

3 О мерах зависимости и корреляционных расстояниях

В рамках этого занятия мы будем заниматься вопросами зависимости и независимости. Простейшей мерой выражения зависимости для нас является коэффициент корреляции

$$\rho_{X,Y} = \frac{\mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}}.$$

Этот коэффициент $\rho \in [-1, 1]$, причем для независимых величин $\rho = 0$, а для линейно зависимых величин (и только для них) $\rho = 1$ или $\rho = -1$. Коэффициент ρ хорошо отражает прямую (линейную зависимость), но не отслеживает более сложной зависимости.

Вопрос 1. Показать, что для $X \sim \mathcal{N}(\mu, \sigma^2)$ коэффициент ρ_{X,X^2} равен нулю, хотя величины зависят.

Тем не менее, отличие коэффициента корреляции от нуля показывает нам наличие зависимости. В связи с этим используют оценку для ρ , имеющую вид

$$\hat{\rho} = \frac{\overline{XY} - \overline{X}\overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}.$$

Мы обсудим эту меру зависимости и ее модификации.

3.1 Независимость дискретных выборок

Для проверки независимости двух дискретных выборок данные удобно представить в виде так называемых таблиц сопряженности (contingency table). Если X принимает значения x_1, \dots, x_k , а Y — y_1, \dots, y_l , то составим таблицу со столбцами y_1, \dots, y_l и строками x_1, \dots, x_k , где в ячейке (i, j) стоит $p_{i,j} = \mathbf{P}(X = x_i, Y = y_j)$. Если гипотеза верна, то $p_{i,j} = p_i q_j$, где $p_i = \mathbf{P}(X = x_i)$, $q_j = \mathbf{P}(Y = y_j)$.

1. Критерий хи-квадрат.

Используем параметрический критерий хи-квадрат на основе частот $\nu_{i,j}$ появления пар (x_i, y_j) .

Тогда гипотеза заключается в том, что наше $kl - 1$ параметрическое семейство вероятностей $p_{i,j}$ выражается через $k + l - 2$ вероятности $p_i, q_j, i < k, j < l$. Функция правдоподобия при выполнении гипотезы имеет вид

$$\prod_{i=1}^k \prod_{j=1}^l (p_i q_j)^{\nu_{i,j}} = \prod_{i=1}^k e^{\ln p_i \sum_{j=1}^l \nu_{i,j}} \prod_{j=1}^l e^{\ln q_j \sum_{i=1}^k \nu_{i,j}}$$

Отсюда ОМП $\hat{p}_i = \nu_{i,.}/N = \sum_{j=1}^l \nu_{i,j}/N$, $\hat{q}_j = \sum_{i=1}^k \nu_{i,j}/N = \nu_{.,j}/N$.

Следовательно, при выполнении гипотезы

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(\nu_{i,j} - N\hat{p}_i\hat{q}_j)^2}{N\hat{p}_i\hat{q}_j} \sim \chi^2_{(k-1)(l-1)}$$

Отсюда возникает критерий хи-квадрат для проверки независимости, который в R задан все тем же chisq.test.

2. Критерий отношения правдоподобий.

Классический подход

$$2 \ln \left(\frac{\sup_{p_{i,j}} L(x, p)}{\sup_{p_i, q_j} L(x, p)} \right) = 2 \sum_{i,j} \ln \left(\frac{\nu_{i,j}^{\nu_{i,j}}}{(\nu_{i,.}\nu_{.,j}/N)^{\nu_{i,j}}} \right) = 2 \sum_{i,j} \nu_{i,j} \ln \left(\frac{\nu_{i,j} N}{\nu_{i,.}\nu_{.,j}} \right) \sim \chi^2_{(k-1)(l-1)}.$$

Тем самым, статистики критериев хи-квадрат и отношения правдоподобий для проверки однородности и независимости одинаковы. Это следствие того, что независимость X и Y равносильна совпадению условных распределений $\mathbf{P}(X = x|Y = y_1), \mathbf{P}(X = x|Y = y_2), \dots, \mathbf{P}(X = x|Y = y_k)$.

В случае двух выборок одинаковы и рассматриваемые предельные распределения.

3.2 Независимость в общем случае

3.2.1 Меры связи

Мерами связи (measure of association) называют коэффициенты, показывающие взаимосвязь параметров.

1. Коэффициент корреляции Пирсона ρ .

Оценим коэффициент корреляции выборочной величиной

$$\rho = \frac{\overline{XY} - \overline{X} \overline{Y}}{\sqrt{S_X^2 S_Y^2}},$$

где S_X^2 , S_Y^2 — смещенные оценки дисперсий $(\overline{(X - \bar{X})^2})$, $(\overline{(Y - \bar{Y})^2})$. Этот коэффициент лежит в диапазоне $[-1, 1]$, причем 1 будет достигаться только при $X = aY + b$ п.н. при некотором $a > 0$, -1 при $X = aY + b$ п.н., $a < 0$. Теоретический коэффициент корреляции измеряет наличие прямой линейной зависимости, поэтому коэффициент Пирсона также будет принимать большие абсолютные значения именно при линейной зависимости, но хуже будет отслеживать другие виды зависимостей.

Коэффициент Пирсона обычно применяют для двумерных нормальных выборок, для которых он является неплохой мерой зависимости.

В Python он реализован в `pearsonr` в `scipy.stats` (включая критерий).

2. Коэффициент Спирмена ρ_S . Этот коэффициент позволяет уйти от условия линейности зависимости, заменив наблюдения X_i на их ранги R_i в ряду X , а Y_i — на их ранги T_i в ряду Y . Тогда

$$\rho_S = \frac{\overline{RT} - \overline{R} \overline{T}}{\sqrt{S_R^2 S_T^2}}.$$

Тогда близость ρ_S по абсолютному значению к 1 означает, что R_i линейно зависят от T_i , т.е. зависимость Y от X монотонна.

В Python он реализован в `spearmanr` в `scipy.stats` (включая критерий).

3. Коэффициент Кенделла τ .

Назовем две пары значений x_i, y_i, x_j, y_j согласованными, если $x_i - x_j, y_i - y_j$ — одного знака. Пусть C — количество согласованных, D — несогласованных пар. Тогда

$$\tau = \frac{C - D}{C + D} = \frac{2(C - D)}{n(n - 1)}$$

называют коэффициентом согласия Кенделла или τ -коэффициентом. Он также лежит в диапазоне $[-1, 1]$. Этот коэффициент сильно коррелирован с коэффициентом ρ_S .

В Python он реализован в `kendallr` в `scipy.stats` (включая критерий).

Можно визуализировать корреляционные взаимосвязи между множеством переменных с помощью тепловой диаграммы, как это описано здесь.

3.2.2 Критерии проверки независимости

Для двумерных нормальных данных удобно использовать так называемое преобразование Фишера $\text{artanh}(z) = 0.5 \ln((1 + z)/(1 - z))$, приводящее коэффициент ρ к асимптотической нормальности со среднем $\text{artanh}(\rho)$ и дисперсией $1/(n - 3)$, откуда можно построить доверительный интервал для ρ . Аналогичный результат верен и для коэффициентов Спирмена и Кенделла.

Можно использовать коэффициент Пирсона, но гипотезу проверять перестановочным подходом. В этом случае критерий будет работать уже не только для нормального распределения.

При выполнении гипотезы независимости справедливы сходимости

$$\frac{\rho_S}{\sqrt{\mathbf{D}\rho_S}} = \rho_S \sqrt{n-1} \xrightarrow{d} Z \sim \mathcal{N}(0, 1), \quad \frac{\tau}{\sqrt{\mathbf{D}\tau}} = \tau \sqrt{\frac{9n(n-1)}{2(2n+5)}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1),$$

откуда вытекают асимптотические критерии для проверки гипотезы независимости.

Вопрос 2. Показать, что для наблюдений без повторений $\mathbf{D}\rho_S = 1/(n-1)$.

3.2.3 Множественное сравнение

Для многих выборок ($k \geq 3$) аналогичную роль может сыграть коэффициент конкордации, подсчитанный по многим выборкам. Одним из таких коэффициентов является коэффициент Кенделла

$$W = \frac{12}{k^2(n^3 - n)} \sum_{i=1}^n \left(\sum_{j=1}^k R_{i,j} - \frac{k(n+1)}{2} \right)^2,$$

где $R_{i,j}$ — ранг i -ого элемента j -ой выборки внутри своей выборки. Эта величина лежит в диапазоне $[0, 1]$. При больших n $k(n-1)W$ близок по распределению к χ^2_{n-1} .

Еще одним важным коэффициентом является частный или очищенный коэффициент корреляции, позволяющий исключить зависимость двух переменных через третью

$$\frac{\rho_{X,Y} - \rho_{X,Z}\rho_{Y,Z}}{\sqrt{(1-\rho_{X,Z}^2)(1-\rho_{Y,Z}^2)}},$$

который оценивается исходя из выборки с помощью статистики

$$\frac{\hat{\rho}_{X,Y} - \hat{\rho}_{X,Z}\hat{\rho}_{Y,Z}}{\sqrt{(1-\hat{\rho}_{X,Z}^2)(1-\hat{\rho}_{Y,Z}^2)}},$$

где $\hat{\rho}$ — выборочные коэффициенты корреляции Пирсона. Этот коэффициент эффективно работает для многомерных нормальных данных, для общих данных можно использовать выборочные коэффициенты корреляции Кенделла.

В более общем виде существуют аналоги, учитывающие несколько устранимых переменных.

В Python частные корреляции реализованы здесь

3.3 Ковариационные расстояния

Работа с ковариацией затрудняется тем, что она измеряет именно линейную зависимость. Для измерения общей зависимости можно использовать другие функционалы — ковариационные расстояния. При этом многомерные критерии однородности позволяют проверять и независимость, ведь независимость проверяет однородность распределений H и $F \times G$, где H — совместное распределение X, Y , а F, G — маргинальные распределения X и Y . Тем самым, мы можем использовать многомерные подходы к однородности из пакета `hyppo` с прошлой пары (все описанные ниже методы есть в том же пакете).

1. Метод Секея-Риццо (Szekely-Rizzo, 2009). Этот метод называют также DCor - Distance Correlation, это аналог Energy distance.

Этот метод предлагает рассматривать в качестве меры зависимости между случайными векторами $X \in \mathbb{R}^k, Y \in \mathbb{R}^l$, имеющими конечное математическое ожидание, величину

$$V_{X,Y}^2 = \int_{t \in \mathbb{R}^k, s \in \mathbb{R}^l} |\psi_{X,Y}(t, s) - \psi_X(t)\psi_Y(s)|^2 w(t, s) dt ds,$$

обнуляющуюся только при независимых векторах X, Y , где ψ — характеристические функции, w

— некоторая (интегрируемая) весовая функция. Тогда

$$\rho_V = \frac{V_{X,Y}^2}{\sqrt{V_{X,X} V_{Y,Y}}}$$

будет принимать значения в отрезке $[0, 1]$. В роли $w(t, s)$ предлагается рассматривать

$$\frac{1}{c_k c_l \|t\|_k^{1+k} \|s\|_l^{1+l}}, \quad c_m = \frac{\pi^{1+m}}{\Gamma((1+m)/2)},$$

где норма рассматривается Евклидова. Для оценки ρ_V на выборке из n выборок (X, Y) введем

$$a_{i,j} = \|X_i - X_j\|_k, \quad a_{i,\cdot} = \frac{1}{n} \sum_{j=1}^n a_{i,j}, \quad a_{\cdot,j} = \frac{1}{n^2} \sum_{i,j} a_{i,j}, \quad b_{i,j} = \|Y_i - Y_j\|_l, \quad b_{\cdot,j} = \frac{1}{n} \sum_{i=1}^n b_{i,j}, \quad b_{\cdot,\cdot} = \frac{1}{n^2} \sum_{i,j} b_{i,j},$$

положим $A_{i,j} = a_{i,j} - a_{i,\cdot} - a_{\cdot,j} + a_{\cdot,\cdot}$, $B_{i,j} = b_{i,j} - b_{i,\cdot} - b_{\cdot,j} + b_{\cdot,\cdot}$. Тогда

$$\hat{\rho}_V = \frac{\sum_{i,j} A_{i,j} B_{i,j}}{\sqrt{\sum_{i,j} A_{i,j}^2 \sum_{i,j} B_{i,j}^2}}.$$

Это сильно состоятельная оценка ρ_V .

При этом можно построить асимптотический критерий, основанный на том, что при выполнении гипотезы независимости

$$\limsup_{n \rightarrow \infty} \mathbf{P} \left(\frac{n \hat{\rho}_V}{S_2} > z_{1-\alpha/2}^2 \right) \leq \alpha,$$

при некотором S_2 , где z — квантиль $\mathcal{N}(0, 1)$.

Дополнительным плюсом метода является то, что он не привязан напрямую к дискретности или непрерывности данных.

В случае многих переменных строят матрицу попарных расстояний и таким образом исследуют, какие переменные сильно зависят.

2. HSIC (Hilbert Schmidt Independence Criterion).

Идея метода в том, чтобы построить все попарные расстояния между парами (X_i, Y_j) в качестве первого набора, и расстояния между (X_i, Y_i) в качестве второго набора, а затем сравнить их друг с другом. Более конкретно, рассматривается матрица расстояний $K = k(x_i, x_j)$ и $L = l(y_i, y_j)$, матрица $H = I - 1/n$, где 1 — матрица из одних единиц, и строится величина

$$\frac{1}{n^2} \text{tr}(HKHL) = \frac{1}{n^2} \sum_{i,j=1}^n k(x_i, x_j) l(y_i, y_j) - \frac{2}{n^3} \sum_{i,i,j=1}^n k(x_i, x_{i'}) l(y_i, y_j) + \frac{1}{n^4} \sum_{i,i',j,j'=1}^n k(x_i, x_{i'}) l(y_j, y_{j'}).$$

Важную роль в этом методе имеет выбор ядер k, l .

3. Подход Хеллера, Хеллера и Горфина (Heller, Heller, Gorfine, 2012) также подходит для многомерных данных и основан на следующей идеи.

Пусть d — расстояние. Давайте фиксируем два индекса i, j и рассмотрим бернуlliевские величины

$$I_1 = I(d(X, x_i) \leq d(x_i, x_j)), \quad I_2 = I(d(Y, y_i) \leq d(y_i, y_j)),$$

где x_i, x_j, y_i, y_j — фиксированные точки выборки, а X, Y — наш случайный вектор. Если X, Y независимы, то индикаторы I_1, I_2 также независимы.

Гипотеза независимости этих двух бернуlliевских величин проверяется с помощью статистики

критерия отношения правдоподобий

$$S_{LR}(i, j) = 2 \sum_{k=0}^1 \sum_{l=0}^1 A_{k,l}(i, j) \ln \left(\frac{A_{k,l}(i, j)(N-2)}{A_{\cdot,l}(i, j)A_{k,\cdot}(i, j)} \right),$$

где $A_{k,l}(i, j)$ – количество наблюдений (x_m, y_m) для которых (то есть при подстановке которых вместо X, Y) индикаторы I_1, I_2 равны k и l соответственно,

$$A_{\cdot,l}(i, j) = A_{0,l}(i, j) + A_{1,l}(i, j), \quad A_{k,\cdot}(i, j) = A_{k,0}(i, j) + A_{k,1}(i, j).$$

Аналогично можно проверить эту гипотезу с помощью величины

$$S_{CS}(i, j) = \frac{(N-2)(A_{0,1}(i, j)A_{1,0}(i, j) - A_{1,1}(i, j)A_{0,0}(i, j))^2}{A_{0,\cdot}(i, j)A_{1,\cdot}(i, j)A_{\cdot,0}(i, j)A_{\cdot,1}(i, j)}.$$

Соответственно, подход HNG предлагает рассматривать в качестве критерия величину

$$T_1 = \sum S_{LR}(i, j)$$

или

$$T_2 = \sum_{i,j} S_{CS}(i, j).$$

Распределений для данных величин, насколько мне известно, получено не было, но можно проверять гипотезу критерием перестановок – переставлять величины x_i , не меняя y_i . При гипотезе независимости величина не изменится, а при альтернативе это существенно поменяет результат. Поэтому если величина T_1 для исходной выборки окажется среди αN самых больших значений по всем перестановкам, гипотезу можно отвергать.

4. Подход MGC (Multiscale Graph Correlation).

Идея в том, чтобы локализовать расстояния, рассматривая для каждой точки x_i ближайшие k точек к ней и для каждой y_j ближайшие l точек к ней. Для i, j точки введем

$$A_{i,j} = d(x_i, x_j) - \frac{1}{n-1} \sum_{l=1}^n d(x_i, x_l)$$

для x_j , являющейся одной из k ближайших точек к x_i (не считая самой точки) и 0 иначе. Аналогично определим $B_{i,j}$ для y_i, y_j . Определим

$$\begin{aligned} Cov &= \frac{1}{n(n-1)} \sum_{i \neq j} A_{i,j} B_{i,j} - \frac{1}{n(n-1)} \sum_{i \neq j} A_{i,j} \frac{1}{n(n-1)} \sum_{i \neq j} B_{i,j}, \\ D_X &= \frac{1}{n(n-1)} \sum_{i \neq j} A_{i,j}^2 - \left(\frac{1}{n(n-1)} \sum_{i \neq j} A_{i,j} \right)^2, \quad D_Y = \frac{1}{n(n-1)} \sum_{i \neq j} B_{i,j}^2 - \left(\frac{1}{n(n-1)} \sum_{i \neq j} B_{i,j} \right)^2. \end{aligned}$$

Положим $Corr = Cov / \sqrt{D_X D_Y}$, если $\min(D_X, D_Y) > \varepsilon$ и 0 иначе. Величина $Corr$ дальше сравнивается с нулем. Подход похож на предыдущие, однако, дополнительно прореживает матрицу расстояний, акцентируясь на локальном поведении расстояний рядом с каждой из точек.

Критерии HNG и MGC мощны, но достаточно медленны. DCorr и HSIC гораздо быстрее, HSIC более гибко настраиваемый подход, включающий в себя DCorr.

3.4 Ответы на вопросы

- Отметим, что коэффициент корреляции не меняется при линейных заменах, то есть $\text{corr}(aX + b, Y) = \text{corr}(X, Y)$. Следовательно, можно считать, что $X \sim \mathcal{N}(0, 1)$. Отсюда

$cov(X, X^2) = \mathbf{E}X^3 - \mathbf{E}X\mathbf{E}X^2 = 0$, поскольку $\mathbf{E}X = 0$, $\mathbf{E}X^3 = 0$ из симметрии.

2. Заметим, что раз совпадений нет, то R_i принимают значения от 1 до n и S_i тоже. Тогда

$$\rho_S = \frac{\frac{1}{n} \sum_{i=1}^n R_i S_i - \left(\frac{n+1}{2}\right)^2}{\frac{1}{n} \sum_{i=1}^n i^2 - \left(\frac{n+1}{2}\right)^2}.$$

При этом,

$$\mathbf{E}R_i^2 = \frac{1}{n} \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6n}, \quad \mathbf{D}R_i = \frac{n+1}{2} \left(\frac{2n+1}{3} - \frac{n+1}{2} \right) = \frac{(n+1)(n-1)}{12}.$$

Аналогично при $i \neq j$

$$\mathbf{E}R_i R_j = \sum_{i \neq j} \frac{ij}{n(n-1)} = \frac{(\sum_{i=1}^n i)^2 - \sum_{i=1}^n i^2}{n(n-1)} = \frac{\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}}{n(n-1)} = \frac{(n+1)(3n+2)(n-1)}{12(n-1)}$$

Отсюда при $i \neq j$

$$cov(R_i S_i, R_j S_j) = (\mathbf{E}R_i R_j)^2 - (\mathbf{E}R_i)^4 = \left(\frac{(n+1)(3n+2)}{12} \right)^2 - \frac{(n+1)^4}{2^4} = \frac{(n+1)^2((3n+2)^2 - 9(n+1)^2)}{12^2},$$

при $i = j$

$$cov(R_i S_i, R_i S_i) = (\mathbf{E}R_i^2)^2 - (\mathbf{E}R_i)^4 = \frac{(n+1)^2(2n+1)^2}{6^2} - \frac{(n+1)^4}{2^4} = \frac{(n+1)^2((4n+2)^2 - (3n+3)^2)}{12^2}.$$

Следовательно,

$$\mathbf{D}\rho_S = \frac{\frac{1}{n^2} \mathbf{D}(\sum_{i=1}^n R_i S_i)}{\frac{(n+1)^2(n-1)^2}{12^2}} = \frac{\sum_{i,j=1}^n cov(R_i S_i, R_j S_j)}{n^2(n+1)^2(n-1)^2/12^2} = \frac{-n(n-1) \frac{(n+1)^2(6n+5)}{12^2} + n \frac{(n+1)^2(n-1)(7n+5)}{12^2}}{n^2(n+1)^2(n-1)^2/12^2} = \frac{1}{n-1}.$$

3. Рассмотрим вектор (X, Y) , принимающий значения $(0, 1)$ и $(1, 0)$. Тогда из условия

$$\int_{\mathbb{R}^2} |F_{X,Y}(x, y) - F_X(x)F_Y(y)| dF_{X,Y}(x, y) = |F_{X,Y}(1, 0) - F_X(1)F_Y(0)| + |F_{X,Y}(0, 1) - F_X(0)F_Y(1)| = \\ |\mathbf{P}((X, Y) = (1, 0)) - \mathbf{P}((X, Y) = (1, 0))| + |\mathbf{P}((X, Y) = (0, 1)) - \mathbf{P}((X, Y) = (0, 1))| = 0,$$

хотя величины зависимы.