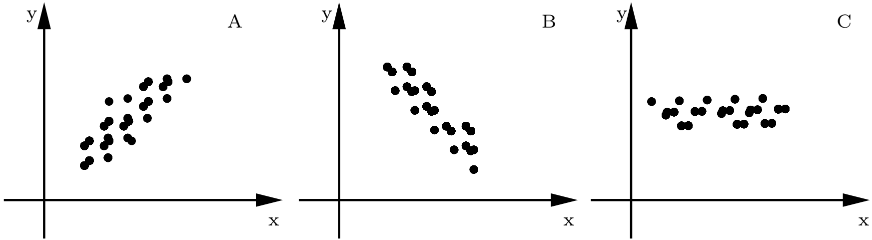
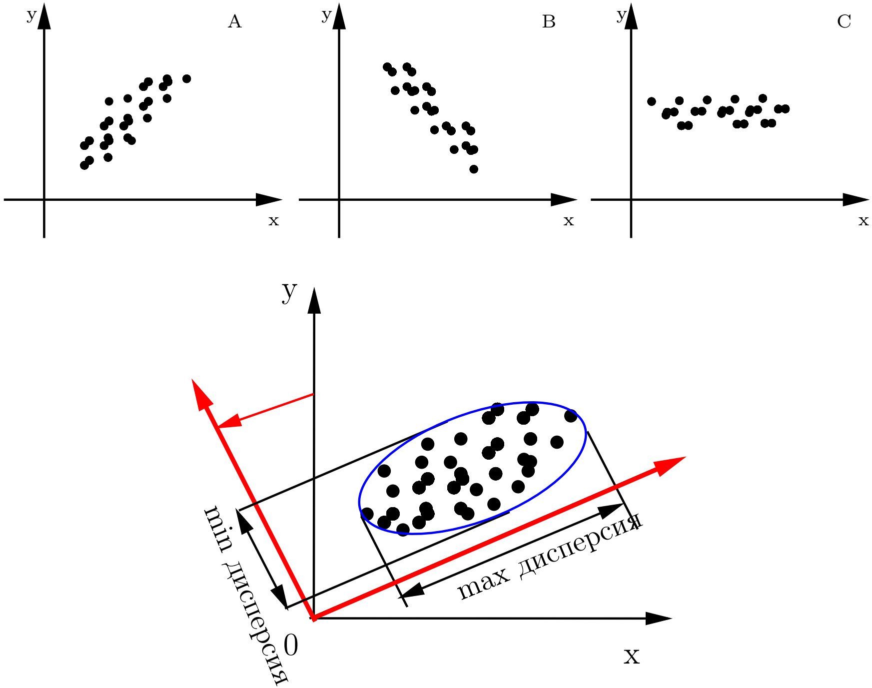
Метод главных компонент

М.И.Зуев, Ю.Л.Калиновский, О.И.Стрельцова

IT SCHOOL JINR
7 - 11 октября 2024



y
A
x
y
B
x
y
C
x


A
B
C
x
y
0
max дисперсия
min дисперсия

Целями PCA являются:

1. извлечь наиболее важную информацию из таблицы данных;
2. сократить размер набора данных, сохранив только эту важную информацию;
3. упростить описание набора данных;
4. Проанализировать структуру наблюдений и переменных.
5. Сжать данные, уменьшив число измерений, без значительной потери информации.

Математические основы

Обсудим статистику, которая рассматривает измерения распределения, как распределяются данные, а также матричную алгебру, вычисляя собственные векторы и собственные значения.

Дисперсия (variance)

- X - случайная величина
- $\mathrm{D}[\mathrm{X}] = \mathrm{M}[\mathrm{X} - \mathrm{M}[\mathrm{X}]]^2$
- $\mathrm{Var}(\mathrm{X}) = \mathrm{E}[(\mathrm{X} - \mathrm{E}(\mathrm{X}))^2]$
- Для вещественных значений: $\mathrm{D}[\mathrm{X}] = \mathrm{M}[\mathrm{X}^2] - (\mathrm{M}[\mathrm{X}])^2$
- Среднеквадратическое отклонение: $\sigma_{\mathrm{X}} = \sqrt{\mathrm{D}[\mathrm{X}]}$

Ковариация (covariance)

- X, Y - случайные величины

$$\mathrm{cov}(X, Y) = M[X - M[X]]M[Y - M[Y]]$$
$$\mathrm{cov}(X, Y) = E[X - E[X]]E[Y - E[Y]]$$

- Для выборки $X_{(n)}, Y_{(n)}$

$$\mathrm{cov}(X_{(n)}, Y_{(n)}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

где $\bar{x}, \bar{y}$ - средние значения выборок

Ковариационная матрица

- Пусть X, Y - случайные векторы размерности m и n соответственно

$$\begin{aligned} C &= \operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)^T] \\ C &= [c_{ij}] \\ c_{ij} &= \operatorname{cov}(X_i, Y_j) = E[(X_i - EX_i)(Y_j - EY_j)], \end{aligned}$$

$i - 1, 2, \ldots, n,\ j - 1, 2, \ldots, m$

- Если $X = Y$
  C - матрица ковариации вектора X
- Для выборок

$$c_{ij} = \frac{1}{m-1} \sum_{k-1}^{m} (x_{ki} - \bar{X}_i)(x_{kj} - \bar{X}_j),$$

где $\bar{X}_i$ и $\bar{X}_j$ - средние соответствующих компонентов векторов

Корреляция (correlation)

- Для выборки $X_{(n)}, Y_{(n)}$
- Коэффициент корреляции Пирсона

$$r_{XY} = \frac{(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

  $r_{XY} \in [-1, +1]$
- Мера линейной зависимости
  $||r_{XY}|| = 1, \Longrightarrow x, y$ - линейно зависимы
  $r_{XY} = 0, \Longrightarrow x, y$ - линейно независимы

Корреляция и коэффициент корреляции

Неравенство Коши-Буняковского

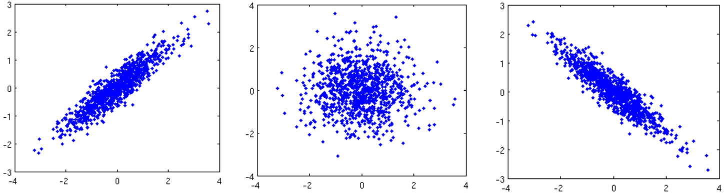
$$\operatorname{cov}^2(X,Y) \le D(X)D(Y)$$

Линейный коэффициент корреляции

$$r_{XY} = \frac{\operatorname{cov}_{XY}}{\sigma_x \sigma_y} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 (Y-\bar{Y})^2}}$$

$$-1 \le r_{XY} \le 1$$

Положительная корреляция | **Отсутствие корреляции** | **Отрицательная корреляция**



Матричная алгебра

Множество $\mathbb{R}^2$ имеет геометрическую интерпретацию как евклидова плоскость, в которой вектор $\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ в $\mathbb{R}^2$ представляет собой точку с координатами (a_1, a_2) на плоскости (рис. 1).

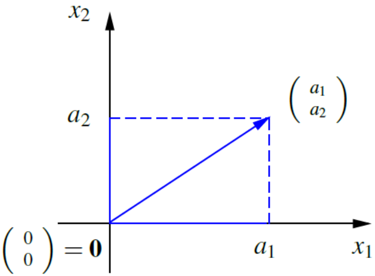


Figure 1: Векторы на плоскости.

Точно также отождествим $\mathbb{R}^3$ с трехмерным пространством, запишем точку с координатами $(\mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3)$ как вектор $\begin{pmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \\ \mathrm{a}_3 \end{pmatrix}$ в $\mathbb{R}^3$, выходящий из начала координат к точке (рис. 2).

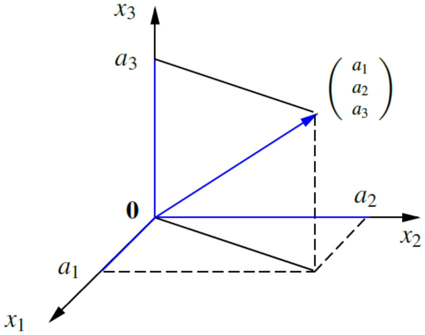


Figure 2: Векторы в пространстве.

Для описания преобразования $T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ необходимо задать вектор $T(x)$ в $\mathbb{R}^m$ для каждого x в $\mathbb{R}^n$. Это называется определением T, или определением действия T.
Говоря, что действие определяет преобразование, мы имеем в виду, что два преобразования $S : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $T : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ являются равными, если их действие одинаково.
Более формально

$$S = T, \text{ если, и только если } S(x) = T(x) \text{ для всех } x \in \mathbb{R}^n.$$

Умножение матриц является важным способом определения преобразований.
Если A - матрица $m \times n$, то умножение на A и дает преобразование

$$T_A : \mathbb{R}^n \longrightarrow \mathbb{R}^m,$$

$$\text{определяемое как } T_A(x) = Ax \text{ для каждого } x \in \mathbb{R}^n.$$

T_A называется преобразованием, индуцированным A.

Метод главных компонент (PCA)

- Один из основных практических способов уменьшить размерность данных
- Дана матрица $X_{m \times n}$ - матрица «объекты - признаки»
- Реализация метода
  - вычисление собственных векторов и собственных значений ковариационной матрицы исходных данных
  - сингулярное разложение центрированной матрицы исходных данных

Изменение базиса

Представим набор исходным данных: каждый столбец представляет собой единичный образец (или момент времени) нашего набора данных (т.е. X).
Пусть Y - другая m×n матрица, связанная линейным преобразованием P.
X - это исходный записанный набор данных. Y - представляет собой новое представление этого набора данных.

$$\mathrm{PX} = \mathrm{Y}$$

$\mathrm{p_i}$ - строки P,
$\mathrm{x_i}$ - столбцы X,
$\mathrm{y_i}$ - столбцы Y.

$$\mathrm{PX} = \begin{bmatrix} \mathrm{p_1} \\ \vdots \\ \mathrm{p_m} \end{bmatrix} \left[\begin{array}{ccc} \mathrm{x_1} & \ldots & \mathrm{x_n} \end{array} \right]$$

$$\mathrm{Y} = \begin{bmatrix} \mathrm{p_1 x_1} & \ldots & \mathrm{p_1 x_n} \\ \vdots & \ddots & \vdots \\ \mathrm{p_m x_1} & \ldots & \mathrm{p_m x_n} \end{bmatrix}$$

$$\mathrm{y_i} = \begin{bmatrix} \mathrm{p_1 x_i} \\ \vdots \\ \mathrm{p_m x_i} \end{bmatrix}$$

Матрица ковариантности

$$\mathrm{C_X} = \frac{1}{\mathrm{n}}\mathrm{X}\mathrm{X}^{\mathrm{T}}$$

$\mathrm{C_X}$ является квадратной симметричной матрицей
Диагональные элементы $\mathrm{C_X}$ являются дисперсиями определенных типов измерений.
Недиагональные термины $\mathrm{C_X}$ являются ковариациями между типами измерений.

- Выбрать нормированное направление в m - мерное пространство, вдоль которого дисперсия в X максимальна. Сохранить этот вектор как p_1.
- Найти другое направление, вдоль которого дисперсия максимальна, однако, из-за условия ортонормальности, ограничить поиск всеми направлениями, ортогональными всем ранее выбранным направлениям. Сохраните этот вектор как p_i
- Повторить эту процедуру до тех пор, пока не выбраны все m векторов.

Полученный упорядоченный набор представляет собой основные компоненты.

Найдем некоторую ортонормальную матрицу P в $\mathrm{Y} = \mathrm{PX}$ такую, что $\mathrm{C_Y} = \frac{1}{\mathrm{n}}\mathrm{YY^T}$ - диагональная матрица.
Строки в P являются основными компонентами для X.

$$
\begin{aligned}
\mathrm{C_Y} &= \frac{1}{\mathrm{n}}\mathrm{YY^T} \\
&= \frac{1}{\mathrm{n}}(\mathrm{PX})(\mathrm{PX^T}) \\
&= \frac{1}{\mathrm{n}}\mathrm{PXX^TP^T} \\
&= \frac{1}{\mathrm{n}}\mathrm{PC_XP^T} \\
\mathrm{C_Y} &= \mathrm{PC_XP^T}
\end{aligned}
$$

Метод SVD

.

Пусть X произвольная матрица $n \times m$ и $X^T X$ будет квадратной симметричной матрицей ранга r.

Определим все интересующие нас величины.

- $\{\hat{v}_1, \hat{v}_2, \ldots \hat{v}_r\}$ - ортонормированный набор $m \times 1$ собственных векторов с соответствующими собственными значениями $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$.
  Для симметричной матрицы $X^T X$:

$$(X^T X)\hat{v}_i = \lambda_i \hat{v}_i.$$

- $\sigma_i = \sqrt{\lambda_i}$ являются положительными действительными и называются сингулярными числами.
- $\{\hat{u}_1, \hat{u}_2, \ldots \hat{u}_r\}$ - набор $n \times 1$ векторов $\hat{u}_i = \frac{1}{n} X \hat{v}_i$.
-

$$\begin{aligned} \hat{u}_i \cdot \hat{u}_j &= 1, i = j \\ \hat{u}_i \cdot \hat{u}_j &= 0, i \neq j \end{aligned}$$

- $||X\hat{v}_i|| = \sigma_i$.

Отсюда следует, что мы можем построить матрицу Σ, у которой на главной диагонали будут стоять собственные значения, а все остальные элементы будут равны нулю .

$$\mathrm{XV} = \mathrm{U}\Sigma \Longrightarrow \mathrm{X} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}.$$