

6 Дополнительная лекция

6.1 Анализ соответствий

Анализ соответствий (correspondence analysis) — это специальный вид анализа таблиц сопряженностей, чтобы понять, какие градации с какими ‘сопряжены’. В основном, используется в маркетинге, социологии, поэтому будем использовать слово человек (правильнее — респондент) вместо индивид. Таблица сопряженностей строится для двух категориальных признаков и состоит из частот n_{ij} , сколько человек выбрало градацию i по первому признаку и градацию j по второму.

Пример: есть банки. У человека спрашивают, каким банком он пользуется, и каким основным свойством из предложенного списка банк обладает. Получается таблица сопряженностей, где по одной оси откладывается название банка, а по другой — свойство банка (например, надежность, доступность, уровень сервиса). Ставится вопрос — с какими свойствами какой банк ассоциируется, в отрицательную или положительную сторону.

6.1.1 Критерий независимости χ^2

Проверим $H_0 : \xi$ независима с η .

По определению, для двумерных дискретных распределений (признаки могут быть категориальными),

$$\begin{aligned} \xi \text{ независима с } \eta &\iff \underbrace{\mathbb{P}(\xi = i, \eta = j)}_{p_{ij}} = \underbrace{\mathbb{P}(\xi = i)}_{p_i} \underbrace{\mathbb{P}(\eta = j)}_{p_j} = \\ &= \underbrace{\sum_{k=1}^K \mathbb{P}(\xi = i, \eta = k)}_{p_{i\cdot}} \cdot \underbrace{\sum_{s=1}^S \mathbb{P}(\xi = s, \eta = j)}_{p_{\cdot j}}. \end{aligned} \quad (19)$$

Эти равенства задают модель для распределения двух независимых признаков.

Можно показать, что ОМП оценками параметров модели будут $\hat{p}_{i\cdot} = n_{i\cdot}/n$ и $\hat{p}_{\cdot j} = n_{\cdot j}/n$. Если

$$\xi \text{ независима с } \eta \Rightarrow \hat{p}_{ij} = \frac{n_{ij}}{n} \approx \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}.$$

Поэтому в качестве статистики критерия возьмем

$$\begin{aligned} \chi^2 &= \sum_{i=1}^K \sum_{j=1}^S \frac{(n_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^K \sum_{j=1}^S \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} = \\ &= n \sum_{i=1}^K \sum_{j=1}^S \frac{(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}} \xrightarrow{d} \chi^2((K-1)(S-1)) \end{aligned}$$

Количество степеней свободы таково, потому что всего KS состояний и $KS-1$ уравнений (-1 , потому что $\sum_{ij} p_{ij} = 1$); если ξ и η независимы, то число параметров $K+S-2$ (-2 , потому что $\sum_i p_{i\cdot} = 1$ и $\sum_j p_{\cdot j} = 1$). Значит, число степеней свободы равно $(KS-1) - (K+S-2) = (K-1)(S-1)$.

“Идеальное” значение ноль, поэтому критическая область справа.

6.1.2 Мера сопряженности между грациями

Для каждой клетки таблицы сопряженностей величина $\hat{p}_{ij}$ характеризует связь между грациями. Но без стандартизации невозможно понять, много это или мало. Поэтому рассматривают центрирование $\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}$ (как видно, получаем меру отклонения от независимости), а потом стандартизацию $c_{ij} = \sqrt{n} \frac{(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})}{\sqrt{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}}$. При такой стандартизации сумма квадратов отклонений, — это в точности значение статистики критерия независимости χ^2 .

Несколько условно, отрицательные значения соответствует низкой ассоциации между категориями, положительные — высокой. Чтобы понять, значим ли этот вывод, пользуются тем, что в условиях независимости c_{ij} имеет примерно стандартное нормальное распределение. Тем самым, условно, если $|c_{ij}| > 2$, то вывод значим.

6.2 SVD для таблицы сопряженностей

Можно сказать, что анализ соответствий — это взвешенный АГК, примененный к матрице из $\hat{p}_{ij}$. Точнее, это взвешенный SVD, примененный к центрированной матрице $\mathbf{X}$ из $(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})$, с весами по строкам $\hat{p}_{i\cdot}$, а по столбцам — $\hat{p}_{\cdot j}$. Замечу, что чисто технически построение такого SVD сводится к обычному не взвешенному SVD матрицы $\tilde{\mathbf{X}}$ из $\frac{(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})}{\sqrt{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}}$. А уже через него левые и правые сингулярные вектора пересчитываются в левые и правые собственные вектора исходного взвешенного SVD.

Утверждение (без доказательства, хотя оно простое, если записать в матричной форме): Пусть $\tilde{\mathbf{X}} = \sum_i \sqrt{\lambda_i} \tilde{U}_i \tilde{V}_i^T$ — обычное не взвешенное SVD матрицы $\tilde{\mathbf{X}}$. Введем $U_i = \text{diag}(1/\sqrt{p_{1\cdot}}, \dots, 1/\sqrt{p_{K\cdot}}) \tilde{U}_i$ и $V_i = \text{diag}(1/\sqrt{p_{\cdot 1}}, \dots, 1/\sqrt{p_{\cdot S}}) \tilde{V}_i$. Тогда $\mathbf{X} = \sum_i \sqrt{\lambda_i} U_i V_i^T$ — взвешенное SVD матрицы $\mathbf{X}$ с весами, указанными выше.

Отсюда, в частности, следует, что $\sum_i \lambda_i = \|\tilde{\mathbf{X}}\|_{1,2}^2 = \chi^2/n$ (значению статистики критерия независимости χ^2 , делённой на n). Каждое λ_i дает вклад i -й главной компоненты.

Во взвешенном SVD сингулярные вектора будут ортогональные не в обычном смысле, а с весами (покоординатная сумма с весами из соответствующего пространства равна нулю).

Если в АГК левые сингулярные вектора составляли базис в пространстве индивидов, а правые — в пространстве признаков, то тут в матрице данных строки и столбцы равноправны. Поэтому мы получаем левые и правые сингулярные вектора, которые составляют базисы в пространстве категорий первого и второго признаков. При интерпретации, мы можем смотреть на результат как будто категории одного признака — это индивиды, а второго — признаки. А можем наоборот. При изображении разница в том, что мы рисуем точками на биплоте, а что стрелочками.

U_i и V_i называются стандартными (стандартизованными?) координатами, а если мы их умножим на корни из лямбд и получим $\sqrt{\lambda_i} U_i$ и $\sqrt{\lambda_i} V_i$, то главными координатами. Именно главные координаты обычно изображаются на скаттерплоте, чтобы показать, какая категория одного признака близка к какой категории второго признака. Называют это картой. (Это как если бы мы на одной картинке изобразили факторные вектора и главные компоненты, что не очень правильно.) В литературе тоже пишут, что это не очень правильно, но делают из соображений симметрии. Мы с вами знаем, что правильнее рисовать $\sqrt{\lambda_i} V_i$ и U_i или $\sqrt{\lambda_i} U_i$ и V_i , чтобы видеть, как выглядят точки и единичные орты в новых координатах.