Последовательное тестирование

1. Введение

Первый метод последовательного анализа данных (англ. sequential probability ratio test) разработал Абрахам Вальд в 1945 году (Wald, 1945). Данный метод, не требующий заранее заданного количества наблюдений, нашёл применение на предприятиях для выявления бракованных партий. В наши дни методы последовательного анализа широко применяются в A/B-тестировании, поскольку в некоторых ситуациях они позволяют заранее останавливать неудачные эксперименты.

2. Ускорения A/B-тестирования

Многие статистические тесты, проверяющие гипотезы о средних значений двух выборок $\overline{X}$ и $\overline{Y}$, требуют заранее заданного числа наблюдений. Если получение необходимого числа наблюдений потребует значительное число денежно-временных ресурсов, то возникает естественное желание получить статистически значимый результат на меньшем числе наблюдений.

Пример 1. Сайт, зарабатывающий на кликах по контекстной рекламе, тестирует новый дизайн страниц. Базовый уровень конверсии составляет 0.05. Хочется убедиться, что новый дизайн поможет повысить конверсию на 20%. Для этого был составлен эксперимент: было собрано контрольное множество пользователей X (пользователи со старым дизайном) и тестовое множество Y (пользователи со старым дизайном). При использовании z-теста для получения мощности не менее 0.8 требуется, чтобы в каждой из двух выборок содержалось 6426 наблюдений

3. Правдоподобие выборки

Рассмотрим выборку $X = \{x_1, x_2, ..., x_n\}$, каждый элемент которой взят из некоторого параметрического распределения с функцией плотности распределения $f(x, \theta)$. Значением правдоподобия данной выборки при фиксированном значении параметра $\theta = \theta_0$ называется значение $\mathcal{L}(X|\theta_0) = \prod_{i=1}^{n} f(x_i, \theta_0)$, где $f(x_i, \theta_0)$ — вероятность генерации наблюдения x_i. Очевидно, что чем ближе значение параметра θ_0 к тому значению, которое использовалось при генерации выборки, тем выше значение функции правдоподобия.

Часто также рассматривают значение функции лог-правдоподобия, вычисляемой как $\log \mathcal{L}(X|\theta_0) = \sum_{i=1}^{n} \log f(x_i, \theta_0)$. Данная функция обладает более простой производной, не меняя значения, при котором достигается максимум.

Оценкой максимального правдоподобия параметра θ называется такое значение $\hat{\theta}$, что $\hat{\theta} = \arg\max_\theta \mathcal{L}(X|\theta) = \arg\max_\theta \log \mathcal{L}(X|\theta)$. Метод нахождения такого значения $\hat{\theta}$ называется методом максимального правдоподобия.

4. Последовательный анализ Вальда

4.1. Тест отношения правдоподобия

Тест отношения правдоподобия является одним из методов сравнения моделей, использующих различные значения параметров. К примеру, при добавлении новой переменной к линейной модели качество предсказаний не ухудшается, из-за чего требуется проверить, что модель с бо́льшим числом параметров (длинная модель) статистически значимо лучше модели с меньшим числом параметров (короткая).

Рассмотрим частный случай метода отношения правдоподобия, когда модель содержит один параметр. Есть выборка $X = \{x_1, x_2, ..., x_n\}$, сгенерированная параметрическим распределением $f(x, \theta)$. Обозначим как Θ множество всех возможных значений параметра θ. Выдвинем гипотезу H_0, что выборка была сгенерирована со значением параметра $\theta_0 \in \Theta_0$. В качестве альтернативной рассмотрим гипотезу $H_1 : \theta_0 \in \Theta \setminus \Theta_0$. Статистика теста отношения правдоподобия считается по следующей формуле:

$$\lambda = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(X|\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(X|\theta)}.$$

Поскольку $\Theta_0 \subset \Theta$, знаменатель будет не меньше числителя, следовательно, $0 \leq \lambda \leq 1$. На основании выдвинутых гипотез H_0 и H_1, а также требуемого уровня значимости α, находится такая константа c, что H_0 отвергается при $\lambda < c$.

Интуитивно данный подход можно объяснить следующим образом: как сильно правдоподобие предполагаемого значения отличается от максимально достижимого правдоподобия для рассматриваемой выборки. Если предполагаемый параметр близок к истинному, то значение λ будет близким к единице. В противном случае λ близка к нулю.

Пример 2. Имеется погнутая монетка. Выдвигается гипотеза, что вероятность выпадения орла в пределах [0; 0.6]. После пяти подбрасываний данной монетки были получены следующие результаты: орёл, орёл, орёл, орёл, решка. Методом максимального правдоподобия можно получить оценку вероятности выпадения орла 0.8. Наиболее правдоподобной вероятностью выпадения орла из [0; 0.6] является значение 0.6, так как оно ближе остальных к 0.8. Используя распределение Бернулли, можно получить $\mathcal{L}(1, 1, 1, 1, 0|p = 0.8) = 0.8^4 * (1 - 0.8) \approx 0.082$ и $\mathcal{L}(1, 1, 1, 1, 0|p = 0.6) \approx 0.052$, что даёт $\lambda \approx 0.63$

Лемма Неймана – Пирсона При проверке простой статистической гипотезы против простой альтернативы метод, основанный на методе отношения правдоподобия, будет являться наиболее общим среди всех методов с уровнем значимости α

Подробности данной леммы не важны для понимания сегодняшнего материала, но она даёт нам важную подсказку для ускорения получения результата: если требуется получить результат как можно скорее, то следует использовать метод отношения правдоподобия.

Говорим, что раз он самый мощный, значит, можно вооружиться им для получения результата с требуемой мощностью и минимальным числом наблюдений.

5. Последовательный анализ

5.1. Последовательный метод отношения правдоподобия

Метод последовательного отношения правдоподобия работает немного не так, как другие тесты с фиксированными выборками. Данный метод следит за получаемыми данными и с каждым новым наблюдением либо принимает решение о принятии/отвержении нулевой гипотезы, либо ожидает следующее наблюдение.

Пусть имеется выборка $X_n = \{x_1, x_2, ..., x_n\}$, каждое наблюдение которой взято из распределения $f(x, \theta)$. Выдвинем гипотезу $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta = \theta_1$. Найдём значение статистики Λ_n, вычисляемой как

$$\Lambda_n = \frac{\mathcal{L}(X_n|\theta_1)}{\mathcal{L}(X_n|\theta_0)}.$$

Критерий остановки работает по следующей схеме:

1. Если $\Lambda_n \geq B$, то принимаем гипотезу H_1;
2. Если $\Lambda_n \leq A$, то принимаем гипотезу H_0;
3. Если $A < \Lambda_n < B$, то ждём следующее наблюдение и проверяем значение Λ_{n+1},

где значения A и B находятся на основе семейства распределения $f(x, \theta)$, а также требуемых значений вероятностей ошибок α и β. Иногда идею SPRT (sequential probability ratio test) объясняют через области пространства: критерий останавливается, когда значение статистики попадает в область действия одной из гипотез. Также можно сказать, что критерий останавливается как только выборка X_n попадает в критическую область одной из гипотез.

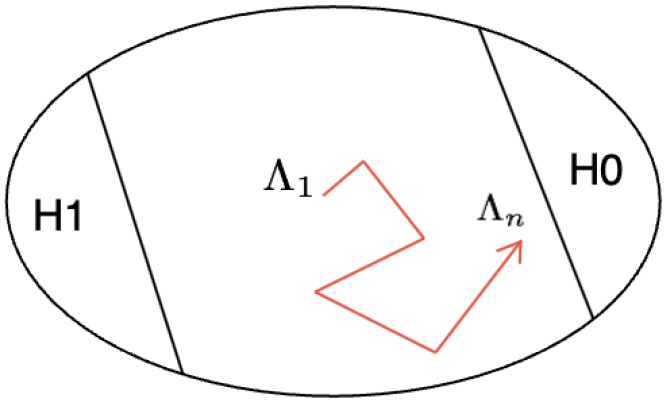

Рис. 1. Визуализация метода SPRT

Задача нахождения точных значений A и B достаточно сложна, поэтому обычно используется их аппроксимация:

$$A \approx \frac{\beta}{1-\alpha}, B \approx \frac{1-\beta}{\alpha}.$$

Данные приближенные значения обеспечивают значения ошибок I и II рода достаточно близкими к требуемым. Также следует отметить, что существует эффект «переоценки»: практически никогда

статистика Λ_n не достигает точных значений A или B, а «перепрыгивает» их, что приводит даже к меньшим вероятностям ошибок.

Также следует отметить, что значения Λ_n можно вычислять не на каждом шаге, а при достижении других условий: каждые сто наблюдений или в начале каждого часа. Однако это может привести к тому, что будет упущен момент, когда статистика впервые преодолеет пороговое значение и тест может быть остановлен.

Для наиболее популярных значений $\alpha = 0.05$ и $\beta = 0.2$ получается, что $A = 0.21$ и $B = 16$.

Также следует отметить, что часто рассматривается логарифм отношения, а именно $\log \Lambda_n$, и в качестве пороговых значений берутся $\log A$ и $\log B$. Это никак не влияет на мощность метода или его результат и делается лишь для удобства, поскольку лог-правдоподобие обычно проще вычислить.

Вернёмся к примеру 2 с погнутой монеткой, но положим, что изначально была выдвинута гипотеза $H0 : p = 0.8$ против $H1 : p = 0.6$. Как было посчитано ранее, для последовательности из 4 орлов и 1 решки имеем $\Lambda_5 \approx 0.63$. Можно заключить, что $0.21 < \Lambda_5 < 16$, следовательно, ещё рано делать какой-либо вывод и надо подкинуть монетку в 6-й раз

5.2. Применение SPRT к двум выборкам

Исходный метод последовательного анализа, сформулированный Вальдом, не очень подходит для A/B-тестирования, поскольку является одновыборочным (ниже будет рассмотрена модификация SPRT, используемая современными A/B-платформами). Однако даже оригинальный одновыборочный метод можно использовать для проверок двухвыборочных гипотез, как это было описано в (Hajnal, 1961). Для этого требуется выбрать некоторую статистику, зависящую от двух выборок, и оценить распределение этой статистики при справедливости той или иной гипотезы.

Для примера рассмотрим t-тест Стьюдента. Пусть имеются две нормальные выборки X и Y со значениями дисперсии $\sigma_x^2 = \sigma_y^2 = \sigma^2$, для которых требуется сравнить средние значения μ_x и μ_y. Выдвинем нулевую гипотезу $H_0 : \mu_x = \mu_y$ против альтернативы $H_1 : \dfrac{(\mu_x - \mu_y)^2}{\sigma^2} > \zeta^2$.

Положим, что $|X| = N_x$ и $|Y| = N_y$. Статистика данного теста будет вычисляться как

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{u}},$$

где

$$s = \sqrt{\frac{\sum_{i=1}^{N_x}(x_i - \overline{X})^2 + \sum_{i=1}^{N_y}(y_i - \overline{Y})^2}{N_x + N_y - 2}}, u = \frac{1}{N_x} + \frac{1}{N_y}.$$

При условии справедливости гипотезы H_0 статистика t имеет распределение Стьюдента с $N_x + N_y - 2$ степенями свободы, а при справедливости H_1 — нецентральное распределение Стьюдента с параметром сдвига $\dfrac{\zeta}{\sqrt{u}}$. Значение статистики Λ вычисляется как

$$\Lambda = \frac{p(t^2|H_1)}{p(t^2|H_0)}.$$

Таким образом, правдоподобие выборок X и Y оценивается через значения статистики t.

Пример 3. На входе в университет поставили камеру, которая записывает рост каждого вошедшего в здание человека. За час в здание вошло 30 мужчин и 27 женщин, при этом их средний рост составил 175 см и 166 см соответственно. Можно полагать, что внутри каждой группы рост имеет нормальное распределение, а значение дисперсии составило 10 см. Таким образом, можно получить, что $s \approx 3.1, u \approx 0.07, t \approx 10.7$. Тогда, выбрав, к примеру, $\zeta = 5$, получаем $\Lambda \approx 11.3$, то есть нужно больше наблюдений для определения разницы

6. Модификации SPRT

В данном разделе будут рассмотрены модификации метода последовательного отношения правдоподобия.

6.1. T-SPRT

Как было сказано выше, SPRT не требует заранее заданного числа наблюдений, позволяя принимать решение по мере поступления наблюдений. Однако это может привести к плохим результатам.

Для решения данной проблемы была разработана очень простая модификация для SPRT, названная "truncated SPRT", или t-sprt. В качестве решения предлагается добавить новое условие остановки, срабатывающее при достижении некоторого числа наблюдений. В качестве такого числа в простом случае можно взять число наблюдений, необходимых для какого-либо теста с фиксированной выборкой. В общем случае предлагается выбирать более тонкие пороги, зависящие от гипотез H_0 и H_1 и предполагаемых распределений (Tantaratana and Thomas, 1977).

Вернёмся к примеру 1 с новым дизайном страниц. Выдвинем две гипотезы о значении конверсии p страниц с новым дизайном: $H_0 : p = 0.05$ против $H_1 : p = 0.07$. Если истинное значение конверсии $p = 0.06$, то получается ситуация, в которой H_0 и H_1 примерно одинаково неверны, поэтому последовательный анализ будет очень долго думать, какой из двух гипотез отдать предпочтение. В случае «классических» подходов с фиксированным размером выборки будет рекомендовано придерживаться гипотезы H_0, поскольку нет свидетельств о справедливости отвержения H_0 в пользу H_1

6.2. mSPRT

Метод mixture-SPRT (mSPRT) был предложен в (**?**). Именно данный вариант SPRT лежит в основе тестирования, используемого такими компаниями, как Uber и Netflix.

Говоря о недостатках оригинального SPRT, можно выделить ещё две проблемы:

1. SPRT работает только с простыми гипотезами, то есть когда каждая гипотеза предполагает одно конкретное значение параметра.
2. SPRT не учитывает априорные знания о распределении параметра.

Первая проблема является достаточно естественной, ведь для проверки параметра часто хочется использовать двустороннюю альтернативу (или, по крайней мере, рассмотреть множество альтернативных параметров). Вторая проблема может помочь с аномалиями, выбросами и получением более надёжного результата. Если известно ожидаемое значение искомого параметра, то можно находить наиболее вероятное значение не только с точки зрения данных, но и с точки зрения предметной области.

Введём функцию плотности распределия параметра $h(\theta)$, которая считает вероятность того или иного значения параметра θ. Значение статистики Λ_n находится следующим образом:

$$\Lambda_n = \int_\Theta \frac{\mathcal{L}(X_n|\theta)}{\mathcal{L}(X_n|\theta_0)} h(\theta) d\theta,$$

где Θ — это множество возможных значений параметра θ. Интуитивно данную формулу можно понять следующим образом: предполагаемое значение параметра θ_0 сравнивается со всеми другими возможными значениями, а дальше эти отношения усредняются в соответствии с вероятностями $h(\theta)$. Таким образом получается средневзвешенное отношение возможных значений к предполагаемому.

Вернёмся к примеру 2 с погнутой монеткой. При помощи mSPRT мы можем проверить следующую гипотезу: $H_0 : p = 0.5$ против $H_1 : p \neq 0.5$. Данный вариант является более удобным, поскольку альтернативная гипотеза является двусторонней, то есть нам не обязательно выбирать какой-то один отрезок для возможных значений p. Однако требуется выбрать распределение $h(p)$. В данном случае, если мы верим, что монетка погнулась не сильно, то можно считать $h(p) = \mathcal{N}(0.5, 0.01)$

6.3. Always valid inference

Как было сказано выше, метод SPRT является одновыборочным, что сильно ограничивало его применение в тестировании. Также проблемой SPRT является отсутствие интуитивно понятной статистики. Если p-value можно легко объяснить даже не специалистам, а то вот значение Λ_n требует понимания формул. Поэтому в работе (Johari et al., 2015) был описан метод, адаптирующий mSPRT к A/B-тестированию.

Одна из проблем применения SPRT в A/B-тестировании состоит в том, что данный подход был разработан для последовательности данных, что не всегда применимо к A/B.

Авторы вводят понятие потоков данных. Предполагается, что пользователи контрольной и тестовой групп составляют отдельные потоки, которые затем объединяются, и итоговая выборка состоит из пар элементов (x_i, y_i), поэтому в каждый момент времени требуется, чтобы количество наблюдений в выборках X_n и Y_n было равно (в крайнем случае из-за этого приходится отбрасывать избыточные наблюдения).

Для расчёта статистики mSPRT можно вывести две основные формы. Если выборки X и Y, а также априорное распределение $h(\theta)$ имеют нормальные распределения, то статистика принимает следующий вид:

$$\Lambda_n = \sqrt{\frac{2\sigma^2}{2\sigma^2 + n\tau^2}} \exp\left(\frac{n^2\tau^2(\overline{Y}_n - \overline{X}_n - \theta_0)^2}{4\sigma^2(2\sigma^2 + n\tau^2)}\right).$$

Если элементы выборок X и Y пришли из распределений Бернулли, то

$$\Lambda_n = \sqrt{\frac{\sigma_n^2}{\sigma_n^2 + n\tau^2}} \exp\left(\frac{n^2\tau^2(\overline{Y}_n - \overline{X}_n - \theta_0)^2}{2\sigma_n^2(\sigma_n^2 + n\tau^2)}\right),$$

$$\sigma_n^2 = \overline{Y}_n(1 - \overline{Y}_n) + \overline{X}_n(1 - \overline{X}_n).$$

В данном случае параметр θ_0 является значением разности средних выборок при условии справедливости H_0.

Значение τ является значением смешанной дисперсии данных выборок. Единой формулы для вычисления τ не существует. Множество исследователей предлагают различные методы вычислений данной

формулы, однако наиболее надёжным является экспериментальный: при помощи симуляций требуется сравнить различные значения τ и выбрать наилучшее.

Значение p-value считается по следующему принципу:

$$p_0 = 1, p_n = \min\{p_{n-1}, 1/\Lambda_n\}.$$

Как и в обычном тестировании, нулевая гипотеза отвергается в момент, когда $p_n < \alpha$.

Вернёмся к примеру с погнутой монеткой. Предположим, используя mSPRT было получено $\Lambda_5 \approx 0.63$. Если $p_4 = 1$, то $p_5 = \min\{p_4, 1/\Lambda_5\} = \min\{1, 1.59\} = 1$. Таким образом, значение p_5 равно 1, значит, нет никаких причин отвергать нулевую гипотезу

7. Альтернативный подход к последовательному анализу

В данном тексте описан альтернативный способ проведения последовательного анализа для биномиального распределения. Его преимущество состоит в простоте реализации, а также в дополнительном ограничении, не позволяющем эксперименту продвигаться бесконечно долго.

Рассмотрим две группы пользователей: контрольную $X \sim Bernulli(p_x)$ и тестовую $Y \sim Bernulli(p_y)$. Требуется заранее определить максимальный размер выборки N, который проще всего получить при помощи данного онлайн-калькулятора. Следует отметить, что значение N зависит от выбранных значений MDE, α и β. В ходе эксперимента субъекты случайным образом относятся либо к тестовой, либо к контрольной группе с вероятностями 50%. Также требуется непрерывно оценивать две величины: количество успехов в тестовой группе T и в контрольной группе C.

Эксперимент останавливается в одном из двух случаев:

- если значение $T - C \geq 2\sqrt{N}$, то эксперимент останавливается с выявлением искомого эффекта;
- если значение $T + C = N$, то эксперимент оснавливается и гипотеза о наличии эффекта отвергается.

Интуитивно данный метод легко проиллюстрировать данным примером.

Пример 4. Пусть имеется две монетки. Предположим, что вероятность выпадения орлов у них одинаковая, и начнём подбрасывать их по очереди, запоминая число выпавших орлов. Если наше предположение верно, то число выпадения орлов будет примерно одинаковым. Если же одна из монеток сильно погнута, то она будет опережать или отставать от второй по числу орлов.
Более конкретно: возьмём две справедливые монетки и погнём одну из них так, чтобы увеличить вероятность выпадения орла. Предположим, что в итоге наших манипуляций вероятность выпадения орла увеличилась на 20%, то есть составляет более 60%. Воспользовавшись онлайн-калькулятором, определяем, что при $\alpha = 0.05$ и $\beta = 0.1$ можно удостовериться, что монетка погнута достаточно сильно, если число орлов на выпавшей монетке будет на 66 больше, чем на обычной монетке. И монетка окажется погнутой недостаточно сильно, если суммарное число выпавших орлов превысит 1115

Стоит также отметить, что последовательный анализ не является альтернативой другим методам ускорения тестов, основанных на методах сокращения дисперсии. Данные методы можно комбинировать для дальнейшего ускорения тестирования.

Список литературы

Hajnal, J. (1961). A two-sample sequential t-test. *Biometrika*, 48(1/2):65–75.

Johari, R., Pekelis, L., and Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Tantaratana, S. and Thomas, J. B. (1977). Truncated sequential probability ratio test. *Information Sciences*, 13(3):283–300.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.