Байесовские методы. Лекция 2.
Оценка параметров в классической статистике.
Байесовский вывод для простейших моделей.

Целищев М.А.

МГУ им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

весна 2021

Список литературы

- J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
- R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan. Chapman and Hall CRC, 2015.
- O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
- K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
- A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
- S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

Оценка параметров в статистике

Одна из задач математической статистики выглядит так:
$\xi \sim F_\xi \in \{\, F_\theta \ : \theta \in \Theta \,\}$. Нужно оценить параметр θ по выборке:

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \xi.$$

А именно, нужно построить статистику $T(X) = T(X_1, \ldots, X_n)$, которая бы служила оценкой неизвестного параметра θ.
Пример. Среди n привитых новой вакциной антитела выработались у k человек. Требуется оценить вероятность выработки антител θ.
Здесь $\xi \sim \mathbf{Be}(\theta)$, $\theta \in [0, 1]$, и X_i — индикатор того, что у i-го испытуемого выработались антитела.
В качестве оценки параметра θ здесь разумно рассмотреть долю успехов:

$$T(X) = \frac{\sum\limits_{i=1}^{n} X_n}{n} = \frac{k}{n}.$$

Как решаются такие задачи в общем случае?

Функция потерь, риск

$\xi \sim F_\xi \in \{ F_\theta \,:\, \theta \in \Theta \}$. $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \xi$.

Вводят так называемую *функцию потерь* $L\left(T(X), \theta\right)$ (loss function), которая имеет физический смысл потерь (или убытков) в случае, когда истинное значение параметра θ оценивается с помощью статистики $T(X)$ для конкретной реализации выборки X.

Если $\Theta \subset \mathbb{R}$, то часто рассматривают квадратичную функцию потерь:

$$L(T(X), \theta) = (\theta - T(X))^2.$$

Функцией риска называют средние потери:

$$R(T, \theta) = \mathbf{E}_\theta\, L(T(X), \theta).$$

Зачем нужна функция риска?

Упорядочение статистик

Обычно фиксируют некоторый класс функций $\mathcal{T}$ и среди его элементов пытаются выбрать в некотором смысле лучшую оценку.

Например, в качестве $\mathcal{T}$ в случае $\Theta \subset \mathbb{R}$ часто берут класс *несмещённых* оценок:

$$\mathcal{T} = \left\{ T\colon \mathbb{R}^n \to \Theta \,\middle|\, \mathbf{E}_\theta T(X) = \theta \quad \forall \theta \in \Theta \right\}.$$

Заметим, что при этом в случае квадратичной функции потерь

$$R(T,\theta) = \mathbf{E}_\theta(\theta - T(X))^2 = \mathbf{D}_\theta T(X) \quad \forall T \in \mathcal{T} \quad \forall \theta \in \Theta.$$

Если есть две функции $T_1, T_2 \in \mathcal{T}$, то можно сравнивать статистики $T_1(X)$ и $T_2(X)$, исходя из их функций риска $R(T_1,\theta)$ и $R(T_2,\theta)$.

Определение

Статистика $T_* \in \mathcal{T}$ называется *оптимальной* оценкой параметра θ в классе $\mathcal{T}$ с функцией потерь L, если

$$R(T_*, \theta) \leqslant R(T, \theta) \quad \forall \theta \in \Theta \quad \forall T \in \mathcal{T}.$$

Например, для задачи $\xi \sim \mathrm{Be}(\theta)$, $\theta \in [0, 1]$, оптимальной в среднем квадратическом среди несмещённых оценок будет как раз доля успехов:

$$T_*(X) = \overline{X} \equiv \frac{\sum_{i=1}^{n} X_i}{n}.$$

Это доказывалось в курсе мат. статистики двумя способами:

- с помощью неравенства Рао-Крамера и понятия эффективной оценки
- с помощью теоремы Колмогорова и понятия полных и достаточных статистик

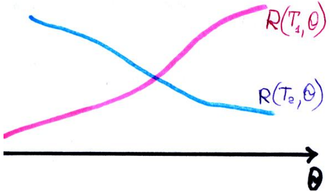
Оптимальные оценки

Определение

Статистика $T_* \in \mathcal{T}$ называется *оптимальной* оценкой параметра θ в классе $\mathcal{T}$ с функцией потерь L, если

$$R(T_*, \theta) \leqslant R(T, \theta) \quad \forall \theta \in \Theta \quad \forall T \in \mathcal{T}.$$

Но при таком подходе есть проблема...
Поскольку в определении неравенство выполняется сразу для всех допустимых θ, то некоторые статистики T_1 и T_2 могут быть несравнимы...



И может такое случиться, что оптимальной оценки вообще не существует...

Оценки максимального правдоподобия

... Но обычно это не останавливает статистиков. На практике чаще всего статистики ищут не оптимальные оценки, а *оценки максимального правдоподобия* (МП):

$$\hat{\theta}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{ML}}(X) = \arg\max_{\theta \in \Theta} \mathcal{L}(X, \theta),$$

где $\mathcal{L}(\mathbf{x}, \theta)$ — функция *правдоподобия* (likelihood function), то есть совместная плотность выборки:

$$\mathcal{L}(\mathbf{x}, \theta) = p_\theta(x_1) \cdot \ldots \cdot p_\theta(x_n), \quad \mathbf{x} \in \mathbb{R}^n,\ \theta \in \Theta,$$

где $p_\theta(x)$ — либо плотность с.в. ξ (в абсолютно непрерывном случае), либо $p_\theta(x) = \mathbf{P}_\theta(\xi = x)$ в дискретном случае.

ДЗ: показать, что МП-оценка появляется естественным образом, если в качестве функции потерь взять дираковскую функцию: $L(T(X), \theta) = -\delta_\theta(T(X))$.

Свойства оценок МП

Из курса статистики известно, что при достаточно мягких ограничениях оценки максимального правдоподобия удовлетворяют свойствам:

- асимптотической несмещённости:

$$\mathbf{E}_\theta\, \hat{\theta}_{\mathrm{ML}}(X) \to \theta \quad \text{при} \quad n \to \infty, \quad \forall \theta \in \Theta,$$

- состоятельности:

$$\hat{\theta}_{\mathrm{ML}}(X) \xrightarrow{\mathbf{P}_\theta} \theta \quad \text{при} \quad n \to \infty, \quad \forall \theta \in \Theta,$$

- асимптотической оптимальной в среднем квадратическом, то есть при больших объёмах выборки её дисперсия — наименьшая.

Кстати, в задаче с антителами оценка МП совпадает с оптимальной в среднем квадратическом (т.е. с долей успехов).

Увы, все эти замечательные свойства полезны только при больших n.
А при малых n любая точечная оценка выглядит сомнительно...

Байесовский подход

С точки зрения байесовского подхода, параметр θ — это не неизвестное число, а случайная величина с *априорным* распределением $p(\theta)$.

При этом $p_\theta(x)$, то есть плотность распределения с.в. ξ, можно рассматривать как условную плотность с.в. ξ при условии, что параметр принимает значение θ:

$$p_\theta(x) \equiv p(x|\theta).$$

Но тогда, зная функцию риска

$$R(T, \theta) = \mathbf{E}_\theta\, L(T(X), \theta) = \int_{\mathbb{R}^n} L\left(T(\mathbf{x}), \theta\right)\, p_\theta(\mathbf{x})\, d\mathbf{x},$$

можно линейно упорядочить ВСЕ допустимые оценки, усредняя риск по априорному распределению параметра θ.

Определение

Байесовским риском оценки T параметра θ называют

$$r(T) = \int_{\Theta} R(T, \theta)\, p(\theta)\, d\theta.$$

Тогда оптимальной в байесовском смысле будет такая оценка, которая минимизирует байесовский риск:

$$T_* = \underset{T \in \mathcal{T}}{\arg\min}\ r(T)$$

Такая оценка точно существует (т.к. оптимизируется числовой функционал, а не функция от θ), но уже зависит от вида априорного распределения. Иными словами, для разных априорных распределений она будет разной...

Заметим, что

$$r(T) = \int_\Theta R(T,\theta)\, p(\theta)\, d\theta = \int_\Theta \Big[\mathbf{E}_\theta L(T(X),\theta)\Big]\ p(\theta)\, d\theta =$$

$$= \int_\Theta \left(\int_{\mathbb{R}^n} L(T(\mathbf{x}),\theta)\, p_\theta(\mathbf{x})\, d\mathbf{x}\right) p(\theta)\, d\theta =$$

$$= \int_\Theta \left(\int_{\mathbb{R}^n} L(T(\mathbf{x}),\theta)\, p(\mathbf{x}|\theta)\, d\mathbf{x}\right) p(\theta)\, d\theta =$$

$$= \int_{\mathbb{R}^n} \int_\Theta L(T(\mathbf{x}),\theta)\, p(\mathbf{x},\theta)\, d\theta\, d\mathbf{x} =$$

$$= \int_{\mathbb{R}^n} \left(\int_\Theta L(T(\mathbf{x}),\theta)\, p(\theta|\mathbf{x})\, d\theta\right) p(\mathbf{x})\, d\mathbf{x},$$

где $p(\theta|\mathbf{x})$ называют апостериорным распределением параметра θ,
а $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)\, d\theta$ называют *обоснованностью* (evidence).

Итак,

$$r(T) = \int_{\mathbb{R}^n} \left(\int_{\Theta} L\left(T(\mathbf{x}), \theta\right)\, p(\theta|\mathbf{x})\, d\theta \right) p(\mathbf{x})\, d\mathbf{x} \quad \longrightarrow \quad \min_{T}.$$

Поэтому

$$T_*(\mathbf{x}) = \arg\min_{t} \int_{\Theta} L(t, \theta)\, p(\theta|\mathbf{x})\, d\theta = $$
$$= \arg\min_{t} \mathbf{E}\left[L(t, \theta) \middle| X = \mathbf{x} \right].$$

Иными словами, в байесовском подходе минимизируют функцию потерь, усреднённую по апостериорному распределению параметра θ.

Байесовский подход

В частности, если $\Theta \subset \mathbb{R}$ и $L(T(X),\theta) = (T(X)-\theta)^2$, то оптимальная байесовская оценка вычисляется совсем просто:

$$T_*(\mathbf{x}) = \mathbf{E}\,(\theta | X = \mathbf{x}) = \int_\Theta \theta\, p(\theta|\mathbf{x})\, d\theta,$$

то есть как среднее по апостериорному распределению параметра θ.

Доказательство.

$$\mathbf{E}\left[(T(X)-\theta)^2 \,\middle|\, X=\mathbf{x}\right] = \mathbf{E}\left[(T(X)-T_*(X)+T_*(X)-\theta)^2 \,\middle|\, X=\mathbf{x}\right] =$$

$$= \mathbf{E}\left[(T(X)-T_*(X))^2 \,\middle|\, X=\mathbf{x}\right] + \mathbf{E}\left[(T_*(X)-\theta)^2 \,\middle|\, X=\mathbf{x}\right] +$$

$$+2\,\mathbf{E}\left[(T(X)-T_*(X))\,(T_*(X)-\theta) \,\middle|\, X=\mathbf{x}\right],$$

где $\mathbf{E}\left[(T(X)-T_*(X))\,(T_*(X)-\theta) \,\middle|\, X=\mathbf{x}\right] =$

$$= (T(\mathbf{x})-T_*(\mathbf{x}))\,\mathbf{E}\left[T_*(X)-\theta \middle| X=\mathbf{x}\right] = (\ldots)\,(T_*(\mathbf{x})-T_*(\mathbf{x})) = 0.$$

Но в качестве оценки параметра θ не обязательно использовать среднее апостериорного распределения. Можно брать:

- моду апостериорного распределения (MAP = maximum a-posteriori):

$$\theta_{\mathsf{MAP}}(\mathbf{x}) = \arg\max_{\theta} p(\theta|\mathbf{x}) = \arg\max_{\theta} \frac{p(\mathbf{x}|\theta)\, p(\theta)}{p(\mathbf{x})} = \arg\max_{\theta} p(\mathbf{x}|\theta)\, p(\theta),$$

  которую можно рассматривать как оценку МП с весом $p(\theta)$
- медиану апостериорного распределения, т.е. такую точку $m(\mathbf{x})$, что $\int\limits_{-\infty}^{m} p(\theta|\mathbf{x})\, d\theta = 1/2$,
- ...

Да и вообще, зачем нужны точечные оценки параметра θ, если мы знаем ВСЁ апостериорное распределение? С его помощью можно строить доверительные интервалы для параметра θ, оценки для функций от параметра, и, самое главное, строить прогнозы для нового наблюдения X_{n+1} при условии, что известны наблюдения $X = (X_1, \ldots, X_n)$.

Байесовский прогноз

Хотим посчитать $p(x_{\text{new}}|\mathbf{x})$ — плотность распределения следующего наблюдения X_{new} в точке $x_{\text{new}} \in \mathbb{R}$ при условии, что значение выборки $X_1, \ldots, X_n$ равно $\mathbf{x} \in \mathbb{R}^n$.

$$p(x_{\text{new}}|\mathbf{x}) = \frac{p(x_{\text{new}}, \mathbf{x})}{p(\mathbf{x})} = \frac{\int\limits_{\Theta} p(x_{\text{new}}, \mathbf{x}|\theta)\, p(\theta)\, d\theta}{p(\mathbf{x})}$$

Поскольку наблюдение X_{new} не зависит от всех предыдущих при известном параметре θ (*условная независимость!*), то $p(x_{\text{new}}, \mathbf{x}|\theta) = p(x_{\text{new}}|\theta)\, p(\mathbf{x}|\theta)$, и

$$p(x_{\text{new}}|\mathbf{x}) = \int\limits_{\Theta} \frac{p(x_{\text{new}}, \mathbf{x}, \theta)}{p(\mathbf{x})}\, d\theta = \int\limits_{\Theta} p(x_{\text{new}}|\theta) \frac{p(\mathbf{x}|\theta) p(\theta)}{p(\mathbf{x})}\, d\theta = \int\limits_{\Theta} p(x_{\text{new}}|\theta)\, p(\theta|\mathbf{x})\, d\theta,$$

где последнее равенство следует из теоремы Байеса.

Таким образом, байесовский прогноз для новых данных есть правдоподобие $p(x_{\text{new}}|\theta)$, усреднённое по апостериорному распределению параметра θ.

Пример с антителами

Вернёмся к задаче $\xi \sim \mathbf{Be}(\theta)$, $\theta \in [0,1]$ с выборкой $X_1, \ldots, X_n$.

В этом случае $p(x_1|\theta) = \theta^{x_1}(1-\theta)^{1-x_1}$, $x_1 \in \{0,1\}$, и потому правдоподобие всей выборки равно:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = \theta^k(1-\theta)^{n-k}.$$

Для байесовского вывода нужно выбрать априорное распределение на параметр θ. Для простоты возьмём равномерное распределение, то есть $p(\theta) = 1$, $\theta \in [0,1]$. Тогда

$$p(\theta|\mathbf{x}) = \frac{p(\theta)\,p(\mathbf{x}|\theta)}{p(\mathbf{x})} = \frac{\theta^k(1-\theta)^{n-k}}{p(\mathbf{x})},$$

где

$$p(\mathbf{x}) = \int_\Theta p(\theta)\,p(\mathbf{x}|\theta)\,d\theta = \int_0^1 \theta^k(1-\theta)^{n-k}\,d\theta.$$

Вспоминая мат. анализ

Факт из курса анализа (wiki). Функция

$$B(\alpha,\beta)=\int_0^1 z^{\alpha-1}\,(1-z)^{\beta-1}\,dz,\quad \alpha,\beta>0,$$

называется бета-функцией Эйлера, и для неё справедливо представление:

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

где гамма-функция Эйлера $\Gamma(\alpha)=\int\limits_0^{+\infty} z^{\alpha-1}e^{-z}\,dz$, $\alpha>0$.

Если гамма-функция есть обобщение факториала, $\Gamma(n+1)=n!$, то бета-функция есть в некотором смысле обобщение биномиального коэффициента:

$$\frac{1}{B(k+1,n-k+1)}=\frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)}=\frac{(n+1)!}{k!(n-k)!}=(n+1)\,C_n^k$$

Пример с антителами

Итак, апостериорное распределение параметра θ задаётся плотностью

$$p(\theta|\mathbf{x}) = \frac{1}{B(k+1, n-k+1)}\,\theta^k(1-\theta)^{n-k}, \quad \theta \in [0,1].$$

где $k = \sum\limits_{i=1}^{n} x_i$ — число привитых с выработанными антителами а $n-k$ — число привитых, у которых не выработались антитела.

Это распределение называется *бета-распределением* (link) с параметрами $k+1$ и $n-k+1$, и записывается в виде

$$\theta|\mathbf{x} \sim \mathbf{Beta}(k+1, n-k+1).$$

CLICK ME

Пример с антителами

Заметим, что если бы мы использовали классический метод мат. статистики и в качестве оценки θ взяли бы $\overline{X} = \frac{k}{n}$, то прогноз был бы такой: вероятность выработки антител для нового привитого равна $\frac{k}{n}$.

Сделаем теперь предсказание по-байесовски.

$$\mathbf{P}(X_{\text{new}} = 1 | X = \mathbf{x}) = \int\limits_0^1 p(1|\theta)\, p(\theta|\mathbf{x})\, d\theta = \int\limits_0^1 \theta \frac{\theta^k (1-\theta)^{n-k}}{B(k+1, n-k+1)}\, d\theta =$$

$$= \frac{1}{B(k+1, n-k+1)} \int\limits_0^1 \theta^{(k+1)} (1-\theta)^{n-k}\, d\theta = \frac{B(k+2, n-k+1)}{B(k+1, n-k+1)} =$$

$$= \frac{(k+1)!\,(n-k)!\,(n+1)!}{k!\,(n-k)!\,(n+2)!} = \frac{k+1}{n+2}.$$

Пример с антителами

В частности, если бы наблюдаемые значения были $\mathbf{x} = (1, 1)$, то классическая оценка давала бы единичную вероятность выработки антител, а байесовская:

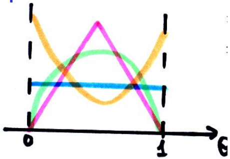
$$\mathbf{P}(X_3 = 1 | X_1 = 1, X_2 = 1) = \frac{2+1}{2+2} = \frac{3}{4},$$

то есть байесовская оценка более консервативная при малых размерах выборки n, хотя при больших n обе оценки асимптотически одинаковы.

Ещё раз подчеркнём, что во всех предыдущих рассуждениях мы использовали равномерное априорное распределение на параметр θ.
А насколько это вообще разумно???

Пример с антителами

В принципе, можно было бы брать в качестве априорного любое вероятностное распределение $p(\theta)$, сосредоточненное на отрезке $[0, 1]$.



Тогда апостериорное распределение считалось бы по формуле Байеса:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\,p(\theta)}{p(\mathbf{x})}.$$

Но на практике при этом возникает сложность с подсчётом evidence:

$$p(\mathbf{x}) = \int_0^1 p(\mathbf{x}|\theta)\,p(\theta)\,d\theta.$$

Во-первых, этот интеграл может оказаться неберущимся. Но что мешает посчитать его численно? В этой задаче - ничего, но в общем случае размерность параметра θ может быть огромной...

Пример с антителами

Хотелось бы предложить такое априорное распределение $p(\theta)$, для которого бы $p(\mathbf{x})$ считался легко. В задаче с антителами:

$$p(\mathbf{x}) = \int_0^1 p(\mathbf{x}|\theta)\, p(\theta)\, d\theta = \int_0^1 \theta^k (1-\theta)^{n-k}\, p(\theta)\, d\theta.$$

Например, это будет выполнено, если $p(\theta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$, $\theta \in [0,1]$, где c — нормировочная константа, равная $\frac{1}{B(\alpha,\beta)}$.

Иными словами, берём $\theta \sim \mathbf{Beta}(\alpha,\beta)$ для произвольных $\alpha, \beta > 0$. При этом

$$p(\mathbf{x}) = \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{(k+\alpha-1)}(1-\theta)^{(n-k+\beta-1)}\, d\theta = \frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)}.$$

Здесь было важно то, что $p(\theta)$ имеет тот же вид, что и $p(\mathbf{x}|\theta)$.

При этом говорят, что бета-распределение *сопряжено* к распределению Бернулли.

Параметры α и β называются *гиперпараметрами*.

Если $\alpha = \beta = 1$, то приходим к случаю равномерного априорного распределения.

Итак, в случае априорного $\theta \sim \mathbf{Beta}(\alpha, \beta)$ имеем

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\, p(\theta)}{p(\mathbf{x})} = \frac{\theta^k (1-\theta)^{n-k}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)\, p(\mathbf{x})} =$$

$$= \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)} \quad ,$$

то есть $(\theta|\mathbf{x}) \sim \mathbf{Beta}(k+\alpha, n-k+\beta)$.

Заметим, что интеграл $p(\mathbf{x})$ считать было и вовсе не обязательно, т.к. это всего лишь нормировочная константа для апостериорного распределения $p(\theta|\mathbf{x})$.
Это записывают в виде:

$$p(\theta|\mathbf{x}) \;\propto\; p(\mathbf{x}|\theta)\, p(\theta) \;\propto\; \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}, \quad \theta \in [0,1],$$

где $\propto$ — значок пропорциональности, и исходя из этого вида сразу ясно, что апостериорное распределение $(\theta|\mathbf{x}) \sim \mathbf{Beta}(k+\alpha, n-k+\beta)$, а нормировочная константа равна $B(k+\alpha, n-k+\beta)$.

Байесовский прогноз, опять

Теперь спрогнозируем результат следующего привитого в предположении, что априорное распределение параметра $\theta \sim \mathbf{Beta}(\alpha, \beta)$.
Поскольку $(\theta|\mathbf{x}) \sim \mathbf{Beta}(k+\alpha, n-k+\beta)$, то

$$\mathbf{P}(X_{\text{new}} = 1 | X = \mathbf{x}) = \int\limits_0^1 p(1|\theta)\, p(\theta|\mathbf{x})\, d\theta = \int\limits_0^1 \theta \frac{\theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}}{B(k+\alpha, n-k+\beta)}\, d\theta =$$

$$= \frac{1}{B(k+\alpha, n-k+\beta)} \int\limits_0^1 \theta^{(k+\alpha)}(1-\theta)^{n-k+\beta-1}\, d\theta = \frac{B(k+\alpha+1, n-k+\beta)}{B(k+\alpha, n-k+\beta)} =$$

$$= \frac{\Gamma(k+\alpha+1)\,\Gamma(n-k+\beta)\,\Gamma(k+n-k+\alpha+\beta)}{\Gamma(n+\alpha+\beta+1)\,\Gamma(k+\alpha)\,\Gamma(n-k+\beta)} = \frac{k+\alpha}{n+\alpha+\beta} \quad .$$

Байесовский прогноз, опять

Байесовский прогноз

$$\mathbf{P}(X_{\text{new}} = 1 | X = \mathbf{x}) = \frac{k + \alpha}{n + \alpha + \beta} \quad .$$

можно рассматривать как компромисс между прогнозом по данным $\frac{k}{n}$ и прогнозом априорного распределения $\mathbf{Beta}(\alpha, \beta)$:

$$\frac{k + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta} \cdot \left(\frac{k}{n}\right) + \frac{\alpha + \beta}{n + \alpha + \beta} \cdot \left(\frac{\alpha}{\alpha + \beta}\right),$$

где, как и ранее, k — число привитых с выработанными антителами, n — размер выборки.

ДЗ: проверить, что $\frac{\alpha}{\alpha+\beta}$ есть как раз математическое ожидание априорного распределения $\mathrm{Beta}(\alpha, \beta)$.

Обновление апостериорного распределения

Пусть последовательно наблюдаются две выборки $\mathbf{x} \in \mathbb{R}^n$ и $\mathbf{y} \in \mathbb{R}^m$, независимые друг от друга при фиксированном θ (*условная независимость!*).

После учёта первой выборки:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\, p(\theta)}{p(\mathbf{x})},$$

а после прихода второй хочется сделать прогноз на θ, учитывая обе выборки:

$$p(\theta|\mathbf{x},\mathbf{y}) = \frac{p(\mathbf{x},\mathbf{y}|\theta)\, p(\theta)}{p(\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\, p(\mathbf{x}|\theta)\, p(\theta)}{p(\mathbf{x},\mathbf{y})} =$$

$$= \frac{p(\mathbf{y}|\theta)\, p(\theta|\mathbf{x})\, p(\mathbf{x})}{p(\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\, p(\theta|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})} \ \propto \ p(\mathbf{y}|\theta)\, p(\theta|\mathbf{x}).$$

Таким образом, нет необходимости хранить значения выборки $\mathbf{x}$ — вся информация о параметре содержится в апостериорном распределении $p(\theta|\mathbf{x})$, которое таким же способом можно обновлять, когда приходят новые наблюдения $\mathbf{y}$.

Обновление апостериорного распределения

Пусть в примере с антителами сначала провели n испытаний, из них антитела выработались у k человек.

А потом провели ещё m вакцинаций, из них антитела выработались у r испытуемых.

Если априорное распредение $\theta \sim \mathbf{U}[0,1] \equiv \mathbf{Beta}(1,1)$, то после первой последовательности наблюдений:

$$\theta|\mathbf{x} \sim \mathbf{Beta}(k+1, n-k+1),$$

а после второй:

$$p(\theta|\mathbf{x},\mathbf{y}) \;\propto\; p(\mathbf{y}|\theta)p(\theta|\mathbf{x}) \;\propto\; \theta^r(1-\theta)^{m-r}\,\theta^k(1-\theta)^{n-k},$$

то есть

$$\theta|\mathbf{x},\mathbf{y} \sim \mathrm{Beta}(k+r+1, n-k+m-r+1).$$

Обновление апостериорного распределения

Из этого, кстати, следует, что когда мы берём в качестве априорного распределения

$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

и получаем

$$\theta|\mathbf{x} \sim \mathrm{Beta}(k + \alpha, n - k + \beta),$$

то неявно подразумеваем, что априорное распределение известно нам из какого-то опыта, в котором $\alpha - 1$ раз выработались антитела и $\beta - 1$ раз не выработались (если, конечно, α и β натуральные).

ДЗ

Сделать точный байесовский вывод для той же задачи с антителами, взяв в качестве априорного распределения *смесь* бета-распределений:

$$p(\theta) = \sum_{j=1}^{s} \gamma_j \, \frac{\theta^{\alpha_j - 1}(1-\theta)^{\beta_j - 1}}{B(\alpha_j, \beta_j)}, \quad \theta \in [0,1], \qquad \text{где } \gamma_j \geqslant 0, \sum_j \gamma_j = 1.$$