

# Прикладной статистический анализ данных Введение в байесовскую статистику

Андрей Грабовой

## Пример

Монетку подбросили 5 раз, и все 5 раз выпал орел. Какова вероятность выпадения решки?

## Пример

Монетку подбросили 5 раз, и все 5 раз выпал орел. Какова вероятность выпадения решки?

Подход на основе ММП (“фреквентистский”): посчитаем вероятность выпадения решки по выборке.

Ответ: 0.

## Пример

Монетку подбросили 5 раз, и все 5 раз выпал орел. Какова вероятность выпадения решки?

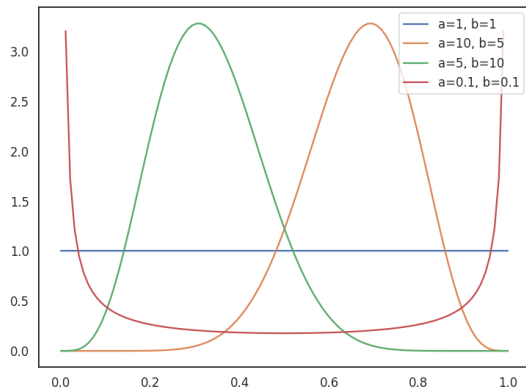
Подход на основе ММП (“фреквентистский”): посчитаем вероятность выпадения решки по выборке.

Ответ: 0.

**Проблема:** выборка слишком мала, чтобы делать такие поспешные выводы о выпадении решки. Кроме того, мы можем предположить что монетка должна давать более-менее равномерные результаты (это наши априорные предположения).

## Бета-распределение

- ▶ соответствует *априорным* ожиданиям о распределении Бернулли
- ▶ при  $n \rightarrow \infty$  сходится к  $\delta$ -распределению в точке ОМП распределения Бернулли.



## Байесовский подход

Введем бета-распределение в качестве *априорного* предположения о распределении нашего параметра. Из общих соображений распределение должно быть симметрично (если у нас нет дополнительной информации):

$$p(w) \sim B(\alpha, \beta).$$

Найдем *апостериорное* распределение параметра  $w$  распределения Бернулли по формуле Байеса:

$$p(w|\mathbf{x}) = \frac{p(\mathbf{X}|w)p(w)}{p(\mathbf{X})} \propto p(\mathbf{X}|w)p(w);$$

$$\log p(w|\mathbf{x}) = \log p(\mathbf{X}|w) + \log p(w) + \text{Const.}$$

**Вывод:** грубая интерпретация априорного распределения — *регуляризатор*.

## Байесовский вывод: первый уровень

Заданы:

- ▶ правдоподобие  $p(\mathbf{X}|\mathbf{w})$  выборки  $\mathbf{X}$  при условии параметра  $\mathbf{w}$ ;
- ▶ априорное распределение  $p(\mathbf{w}|\mathbf{h})$
- ▶ параметры априорного распределения  $\mathbf{h}$  (В примере с монеткой:  $\mathbf{h} = [\alpha, \beta]$ );

Тогда апостериорное распределение параметров  $\mathbf{w}$  при условии выборки  $\mathbf{X}$ :

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

Точечная оценка параметров находится как максимум апостериорной вероятности (MAP):

$$\hat{\mathbf{w}} = \arg \max p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{h}).$$

MAP-оценка схожа с оценкой методом максимального правдоподобия, если

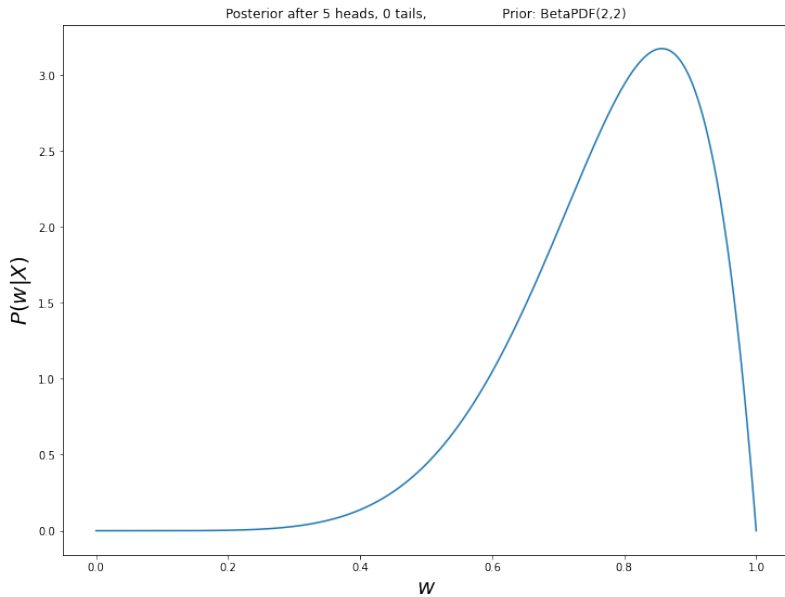
- ▶ Мощность выборка велика
- ▶ Априорное распределение — равномерное на очень большой области

## Байесовская статистика: пример

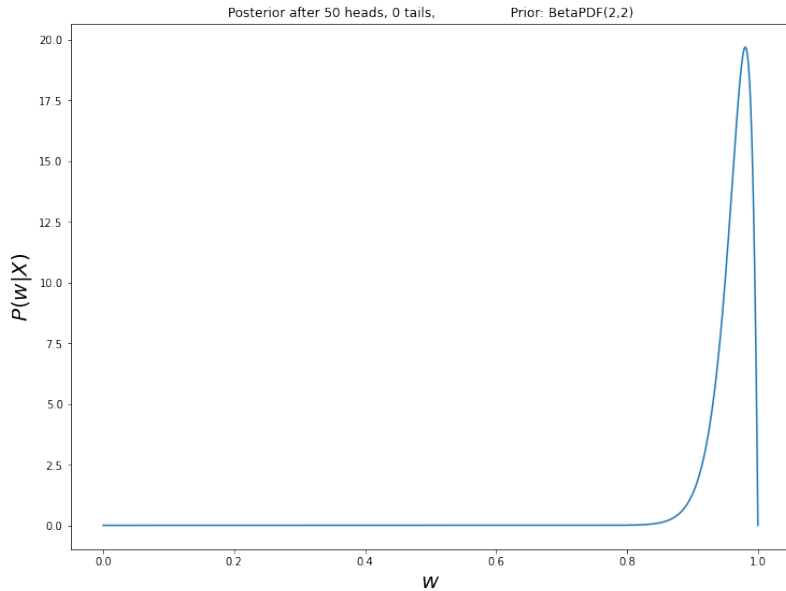
Предположим, что параметр нашей модели (монетки) — случайная величина.  
Возьмем в качестве распределения модели — бета-распределение с параметрами 2,2.

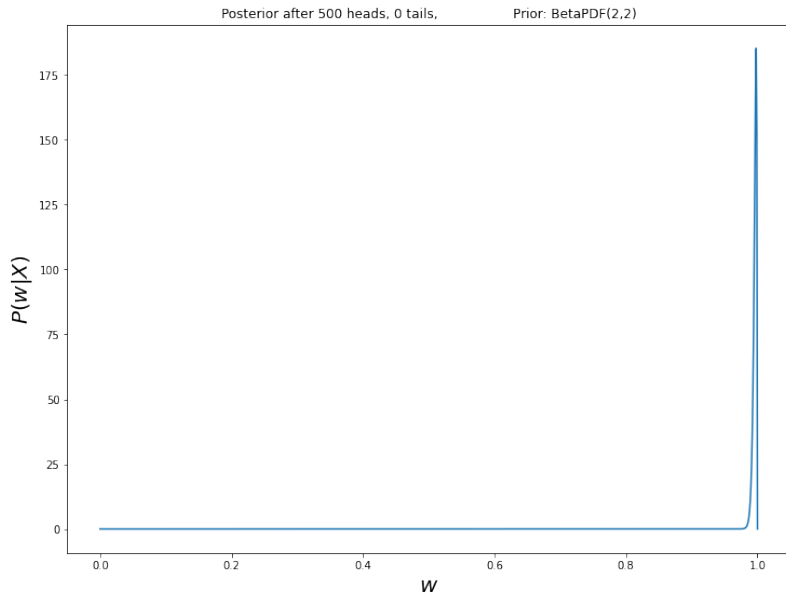
- ▶  $p(w) \sim \mathcal{B}(2, 2)$  — априорное распределение
- ▶  $p(X|w)$  — правдоподобие
- ▶  $p(w|X) = \frac{p(w)p(X|w)}{p(X)}$  — апостериорное распределение параметра.





В качестве ответа на вопрос из примера мы можем взять максимум апостериорной вероятности.





Почему для монетки подошло бета-распределение?

$$\begin{aligned} p(w|\mathbf{x}, \alpha, \beta) &\propto p(\mathbf{X}|w)p(w|\alpha, \beta) \propto \\ &\propto w^{\sum x} (1-w)^{m-\sum x} \times w^{\alpha-1} (1-w)^{\beta-1} = \\ &= w^{\alpha-1+\sum x} (1-w)^{m+\beta-\sum x-1} \sim B(\alpha + \sum x, \beta + m - \sum x). \end{aligned}$$

Семейство распределений называется сопряженным к распределению правдоподобия, если апостериорное распределение принадлежит этому же семейству.

## Формальная постановка

$$\hat{w} = \arg \max \frac{p(X|w)p(w)}{p(X)},$$

- ▶  $p(w) \sim \mathcal{B}(2, 2)$  — априорное распределение, соответствующие нашим ожиданиям относительно параметра.
- ▶  $p(X|w)$  — правдоподобие.
- ▶  $p(w|X) = \frac{p(w)p(X|w)}{p(X)}$  — апостериорное распределение параметра.
- ▶  $\hat{w}$  — оценка, полученная методом максимума апостериорной вероятности (MAP).
- ▶  $p(X)$  — обоснованность модели (“Evidence”) — насколько модель хорошо описывает выборку при разных значениях параметров.

## Как назначаются априорные распределения

Априорные распределения назначаются на основе априорных ожиданий от поведения модели.

**Назначение априорного распределения, которое противоречит гипотезе о порождении данных — некорректно**

Некоторые виды априорного распределения:

- ▶ Равномерное
- ▶ Равномерное неограниченное
- ▶ На основе предыдущих экспериментов
- ▶ Для сдвигов
  - ▶ Нормальное распределение
  - ▶ Распределение Лапласа
- ▶ Для масштаба
  - ▶ Гамма и обратное гамма-распределение
  - ▶ Коши (и производные)

## Распределение Джеффирса

Распределение соответствует объему информации, хранимому в выборке относительно параметров:

$$p(w) \propto \sqrt{\det I(w)},$$

$I(w)$  — информация Фишера:

$$I(w) \equiv -\frac{\partial^2}{\partial w^2} \log L(w).$$

- ▶ Инвариантно относительно замены переменных;
- ▶ Для среднего в нормальном распределении:  $p(w) \propto 1$ ;
- ▶ Для отклонения в нормальном распределении:  $p(w) \propto \frac{1}{w}$ ;
- ▶ Для параметра в распределении Бернулли:  $p(w) \propto \frac{1}{\sqrt{p(1-p)}}$ .

## Проблемы настройки параметров

Если матрица  $X$  вырождена, некоторые коэффициенты модели не будут определены.

Если наблюдения  $y = 0$  и  $y = 1$  линейно разделимы в пространстве  $X$ , то:

- ▶ в теории коэффициенты бесконечно возрастают
- ▶ на практике коэффициенты и их дисперсии получаются большими, а почти все вероятности в обучающей выборке близки к 0 или 1.

Можно использовать регуляризацию Фирта. Функция меток исходной модели для коэффициента  $\beta_j$ :

$$\sum_{i=1}^n (y_i - \pi(x_i)) x_{ij}.$$

Регуляризованная версия:

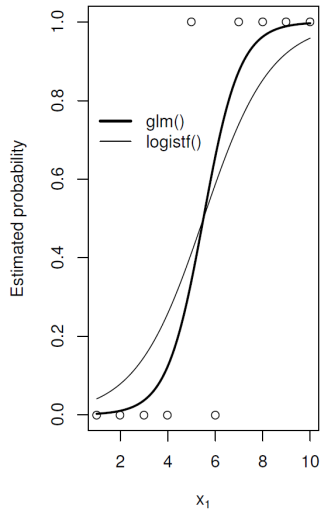
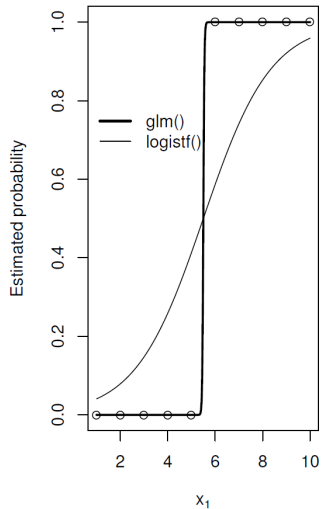
$$\sum_{i=1}^n (y_i - \pi(x_i) + h_i (0.5 - \pi(x_i))) x_{ij},$$

$h_i$  — диагональный элемент hat matrix:

$$H = V^{1/2} X \left( X^T V X \right)^{-1} X^T V^{1/2}.$$



## Проблемы настройки параметров

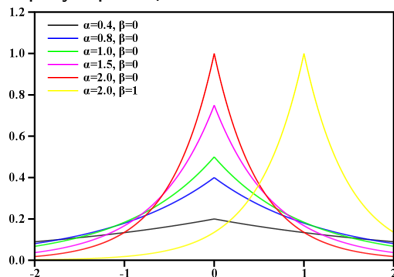
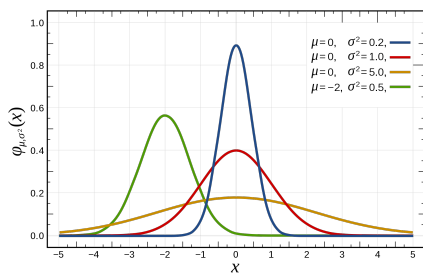


## Нормальное распределение

$$\log p(w|X) \propto \log p(w)p(X|w) \propto \log p(X|w) - \frac{(w - \mu)^2}{2\sigma^2}$$

Получили  $l_2$ -регуляризацию.

При распределении Лапласа получаем  $l_1$ -регуляризацию.



## Informative prior vs Uninformative prior

- ▶ Informative prior: соответствует экспертным знаниям о наблюдаемой переменной
  - ▶ Пример: температура воздуха: нормальная величина с известным средним и дисперсией, соответствующими прошлым наблюдениям.
- ▶ Uninformative prior: соответствует базовым предположениям о распределении переменной
  - ▶ Пример: температура воздуха: равномерное распределение (improper).
- ▶ Weakly-informative prior: где-то по середине
  - ▶ Пример: температура воздуха: равномерное распределение от -50 до +50.

## Напоминание: Интервальные оценки

Доверительный интервал:

$$\mathbf{P}(\theta \in [C_L, C_U]) \geq 1 - \alpha,$$

$1 - \alpha$  — уровень доверия,

$C_L, C_U$  — нижний и верхний доверительные пределы.

**Неверная интерпретация:** неизвестный параметр лежит в пределах построенного доверительного интервала с вероятностью  $1 - \alpha$ .

**Верная интерпретация:** при бесконечном повторении процедуры построения доверительного интервала на аналогичных выборках в  $100(1 - \alpha)\%$  случаев он будет содержать истинное значение  $\theta$ .

Напоминание: для нормального распределения

$$X \sim N(\mu, \sigma^2), \quad X^n = (X_1, \dots, X_n),$$

$\bar{X}_n$  — оценка  $\mathbb{E}X = \mu$ ,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow$$

$$\mathbf{P}\left(\mu - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \Rightarrow$$

доверительный интервал для  $\mu$ :

$$\mathbf{P}\left(\bar{X}_n - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

$z_{1-\frac{\alpha}{2}}$  — квантиль стандартного нормального распределения.

## Байесовская интервальные оценка (credible interval)

Доверительный интервал:

$$\int_{w \in [C_L, C_U]} p(w|X) \geq 1 - \alpha,$$

$1 - \alpha$  — уровень доверия,

$C_L, C_U$  — нижний и верхний доверительные пределы.

**Интерпретация:** неизвестный параметр, породивший выборку, лежит в пределах построенного доверительного интервала с вероятностью  $1 - \alpha$ .

- Доверительные интервалы совпадают для параметров сдвига с равномерным распределением и масштаба с распределением Джеффриса.

## Пример

$$X \sim \mathcal{N}(0, 1), |X| = 10, \bar{X} = 0.17.$$

- ▶ Доверительный интервал:  $[-0.45, 0.78]$
- ▶ Prior:  $\mu \sim \mathcal{N}(0, 0.01)$ , доверительный интервал:  $[-0.002, 0.03]$ .

$$X \sim \mathcal{N}(0, 1), |X| = 100000, \bar{X} = 0.17.$$

- ▶ Доверительный интервал:  $[0.1638, 0.1763]$
- ▶ Prior:  $\mu \sim \mathcal{N}(0, 0.01)$ , доверительный интервал:  $[0.16981, 0.16984]$ .

Задан набор моделей, требуется определить, какой из них лучше подходит для работы с выборкой.

- ▶ Линейная модель —  $R^2$  и пр.
- ▶ Обобщенно-линейная модель — остаточная аномальность.
- ▶ Что делать, если модель нелинейная?
- ▶ Что делать с параметрами априорного распределения, как их выбирать?



$$AIC = -2L + 2(k + 1),$$

где  $k$  — количество параметров.

Критерий соответствует потере в информации относительно истинного распределения данных:

$$AIC \approx KL(f|f_i),$$

где  $f$  — истинное распределение,  $f_i$  — модель-кандидат, описывающий распределение.

## Связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

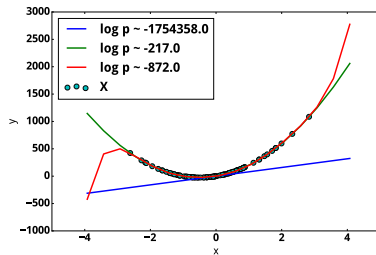
$$w = \arg \max \frac{p(X|w)p(w)}{p(X)},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели. Обоснованность модели (“Evidence”):

$$p(X) = \int_w p(X|w)p(w)dw.$$



(a) Схема выбора модели



(b) Пример: полиномы

## Принцип минимальной длины описания

$$\text{MDL}(\mathbf{f}, \mathfrak{D}) = L(\mathbf{f}) + L(\mathfrak{D}|\mathbf{f}),$$

где  $\mathbf{f}$  — модель,  $\mathfrak{D}$  — выборка,  $L$  — длина описания в битах.

Аппроксимация этой величины для достаточно большой мощности выборки  $n$ :

$$BIC = -2L + \log n(k + 1).$$

$$\text{MDL}(\mathbf{f}, \mathfrak{D}) \sim L(\mathbf{f}) + L(\mathbf{w}^*|\mathbf{f}) + L(\mathfrak{D}|\mathbf{w}^*, \mathbf{f}),$$

$\mathbf{w}^*$  — оптимальные параметры модели.

$\mathbf{f}_1$	$L(\mathbf{f}_1)$	$L(\mathbf{w}_1^* \mathbf{f}_1)$	$L(\mathfrak{D} \mathbf{w}_1^*, \mathbf{f}_1)$
$\mathbf{f}_2$	$L(\mathbf{f}_2)$	$L(\mathbf{w}_2^* \mathbf{f}_2)$	$L(\mathfrak{D} \mathbf{w}_2^*, \mathbf{f}_2)$
$\mathbf{f}_3$	$L(\mathbf{f}_3)$	$L(\mathbf{w}_3^* \mathbf{f}_3)$	$L(\mathfrak{D} \mathbf{w}_3^*, \mathbf{f}_3)$

**Колмогоровская сложность** — длина минимального кода для выборки на предварительно заданном языке.

### **Теорема инвариантности**

Для двух сводимых по Тьюрингу языков колмогоровская сложность отличается не более чем на константу, не зависящую от мощности выборки.

### **Отличия от MDL:**

- ▶ Колмогоровская сложность невычислима.
- ▶ Длина кода может зависеть от выбранного языка. Для небольших выборок теорема инвариантности не дает адекватных результатов.

## Evidence vs MDL

Evidence	MDL
Использует априорные знания	Независима от априорных знаний
Основывается на гипотезе о порождении выборки вне зависимости от их природы	Минимизирует длину описания выборки

## Как считать Evidence?

- ▶ Для линейных моделей: аналитическая формула
- ▶ Для нелинейных моделей аппроксимация Лапласа:

$$p(X) = \int_w p(X, w) dw$$

- ▶ Разложим  $\log p(X|w)$  в ряд Тейлора:

$$\log p(X|w) \approx \log p(X, w_0) - \frac{\partial^2}{2\partial w^2} \log p(X, w_0)(w - w_0)^2.$$

- ▶ Вычислим интеграл для ненормированной гауссовой величины

$$p(x, w_0) \exp\left(-\frac{\partial^2}{2\partial w^2} \log p(X, w_0)(w - w_0)^2\right).$$

- ▶ MCMC, вариационный вывод и пр.

## Пример: линейная регрессия

Линейный случай с  $m$  объектами и  $n$  признаками:  $\mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$ ;

$\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1})$ ,  $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$ .

Запишем интеграл:

$$\begin{aligned} p(\mathcal{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) &= \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp\left(-0.5\beta(\mathbf{y} - \mathbf{f})^\top(\mathbf{y} - \mathbf{f})\right) \exp\left(-0.5\mathbf{w}^\top \mathbf{A} \mathbf{w}\right) d\mathbf{w} = \\ &= \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w} \end{aligned}$$

Для линейного случая интеграл вычисляется аналитически:

$$\int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w} = (2\pi)^{\frac{n}{2}} \exp(-S(\hat{\mathbf{w}})) |\mathbf{H}^{-1}|^{0.5},$$

где

$$\mathbf{H} = \mathbf{A} + \beta \mathbf{X}^\top \mathbf{X},$$

$$\hat{\mathbf{w}} = \beta \mathbf{H}^{-1} \mathbf{X}^\top \mathbf{y}$$

**Вывод:** для линейных моделей Evidence считается аналитически.

## Пример: аппроксимация Лапласа

Нелинейный случай с  $m$  объектами и  $n$  признаками:  $\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1})$ ,  $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$ .  
Запишем интеграл:

$$p(\mathcal{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) = \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w}.$$

Разложим  $S$  в ряд Тейлора:

$$S(\mathbf{w}) \approx S(\hat{\mathbf{w}}) + \frac{1}{2} \Delta \mathbf{w}^T \mathbf{H} \Delta \mathbf{w}$$

Интеграл приводится к виду:

$$\frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} S(\hat{\mathbf{w}}) \int_{\mathbf{w}} \exp(-\frac{1}{2} \Delta \mathbf{w}^T \mathbf{H} \Delta \mathbf{w}) d\mathbf{w}$$

Выражение под интегралом соответствует плотности ненормированного нормального распределения.

**Вывод:** для нелинейных моделей можно использовать аппроксимацию Лапласа для получения оценок Evidence.



- ▶ MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- ▶ Bishop C. M. Pattern recognition and machine learning. – springer, 2006.
- ▶ <https://www.thomasjpfan.com/2015/09/bayesian-coin-flips/>
- ▶ <https://people.stat.sc.edu/Hitchcock/stat535slidesday3.pdf>
- ▶ Лекции Д. П. Ветрова на <http://www.machinelearning.ru>
- ▶ Kuznetsov M., Tokmakova A., Strijov V. Analytic and stochastic methods of structure parameter estimation // Informatica. – 2016. – Т. 27. – №. 3. – С. 607-624.
- ▶ Пример с монеткой: <https://towardsdatascience.com/visualizing-beta-distribution-7391c18031f1>
- ▶ Немного про распределение Джеффриса:  
<https://medium.datadriveninvestor.com/firths-logistic-regression-classification-with-datasets-that-are-small-imbalanced-or-separated-49d7782a13f1>