



Прикладная статистика и анализ данных

Съезд VI



0. Про задачи



Одна выборка

X_1, \dots, X_n — выборка из неизвестного распределения.

Задачи:

1. Подобрать класс распределений

Критерии согласия

2. Высказать суждение о параметре распределения

Точечные оценки

Доверительные интервалы

Гипотезы о значении параметра

3. Оценить распределение непараметрически

Ядерные оценки плотности



Две выборки

X_1, \dots, X_n и Y_1, \dots, Y_m — две выборки из неизвестных распределений.

Задачи:

1. Что-либо про одну из выборок — см. пред.
2. Исследовать зависимость выборок ($n = m$)

Анализ зависимостей

3. Сравнить распределения выборок

Дисперсионный анализ

4. Восстановить зависимость ($n = m$)

Регрессионный анализ

Классификация



Две выборки

X_1, \dots, X_n и Y_1, \dots, Y_m — две выборки из неизвестных распределений.

Задачи:

1. Что-либо про одну из выборок — см. пред.
2. Исследовать зависимость выборок ($n = m$)

Анализ зависимостей ← повестка съезда

3. Сравнить распределения выборок

Дисперсионный анализ

4. Восстановить зависимость ($n = m$)

Регрессионный анализ

Классификация



Анализ зависимостей

Даны **парные** выборки:

$$X = (X_1, \dots, X_n)$$

$$Y = (Y_1, \dots, Y_n)$$

Задачи:

- ▶ Зависимы ли выборки?

H_0 : выборки независимы vs. H_1 : выборки зависимы

- ▶ Количественная оценка степени

неслучайности их совместного изменения.



1. Коэффициенты корреляции



Коэффициент корреляции

Пусть ξ, η — случайные величины.

$$\text{corr}(\xi, \eta) = \frac{\text{cov}(\xi, \eta)}{\sqrt{D\xi D\eta}} — \text{коэффициент корреляции}$$

Свойства:

- ▶ $|\text{corr}(\xi, \eta)| \leq 1$;
- ▶ $|\text{corr}(\xi, \eta)| = 1 \Leftrightarrow \xi$ и η линейно зависимы п.н.;
- ▶ ξ и η независимы $\rightarrow \text{corr}(\xi, \eta) = 0$. Обратное не верно;
- ▶ Является мерой линейной зависимости.



Коэффициент корреляции Пирсона

Метод подстановки: подставим в $\text{corr}(X_1, Y_1)$ эмпир. распр.

$$\hat{\rho} = \frac{\text{cov}_{P^*}(X_1, Y_1)}{\sqrt{D_{P^*}X_1 D_{P^*}Y_1}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

H_0 : выборки независимы

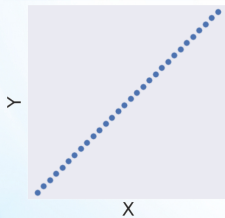
Если H_0 верна и выборки нормальные, то

$$T(X, Y) = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim T_{n-2}.$$

Критерий $\{|T(X, Y)| > t_{n-2, 1-\alpha/2}\}$.

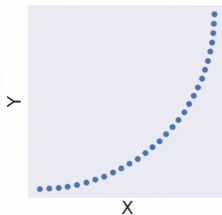


Коэффициент корреляции Пирсона



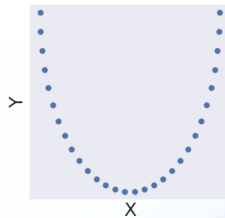
$$\hat{\rho} = 1$$

$$pvalue = 0$$



$$\hat{\rho} = 0.91$$

$$pvalue = 2 \cdot 10^{-12}$$



$$\hat{\rho} = 0$$

$$pvalue = 1$$



Коэффициент корреляции Пирсона

Свойства:

- ▶ $|\hat{\rho}| \leq 1$;
- ▶ $|\hat{\rho}| = 1 \Leftrightarrow$ точки лежат на одной прямой;
- ▶ Работает только для нормальных выборок для линейной зависимости;
- ▶ Не устойчив к выбросам.



Коэффициент корреляции Спирмена

Пусть R_i — ранг наблюдения X_i в выборке X , то есть $X_{(R_i)} = X_i$.

Пусть S_i — ранг наблюдения Y_i в выборке Y , то есть $Y_{(S_i)} = Y_i$.

Пример X_i : 7.3, 2.2, 0.3, 6.2, 1.6, 6.2, 9.6

R_i : 6, 3, 1, 4.5, 2, 4.5, 7

К.к. Спирмена = к.к. Пирсона по выборкам (R_1, \dots, R_n) и (S_1, \dots, S_n) .

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2$$



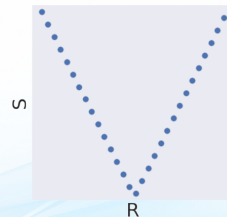
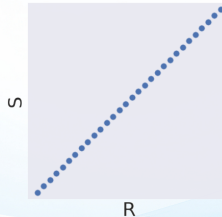
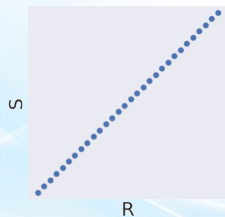
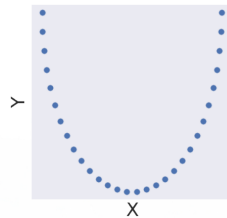
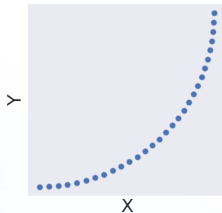
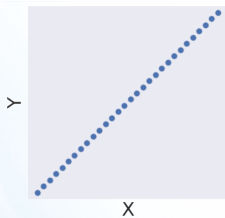
Коэффициент корреляции Спирмена

Свойства:

- ▶ $|\rho_S| \leq 1$;
- ▶ $|\rho_S| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- ▶ Если H_0 верна, то $E\rho_S = 0, D\rho_S = \frac{1}{n-1}$;
- ▶ Если H_0 верна, то $\rho_S / \sqrt{D\rho_S} \xrightarrow{d} \mathcal{N}(0, 1)$.
Критерий $\{|\rho_S / \sqrt{D\rho_S}| > z_{1-\alpha/2}\}$;
- ▶ Устойчив к выбросам.



Коэффициент корреляции Спирмена



$\rho_S = 1$
 $pvalue = 0$

$\rho_S = 1$
 $pvalue = 0$

$\rho_S = 0$
 $pvalue = 1$



Коэффициент корреляции Кендалла

Пары (X_i, Y_i) и (X_j, Y_j) согласованы, если $\text{sign}(X_i - X_j) \text{sign}(Y_i - Y_j) = 1$.

Пусть S — число согласованных пар,
 R — число несогласованных.

$$\tau = \frac{S - R}{S + R} = 1 - \frac{4}{n(n-1)} R$$



Коэффициент корреляции Кендалла

Свойства:

- ▶ $|\tau| \leq 1$;
- ▶ $|\tau| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- ▶ Если H_0 верна, то $E\tau = 0$, $D\tau = \frac{2(2n+5)}{9n(n-1)}$;
- ▶ Если H_0 верна, то $\tau/\sqrt{D\tau} \xrightarrow{d_0} \mathcal{N}(0, 1)$.
Критерий $\{|\tau/\sqrt{D\tau}| > z_{1-\alpha/2}\}$;
- ▶ Если H_0 верна, то $\text{corr}(\rho_S, \tau) = \frac{2n+2}{\sqrt{4n^2+10n}}$;
- ▶ Менее чувствителен к большим различиям между рангами, чем ρ_S ;
- ▶ Точнее оценивается по выборкам малых размеров.

Еще раз формулы

Пирсон:

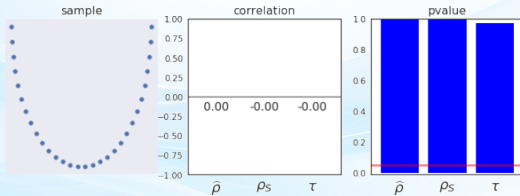
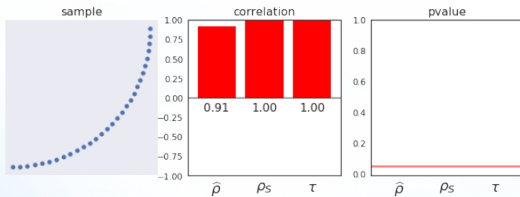
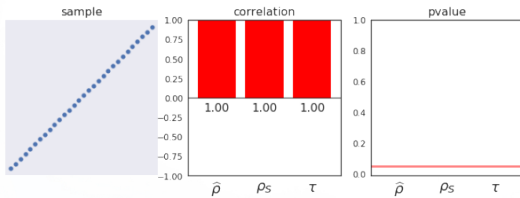
$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

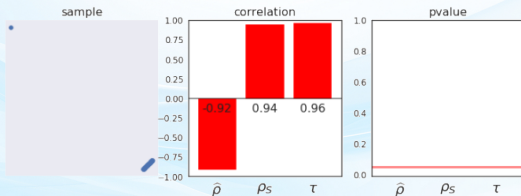
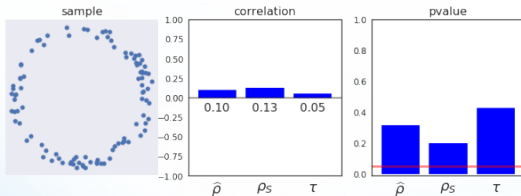
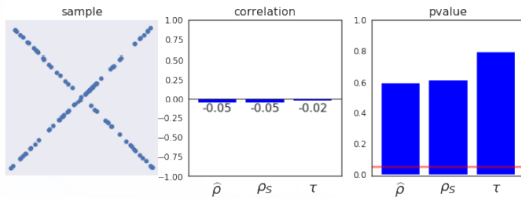
Спирмен: R и S — ранги наблюдений в выборках X и Y

$$\rho_S = \frac{\sum_{i=1}^n (R_i - \bar{R}) (S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2$$

Кендалл: S — число соглас. пар, а R — число несоглас.

$$\tau = \frac{S - R}{S + R} = 1 - \frac{4}{n(n-1)} R$$







Реализация в scipy.stats

`pearsonr` — коэффициент корреляции Пирсона

`spearmanr` — коэффициент корреляции Спирмена

`kendalltau` — коэффициент корреляции Кендалла

Общий принцип: $f(x_1, x_2) = (\text{correlation}, \text{pvalue})$



2. Таблицы сопряженности 2×2



Осенний семестр (2018)

Результаты решения задачи 7 из ДЗ-12:

Семинарист	Никита	Оля	Лена	Настя
Справились	0	5	3	2
Не справились	8	2	4	5

Факты:

1. Случайное разбиение на группы;
2. Задача на алгоритмы и методы оптимизации
 \implies не должна зависеть от семинариста по статистике;
3. На первом семинаре задача была разобрана.



Хотим воспользоваться методом проверки статистических гипотез.

Какие взять H_0 и H_1 ?

Презумпция невиновности: не виновны пока нет доказательств.

H_0 : решаемость задачи не зависит от семинара

H_1 : решаемость задачи зависит от семинара

Упростим данные

Разбиралась ли задача до семинара?	Нет	Да
Справились	0	10
Не справились	8	11



Математическая формулировка

Даны **парные** выборки

$$X = (X_1, \dots, X_n) \sim \text{Bern}(p_1)$$

$$Y = (Y_1, \dots, Y_n) \sim \text{Bern}(p_2)$$

H_0 : выборки X и Y независимы

H_1 : выборки X и Y зависимы

	$Y_i = 0$	$Y_i = 1$	Σ
$X_i = 0$	a	b	$a + b$
$X_i = 1$	c	d	$c + d$
Σ	$a + c$	$b + d$	n

Вероятность таблицы с фиксированными суммами задается гипергеометрическим распределением:

$$P(\text{table}) = \frac{C_{a+b}^a C_{c+d}^c}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$



Точный тест Фишера

p-value = сумма вероятностей по всем возможным вариантам таблицы с такими же суммами по строкам и столбцам, имеющим вероятность не больше, чем у полученной таблицы.

```
scipy.stats.fisher_exact(table)
```

Особенности:

1. Критерий является точным (неасимптотическим);
2. Вычислительно затратный \Rightarrow используется для малых выборок;
3. Что в сложных случаях? \Rightarrow Увидим далее!



Пример про задачу 7 из ДЗ-12

Разбиралась ли задача до семинара?	Нет	Да
Справились	0	10
Не справились	8	11

`scipy.stats.fisher_exact([[0, 8], [10, 11]])`

вернет $p\text{-value} = 0.0265$.

Вывод: гипотеза о независимости отвергается.



Численные характеристики взаимосвязи

$$Q = \frac{ad - bc}{ad + bc} \text{ — коэффициент ассоциации}$$

$$V = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \text{ — коэффициент контингенции}$$

В обоих случаях:

$0 \Rightarrow$ полное отсутствие взаимосвязи

$\pm 1 \Rightarrow$ полная связь



Определение числа наблюдений (при $a + b = c + d$)

Задаем:

α — ур. значимости

β — мощность

$\left. \begin{array}{l} p_1 = a/b \\ p_2 = c/d \end{array} \right\} \text{значимый эффект}$

	$Y_i = 0$	$Y_i = 1$	Σ
$X_i = 0$	a	b	$a + b$
$X_i = 1$	c	d	$c + d$
Σ	$a + c$	$b + d$	n

Тогда необходимое число наблюдений в каждой строке равно

$$K / (\arcsin \sqrt{p_1} - \arcsin \sqrt{p_2})^2$$

K	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
$\alpha = 0.05$	12885	17250	30161
$\alpha = 0.01$	16474	21369	35537
$\alpha = 0.001$	19172	24426	43945



3. Таблицы сопряженности (общий случай)



Категориальные признаки

Даны **парные** выборки

$X = (X_1, \dots, X_n)$, причем $X_i \in \{1, \dots, k_1\}$

$Y = (Y_1, \dots, Y_n)$, причем $Y_i \in \{1, \dots, k_2\}$

Таблица сопряженности:

	1	...	j	...	k_2	Σ
1	n_{11}	...	n_{1j}	...	n_{1k_2}	$n_{1\bullet}$
...
i	n_{i1}	...	n_{ij}	...	n_{ik_2}	$n_{i\bullet}$
...
k_1	$n_{k_1 1}$...	$n_{k_1 j}$...	$n_{k_1 k_2}$	$n_{k_1 \bullet}$
Σ	$n_{\bullet 1}$...	$n_{\bullet j}$...	$n_{\bullet k_2}$	n

Элементы таблицы:

$$n_{ij} = \#\{s \mid X_s = i, Y_s = j\}$$

$$n_{i\bullet} = \#\{s \mid X_s = i\}$$

$$n_{\bullet j} = \#\{s \mid Y_s = j\}$$



Вероятностные модели

Случай 1: X и Y случайны.

$\pi_{ij} = P(X_1 = i, Y_1 = j) \implies \{\pi_{ij}\}_{ij}$ — совместное распределение;

$\pi_{i\bullet} = P(X_1 = i) \implies P = \{\pi_{i\bullet}\}_i$ — распределение X ;

$\pi_{\bullet j} = P(Y_1 = j) \implies Q = \{\pi_{\bullet j}\}_j$ — распределение Y ;

Определение: X и Y **независимы**, если $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall i, j$.

Случай 2: X неслучаен, Y случаен.

\implies суммы по строкам $n_{i\bullet}$ фиксированы.

$\pi_{j|i} = P_i(Y_1 = j)$ — вероятность события $Y_1 = j$ если $X_1 = i$;

$P_i = \{\pi_{j|i}\}_j$ — распределение Y если $X_1 = i$, т.е. X — параметр.

Определение: X и Y **независимы**, если $P_1 = \dots = P_{k_1}$.



Критерий хи-квадрат (обе вер. модели)

H_0 : выборки X и Y независимы

$$\chi^2(X, Y) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}$$

Если H_0 верна, то $\chi^2(X, Y) \xrightarrow{d} \chi^2_{(k_1-1)(k_2-1)}$
 \Rightarrow критерий $\left\{ \chi^2(X, Y) > \chi^2_{(k_1-1)(k_2-1), 1-\alpha} \right\}$.

Условия применимости:

1. $n \geq 40$;

2. $\frac{n_{i\bullet} n_{\bullet j}}{n} < 5$

не более чем в 5% ячеек.

Коэффициент корреляции Крамера

$$\varphi_C(X, Y) = \sqrt{\frac{\chi^2(X, Y)}{n(\min(k_1, k_2) - 1)}}$$

$0 \Rightarrow$ полное отсутствие взаимосвязи;

$1 \Rightarrow$ совпадение переменных.



Пример (влияние нового препарата на выздоровление)

Испытуемые делятся случайно на две группы:

1. *Исследуемая группа* — принимает новый препарат;
2. *Контрольная группа* — принимает плацебо.

	Выздоровели	Нет
Препарат	850	870
Плацебо	380	410

H_0 : препарат не отличим от плацебо (т.е. связи нет);

H_1 : эффект препарата отличается от эффекта плацебо.

Критерий хи-квадрат: $\chi^2(X, Y) = 0.325$, $pvalue = 0.569$,

Численные характеристики: $\varphi_C(X, Y) = 0.008$, $Q = 0.026$, $V = 0.012$



4. Влияние признаков на таргет.

Важности признаков.



Постановка задачи

Наблюдается зашумленная зависимость

$$Y_i = f(X_{i1}, \dots, X_{id}, \varepsilon_i),$$

где X_{ij} — значение j -признака;

Y_i — целевая переменная;

ε_i — шум.

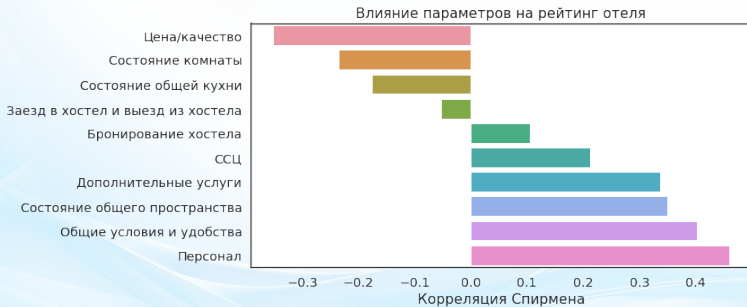
Задача:

Оценить степень влияния каждого признака на целевую переменную.



Корреляционный метод

Посчитать $\widehat{corr}(Y, X_j)$ для всех j .





1. Mean Decrease in Impurity (MDI)

Случай дерева.

Пусть m — узел дерева и X_m — подвыборка, дошедшая до m .

H — выбранный критерий информативности.

При разбиении вершины m решается задача:

$$Q(X_m, j, t) = \frac{|X_l|}{|X_m|} \cdot H(X_l) + \frac{|X_r|}{|X_m|} \cdot H(X_r) \rightarrow \min_{j, t}$$

Уменьшение ошибки относительно вершины m составляет

$$H(X_m) - \frac{|X_l|}{|X_m|} \cdot H(X_l) - \frac{|X_r|}{|X_m|} \cdot H(X_r)$$

Общее уменьшение ошибки на этапе разбиения вершины m по признаку j и порогу t по отношению ко всей выборке:

$$\Delta I_j^m = \frac{|X_m|}{|X|} H(X_m) - \frac{|X_l|}{|X|} \cdot H(X_l) - \frac{|X_r|}{|X|} \cdot H(X_r)$$



1. Mean Decrease in Impurity (MDI)

⇒ При построении дерева можем посчитать, какой вклад каждый признак вносит в уменьшение ошибки:

$$\Delta I_j = \sum_m \Delta I_j^m \cdot I \left\{ \begin{array}{l} \text{разбиение в вершине } m \\ \text{происходит по признаку } j \end{array} \right\}$$

Отнормируем данные значения:

$$\widetilde{\Delta I_j} = \frac{\Delta I_j}{\sum_{j=1}^d \Delta I_j}$$

Случай леса.

Пусть \mathcal{T} — набор деревьев в лесу.

$\Delta I_j(T)$ — важность признака j для дерева T .

$$\Delta I_j = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \Delta I_j(T)$$



1. Mean Decrease in Impurity (MDI)

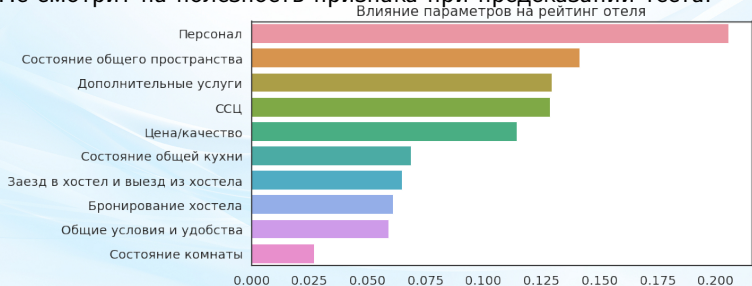
Плюсы:

- ▶ Поле `feature_importances_` в `sklearn`
— важности признаков, посчитанные этим методом.
- ▶ Быстро считается, обучение происходит один раз.

Минусы:

- ▶ Важность признаков смещена в сторону признаков с большим количеством значений.
- ▶ Считается при использовании лишь обучающей выборки.

Не смотрит на полезность признака при предсказании теста.





2. Permutation feature importance

1. Обучим модель и измерим метрику на валидации/OOB/прочее .
2. Для одного выбранного признака перемешаем все его значения в датасете, на котором до этого измерили метрику.
3. Измерим метрику на видоизмененном датасете.
4. Посчитаем важность данного признака:
Разница между исходным и новым значением метрики.
5. Сделаем пункты 2-4 для всех признаков.



2. Permutation feature importance

Плюсы:

- ▶ Подходит для любых моделей.
- ▶ Требуется только обучение модели.
- ▶ Использует тестовое множество и является более надежным, чем MDI.

Минусы:

- ▶ Более вычислительно затратно, чем MDI.
 - ▶ Переоценивает важность для скоррелированных признаков.
- (Strobl et al (2008))



3. Drop Column feature importance

Сравним 2 модели:

- ▶ Модель, обученная на датасете со всеми признаками
- ▶ Модель, обученная на данных без одного признака.

Важность этого признака — разница метрик на тесте/валидации/OOB для этих моделей.

Плюсы:

- ▶ Самая точная важность признаков.
- ▶ Подходит для любых моделей.

Минусы:

- ▶ Вычислительно сложно.
Требует обучения многих разных моделей.



ВСЁ!