Корреляционный анализ данных

Случайная векторная переменная

Случайную величину, как это принято в теории вероятностей, будем обозначать так $X_j = [x_{ji}, \ldots, x_{jn}]$,

x_{ij} – *i*-я реализацией *j*-го признака. Матрица данных «объект-признак» является случайной векторной переменной (СВП)

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix}$$

компонентами которой являются случайные признаки X_j. Теперь каждый объект описывается *i*-й реализацией СВП и определяется как неслучайный *m*-мерный вектор-столбец $\mathbf{x}_i = [x_{1i}, \ldots, x_{mi}]', \quad i = 1, \ldots, n$

Отсюда наблюдения над n-объектами можно представить в виде n-реализаций СВП $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$

СВП является (*m* × *n*)-матрицей, равной транспонированной матрице «объект-признак».

Ковариационная и корреляционные матрицы

Ковариационная матрица $\mathbf{K}_x$ для матрицы $\mathbf{X}$ определяется так

$$\mathbf{K}_x = \mathbf{M}[(\mathbf{X} - \mathbf{MX})'(\mathbf{X} - \mathbf{MX})]$$

где $\mathbf{MX} = (\mathbf{M}X_1, \ldots, \mathbf{M}X_m)$ – вектор средних значений.

Компоненты *ковариационной матрицы*

$$k_{ij} = \mathbf{M}[(X_i - \mathbf{M}X_i)(X_j - \mathbf{M}X_j)] = \operatorname{cov}(X_i, X_j), i, j = 1, \ldots, m$$

при $i = j$ совпадают с дисперсией величины Xi.

Если дисперсии признаков $X_1, \ldots, X_m$ равны 1, то ковариационная матрица называется *корреляционной матрицей.*

Часто используется и оценка $$\hat{k}_{ij} = \frac{1}{n-1}\left[\sum_{s=1}^{n} x_{si} x_{sj} - \frac{1}{n}\bar{x}_i \bar{x}_j\right]$$

Оценки элементов r_{ij} корреляционной матрицы $$\hat{r}_{ij} = \hat{k}_{ij} / S_i S_j,$$

$S_i = (\hat{k}_{ii})^{1/2}$ – среднеквадратические отклонения признаков Xi *и* Xj.

$S_j = (\hat{k}_{jj})^{1/2}$

Свойства оценок при наличии пропусков в наблюдениях

1)вычисление оценки только по данным объектов без пропусков

2)вычисление оценки по всем доступным наблюдениям

3)вычисление оценки по данным, пропуски которых заполнены одним из существующих методов

Матрицы близостей

Метрика объектов Xi, Xj и Xk должна удовлетворять аксиомам:

1) $d_{ij} = 0 \Leftrightarrow X_i = X_j$ **для всех** $i, j = 1, \ldots, n$ **(аксиома тождества)**

2. $d_{ij} \le d_{ik} + d_{jk}$ **для всех** $i, j = 1, \ldots, n$ **(аксиома треугольника)**

3. $d_{ij} = d_{ji}$ **для всех** $i, j = 1, \ldots, n$ **(аксиома симметрии)**

Корреляционная матрица «объект-объект»

Мера сходства двух объектов – квадратный корень из коэффициента корреляции

$$\hat{r}_{ik} = \frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k)}{\{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2 \sum_{j=1}^{m}(x_{kj} - \bar{x}_k)^2\}^{1/2}}$$

xij– значение *j*-ой переменной для *i*-го объекта, $\bar{x}_i$ – среднее для всех переменных *i*-го объекта, *n* – число объектов, *m* – число признаков

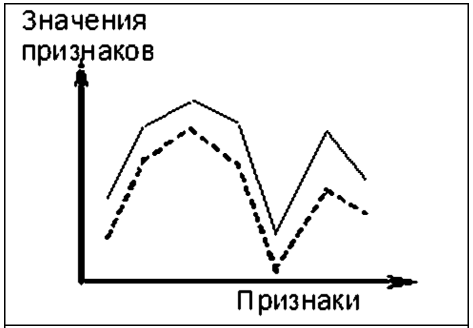

Рис. 1. Профили двух объектов

Меры расстояния

расстояние Минковского $d_{ij} = d(X_i, X_j) = \left(\sum_{k=1}^{m} w_k \,|x_{ik} - x_{jk}|^q\right)^{1/q},$

Частные случааи расстояния Минковского:
а) евклидово расстояние $d_{ij} = \left(\sum_{k=1}^{m} (x_{ik} - x_{jk})^2\right)^{1/2}, \quad w_k = 1,\ q = 2$

б) манхеттингское расстояние (расстояние «Сити-блок»

$$d_{ij} = \sum_{k=1}^{m} |x_{ik} - x_{jk}|, \quad w_k = 1,\ q = 1$$

в) доминантное расстояние (чебышевская метрика)

$$d_{ij} = \max_k |x_{ik} - x_{jk}|, \quad w_k = 1,\ q = \infty$$

Расстояние Махаланобиса

$$d_{ij} = (\mathbf{x}_i - \mathbf{x}_j)' \mathbf{K}_x^{-1} (\mathbf{x}_i - \mathbf{x}_j)$$

где $\mathbf{x}_i$, $\mathbf{x}_j$ – векторы переменных объектов *i* и *j*, $\mathbf{K}_x^{-1}$ – обратная корреляционная матрица признаков.

Определение статистической значимости коэффициентов корреляции

1. Коэффициент корреляции

(9)
$$r = \frac{\sum\left(x_{1,i} - \overline{x_1}\right)\cdot\left(x_{2,i} - \overline{x_2}\right)}{\sqrt{\sum\left(x_{1,i} - \overline{x_1}\right)^2}\cdot\sqrt{\sum\left(x_{2,i} - \overline{x_2}\right)^2}}$$

2. Значение тестовой статистики

$$\xi = \left(0.5\cdot\ln\left(\frac{1+r}{1-r}\right) - \frac{|r|}{2\cdot(n-1)}\right)\cdot\sqrt{n-3}$$

3. Если тестовая статистика больше критерия Стьюдента, то коэффициент корреляции значимо отличается от 0, в противном случае – не значимо

Списки использованной литературы и источников:

- А.А.Большаков, Р.Н.Каримов «Методы обработки многомерных данных и временных рядов» Москва 2007 г.
- Электронный учебник StatSoft по анализу данных.