Прикладной статистический анализ данных

Введение в байесовскую статистику

Олег Бахтеев
psad@phystech.edu

2022

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 (“фреквентисткий”): посчитаем вероятность выпадения решки по выборке.
Ответ: 0.

Пример

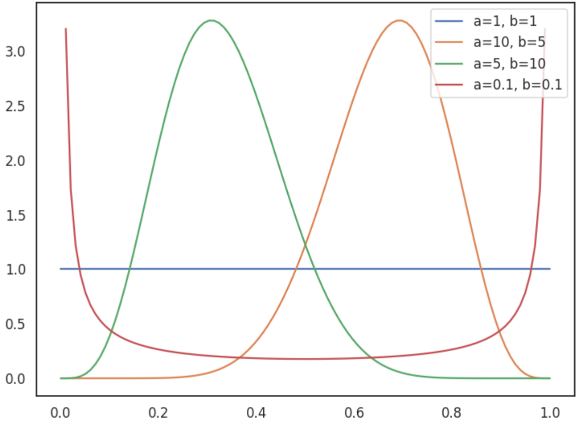
Монетку подбросили 5 раз, и все 5 раз выпал орел. Какова вероятность выпадения решки?

Подход на основе ММП (“фреквентисткий”): 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 $n \to \infty$ сходится к δ-распределению в точке ОМП распределения Бернулли.



Байесовский подход

Введем бета-распределение в качетсве *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* распрделение параметра w распределения Бернулли по формуле Байеса:

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

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

Вывод: грубая интерпретация априорного распределения — регуляризатор.

Байесовский вывод: первый уровень

Заданы:

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

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

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

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

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

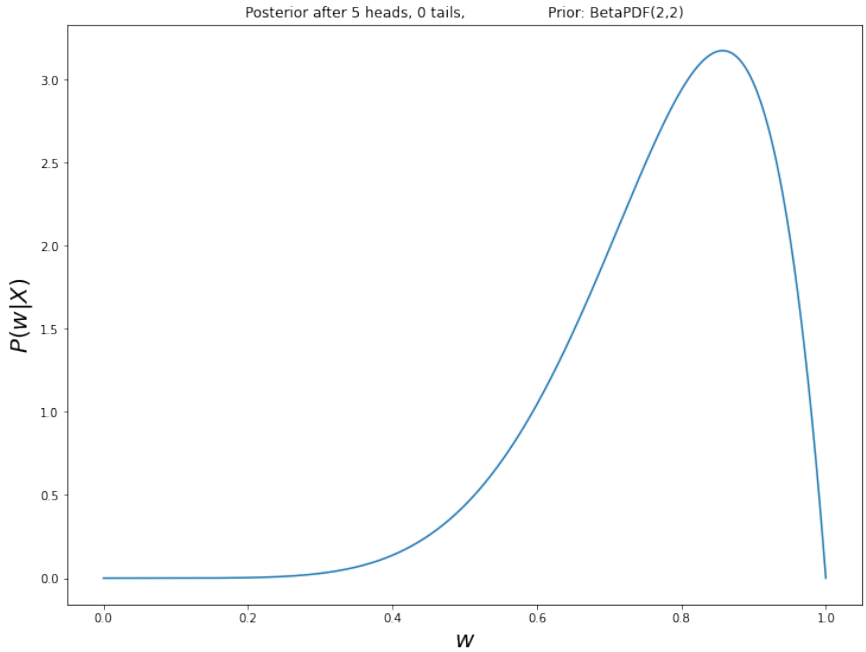
MAP-оценка схожа с оценкой методом максимального правдоподобия, если

- Мощность выборка велика
- Априорное распределенеие — равномерное на очень большой области

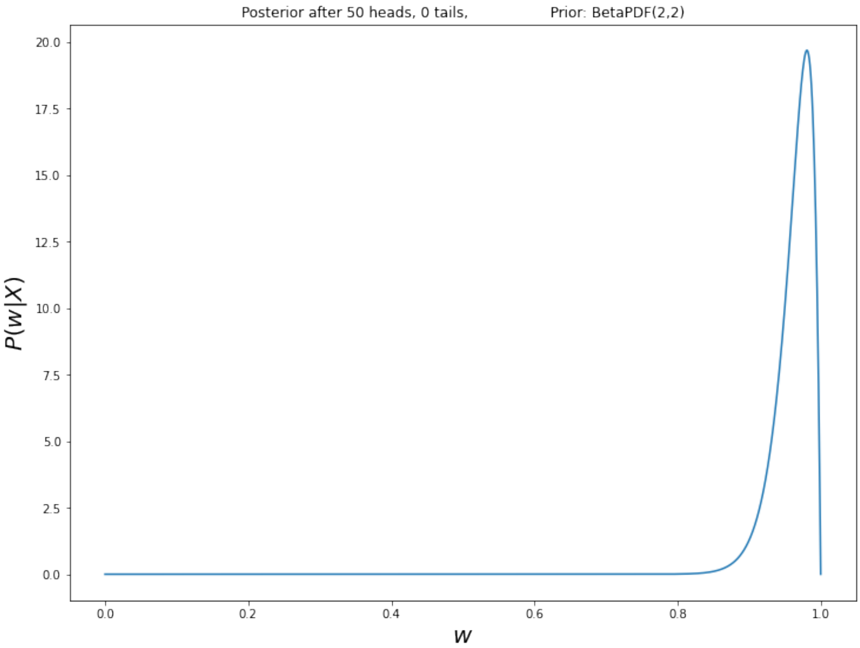
Байесовская статистика: пример

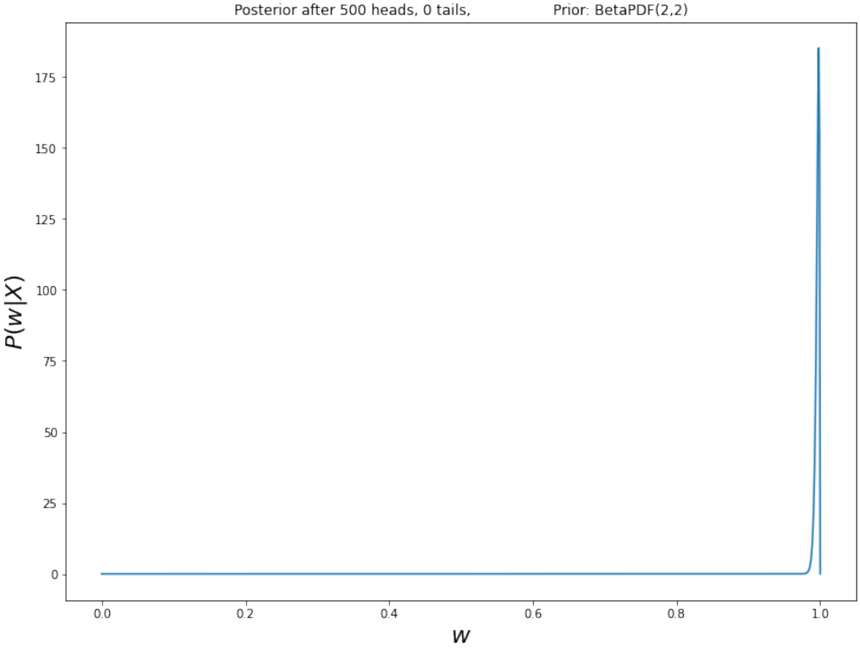
Предположим, что параметр нашей модели (монетки) — случайная величина.
Возьмем в качестве распределения модели — бета-распределение с параметрами 2,2.

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



Posterior after 5 heads, 0 tails,
Prior: BetaPDF(2,2)
P(w|X)
w
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0


Posterior after 50 heads, 0 tails,
Prior: BetaPDF(2,2)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
P(w|X)
0.0
0.2
0.4
0.6
0.8
1.0
w


Posterior after 500 heads, 0 tails, Prior: BetaPDF(2,2)
175
150
125
100
75
50
25
0
P(w|X)
0.0
0.2
0.4
0.6
0.8
1.0
w

Почему для монетки подошло бета-распределение?

$$p(w|\boldsymbol{x}, \alpha, \beta) \propto {\color{blue}p(\boldsymbol{X}|w)}{\color{red}p(w|\alpha, \beta)} \propto$$
$$\propto w^{\sum x}(1-w)^{m-\sum x} \times w^{\alpha-1}(1-w)^{\beta-1} =$$
$$= w^{\alpha-1+\sum x}(1-w)^{m+\beta-\sum x-1} \sim B(\alpha + \sum x, \beta + m - \sum x).$$

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 выбокре 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 β_j:

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

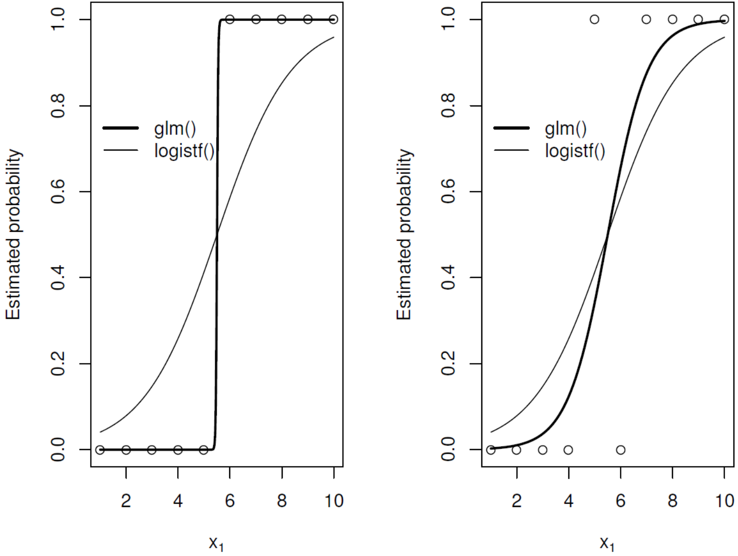
Регуляризованная версия:

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

h_i — диагональный элемент hat matrix:

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

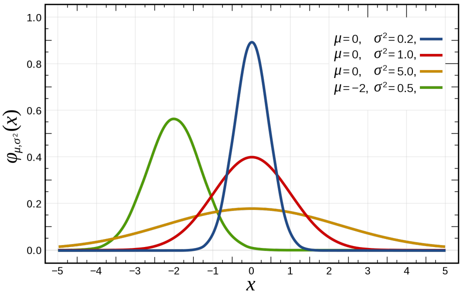
Проблемы настройки параметров

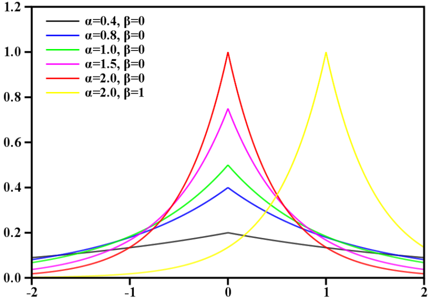


Нормальное распределение

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

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:

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

$1 - \alpha$ — 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 θ.

Напоминание: для нормального распределения

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

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

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

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

доверительный интервал для μ:

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

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

Байесовская интервальные оценка (credible interval)

Доверительный интервал:

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

$1-\alpha$ — 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].

Выбор моделей

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

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

где k — количество параметров.
Критерий соответствует потерю в информации относительно истинного распределения данных:

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

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

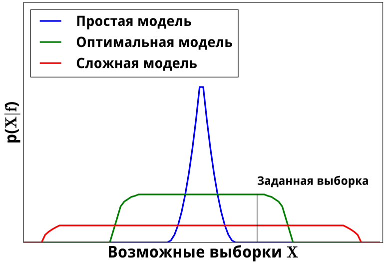
Связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

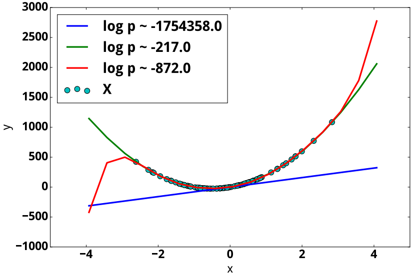
$$w = \arg\max \frac{p(X|w)p(w)}{p(X)},$$

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

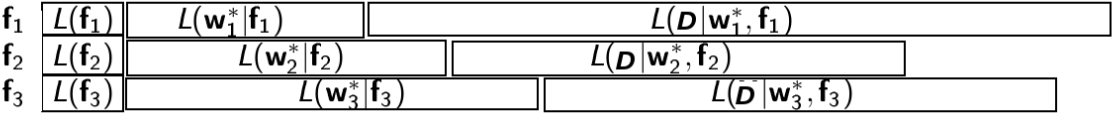
$$\text{MDL}(\mathbf{f}, \mathfrak{D}) = L(\mathbf{f}) + L(\mathfrak{D}|\mathbf{f}),$$

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



MDL и Колмогоровская сложность

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$(-\frac{\partial^2}{2\partial w^2} \log p(X, w_0)(w - w_0)^2)$.
- MCMC, вариационный вывод и пр.

Пример: линейная регрессия

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

$$p(\mathfrak{D}|\boldsymbol{h}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{A}, \beta) = \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \text{exp}\left(-0.5\beta(\boldsymbol{y} - \boldsymbol{f})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{f})\right) \text{exp}\left(-0.5\boldsymbol{w}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{w}\right) d\boldsymbol{w} =$$

$$= \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \text{exp}(-S(\boldsymbol{w}))d\boldsymbol{w}$$

Для линейного случая интеграл вычисляется аналитически:

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

где

$$\boldsymbol{H} = \boldsymbol{A} + \beta\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X},$$
$$\hat{\boldsymbol{w}} = \beta\boldsymbol{H}^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

Вывод: для линейных моделей Evidence считается аналитически.

Пример: аппроксимация Лапласа

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

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

Разложим S в ряд Тейлора:

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

Интеграл приводится к виду:

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

Выражение под интегралом соответствует плотности ненормированного нормального распределения.
Вывод: для нелинейных моделей можно использовать аппроксимацию Лапласа для получения оценок Evidence.

Литература

- 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. – T. 27. – №. 3. – C. 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