Заметки про метод главных компонент

Винни-Пух

6/3/2017

Обозначения:

z_j — вектор столбец j-ой исходной переменной

Z — матрица исходных переменных

x_j — центрированный и, возможно, нормированный вектор столбец исходных переменных

X — матрица центрированных (нормированных) переменных

pc_j — главная компонента номер j

PC — матрица главных компонент

v_j — вектор столбец весов, с которыми x_1, x_2, ..., x_k входят в компоненту pc_j.

V — матрица всех весов

Статистический смысл некоторых объектов

- Длина центрированного вектора пропорциональна стандартному отклонению

Вектор x_j имеет нулевое среднее, $\bar{x}_j = 0$, поэтому его выборочное стандартное отклонение равняется

$$\hat{\sigma}(x_j) = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}} = \frac{\sqrt{\sum_{i=1}^n x_{ij}^2}}{\sqrt{n-1}} = \frac{|x_j|}{\sqrt{n-1}}$$

То есть, если среднее значение по вектору равняется нулю, то длина вектора в $\sqrt{n-1}$ раз больше, чем выборочное стандартное отклонение.

- Смысл матрицы $X^T X$

В матрице $X^T X$ на месте (i, j) находится произведение i-ой строки из X^T и j-го столбца из X, то есть

$$(X^T X)_{ij} = x_i^T \cdot x_j = \sum_{k=1}^n x_{ki} \cdot x_{kj}$$

Допустим, исходные переменные были только центрированы. Тогда

$$\sum_{k=1}^n x_{ki} \cdot x_{kj} = \sum_{k=1}^n (z_{ki} - \bar{z}_i) \cdot (z_{kj} - \bar{z}_j)$$

Если эту величину поделить на $n-1$, то получится выборочная ковариация между векторами z_i и z_j. Значит, $\frac{1}{n-1} X^T X$ — это выборочная ковариационная матрица векторов z_1, z_2, ..., z_k.

Если исходные переменные были ещё и нормированы, то $\frac{1}{n-1} X^T X$ — это выборочная корреляционная матрица векторов z_1, z_2, ..., z_k.

Алгоритм 1

1. Центрируем переменные
2. Если переменные в разном масштабе, то приводим переменные к общему масштабу
3. Находим главные компоненты

3.1. Первая главная компонента:

[

]