_{i=1}^n x_i^2)$.

Теорема Фишера-Неймана о факторизации. $T(\mathbf{x})$ достаточна относительно параметра $\Theta \iff p(\mathbf{x}|\Theta) = h(\mathbf{x})g(\Theta, T(\mathbf{x}))$.

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

Свойство: $E\mathbf{u}(\mathbf{x}) = \nabla \log Z(\Theta)$, $E\mathbf{u}\mathbf{u}^T = \nabla \nabla \log Z(\Theta)$.

Пример (нормальное распределение): $Z(\Theta) = \sqrt{-\pi/\theta_2} e^{-\frac{\theta_1^2}{4\theta_2}}$.

$$Eu_1(x) = Ex = -\frac{\theta_1}{2\theta_2} = m, \quad Ex^2 = \frac{\theta_1^2}{4\theta_2^2} - \frac{1}{2\theta_2} = m^2 + \sigma^2;$$

$$Eu_1^2 = Dx^2 = \frac{1}{2\theta_2^2} - \frac{\theta_1^2}{2\theta_2^3} = 2\sigma^4 + 4m^2\sigma^2.$$

Пример (гамма-распределение): $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

$$\log Z(\Theta) = \log \frac{\Gamma(\alpha)}{\beta^\alpha} = \log \Gamma(\alpha) - \alpha \log \beta;$$

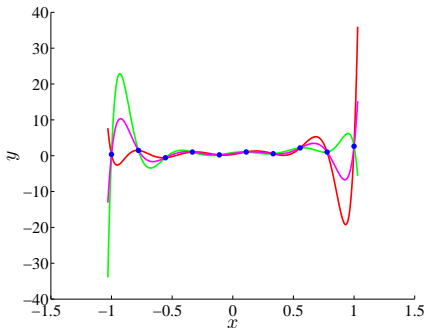
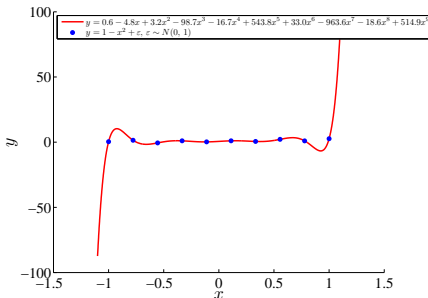
$$E \log x = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log \beta = \psi(\alpha) - \log \beta; \quad Ex = \frac{\alpha}{\beta}.$$

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

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

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

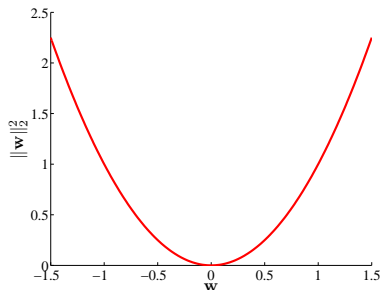
$$p(y|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2} = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2}.$$

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

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

Квадратическая регуляризация

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

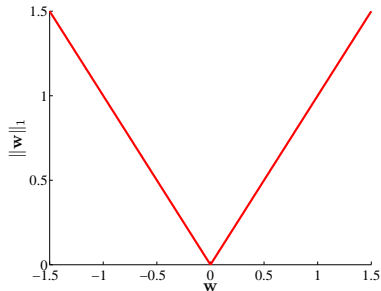


Свойства:

- + 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

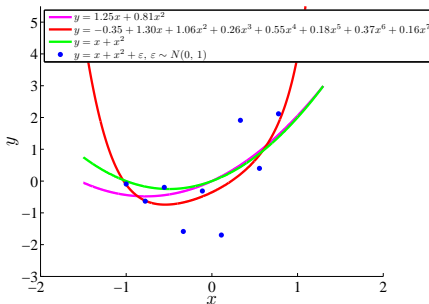
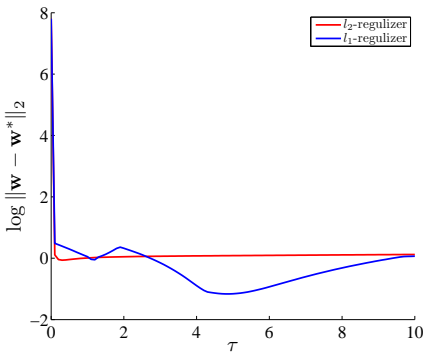
Пример с регрессией на полиномы

Данные

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

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

выбраны равномерно на $[-1, 1]$.



Зависимость точности от параметра регуляризации τ Наилучшие полиномы

Пример “томография”

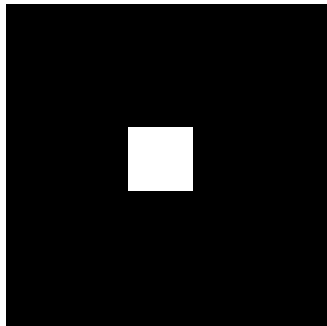
Постановка задачи

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

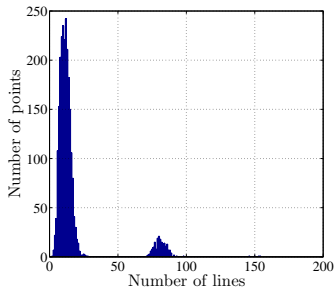
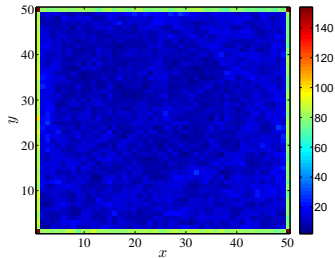
$$\mathbf{y} \in \mathbb{R}^m, \mathbf{X} \in \mathbb{R}^{m \times n^2}, m < n^2.$$

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

Параметры: $m = 1000$, $n = 50$.

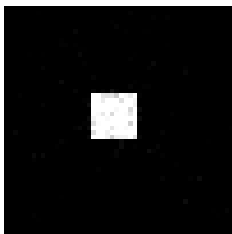


Настоящий \mathbf{w}

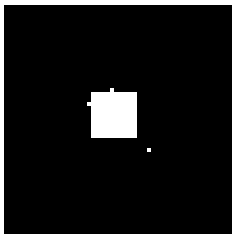


Распределение точек по числу линий

l_1 -регуляризация

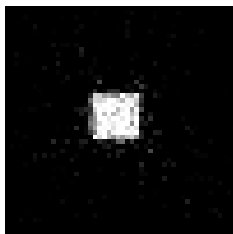


\hat{w}

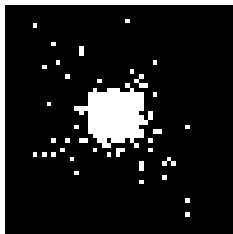


$[\hat{w} > 0.05]$

Квадратическая регуляризация

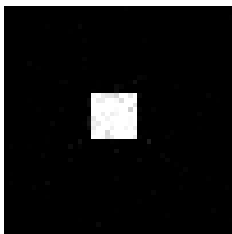


\hat{w}

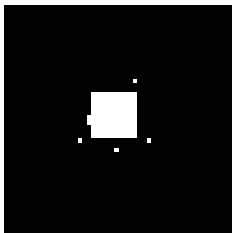


$[\hat{w} > 0.05]$

l_1 -регуляризация

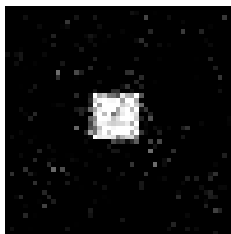


\hat{w}

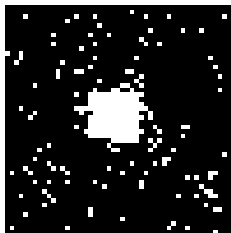


$[\hat{w} > 0.05]$

Квадратическая регуляризация



\hat{w}



$[\hat{w} > 0.05]$

Линейная регрессия: байесовский подход

Вероятностная модель линейной регрессии

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

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

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:

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

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

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

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

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

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

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

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

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 $\sum_k \pi_k \mathcal{N}(\mathbf{m}_k, \Sigma_k)$?

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

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

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

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 $p(M_i)$, $i = 1, \dots, K$.
- 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 $p(d|a = 0, r = 0) = \text{Be}(w_1)$, $p(d|a = 0, r = 1) = \text{Be}(w_2)$,

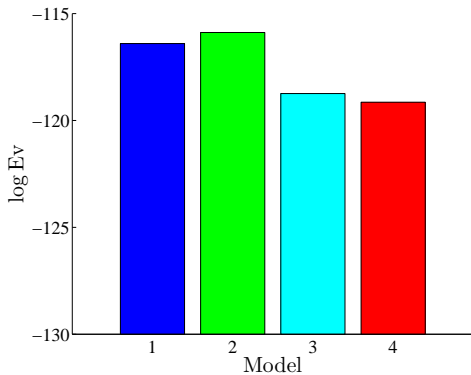
$p(d|a = 1, r = 0) = \text{Be}(w_3)$, $p(d|a = 1, r = 1) = \text{Be}(w_4)$.

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

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

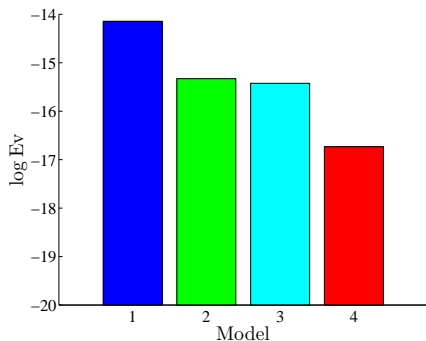
$$p(\mathbf{y}|\mathbf{X}) = 0.18 \cdot 10^{-51} CCCCC$$

Пример выбора модели

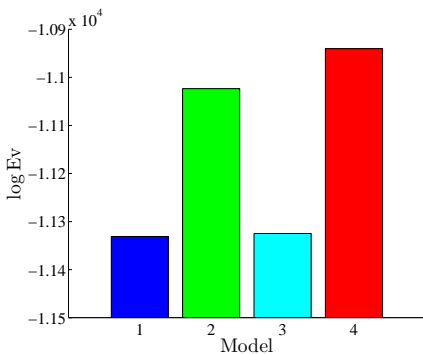


Сравнение обоснованностей, 326 объектов в выборке

Выбор модели: зависимость от размера выборки



Сравнение обоснованностей, 33
объекта в выборке



Сравнение обоснованностей, 32600
объектов в выборке

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

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

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

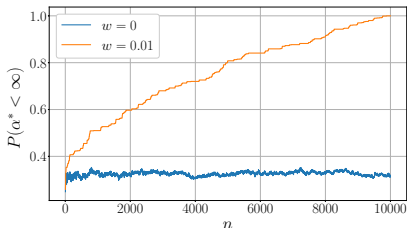
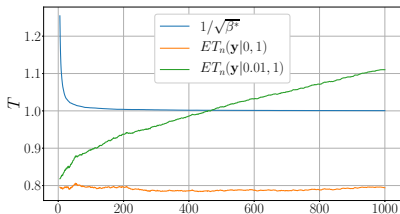
Пример оптимизации evidence

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

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

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

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



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

- 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>