Математические и инструментальные методы машинного обучения

5. Анализ связей

Понятие статистической связи

Статистическая связь — это объективная количественная закономерность изменения массовых явлений и процессов, то есть статистическая закономерность является количественной формой проявления причинной связи. Она устанавливается на основе анализа массовых данных и проявляется только на уровне статистической совокупности.

Закономерность возникает как результат воздействия большого числа постоянно действующих причин и причин случайных, действующих временами. Постоянно действующие причины придают изменениям в явлениях регулярность и повторяемость, случайные - вызывают отклонения в этой регулярности.

Виды анализа статистических связей

Анализ связи между номинальными признаками

- Таблицы сопряженности
- Ранговый корреляционный анализ

Анализ связи между количественными признаками

- Корреляционный анализ
- Регрессионный анализ

Анализ связи между номинальными и количественными признаками

- Номинальный регрессионный анализ
- Дисперсионный анализ (ANOVA)

Задачи корреляционно-регрессионного анализа

Корреляционный анализ

- Есть ли связь между явлениями?
- Насколько сильная связь между явлениями?

Регрессионный анализ

- Какой характер носит связь между явлениями?
- Каковы модель связи между явлениями и её свойства?

Условия применения и задачи корреляционного анализа

К условиям использования корреляционных зависимостей относятся:

- массовость исследуемых явлений и процессов и (или) их проявлений;
- достаточная качественная однородность используемых статистических совокупностей;
- наличие реальных связей между изучаемыми признаками.
- задачами корреляционного анализа считаются:
- выявление неизвестных причинно-следственных связей между исследуемыми признаками;
- количественное измерение тесноты связи между наблюдаемыми признаками;
- установление важнейших факторов, влияющих на результат;
- оценка уравнения регрессии.

Линейный коэффициент корреляции

Линейный коэффициент корреляции (коэффициент Пирсона) показывает тесноту линейной связи между парой переменных – случайными величинами x и y:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}$$

$$\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2}$$

$$r = \frac{\sigma_x}{\sigma_y} a_1$$

Свойства линейного коэффициента корреляции

Коэффициент корреляции представляет собой стандартизованное значение коэффициента регрессии.

Коэффициент корреляции обладает свойством симметрии – от перестановки переменных значение не меняется.

Умножение или деление всех значений переменных на одно и тоже число не приводит к изменению значения коэффициента.

Значения коэффициента лежат в пределах от -1 до 1. Отрицательные значения соответствуют обратной связи между переменными, положительные – прямой.

Значение коэффициента по модулю равное единице означает функциональную линейную связь между переменными.

Значения коэффициента близкие к нулю могут означать слабость линейной зависимости, но не отрицают наличие другой функциональной зависимости, например, полиномиальной.

Классификация тесноты корреляционной связи

Коэффициент *r*	Сила связи
0.1 – 0.3	Слабая
0.3 – 0.5	Умеренная
0.5 – 0.7	Заметная
0.7 – 0.9	Высокая
0.9 – 0.999	Очень высокая

Доверительный интервал для оценки коэффициента корреляции

$$r - tm_r \le r \le r + tm_r$$

$$m_r = \sqrt{(1 - r^2)/(n-2)}$$

$t(\propto, n-2)$ — критическая точка распределения Стьюдента

Проверка значимости коэффициента корреляции

$$t_{факт} = \frac{r}{m_r}$$

Коэффициент значим, если:

$t_{факт} \geq t(\propto, n-2)$ — **критическая точка распределения Стьюдента**

Коэффициент Фехнера для нелинейной функциональной связи

Для расчёта коэффициента Фехнера вычисляют средние выборочные значения $\bar{x}$ и $\bar{y}$ затем определяют знаки совпадений $x-\bar{x}$ и $y-\bar{y}$ при этом возможны сочетания:

(+,+), (-,-), (+,-), (-,+), (0,+), (0, -), (-, 0), (+, 0), (0,0).

Далее вводятся обозначения:

- V – количество совпадений знаков;
- W – количество несовпадений знаков.

Индекс корреляции

$$R_{yx} = \sqrt{\frac{\delta_y^2}{\sigma_y^2}}$$

$$\delta_y^2 = \frac{\Sigma_{i=1}^{m}(y_i - \bar{y})^2 n_i}{n}, y_i = \mathrm{f}(x_i)$$

Корреляционное отношение

$$\eta_{yx} = \sqrt{\frac{\delta_y^2}{\sigma_y^2}}$$

$$\delta_y^2 = \frac{\Sigma_{i=1}^{m} (\overline{y_i} - \overline{y})^2 n_i}{n}$$

Свойства корреляционного отношения и индекса

1. $0 \le R \le 1, 0 \le \eta \le 1;$
2. $R = 1, \eta = 1$ — корреляционная связь есть
3. $R = 0, \eta = 0$ — корреляционная связь отсутствует
4. $R_{xy} \ne R_{yx}, \eta_{xy} \ne \eta_{yx}$
5. $0 \le |r| \le R \le \eta \le 1$

Понятия ранжирования и ранговой корреляции

В случае, если исследуемые признаки измерены в номинальной или порядковой шкалах (доступны только отношения «больше-меньше» или «совпадение-несовпадение»), то для вычисления корреляции между ними используется приём ранжирования. Для этого значения переменной располагают в порядке возрастания или убывания и каждому значению присваивают ранг — номер в данной упорядоченной совокупности.

Если среди элементов ранжированного ряда есть совпадающие, то такие ранги называются связанными, и каждому из группы совпадающих элементов присваивают ранг, равный среднему арифметическому рангов этой группы.Корреляцию между ранжированными признаками называют ранговой.

Коэффициент ранговой корреляции Спирмена

Для несвязных рангов

$$\rho = 1 - \frac{\Sigma_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}$$

Для связных рангов

$$\rho = 1 - \frac{\frac{n^3 - n}{6} - \Sigma_{i=1}^{n}(r_i - s_i)^2 - T_x - T_y}{\sqrt{\left[\frac{n^3 - n}{6} - 2T_x\right]\left[\frac{n^3 - n}{6} - 2T_y\right]}}$$

$$T_x = \frac{1}{12}\Sigma_{i=1}^{k_r}(t_i^3 - t_i), \quad T_y = \frac{1}{12}\Sigma_{i=1}^{k_s}(t_i^3 - t_i)$$

Проверка значимости коэффициента Спирмена

$$t_{факт} = \rho \sqrt{\frac{n-2}{1-\rho^2}}$$

Коэффициент значим, если:

$$\left| t_{факт} \right| \geq t_{кр}(\alpha, n-2)$$

Свойства коэффициента Спирмена

1. $|\rho| \leq 1$ — нет полной связи
2. $\rho = 1$ — полная прямая корреляционная связь
3. $\rho = -1$ — полная обратная корреляционная связь
4. $\rho_{xy} \neq \rho_{yx}$

Коэффициент ранговой корреляции Кендалла (несвязные ранги)

$$\tau = 1 - \frac{4K}{n^2 - n}$$

1. Ранжируют объекты по признаку X и располагают объекты согласно рангам;
2. ранжируют объекты по Y и получают ранги r_i;
3. для каждого ранга находят число инверсий, т.е. число значений рангов признака Y, стоящих справа от r_i и меньше его;
4. находится сумма всех инверсий K.

Коэффициент ранговой корреляции Кендалла. Пример поиска числа инверсий

	Обьект3	Обьект1	Обьект2	Обьект5	Обьект4
X	1	2	3	4	5
Y	4	3	2	5	1
Число инверсий	3	2	1	1	0

Для ранга 4 (Y) количество рангов, меньших 4 и находящихся правее в таблице, равно 3.

Коэффициент ранговой корреляции Кендалла (связные ранги)

$$\tau = \frac{n^2 - n - 4K}{\sqrt{\left[n^2 - n - 2V_x\right]\left[n^2 - n - 2V_y\right]}}$$

$$V_x = \frac{1}{2}\sum_{i=1}^{k_r}(t_i^2 - t_i), \quad V_y = \frac{1}{2}\sum_{i=1}^{k_s}(t_i^2 - t_i)$$

Коэффициент конкордации (согласованности) Кендалла

$$w = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} r_{ij} - \frac{m(n+1)}{2} \right)^2$$

Свойства коэффициента конкордации

1. $0 \leq w \leq 1$, нет полной связи
2. $W = 1$, полная прямая корреляционная связь
3. При $m = 2$, коэффициент W пропорционален коэффициенту Спирмена

Альтернативные признаки. Коэффициент контингенции Пирсона

$$K_c = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

	A - да	A - нет	Всего
B – да	a	b	a+b
B - нет	c	d	c+d
Всего	a+c	b+d	a+b+c+d

Альтернативные признаки. Коэффициент ассоциации Юла

$$K_c = \frac{ad - bc}{(ad + bc)}$$

	A - да	A - нет	Всего
B – да	a	b	a+b
B - нет	c	d	c+d
Всего	a+c	b+d	a+b+c+d

Коэффициент взаимной сопряжённости Пирсона

$$K_{\text{п}} = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \qquad \varphi^2 = \sum_{i=1}^{k1} \sum_{j=1}^{k2} \frac{n_{ij}}{n_i n_j} - 1$$

	Xb1	Xb2	…	Xbk2	
Xa1	n11	n12	…	n1k2	n1
Xa2	n21	n22	…	n2k2	n2
…	…	…	…	…	…
Xak1	nk11	nk12	…	nk1k2	nk1
Всего	n1	n2	…	nk2	n

Коэффициент взаимной сопряжённости Чупрова

$$K_{\text{п}} = \sqrt{\frac{\varphi^2}{\sqrt{(k1-1)(k2-1)}}} \qquad \varphi^2 = \sum_{i=1}^{k1} \sum_{j=1}^{k2} \frac{n_{ij}}{n_i n_j} - 1$$

	Xb1	Xb2	…	Xbk2	
Xa1	n11	n12	…	n1k2	n1
Xa2	n21	n22	…	n2k2	n2
…	…	…	…	…	…
Xak1	nk11	nk12	…	nk1k2	nk1
Всего	n1	n2	…	nk2	n

Частный коэффициент корреляции

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии. Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель. Частные коэффициенты корреляции, измеряющие влияние на y фактора xi при неизменном уровне др. факторов можно определить по формуле:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}}$$

Множественный коэффициент корреляции

Множественный коэффициент корреляции характеризует степень линейной статистической связи (зависимости) результативного y **и линейной комбинации факторных** $x1, x2 \ldots xn$ **признаков:**

$$r_{yx_i} = \frac{\sum_{j=1}^{n} (x_{ij} - \frac{\sum_{j=1}^{n} x_{ij}}{n})(y_{ij} - \frac{\sum_{j=1}^{n} y_{ij}}{n})}{\sqrt{\sum_{j=1}^{n} (x_{ij} - \frac{n\sum_{j=1}^{n} x_{ij}}{n})^2 (y_{ij} - \frac{\sum_{j=1}^{n} y_{ij}}{n})^2}}$$

Виды регрессионных моделей

		Уровень измерения *x*		
		Интервальный или абсолютный для всех x	Порядковый для всех x	Для некоторых x интервальный или абсолютный, для некоторых – порядковый, либо номинальный
Уровень измерения y	Интервальный или абсолютный	Классическая регрессионная модель	Классическая регрессионная модель с использованием фиктивных переменных	Классическая регрессионная модель с использованием фиктивных переменных
	Порядковый	Множественная логистическая регрессия	Порядковая регрессия	Множественная логистическая регрессия с использованием фиктивных переменных
	Номинальный с несколькими значениями	Множественная логистическая регрессия	Множественная логистическая регрессия с использованием фиктивных переменных	Множественная логистическая регрессия с использованием фиктивных переменных
	Номинальный с двумя значениями	Бинарная логистическая регрессия	Бинарная логистическая регрессия с использованием фиктивных переменных	Бинарная логистическая регрессия с использованием фиктивных переменных

Проблема мультиколлинеарности

Мультиколлинеарность — тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат, которая затрудняет оценивание регрессионных параметров. Среди последствий выделить следующие:

- увеличение дисперсий оценок параметров
- уменьшение значений t-статистик для параметров, что приводит к неправильному выводу об их статистической значимости
- получение неустойчивых оценок параметров модели и их дисперсий
- возможность получения неверного с точки зрения теории знака у оценки параметра

Дисперсия в регрессионной модели

Общее отклонение есть сумма объяснимой и необъяснимой вариации:

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

Общая вариация	Объяснимая вариация	Необъяснимая вариация
Общая сумма квадратов	**Объяснимая моделью вариация**	**Сумма квадратов остатков**
TSS Total Sum of Squares	**ESS** Explained Sum of Squares	**RSS** Residual Sum of Squares