

Прикладной статистический анализ данных

Введение в байесовскую статистику

Олег Бахтеев
psad@phystech.edu

2021

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?

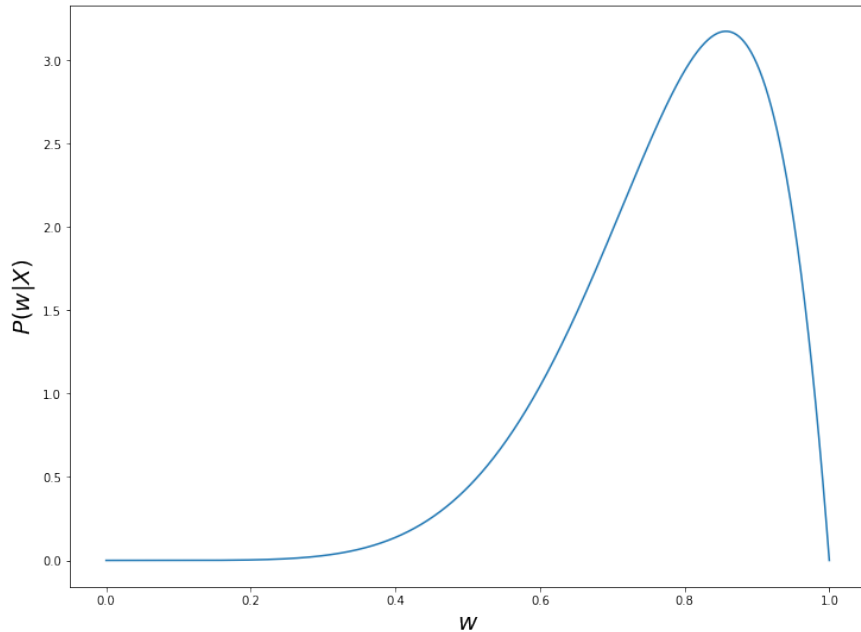
Подход на основе ММП (“фреквентистский”): посчитаем вероятность выпадения решки по выборке.
Ответ: 0.

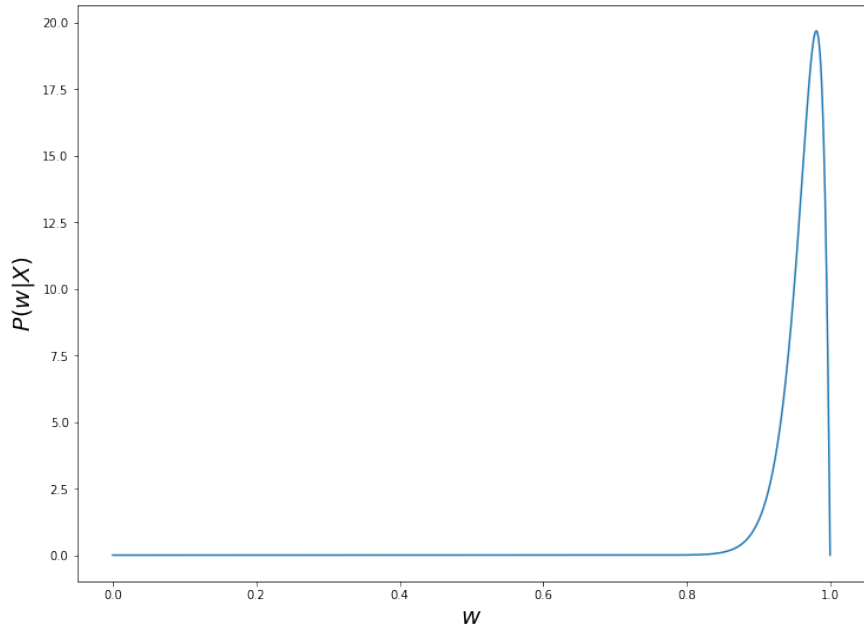
Проблема: выборка слишком мала, чтобы делать такие поспешные выводы о выпадении решки. Кроме того, мы можем предположить что монетка должна давать более-менее равномерные результаты (это наши априорные предположения).

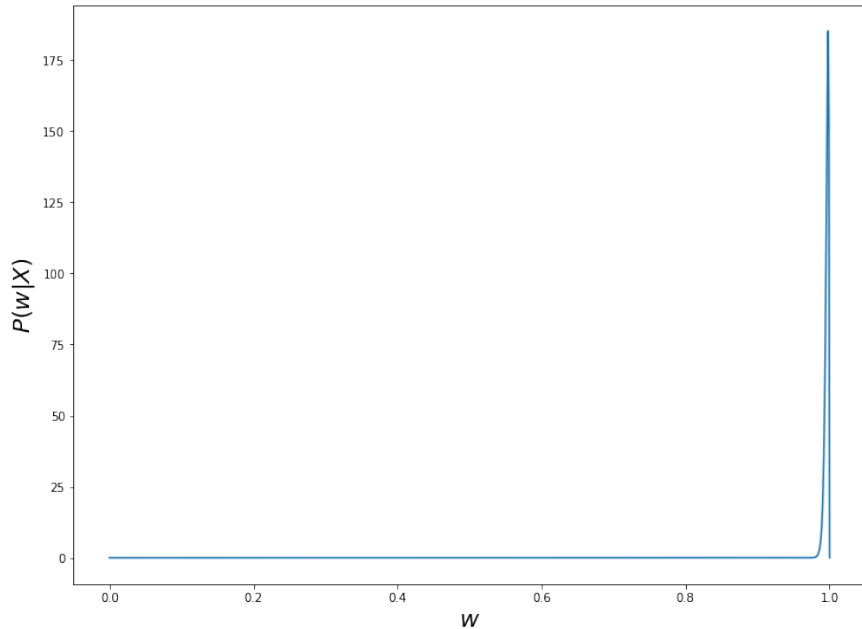
Байесовская статистика: пример

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.







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:

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

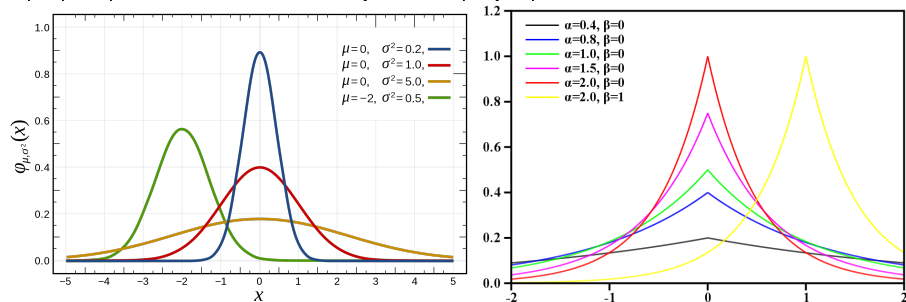
- Инвариантно относительно замены переменных;
- Для среднего в нормальном распределении: $p(w) \propto 1$;
- Для отклонения в нормальном распределении: $p(w) \propto \frac{1}{w}$;
- Для параметра в распределении Бернулли: $p(w) \propto \frac{1}{\sqrt{p(1-p)}}$.

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 θ .

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 μ :

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

$z_{1-\frac{\alpha}{2}}$ — 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: $[0.16981, 0.16984]$.

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 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

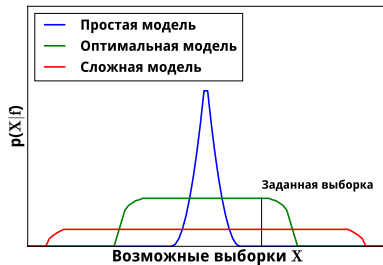
Первый уровень: выбираем оптимальные параметры:

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

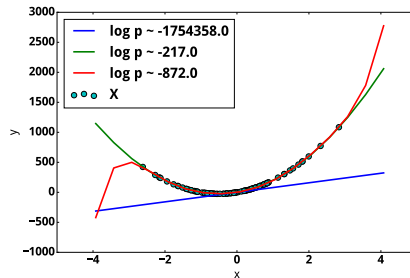
Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.

Обоснованность модели ("Evidence"):

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



(а) Схема выбора модели



(б) Пример: полиномы

Принцип минимальной длины описания

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

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

Аппроксимация этой величины для достаточно большой мощности выборки n :

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

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

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

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

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.

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>

Сэмплирование по значимости

Пусть мы не можем сэмплировать из $p(z)$, но можем оценивать правдоподобие в каждой точке, и хотим получить интерал

$$\mathbb{E}f = \int f(z)p(z)dz.$$

Тогда введем распределение q :

$$\mathbb{E}f = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{L} \sum_{l=1}^L \frac{p(z^l)}{q(z^l)} f(z^l).$$

Орг моменты

Алгоритм получения зачета:

- ① Заполнить форму (будет на гитхабе)
- ② Посмотреть свою предварительную оценку в таблице
- ③ Варианты развития событий
 - ▶ Оценка \geq “Хор” и устраивает: ничего не делать (автомат)
 - ▶ Оценка не устраивает: сдавать зачет
 - ▶ Оценка $<$ “Хор”: сдавать зачет
 - ▶ Оценка $<$ “Хор” и зачет не сдан: неявка

Запись на зачет будет осуществляться через гугл-диск с расписанием