Вероятность и статистика

$\delta\alpha\alpha'$

Nurlan Sadykov

5 сентября 2023 г.

Оглавление

Часть I

Probability

0.1 Полная вероятность

Формула полной вероятности показывается как считается вероятность события, если это событие может случиться при различных условиях. Само по себе правило обычное, но очень полезны мат.ожидания и дисперсия полной вероятности.

$$P(A) = \sum_i P(A|B_i) \cdot P(B_i) \tag{2}$$

0.1.1 Правило полного мат.ожидания

В формуле полного мат.ожидания 4 величина $E[X|Y]$ случайная по Y, поэтому в правой части формулы под внешним мат.ожиданием лежит не константа.

$$E[X] = E[E[X|Y]] \tag{4}$$

Доказательство формулы, такое:

$$E[E[X|Y]] = E\left[\sum_x xP(X=x|Y)\right] =$$

$$\sum_y \left[\sum_x x\,P(X=x|Y)\right] p(Y=y) = \sum_y \sum_x xP(X=x,Y=y) =$$

$$\sum_x x \sum_y P(X=x,Y=y) = \sum_x x\,P(X=x) = E[X]$$

0.1.2 Правило полной дисперсии

Разложим мат.ожидание $E[Y^2]$ по формуле полного мат.ожидания 4 используя свойство $E[Y^2] = Var[Y] + E[Y]^2$

$$Var[Y] = E[Var[Y|X]] + Var[E[Y|X]] \tag{6}$$

$$E[Y^2] = E\left[E[Y^2|X]\right] = E[Var[Y|X] + E[Y|X]^2]$$

Теперь если разложить дисперсию, получим

$$Var[Y] = E[Y^2] - E[Y]^2 =$$

$$E\left[Var[Y|X]\right] + E\left[E[Y|X]^2\right] - E\left[E[Y|X]\right]^2 =$$

$$E[Var[Y|X]] + Var[E[Y|X]]$$

Дисперсия при стратификации. Предположим, X имеет мультиномиальное распределение и принимает значения от 1 до K. Обозначим соответствующие вероятности p_i для удобства. Мы хотим рассчитать вероятность величины Y которая зависит от X. В формуле 6 участвует два слагаемых, которые можно вычислить напрямую

Пример 1: Заработная плата в стране. Стратификация по регионам.
Пример 2: Рост подростков. Стратификация по возрасту и/или полу.

μ_k, σ_k — параметры распределения Y внутри страты k.

$$E[Var[Y|X]] = \sum_{k=1}^{K} \sigma_k^2\, p_k,$$

$$Var[E[Y|X]] = \sum_{k=1}^{K} (\mu_k - \mu)^2\, p_k.$$

Откуда и получим

$$Var[Y] = \sum_{k=1}^{K} \sigma_k^2\, p_k + \sum_{k=1}^{K} (\mu_k - \mu)^2\, p_k \tag{7}$$

0.2 Комбинации случайных величин

$\xi_1,\ \xi_2$ — случайные величины
$f_{\xi_1,\xi_2}(x_1,\ x_2)$ — плотность совместного распределения

Допустим есть функция $g\colon \mathbb{R}^2 \to \mathbb{R}$ и мы хотим найти комбинацию двух случайных величин ξ_1 и ξ_2, то есть мы хотим узнать функцию распределения для случайной величины $\eta = g(\xi_1,\ \xi_2)$. Посчитать величину η можно через функцию распределения

$$F_\eta(x) = P(g(\xi_1, \xi_2) < x) = \iint\limits_{g(x_1,x_2)<x} f_{\xi_1,\xi_2}(x_1,\ x_2)\,dx_1\,dx_2 \tag{8}$$

В случае если случайные величины ξ_1 и ξ_2 независимы и $g(x_1, x_2) = x_1 + x_2$ можно показать, что плотность суммы выражается в виде свертки плотностей

$$f_{\xi_1+\xi_2}(t) = \int\limits_{-\infty}^{\infty} f_{\xi_1}(u)\,f_{\xi_2}(t-u)\,du \tag{9}$$

0.3 Характеристические функции

Характеристической функцией случайной величины ξ называется функция

$$\varphi_\xi(t) = M\left(e^{it\xi}\right) = \int\limits_{-\infty}^{\infty} e^{itx} f(x)\,dx \tag{10}$$

Характеристические функции это преобразование Фурье для случайных величин и по функции можно всегда восстановить распределение применив обратное преобразование.

Свойства:

- $\varphi_{a+b\xi}(t) = e^{ita}\varphi_\xi(tb)$
- $\varphi_{\xi+\eta} = \varphi_\xi(t)\cdot\varphi_\eta(t)$
- $\varphi_\xi^{(k)}(0) = i^k M\left(\xi^k\right)$ - k-ый момент это k-ая производная функции в нуле

Глава 1

Распределения

1.1 Распределение хи-квадрат χ^2

Пусть $\xi_1, \cdots, \xi_k$ - независимые, стандартизованные, нормально распределенные случайные величины. Величина

Стандартизованный: $\mu = 0$, $\sigma = 1$.

$$\chi_k^2 = \xi_1^2 + \ldots + \xi_k^2 \tag{1.1}$$

подчиняется распределению χ_k^2 с k степенями свободы.

Плотность распределения

$$f_{\chi_k^2}(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{1.2}$$

1.2 Распределение Фишера-Снедекора

Положим есть две случайные величины U и V распределенные по χ^2 с n и k степенями свободы соответственно. Величина

$$F = \frac{U/n}{V/k}$$

имеет распределение Фишера-Снедекора.

Плотность распределения

$$f(x) = \begin{cases} 0, & x \le 0 \\ C_0 \frac{x^{n-2}/2}{(k+nx)^{(n+k)/2}}, & x > 0 \end{cases} \qquad C_0 = \frac{\Gamma(\frac{k+n}{2}) n^{n/2} k^{k/2}}{\Gamma(n/2)\Gamma(k/2)}$$

Глава 2

Большие числа

2.1 Закон больших чисел

Есть последовательность независимых наблюдений $X_1,\ X_2,\ \cdots$ выбранных из одного распределения с конечными мат.ожиданием μ и дисперсией σ^2. Среднее значение выборки первых n элементов будет тоже равно μ.

Неравенство Чебышева:

$$P(\mid x-\mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

$$m_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$M(m_n) = \frac{1}{n}(M(X_1) + M(X_2) + \cdots + M(X_n)) = \mu$$

А еще можем посчитать дисперсию:

$$D(m_n) = \frac{1}{n^2}\left(D(X_1) + D(X_2) + \cdots + D(X_n)\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \qquad (2.1)$$

Тогда если воспользоваться неравенством Чебышева, получим

$$P\left(|m_n - \mu| \geq \epsilon\right) = P\left(|m_n - \mu| \geq \epsilon \frac{\sqrt{n}}{\sigma}\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\sigma^2}{n\,\epsilon^2}$$

Теперь можем посчитать вероятность малых различий в разности

$$P\left(|m_n - \mu| < \epsilon\right) = 1 - P\left(|m_n - \mu| \geq \epsilon\right) \geq 1 - \frac{\sigma^2}{n\,\epsilon^2}.$$

Откуда можем вывести предел по вероятности

$$\lim_{n\to\infty} m_n = \mu \qquad (2.2)$$

2.2 Центральная предельная теорема

В общем виде центральную предельную теорему можно выразить так: *мат.ожидание выборки имеет нормальное распределение.*

$X_{i\geq 1}$ - произвольные случайные величины из одного распределения с мат.ожиданием μ и дисперсией σ^2.

Для z-score

$$z_n = \frac{\sqrt{n}}{\sigma}\left(m_n - \mu\right)$$

теорема выражается так

$$\lim_{n\to\infty} P(z_n \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp^{-x^2/2}\, dx = \Phi(z). \qquad (2.3)$$

Для биномиального распределения $\mu = p$ и $\sigma = \sqrt{pq}$ откуда
$z_n = \frac{\sqrt{n}}{\sqrt{p(1-p)}} \left(\frac{k}{n} - p\right) = \frac{k-np}{\sqrt{np(1-p)}}$

Часть II

Statistics

2.3 Cookbook

1. *Репозиторий.* Анализ самостоятельного датасета лучше поместить в отдельную папку.
2. *Описальные статистики* стоит просмотреть чтобы понять качество данных и исключить сильные аномалии.
3. Построить *гистограмму* чтобы понять распределение и проверить на выбросы.

Иногда удобно просматривать логарифмированную фичу. Если фича сильно отстоит от нуля, ее лучше переместить в ноль.

2.4 Train, Test, Validation

Общая рекомендация относить в тестовую выборку до 30% всех данных. Методику с train-test'ом используют чтобы выбрать модель которой. Например, в задаче регрессии нужно понять многочленами какой степени стоит аппроксимировать данные. Отбирается та модель которая на тестовой выборке показывает минимальную ошибку. В цлеом это так или иначе упирается в кросс-валидацию.

- Обучащающая выборка - для обучения модели
- Тестовая выборка - для замера качества модели
- Валидационная выборка - для подбора гиппер-параметров

Альтернативой к валидации является подход с регуляризацией.

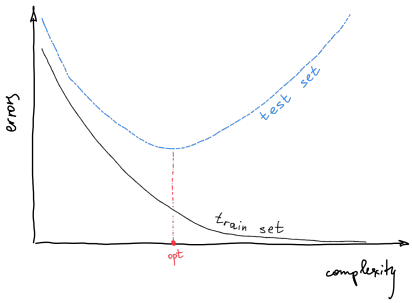


Рис. 2.1: *Выбор модели:* по оси абсцисс лежат модели (например, это могут быть степени аппроксимирующих многочленов), по оси ординат отмечается ошибка. Выбирается та модель у которой достигается наименьшая ошибка на тестовой выборке (на рисунке точка opt).

2.5 Ящик с усами

Есть набор наблюдений $x_1, \ldots, x_n$, По набору можно вычислить:

- min - минимальное наблюдение
- max - максимальное наблюдение
- M - медиана
- Q_1 - первый квартиль (25%)
- Q_3 - третий квартиль (75%)

Межквартильный размах задается разностью $\Delta Q = Q_3 - Q_1$.

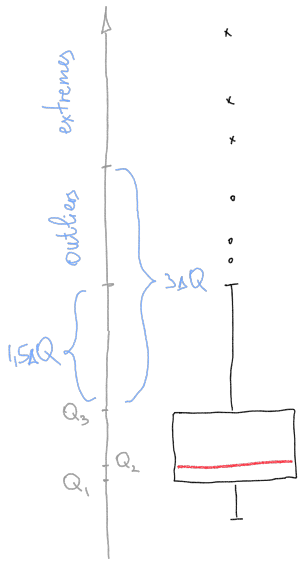
Длина правого усика, это минимум между расстоянием от квартиля Q_3 до максимального элемента и полутора полуторным межквартильным расстоянием. Аналогично расчитывается длина правого усика.

$$l = \min(\max(x_i) - Q_3, 1.5\Delta Q)$$

$$r = \min(Q_1 - \max(x_i), 1.5\Delta Q)$$

Данные за усиками это *выбросы*, а если они вышли за три межквартильных расстояния это *чрезвычайные выбросы*.

В ящике с усами используется медиана вместо матожидания в качестве описания потому, что она устойчива к выбросам.



Усеченное среднее. Если из выборки удалить 2,5% самых маленьких наблюдений и 2,5% самых больших, то выбросы уже не так сильно будут влиять на вычисленное матожидание.

Глава 3

Статистические гипотезы. Базовые методы

3.1 Ошибки, значимость, p-value

Если мы хотим опровергнуть какое-то утверждение, мы должны сформулировать альтернативную гипотезу. Основная гипотеза всегда обозначается H_0, альтернативная H_1. Основная гипотеза всегда должна быть простой, каким-то конкретным утверждением, альтернативная может быть каким угодно. Такое требование создается потому, что у нас нет мат.аппарата на сложные условия.

Влиять мы можем только на ошибку I-го рода. Ее обычно фиксируют на уровне 0,05, 0,01 или 0,005, но этот уровень можно менять на любой другой если этого требует задача.

Ошибки второго рода контролировать сложнее. Для их уменьшения стараются пользоваться состоятельными критериями.

Критическая точка критерия это такое значение вероятность превысить которое равняется α то есть вероятности ошибки первого рода. На рисунке 3.1 это площадь под графиком выше критической точки окрашенная в красный. Синим обозначена вероятность ошибки II-го рода. Как видно это вероятность получить выборку при которой не отвергается нулевая гипотеза, но была верна альтернативная гипотеза.

Критическая точка не обязана лежать на пересечении плотностей распределений двух гипотез. Она определяется исходя только из плотности для нулевой гипотезы.

Альтернативой к механике проверки гипотез с критической точкой является проверка гипотез через p-value. P-value вычисляется из предположения, что нулевая гипотеза верна и определяет вероятность получить такое же значение критерия или более экстремальное. Нам все равно нужно фиксировать уровень значимости α до эксперимента, но не нужно вычислять критическую точку. Если p-value оказалось меньше α то нулевая гипотеза должна быть отвергнута, в противном случае нет оснований отвергать нулевую гипотезу.

Таблица 3.1: Ошибки I-го и II-го рода

гипотеза	верна	не верна
принята	+	*II род*
отвергнута	*I род*	+

Критерий проверки гипотезы называется состоятельным, если ошибка второго рода уменьшается с ростом числа наблюдений.

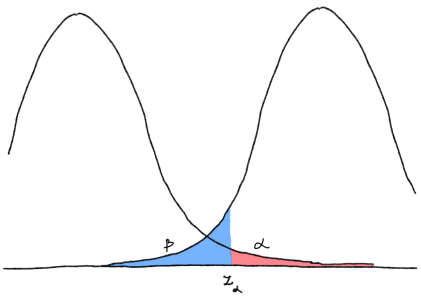


Рис. 3.1: Левое распределение соответствует гипотезе H_0, правое - гипотезе H_1.

Проверять статистические гипотезы можно только в случае массовых явлений. События которые редки стат.гипотезами не проверяются. Число в 30 наблюдений считается достаточным только для нормальных распределений, для всех остальных надо больше данных.

3.1.1 Ошибка подглядывания

Ошибка подглядывания возникает когда мы заглядываем в результаты теста до окончания самого теста (в промежуточное время). В этот момент человек может увидеть промежуточный результат p-value $< \alpha$. Во время подглядывания может возникнуть соблазн дождаться когда p-value упадет ниже уровня стат.значимости и остановить эксперимент. На самом деле график p-value в процессе эксперимента может несколько раз опускаться ниже стат.значимости. Это будет происходить случайно.[1]

[1] Анатолий Карпов. *Тонкости A/B тестирования: проблема подглядывания.* URL: `https://www.youtube.com/live/jnFVmtaeSA0`

В случае если *нулевая гипотеза верна*, вероятность получить p-value на определенном уровне имеет равномерное распределение. Например, вероятность получить p-value из первого дециля равна 10%, вероятность получить p-value из второго дециля это $P(\text{p-value} < 0.2) - P(\text{p-value} < 0.1) = 0.1$ тоже 10%. Это означает, что отвергать нулевую гипотезу, то есть совершать ошибку I рода, мы будем с вероятностью α.

Если же следить за p-value пока он не опустится ниже α мы увеличим вероятность ошибки I рода и перестанем ее контролировать. И по сути будем выдавать желаемое за действительное.

3.1.2 A/A-test

Когда проверяются гипотезы статистикой, лучше провести A/A-тестирование. Во время тестирования нужно просимулировать что-нибудь и проверить как работают критерии которыми будет проверяться будущий эксперимент.

Может оказаться так, что какие-то критерии не работают потому, что нарушаются условия (каким-то неизвестным нам образом) применения этих критериев. В этом случае A/A-тесты не сойдутся и это повод выяснить причины и устранить их до запуска полноценного A/B-теста.

Процедура статистического тестирования:

1. Определить метрику которую проверяем.
2. Сформулировать нулевую и альтернативную гипотезы.
3. Зафиксировать уровень стат.значимости α.
4. Предрасчитать объем необходимой выборки для теста.
5. Предрасчитать длительность теста.
6. Запуск теста. В промежуточные результаты не заглядываем.
7. По истечению времени, интерпретация результатов.

Такая процедура обеспечит, что из всей массы запущенных экспериментов ошибки I рода будут наблюдаться с вероятностью α.

3.2 bootstrap

Довольно мощный тест который позволяет вычислить параметры распределения или какие-то дополнительные метрики по выборке.

Положим, что у нас какая-то *репрезентативная* выборка S и мы хотим оценить некоторый параметр ν. Например, это может быть первый квартиль или разность матожиданий.

Bootstrap предполагает, что начальная выборка из которой мы производим сэмплирования является репрезентативной.

Чтобы провести bootstrap проведем $X(> 1000)$ раз процедуру:

1. Просэмплируем из множества S подвыборку N элементов с возвращениями. На практике достаточно $50 < N < 100$.
2. Посчитаем на полученной подвыборке параметр ν.

В итоге мы получаем новую выборку из X оценок параметра ν и теперь можем оценить и как матожидание этого параметра, так и его доверительный интервал.

3.3 Нормальность

μ_k — k-ый момент
σ — ср.кв.отклонение

Проверять данные на нормальность стоит не слишком тщательно. Данным достаточно только *напоминать* нормальные, чтобы использовать большинство критериев в которых подразумевается, что данные нормальные. Если подходить к проверкам строго, то ничто и никогда невозможно будет проверить.

Точно необходимо проверить:

1. Выбросы. Отсутствие выбросов достаточно критичное условие. Если в данных есть выбросы, то их нельзя считать нормальными. С другой стороны если эти выбросы убрать, возможно уже можно будет работать.

2. Асимметрия. Нормальные данные должны быть симметричными, поэтому если наблюдается асимметрия, то данные не нормальные. Возможно сможет помочь логарифмирование.

$$A = \frac{\mu_3}{\sigma^3}$$

3. Эксцесс (kurtosis) — отклонение от колоколообразности. Пограничным случаем колоколообразности можно считать равномерное распределение, а вот бимодальность гистограммы уже явный признак, что данные не нормальны. Еще чем больше наблюдений в данных, тем все-таки ближе должна быть гистограмма к нормальному распределению. За критерий можно взять выборку в $n = 150$, чей график должен быть похож на колокол.

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Если в выборке менее 2000 объектов, то применяется метод Шапиро-Уилка, если же объектов больше, то используется критерий согласия Колмогорова-Смирнова.

3.3.1 Удаление выбросов

Преобразование Бокса-Кокса это преобразование исходного набора данных. Применяют в основном с целью устранения выбросов. Например, при $\lambda = 0$ это обычное логарифмическое преобразование. Параметр λ подбирается методом оптимизации правдоподобия.

$$x_{i,\lambda} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{если } \lambda \neq 0, \\ \log(x_i) & \text{если } \lambda = 0. \end{cases} \tag{3.2}$$

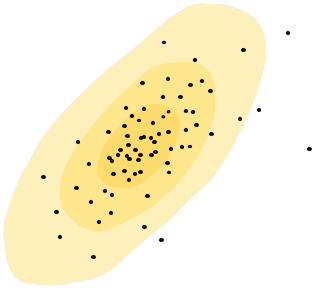
Почему такая формула? Логарифмическое преобразование хорошо себя зарекомендовало, но оно не всегда справляется. Замечено, что $(\log x)' = 1/x$. Преобразование Бокса-Кокса это обобщение до функций чьи производные равны $1/x^\lambda$.

3.3.2 Расстояние Махалонобиса

m — выборочный вектор мат.ожидания
S — выборочная матрица ковариации

Для многомерных данных можно предположить, что они распределены по какому-то многомерному нормальному закону распределения. Если центр (мат.ожидание) и матрицу ковариации для датасета, то можно определить расстояние Махалонобиса. Так мы получим аналог многомерной дисперсии и сможем вычислить отклонение от среднего

$$d = \sqrt{(x - m)S^{-1}(x - m)^T} \tag{3.3}$$



Мы можем использовать это расстояния для определения 95% границы (или любой другой) чтобы построить область допустимых значений. Сэмплы за вычисленной границей будут считаться аутлайерами. Поскольку вычисление границы может оказаться дорогим, можно провести упрощенное действие: рассчитать для каждого сэмпла расстояние 3.3 и отрезать 5% граничных квантилей.

3.4 Сравнение мат.ожиданий

Для сравнения мат.ожиданий используются:

- t-критерий Стюдента
- Fligner-Killeen test
- Brown-Forsythe test

3.4.1 t-test, критерий Стьюдента

Критерий Стюдента предполагает что метрика (н-р, матожидание) нормально распределена. Для двухвыборочного критерия важно равенство дисперсий. Так же тест не устойчив к выбросам в данных.

Одновыборочный тест проводится когда у нас есть точное предположение о значении матожидания.

Статистика для одновыборочного теста:
$\overline{x}$ - выборочное матожидание
s^2 - выборочная несмещенная оценка дисперсии
n - объем выборки

$$H_0 : M(x) = m$$

$$t = \frac{\overline{x} - m}{s/\sqrt{n}}$$

Степеней свободы $n - 1$.

Двухвыборочный тест Предполагается, что есть две независимые выборки объема n_1 и n_2. Из которых извлечены матожидания m_1 и m_2 и несмещенные дисперсии s_1^2 и s_2^2.

$$H_0 : m_1 = m_2$$

Если нулевая гипотеза верна, то разность $\Delta = m_1 - m_2$ имеет матожидание $M(\Delta) = 0$ и исходя из независимости выборок $D(\Delta) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, так же получаем несмещенную оценку разности выборочных средних $D(\Delta) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. Теперь можем определить статистику

Превращаем двувыборочный тест в одновыборочный!

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.4}$$

Степени свободы вычисляются по формуле

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{(\frac{s_1^2}{n_1})^2/(n_1 - 1) + (\frac{s_2^2}{n_2})^2/(n_2 - 1)} \tag{3.5}$$

3.4.2 Доверительный интервал

Оценка показателей по выборке из генеральной совокупности всегда будет точечной оценкой. И если выбрать другую выборку, то оценка изменится. Чтобы учесть этот эффект вводят интервальную оценку показателя.

Интервал строится исходя из препосылки, что распределение измерений матожиданий по различным выборкам из генеральной совокупности должно уложиться в нормальное распределение. Отсюда мы можем вычислить радиус доверительного интервала такими формулами

Доверительный интервал для показателя ν это такой интервал (ν_l, ν_r) в который истинный показатель входит с вероятностью $1 - \alpha$.

$$\Delta = \frac{s}{\sqrt{n}}\, z_\alpha \qquad \text{или} \qquad \Delta = \frac{s}{\sqrt{n}}\, t_\alpha(n-1)\ . \tag{3.6}$$

Критерий выбирается исходя из количества наблюдений или каких-то других особенностей прода. Также вместо выборочного ср.кв.отклонения можно использовать и истинную дисперсию если она известна, как правило это большая редкость.

достаточного объема выборки. Если мы знаем максимальное расстояние которое может достигать доверительный интервал, то остается только обернуть формулу. Радиус подбирают таким, чтобы конкурирующая гипотеза не входила в интервал.

3.4.3 Объем выборки

$1-\beta$ — мощность (power)
$\delta = m_1 - m_0$ — MDE

Чтобы рассчитать объем необходимой выборки для проведения теста нужно зафиксировать уровень ошибок второго рода β. Критическая точка определяет границу между принятием нулевой гипотезы или альтернативной. Если посчитать границу исходя из каждой гипотезы получим такое уравнение

$$\cancel{m_0} + z_{1-\alpha/2}\, s_0 \sqrt{\frac{1}{n_0} + \frac{1}{n_1}} = \cancel{m_0} + \delta - z_{1-\beta} \sqrt{\frac{s_0}{n_0} + \frac{s_1}{n_1}}. \tag{3.7}$$

$$z = \frac{x-m}{s} \quad \longrightarrow \quad x = m + z \cdot s$$

Уравнение 3.7 основа для расчета минимального объема выборки. Если предположить $n_0 = n_1 = n$, что обеспечит одинаковый объем в тестовом и контрольном сплитах, и $\sigma_0 = \sigma_1 = \sigma$ получим формулу

Chapter 2 in Gerald Van Belle. *Statistical rules of thumb*. T. 699. John Wiley & Sons, 2011

$$n = \frac{s^2 (z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} \tag{3.8}$$

Величина δ определяет минимальный обнаруживаемый эффект MDE (minimal detectable effect). В формуле (3.8) он присутствует в виде делителя. Это значит, что за обнаружение минимального профита нам придется заплатить увеличением выборки.

3.5 Сравнение медиан

Для сравнения медиан используются:

- Mood's median test. Только для гигантских данных. На малых данных будет большая ошибка 2 рода.
- критерий Мана-Уитни. Есть критика критерия, но крайние случаи довольно редкие и ими пренебрегают и все равно используют критерий.

3.5.1 Критерий Манна-Уитни. U-test.

Чтобы применить критерий Манна-Уитни данные должны обладать порядком (числа подходят). Это непараметрический тест поэтому область его применения может быть шире, чем просто над числовыми показателями.

Положим есть две выборки $x_{i=\overline{1,n_1}}$ и $y_{j=\overline{1,n_2}}$. Тогда мы можем составить слово из букв X и Y последовательно вынимания из объединения $x \cup y$ минимальный элемент. Буква определяется принадлежностью множеству. Такую последовательность уже можно конвертировать в ROC кривую. Если обе выборки принадлежат одному распределению, то буквы в слове перемешаются равномерно, а значит кривая будет диагональю, площадь под которой равна $S = \frac{n_1 n_2}{2}$. Так мы свели задачу к проверки гипотезы о мат.ожидании.

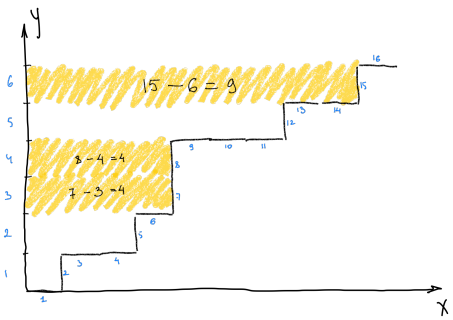


Рис. 3.2: Пояснение формулы U_y. По выборкам была составлена последовательность $XYXXYXYYXXXYXXYX$. Числа на ступенчатой кривой — ранги в общей последовательности, числа на оси Oy — ранги в выборке y.

H_0: Обе выборки из одной генеральной совокупности

H_1: Выборки из разных генеральных совокупностей

Обозначим через T_* сумму рангов элементов из разных выборок в объединении. Теперь мы можем посчитать

$$\left.\begin{array}{ll} U_x & = T_x - \frac{n_x(n_x+1)}{2} \\ U_y & = T_y - \frac{n_y(n_y+1)}{2} \end{array}\right] \quad \Rightarrow \quad U = \min(U_x,\, U_y) \tag{3.9}$$

$$\left.\begin{array}{ll} m_U & = n_x\, n_y/2 \\ s_U & = \sqrt{\frac{n_x\, n_y\, (n_x+n_y+1)}{12}} \end{array}\right] \quad \Rightarrow \quad z = \frac{U - m_U}{s_U} \tag{3.10}$$

Joseph C Watkins. “An introduction to the science of statistics: From theory to implementation”. B: (2019)

MDE можно вычислять как для одновыборочного теста на мат.ожидание.

3.6 Сравнение вероятностей (частот)

3.6.1 Классическая схема

Количество появления события в выборке подчиняется биномиальному закону распределения $Bin(n, \pi)$ При больших n биномиальное распределение становится близким к нормальному распределению с параметрами

$$m = p = \frac{k}{n} \qquad s = \sqrt{\frac{p(1-p)}{n}} \tag{3.11}$$

n - количество испытаний
k - количество появлений события
π - *истинная* вероятность события
p - *оценка* вероятности (частоты)

Задача снова сведена к сравнению через нормальное распределение.

3.6.2 Доверительный интервал

Относительная частота события A вычисляется отношением $p = k/n$, где k — число появлений события A, n — объем выборки. Тогда если учесть, что

$$M[p] = p \qquad \text{и} \qquad D[p] = pq/n\,, \tag{3.12}$$

можем воспользоваться формулой на вычисления вероятности интервала добавив еще уровень стат.значимости

$$P(|X - k| < r) = 2\Phi(r/\sigma) = \alpha\,, \tag{3.13}$$

откуда получим

$$r = z_\alpha \cdot \sqrt{\frac{p(1-p)}{n}} \tag{3.14}$$

Если нужно достать минимальный объем выборки, нужно знать заранее какой радиус r мы хотим обеспечить. Это можно определить исходя из потребности задачи (безнеса). Однако, есть еще один параметр который нужно знать заранее, это вероятность p. Поскольку мы не знаем ее до эксперимента (получения выборки), можно положить $p = 0.5$. В этом случае выражение $p(1-p)$ примет наибольшее значение.

Поскольку каждое изменение в выборке независимо, а объем выборки можно считать константой, можем сказать, что

$$M[p] = M[k/n] = M[k/n] = np/n = p.$$

Если учесть, что $D[k] = npq$ и независимость выборки, получим

$$D[p] = D[k/n] = D[k]/n^2 = npq/n^2 = pq/n.$$

3.6.3 Проверка гипотезы через beta-распределение

Индекс $i = 0$ - контрольный сплит
Индекс $i = 1$ - тестируемый сплит
k_i - наблюдаемые положительные эвенты
n_i - общий объем наблюдений
$r_i = k_i/n_i$ - наблюдаемые доли
ρ_i - истинные доли

Если бы мы знали истинные доли в каждом из сплитов, то смогли бы рассчитать вероятность наблюдать k_i используя биномиальное распределение

$$p(k_i|n_i, \rho_i) = Bin(n_i, \rho_i) \tag{3.15}$$

Однако наша цель определить вероятность доли. Для этого воспользуемся формулой Байеса

$$p(\rho_i = c|n_i, k_i) = \frac{p(k_i|n_i, \rho_i = c)p(\rho_i = c)}{\int_0^1 p(k_i|n_i, x)p(x)dx} \tag{3.16}$$

Поскольку в качестве апостериорного распределения мы используем биномиальное, можем сразу указать априорное - beta-распределение.

Причем вначале будем считать, что $k_i = 0$ и $n_i = 0$ при котором beta-распределение будет равномерным.

p-value вероятность наблюдать значение из тестируемого сплита (или хуже) при условии, что верна нулевая гипотеза. На формулах это будет выражено так

$$p_{value} = \int\limits_{k_1/n_1}^{1} Beta(x, k_0, n_0)dx$$

Доверительный интервал задается через уровень статзначимости α и формально может быть определен двумя уравнениями

$$\int\limits_{0}^{r_l} Beta(x, k_i, n_i)dx = \frac{\alpha}{2} \tag{3.17}$$

$$\int\limits_{r_r}^{1} Beta(x, k_i, n_i)dx = \frac{\alpha}{2} \tag{3.18}$$

r_l - левая граница интервала, r_r - правая. Построенный таким образом интервал будет несимметричным, что бьется с тем, что частота всегда должна лежать в интервале $[0,\ 1]$ потому, что в противном случае мы бы вышли за пределы интервала.

В классической схеме интервал всегда получается симметричным.

MDE. В классическое схеме используется трюк в котором любое нормальное распреление можно масштабировать в стандартное простым линейным преобразованием. К сожалению, для beta-распределения такого трюка нет и придется все вычисления проводить напрямую. Воспользуемся идей из формулы 3.7, то есть при фиксированных значениях α и β вычислить минимальное δ при котором выполняется

$$\begin{cases} \int_{r_i+\delta}^{1} Beta(x, k_0, n_0)\,dx & = \alpha \\ \int_{0}^{r_i+\delta} Beta(x, k_1, n_1)\,dx & = \beta \end{cases}$$

3.7 Исследование дисперсии

Предположим имеется выборка $x_1, \dots, x_n \sim N(\mu, \sigma^2)$. Для исследования дисперсии выбирается одна из статистик в зависимости от того, что мы знаем о выборке.

μ, σ^2 — истинные мат.ожидание и дисперсия
m, s^2 — выборочные мат.ожидание и дисперсия

- При *известном* мат.ожидании считаем величину

$$H = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma^2} \right)^2 \tag{3.19}$$

См. следствия из леммы Фишера в Н.И. Чернова. "Лекции по математической статистике". В: *Новосибирск: НГУ* (2003)

которая подчинена распределению χ^2_n с n-степенями свободы.

- При *неизвестном* мат.ожидании считаем величину

$$H = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma^2} \right)^2 = \frac{(n-1)s^2}{\sigma^2} \tag{3.20}$$

которая имеет распределение χ^2_{n-1} с $(n-1)$-степенью свободы.

3.7.1 Доверительный интервал

Поскольку мы знаем распределение статистики H, можем определить границы в которых выполняется условие $P(q_{\alpha/2} < H < q_{1-\alpha/2}) = 1 - \alpha$ и использовать их для определения доверительного интервала для дисперсии.

q_ψ - ψ-квантиль распределения χ^2

Например, при неизвестном мат.ожидании

Обрати внимание, что левый квантиль определяет правую границу, а правый квантиль - левую.

$$1-\alpha = P\left(q_{\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < q_{1-\alpha/2} \right) = P\left(\frac{(n-1)s^2}{q_{1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{q_{\alpha/2}} \right)$$

остается только посчитать квантили $q_{1-\alpha/2}$ и $q_{\alpha/2}$ и подставить в формулу.

3.7.2 Гипотеза о числовом значении дисперсии

В случае простой гипотезы, когда у нас есть предположение о значении дисперсии можно посчитать критические точки через интервальную оценку. Однако, проще воспользоваться оценкой p-value

$$\begin{aligned} F &= \int_0^H f_{\chi^2}(x)dx \\ p_{value} &= 2\min\{F, 1-F\} \end{aligned} \tag{3.21}$$

Простая гипотеза:

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma \neq \sigma_0 \end{cases}$$

Сложная гипотеза:

$$\begin{matrix} x_1, \dots, x_n \sim N(\mu_1, \sigma_1) \\ y_1, \dots, y_k \sim N(\mu_2, \sigma_2) \end{matrix} \quad \begin{cases} H_0 : \sigma_1 = \sigma_2 \\ H_1 : \sigma_1 \neq \sigma_2 \end{cases}$$

3.7.3 Гомогенность дисперсии

Предположим, что есть вторая выборка Y и мы хотим проверить, что в обеих выборках дисперсии одинаковы. Величина $F = s_x^2/s_y^2$ имеет распределение Фишера-Снедкора со степенями свободы $n-1$ и $k-1$. В случае если дисперсии равны, значение статистики F должно находиться в районе 1. Зная распределение можно вычислить критическую точку или посчитать p-value.

s_x^2 — исправленная дисперсия выборки X
s_y^2 — исправленная дисперсия выборки Y

3.8 ANOVA: Дисперсионный анализ

ANOVA: ANalysis Of VAriance

Тест предназначен для экспериментов когда имеется более чем 2 группы. ANOVA проверяет, что все группы имеют одинаковое мат.ожидания. Альтернативой будет наличие хотя бы одной пары в которой не совпадают мат.ожидания.

q — количество групп
n_j — количество наблюдений в группе j
y_{ij} — i-ое наблюдение в группе j
μ_j — мат.ожидание в группе j

$$\begin{cases} H_0: & \text{для всех } j,\ k, \quad \mu_j = \mu_k, \\ H_1: & \text{есть пара } j,\ k, \quad \mu_j \neq \mu_k \end{cases} \tag{3.22}$$

$$n = \sum_{j=1}^{q} n_j$$

Построим общую сумму квадратов

$$SS_{total} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}}) \tag{3.23}$$

$$\begin{aligned} \overline{y}_j &= \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \\ \overline{\overline{y}} &= \frac{1}{n}\sum_{j=1}^{q} n_j \overline{y}_j \end{aligned}$$

и разложим ее не две составляющие

$$\sum_{i=1}^{n_j}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2 = (n_j - 1)s_j^2 + n_j(\overline{y}_j - \overline{\overline{y}})^2$$

Используя эти формулы сможем разложить SS_{total} на сумму внутригрупповых отклонений и на межгрупповых отклонений

$$SS_{total} = SS_{residual} + SS_{between} \tag{3.24}$$

$SS_{residual}$ имеет $n-q$ степеней свободы
$SS_{between}$ имеет $q-1$ степеней свободы

$$SS_{residual} = \sum_{j=1}^{q}(n_j - 1)s_j^2 \qquad SS_{between} = \sum_{j=1}^{q} n_j(\overline{y}_j - \overline{\overline{y}})^2 \tag{3.25}$$

Если мат.ожидания во всех группах действительно равны, то межгрупповые ошибки не отличаются от внутригрупповых ошибок, а значит статистика F подчинена распределению Фишера

$$F = \frac{SS_{between}/(q-1)}{SS_{residual}/(n-q)}\,. \tag{3.26}$$

3.8.1 Метод контрастов

Если мы отвергли нулевую гипотезу 3.27 и есть пара групп с различ-

$$\begin{aligned} y &= c_1\overline{y}_1 + \ldots + c_q\overline{y}_q \\ M(y) &= c_1\mu_1 + \ldots + c_q\mu_q \\ D(y) &= \frac{c_1^2\sigma_1^2}{n_1} + \ldots + \frac{c_q^2\sigma_q^2}{n_q} \end{aligned}$$

для формулы дисперсии смотри 2.1

ными мат.ожиданиями, следующим шагом будет указать эту пару. Рассмотрим более общую задачу, предположим, что есть линейная зависимость между значениями μ_j, то есть существуют такие коэффициенты *(и мы их знаем)*, что $\sum_{j=1}^{q} c_j \overline{y}_j = 0$. Например, мы можем проверять гипотезу, что $\mu_2 = \mu_3$, что эквивалентно $\mu_2 - \mu_3 = 0$. Можем сформировать нулевую гипотезу и альтернативу

$$\begin{cases} H_0: & \sum_{j=1}^{q} c_j \overline{y}_j = 0 \\ H_1: & \sum_{j=1}^{q} c_j \overline{y}_j \neq 0 \end{cases} \tag{3.27}$$

Изначально, мы ничего не знаем про дисперсии в группах. Поэтому делаем предположение, что дисперсии одинаковы для всех групп $\sigma_1 = \ldots = \sigma_q$. Тогда нам нужна оценка внутригрупповых дисперсий

$$s^2_{residual} = SS_{residual}/(n-q) \tag{3.28}$$

Тогда оценить гипотезу H_0 мы можем используя доверительный интервал для мат.ожидания

$$y \pm t_{1-\alpha,\, n-q}\, s_{residual} \sqrt{\sum_{j=1}^{q} \frac{c_j^2}{n_j}}. \tag{3.29}$$

Если интервал не покрывает ноль, то нулевая гипотеза отвергается.

Для подбора коэффициентов нужно привлекать внешние предположения о природе данных. Нельзя просто подобрать коэффициенты c_j исходя из выборки, поскольку в этом случае ничто не мешает подобрать коэффициенты при которых выборочные средние в линейной комбинации дадут точный ноль.

3.9 Проверка зависимости

При вычислении корреляций надо понимать, что они ищут только линейные зависимости, коэффициенты могут быть очень чувствительными к выбросам и иногда могут показывать ложную корреляцию, например, для монотонных временных рядов.

При проверке зависимости между двумя случайными величинами используют коэффициенты корреляции:

- Пирсона, если данные нормальные.
- Спирмена, в остальных случаях.
- Кендела, очень редко.

3.10 CUPED

Controlled-experiment Using Pre-Experiment Data

Alex Deng и др. “Improving the sensitivity of online controlled experiments by utilizing pre-experiment data”. В: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, с. 123—132

Предположим мы хотим отловить изменения в случайной величине Y, но она обладает очень большой дисперсией. Высокая дисперсия увеличивает время эксперимента, следовательно, если нам удастся понизить дисперсию мы сможем либо ускорить эксперимент, либо выиграть в MDE, смотри формулу (3.7).

Изменения в величине Y мы будем оценивать через среднее значение $M(Y)$, а значит будем использовать t-test. В знаменателе формулы (3.5) для подсчета t-статистики находится сумма дисперсий экспериментальной и контрольной выборок. Цель CUPED уменьшить знаменатель чтобы повысить чувствительность метода.

Y — случайная величина
$M(Y)$ — среднее значение величины Y
n — объем выборки
$\overline{Y}$ — выборочная оценка среднего
$var(\overline{Y}) = var(Y)/n$

Предположим имеется случайная величина X коррелированная с величиной Y. Рассмотрим величину

Корреляция между X и Y, а так же $M(X)$ и есть пред экспериментальные данные необходимы для метода CUPED.

$$\hat{Y}_{cv} = \overline{Y} - \theta\overline{X} + \theta M(X) \tag{3.30}$$

Мат.ожидание величины $\hat{Y}_{cv}$ совпадает с мат.ожиданием величины Y, это легко показать если воспользоваться свойством $M(\overline{Y}) = M(Y)$. В свою очередь дисперсия величины выражается сложнее

$$var(\hat{Y}_{cv}) = var(\overline{Y} - \theta\overline{X}) = \frac{1}{n} var\left(Y - \theta X\right) = \frac{1}{n}\left(var(Y) + \theta^2 var(X) - 2\theta\, cov(X,Y)\right) \tag{3.31}$$

Теперь мы можем подобрать параметр θ при котором достигается минимальное значение $var(\hat{Y}_{cv})$. Собственно функция в правой части формулы (3.31) — парабола, значит минимум функции достигается при $\theta = cov(X,Y)/var(X)$. Минимальная дисперсия которую может достичь $\hat{Y}_{cv}$ равна

$$var(\hat{Y}_{cv}) = var(\overline{Y})(1 - \rho^2), \tag{3.32}$$

где $\rho = cor(X,Y)$ это корреляция между величинами X и Y. Чем сильнее скоррелированы величины тем более низкой дисперсии мы можем достичь.

3.10.1 Стратификационное сэмплирование

Один из методов снизить дисперсию которым часто пользуются — стратифицировать метрику. Метод может использоваться как альтернатива CUPED или вместе с ним.

Stratification sampling — метод при котором мы сперва считаем метрику по отдельности для каждой страты, а общую среднюю считаем через взвешенное среднее. Это отличается от подхода с вычислением выборочной средней тем, что исчезает вклад дисперсии между стратами. Покажем это на формулах.

K — количество страт
Y_{kj} — наблюдение j в страте k
n_k — количество наблюдений в страте k

Выборочное среднее

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}$$

Взвешенное среднее

$$\hat{Y} = \sum_{k=1}^{K} p_k \overline{Y}_k$$

Воспользуемся формулой (7) полной дисперсии с учетом страт и рассчитаем дисперсию выборочной средней

$$var(\overline{Y}) = \frac{1}{n}\, var(Y) = \frac{1}{n}\left(\sum_{k=1}^{K}\sigma_k^2\, p_k + \sum_{k=1}^{K}(\mu_k - \mu)^2\, p_k\right)$$

Теперь рассчитаем дисперсию взвешенной средней. Положим, что вероятность попасть в страту k рассчитывается в виде отношения $p_k = n_k/n$, тогда

$$var(\hat{Y}) = \sum_{k=1}^{K} p_k^2 var(\overline{Y}_k) = \sum_{k=1}^{K}\frac{n_k^2}{n^2}\,\frac{1}{n_k}\,\sigma_k^2 = \frac{1}{n}\sum_{k=1}^{K}\sigma_k^2\, p_k.$$

Из формул явно следует, что $var(\overline{Y}) \geq var(\hat{Y})$ и стратификационное сэмплирование может использоваться как метод снижения дисперсии. Эффект от стратификации тем выше, чем большая разница в средних наблюдается между стратами.

Если выбрать дискретную ковариату X то дисперсия по CUPED совпадает с дисперсией по стратификации.

3.10.2 Рекомендации по запуску

Ковариата — дополнительная случайная величина X коррелированная с Y

В качестве ковариаты часто используют поведение метрики Y, но до эксперимента. Например, если Y — количество запросов юзера, то в качестве ковариаты X удобно взять эту же метрику на исторических данных до эксперимента. Период до эксперимента на котором мы считаем метрику не обязан совпадать с длительностью эксперимента, поскольку важна только коррелированность X и Y.

Например, если метрика Y это CTR web-страницы, удобно делать группировку по страницам и брать исторические данные по CTR страниц до эксперимента.

С поюзерными метриками есть проблема, что во время эксперимента могут появиться пользователи которых не было до эксперимента. Этих ребят можно рассматривать отдельно.

$X^{(t)}$ — метрика на тесте
$X^{(c)}$ — метрика на контроле

В целом для запуска CUPED не обязательно концентрироваться только на юзерских метриках или подобных. Можно выбрать любую коррелированную метрику X, однако помимо высокой корреляции важно чтобы выполнялось условие

$$M\left(X^{(t)}\right) = M\left(X^{(c)}\right) \tag{3.33}$$

Как ковариату можно использовать *день недели* в который юзер впервые появился в эксперименте не зависит от эксперимента как такового.

Если это условие не выполнено, то мы получим смещенную оценку на $\Delta_{cv} = \hat{Y}_{cv}^{(t)} - \hat{Y}_{cv}^{(c)}$. Например, известно, что скорость загрузки страницы влияет на CTR. Если эксперимент заключается в ускорении загрузки страницы и в качестве ковариаты берем CTR, то оценка Δ_{cv} окажется смещенной.

3.10.3 Обобщения метода

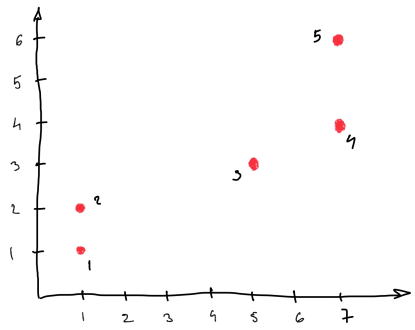
Question: Если оптимизировать по классам функций f, то *вероятно* оптимум достигается когда f — линейная регрессия.

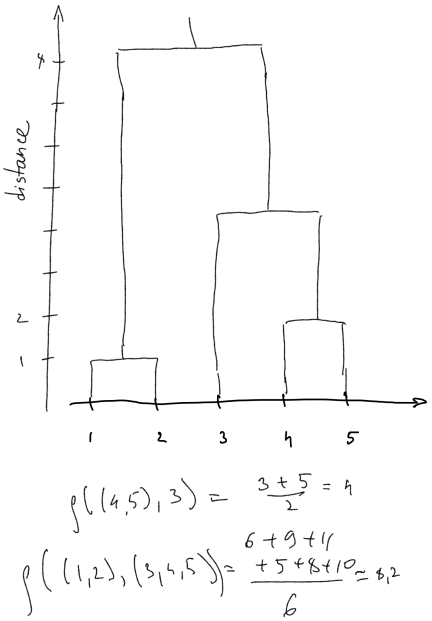
У метода есть понятный вектор обобщений. Если предположить, что есть вектор-ваиант $\mathbf{X} = (X^1, \ldots, X^n)$ и некоторая функция $f : \mathbb{R}^n \to \mathbb{R}$, то мы можем минимизировать дисперсию для

$$\hat{Y}_{cv} = \overline{Y} - \overline{f(\mathbf{X})} + M(f(\mathbf{X})) \tag{3.34}$$

Глава 4

Кластеризация





4.1 Кластерный анализ

Расстояния между кластерами могут быть заданы по разному и стоит обращать внимание на решаемую задачу. Чтобы задать расстояние между кластерами надо определить расстояние между точками и на ее основе определить метод построения кластерного расстояния. Чаще всего используют такие методы:

- среднее невзвешенное расстояние (average linkage clustering) - среднее расстояние между парами из различных кластеров
- центроид метод (устаревший) - расстояние между кластерами определяется как расстояние между центрами масс.
- метод дальнего соседа (complete linked clustering) - по расстоянию между максимально удаленными точками кластеров
- метод ближайшего соседа (single linkage clustering) - по расстоянию между самыми близкими точками
- метод Варда (Ward's method)

Дендрограмма выращивается снизу вверх. Сперва каждая точка представляет свой собственный кластер. Если при достижении уровня h расстояние между точками равно h то точки объединяются в один кластер (на рисунке это спайка). Процесс повторяется до тех пор пока не останется один единственный кластер.

Для определения количества кластеров используют метод *плеча*. Для этого строится график где по оси абсцисс отмечаются шаги объединения (первое объединение, второе и так далее), по оси ординат откладывается высота на котором произошло объединение кластеров. На графике ищется точка перелома, когда расстояние резко увеличивается. Это количество кластеров в данных по методу локтя.

Недостатки иерархической кластеризации:

- Плохо справляется с данными которые вытянуты в длинные ленты в пространстве

Результаты кластерного анализа нужно интерпретировать: анализ всегда должен давать что-то новое о данных, что общего у объектов в кластере и чем различаются кластеры.

Кривую локтя можно построить и для метода k-средних. В этом случае по оси абсцисс будет лежать выбранный k по оси ординат качество получившейся кластеризации.
См. Федотов Станислав и Синицин Филипп. *Учебник по машинному обучению*. URL: `https://academy.yandex.ru/handbook/ml`, глава 7.

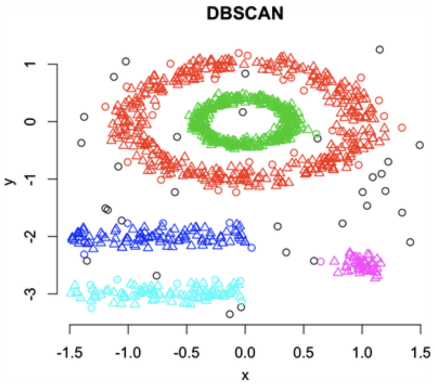
- Для вычисления требуется хранить попарные расстояния между объектами.

4.2 DBSCAN

Density-based spatial clustering of applications with noise. Сильный топологический алгоритм. Основан на поиске связных компонент в покрытии данных ϵ-шарами. Количество кластеров определяет автоматически.

Все объекты в наблюдениях делятся на три категории:

- core point (внутренние / основные) - если в окрестности есть более N_0 соседей
- border point (граничные) - если в окрестности меньше N_0 внутренних
- noise point (шумовые) - если в окрестности нет внутренних точек. Автоматически содержат меньше N_0 объектов.



Алгоритм DBSCAN:

1. Шумовые точки удаляются из рассмотрения и не приписываются ни какому кластеру.
2. Основные точки у которых есть общая окрестность соединяются ребром. Строится граф.
3. В полученном графе вычисляются компоненты связности.
4. Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней внутренняя точка.

Глава 5

Классификация и регрессия

5.1 Линейная регрессия

Для Определения на сколько модель хороша используют коэффициент детерминации. Чтобы его вычислить нужна обученная модель.

$$R^2 = 1 - \frac{D[y|x]}{D[y]} \tag{5.1}$$

В сущностной части это на сколько дисперсия предсказания отличается от дисперсии наблюдений.

Во время построения модели регрессии самая большая проблема - наличие коллинеарных векторов (зависимых переменных в фичах). Чтобы очистить список фичей, можно проверить гипотезу что вычисленная фича нулевая (для каждой фичи индивидуально).

$$\begin{cases} H_0: & \hat{\beta}_i = 0 \\ H_1: & \hat{\beta}_i \neq 0 \end{cases}$$

Если предположить что коэффициент имеет нормальное распределение, то мы можем воспользоваться t-тестом Стьюдента и посчитать p-value. Если гипотеза не будет отвергнута, это будет означать, что мы фичу стоит выкинуть потому что ее коэффициент равен нулю. Соответственно там где гипотеза не подтвердилась фичи являются ценными.

Выкидывать фичи стоит по одному. Начиная с самой слабой по p-value, каждый раз с новым набором фичей анализ лучше проводить заново.[2]

5.1.1 Для временных рядов

Линейная регрессия может быть очень полезной во временных рядах, когда мы наблюдаем в них несколько сезонностей или когда данных относительно мало, нет нескольких полных сезонов в данных.

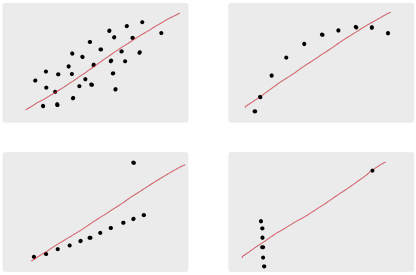


Рис. 5.1: Квартет Анскомба. Иллюстрация как линейная регрессия может лажать. Для данных подобрали не ту модель или в данных есть сильные выбросы - все приводит к некорректному построению регрессии.

Наивная мультивариантная линейная регрессия:

$$y = X\beta + \epsilon$$
$$\epsilon \sim N(0, \sigma^2 I)$$

ϵ - коэффициент *невязки*, определяет шум системы.
Коэффициенты находятся формулой

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Вариация вычисленных коэффициентов

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Это легко вывести если воспользоваться формулой дисперсии произведения

$$Var(AX) = A \cdot Var(X) \cdot A^T.$$

[2] Valerie Watts. *Introduction to Statistics. An Excel-Based Approach.* URL: `https://ecampusontario.pressbooks.pub/introstats/`

Чтобы учесть сезонность во временных рядах создают фичи:

- 1 - свободный коэффициент, bias
- $\delta_s(t)$ - учет сезонности, бинарный признак, что дата/время t попало во временной интервал s (час, день, неделя, месяц и пр).

При этом очевидно, что если сложить все фичи $\delta_s(t)$ сезонности мы получим свободную переменную, то есть на лицо линейная зависимость.

5.2 Деревяшки (решающие деревья)

Обозначения:

N — объем выборки

D — размерность признаков

$y = \{y_i\}_{i=1}^N \subset \mathbb{R}^N$ — вектор таргетов

$X = \{x_i\}_{i=1}^N \in \mathbb{R}^{N \times D}$ — матрица признаков

Решающее дерево имеет в узлах предикаты, а в листьях метки классов или другие значения в зависимости от задачи.

Чаще всего строят бинарные решающие деревья и используют такие предикаты:

$$B(x, j, t) = [x^j \leq t] = B_{j,t}(x) \tag{5.2}$$

Если в вершине выбрано какое-то разбиение и множество R было разбито на два R_l и R_r то можно почитать на сколько мы сделали лучше этим разбиением.[3]

$$Q(R, B_{j,t}) = H(R) - \frac{|R_l|}{|R|} H(R_l) - \frac{|R_r|}{|R|} H(R_r) \tag{5.3}$$

Impurity criterion (критерий информативности) $H(R)$ оценивает качество распределения целевой переменной в множестве R.
Поскольку это про оценку распределения часто используют энтропию или индекс Джини, но также может использоваться и квадрат/модуль ошибок и прочее.

Чтобы выбрать лучший вариант разбиения условия разбиения по признаку j можно упорядочить все наблюдения в узле по этому признаку постепенно перебрать все варианты разбиений и выбрать тот в котором функция качества Q достигает максимума. Эту процедуру повторяем для всех $j = \overline{1, D}$ и выбираем лучшее разбиение.

$$(j_{opt}, t_{opt}) = \arg\max_{j,t} Q(R, B_{j,t}) \tag{5.4}$$

Для процедуры критерий останова можно вводить разными способами. Самые популярные такие:

- максимальная глубина дерева,
- максимальное количество листьев,
- минимальное число объектов в листе,
- все объекты в листе лежат в одном классе,
- качество улучшилось не более чем на x процентов.

Деревья дают кусочно-линейные решения, однако у них есть преимущества в том, что они могут вытащить нелинейные зависимости; а также и недостатки, за пределами наблюдений регрессия будет предсказываться константой. Другими словами регрессия на деревьях не умеет экстраполировать, зато может быть полезна в интерполяции.

Если на тестовой выборке результат заметно хуже, а мы очень хотим обучить дерево, можно попробовать упростить модель, сильнее ограничить глубину дерева.

Для решения *задачи регрессии* достаточно подобрать критерий информативности. Например, это может быть квадрат отклонения от среднего

$$H(R) = \frac{1}{|R|} \sum_{(x_i,\, y_i) \in R} \left(y_i - \frac{1}{|R|} \sum_{(x_i,\, y_i) \in R} y_i \right)^2$$

[3] Е.А. Соколов. *Решающие деревья.* URL: https://www.hse.ru/mirror/pubs/share/215285956

Список литературы

[1] Alex Deng и др. “Improving the sensitivity of online controlled experiments by utilizing pre-experiment data”. В: *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013, с. 123—132.

[2] Gerald Van Belle. *Statistical rules of thumb*. Т. 699. John Wiley & Sons, 2011.

[3] Joseph C Watkins. “An introduction to the science of statistics: From theory to implementation”. В: (2019).

[4] Valerie Watts. *Introduction to Statistics. An Excel-Based Approach*. URL: `https://ecampusontario.pressbooks.pub/introstats/`.

[5] Анатолий Карпов. *Тонкости A/B тестирования: проблема подглядывания*. URL: `https://www.youtube.com/live/jnFVmtaeSA0`.

[6] Е.А. Соколов. *Решающие деревья*. URL: `https://www.hse.ru/mirror/pubs/share/215285956`.

[7] Федотов Станислав и Синицин Филипп. *Учебник по машинному обучению*. URL: `https://academy.yandex.ru/handbook/ml`.

[8] Н.И. Чернова. “Лекции по математической статистике”. В: *Новосибирск: НГУ* (2003).