Статистика

DS-поток

Лекция 11

6.7 Анализ зависимостей

Анализ зависимостей

Даны **парные** выборки:

$X = (X_1, ..., X_n)$

$Y = (Y_1, ..., Y_n)$

Задачи:

- Зависимы ли выборки?

  $H_0:$ выборки независимы vs. $H_1:$ выборки зависимы

- Количественная оценка степени

  неслучайности их совместного изменения.

6.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Важность признаков

Коэффициент корреляции

Пусть ξ, η — случайные величины.

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

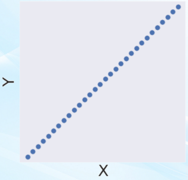
Свойства:

- $|corr(\xi, \eta)| \leqslant 1$;
- $|corr(\xi, \eta)| = 1 \Leftrightarrow \xi$ и η линейно зависимы п.н.;
- ξ и η независимы $\to corr(\xi, \eta) = 0$. Обратное не верно;
- Является мерой линейной зависимости.

Коэффициент корреляции Пирсона

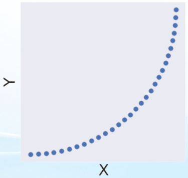
Метод подстановки: подставим в $corr(X_1, Y_1)$ эмпир. распр.

$$\widehat{\rho} = \frac{cov_{\mathrm{P}^*}(X_1, Y_1)}{\sqrt{\mathrm{D}_{\mathrm{P}^*} X_1 \mathrm{D}_{\mathrm{P}^*} Y_1}} = \frac{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum\limits_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}}$$



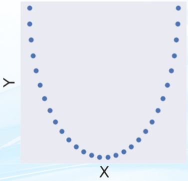
$\widehat{\rho} = 1$

$pvalue = 0$



$\widehat{\rho} = 0.91$

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



$\widehat{\rho} = 0$

$pvalue = 1$

Коэффициент корреляции Пирсона

Свойства:

- $|\widehat{\rho}| \leqslant 1$;
- $|\widehat{\rho}| = 1 \Leftrightarrow$ точки лежат на одной прямой;
- Работает только для нормальных выборок

  для линейной зависимости;
- Не устойчив к выбросам.
- $\mathrm{H}_0:$ выборки независимы

  Если H_0 верна и выборки нормальные, то

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

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

Коэффициент корреляции Спирмена

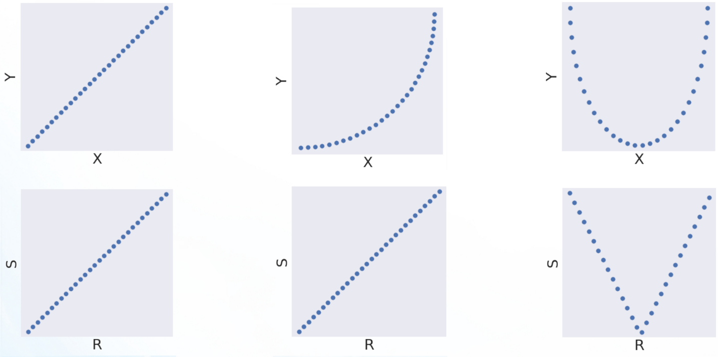
Пусть R_i — ранг наблюдения X_i в выборке X, то есть $X_{(R_i)} = X_i$.

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

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 $(R_1, \ldots, R_n)$ и $(S_1, \ldots, S_n)$.

$$\rho_S = \frac{\sum\limits_{i=1}^{n} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(R_i - \overline{R}\right)^2 \sum\limits_{i=1}^{n} \left(S_i - \overline{S}\right)^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - S_i)^2$$



$\rho_S = 1$
$pvalue = 0$

$\rho_S = 1$
$pvalue = 0$

$\rho_S = 0$
$pvalue = 1$

Свойства:

- $|\rho_S| \leqslant 1$, причем $|\rho_S| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- Если H_0 верна, то $\mathsf{E}\rho_S = 0, \mathsf{D}\rho_S = \frac{1}{n-1}$;
- Если H_0 верна, то $\rho_S / \sqrt{\mathsf{D}\rho_S} \xrightarrow{d_0} \mathcal{N}(0, 1)$.

  Критерий $\{|\rho_S / \sqrt{\mathsf{D}\rho_S}| > z_{1-\alpha/2}\}$;
- Устойчив к выбросам.

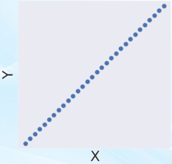
Коэффициент корреляции Кендалла

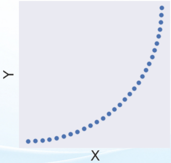
Пары (X_i, Y_i) и (X_j, Y_j) согласованы, если

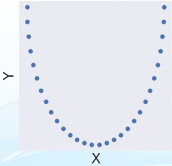
$\operatorname{sign}(X_i - X_j)\operatorname{sign}(Y_i - Y_j) = 1.$

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

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







$S = \frac{n(n-1)}{2}, R = 0$
$\tau = 1$
$pvalue = 0$

$S = \frac{n(n-1)}{2}, R = 0$
$\tau = 1$
$pvalue = 0$

$S = \frac{n(n-1)}{4}, R = \frac{n(n-1)}{4}$
$\tau = 0$
$pvalue = 1$

Коэффициент корреляции Кендалла

Свойства:

- $|\tau| \leqslant 1$;
- $|\tau| = 1 \Leftrightarrow$ точки лежат на монотонной кривой;
- Если H_0 верна, то $\mathrm{E}\tau = 0, \mathrm{D}\tau = \frac{2(2n+5)}{9n(n-1)}$;
- Если H_0 верна, то $\tau/\sqrt{\mathrm{D}\tau} \xrightarrow{d_0} \mathcal{N}(0,1)$.

  Критерий $\{|\tau/\sqrt{\mathrm{D}\tau}| > z_{1-\alpha/2}\}$;
- Если H_0 верна, то $corr(\rho_S, \tau) = \frac{2n+2}{\sqrt{4n^2+10n}}$;
- Менее чувствителен к большим различиям

  между рангами, чем ρ_S;
- Точнее оценивается по выборкам малых размеров.

Еще раз формулы

Пирсон:

$$\widehat{\rho} = \frac{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum\limits_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}}$$

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

$$\rho_S = \frac{\sum\limits_{i=1}^{n} \left(R_i - \overline{R}\right)\left(S_i - \overline{S}\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(R_i - \overline{R}\right)^2 \sum\limits_{i=1}^{n} \left(S_i - \overline{S}\right)^2}} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i - S_i)^2$$

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

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

6.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Важность признаков

Осенний семестр (2018)

Результаты решения теор. задачи:

Семинар	I	II	III	IV
Справились	0	5	3	2
Не справились	8	2	4	5

Факты:

1. Случайное разбиение на группы;
2. Задача на алгоритмы и методы оптимизации

   $\Longrightarrow$ не должна зависеть от семинариста по статистике;
3. Дедлайн перед семинаром;
4. 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

$X = (X_1, ..., X_n) \sim Bern(p_1)$

$Y = (Y_1, ..., Y_n) \sim Bern(p_2)$

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

H_0: выборки X и Y независимы

H_1: выборки X и Y зависимы

Вероятность таблицы с фиксированными суммами задается гипергеометрическим распределением:

$$\mathrm{P}(table) = \frac{C^a_{a+b} C^c_{c+d}}{C^{a+c}_n} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}.$$

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.

Точный тест Фишера

Особенности:

1. Критерий является точным (неасимптотическим);
2. Вычислительно затратный $\Rightarrow$ используется для малых выборок;
3. Что в сложных случаях? 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-value $= 0.0265$.

Вывод: гипотеза о независимости отвергается.

Численные характеристики взаимосвязи

Даны **парные** выборки

$X = (X_1, ..., X_n) \sim Bern(p_1)$

$Y = (Y_1, ..., Y_n) \sim Bern(p_2)$

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

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

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

В обоих случаях:

$0 \Longrightarrow$ полное отсутствие взаимосвязи

$\pm 1 \Longrightarrow$ полная связь

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

Задаем:

α — ур. значимости

β — мощность

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

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

Тогда необходимое число наблюдений в каждой строке равно:

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

Где K определяется из таблицы:

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

6.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Важность признаков

Категориальные признаки

Даны **парные** выборки

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

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

Таблица сопряженности:

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

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

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

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

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

Случай 2: X неслучаен, Y случаен.

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

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

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

Определение: X и Y **независимы**, если $\mathsf{P}_1 = \ldots = \mathsf{P}_{k_1}$.

Критерий хи-квадрат (обе вер. модели)

$\mathrm{H}_0:$ 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 $\chi^2(X,Y) \xrightarrow{d} \chi^2_{(k_1-1)(k_2-1)}$
$\Longrightarrow$ критерий $\left\{\chi^2(X,Y) > \chi^2_{(k_1-1)(k_2-1),1-\alpha}\right\}$.

Условия применимости:

1. $n \geqslant 40$;
2. $\frac{n_{i\bullet}n_{\bullet j}}{n} < 5$

   не более чем в 20% ячеек.

Коэффициент корреляции Крамера

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

$0 \Longrightarrow$ полное отсутствие взаимосвязи;

$1 \Longrightarrow$ совпадение переменных.

Пример

	Вернул кредит	Не вернул кредит
Android	850	870
iOS	380	410

H_0: зависимости возвращаемости кредита от типа ОС нет;

H_1: зависимость есть.

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

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

6.7 Анализ зависимостей

Коэффициенты корреляции

Таблицы сопряженности 2×2

Таблицы сопряженности, общий случай

Важность признаков

Постановка задачи

Задача:

Оценить степень влияния каждого признака на предсказание.

Иначе говоря, определить числа $\mathrm{I}_j \geqslant 0$, $\mathrm{I}_1 + \ldots + \mathrm{I}_d = 1$, такие что если $\mathrm{I}_{j_1} > \mathrm{I}_{j_2}$, то признак j_1 важнее признака j_2 для данной модели.

Замечание:

Данный функционал у моделей на основе решающих деревьев часто интерпретируют как степень влияния признаков на сам таргет.

Это не верно. Получаемые величины позволяют оценить, насколько полезен оказался каждый из признаков **для данной конкретной модели**.

Mean Decrease in Impurity (MDI)

Пусть X_m — подвыборка, дошедшая до узла m,
X_ℓ, X_r — подвыборки, идущие в левое и правое поддеревья,
H — выбранный критерий информативности.

При разбиении вершины m решается задача:

$$Q(X_m, j, t) = \frac{|X_\ell|}{|X_m|} H(X_\ell) + \frac{|X_r|}{|X_m|} H(X_r) \longrightarrow \min_{\ell, r}.$$

Уменьшение ошибки *относительно* вершины m составляет

$$H(X_m) - \frac{|X_\ell|}{|X_m|} H(X_\ell) - \frac{|X_r|}{|X_m|} H(X_r).$$

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

$$\Delta \mathrm{I}_j^m = \frac{|X_m|}{|X|} H(X_m) - \frac{|X_\ell|}{|X|} H(X_\ell) - \frac{|X_r|}{|X|} H(X_r).$$

Пример

Пусть X_m — подвыборка, дошедшая до узла m,
X_ℓ, X_r — подвыборки, идущие в левое и правое поддеревья,
$H = MSE$ — выбранный критерий информативности.

При разбиении вершины m решается задача:

$$\frac{1}{|X_m|}\left\|Y_\ell - \widehat{Y}_\ell\right\|^2 + \frac{1}{|X_m|}\left\|Y_r - \widehat{Y}_r\right\|^2 \longrightarrow \min_{\ell,r}.$$

Уменьшение ошибки *относительно* вершины m составляет

$$\frac{1}{|X_m|}\left\|Y_m - \widehat{Y}_m\right\|^2 - \frac{1}{|X_m|}\left\|Y_\ell - \widehat{Y}_\ell\right\|^2 - \frac{1}{|X_m|}\left\|Y_r - \widehat{Y}_r\right\|^2.$$

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

$$\Delta \mathrm{I}_j^m = \frac{1}{|X|}\left\|Y_m - \widehat{Y}_m\right\|^2 - \frac{1}{|X|}\left\|Y_\ell - \widehat{Y}_\ell\right\|^2 - \frac{1}{|X|}\left\|Y_r - \widehat{Y}_r\right\|^2.$$

Mean Decrease in Impurity (MDI)

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

$$\Delta \mathrm{I}_j = \sum_m \Delta \mathrm{I}_j^m \cdot I \left\{ \begin{matrix} \text{разбиение в вершине } m \\ \text{происходит по признаку } j \end{matrix} \right\}.$$

Отнормируем данные значения, получаем оценку важности:

$$\mathrm{I}_j = \frac{\Delta \mathrm{I}_j}{\sum\limits_{j=1}^{d} \Delta \mathrm{I}_j}$$

Случай леса.

Пусть $\mathscr{T}$ — набор деревьев в лесу.

$\mathrm{I}_j(T)$ — важность признака j для дерева T.

$$\mathrm{I}_j = \frac{1}{|\mathscr{T}|} \sum_{T \in \mathscr{T}} \mathrm{I}_j(T)$$

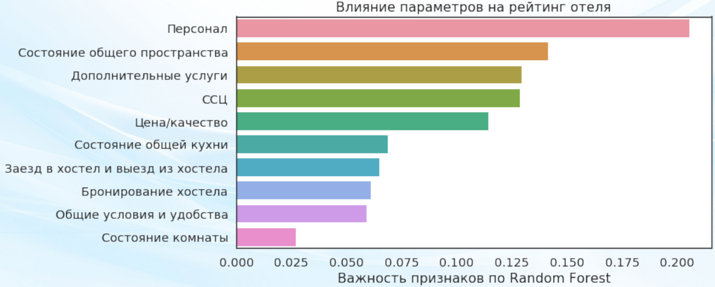
Mean Decrease in Impurity (MDI)

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.




ВСЁ!