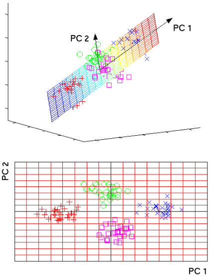
Метод главных компонент

Виктор Китов

victorkitov.github.io

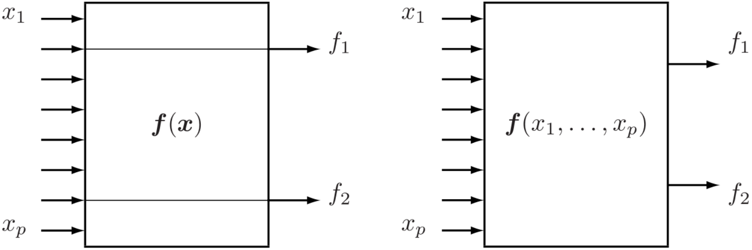


Курс поддержан фондом 'Интеллект'

Победитель конкурса VK среди курсов по IT

Содержание

Задача снижения размерности



(a) feature selector (b) feature extractor

Снижение размерности: трансформация признаков в уменьшенное число признаков, зависящих от всех входных в общем случае.

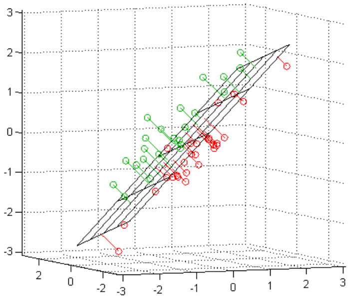
Применения снижения размерности

Применения снижения размерности:

- Визуализация многомерных данных в 2D или 3D
- Снижение вычислительных ресурсов при обучении и применении
  - процессор, память, хранение на диске, пересылка
- Повышение интерпретируемости модели
  - если извлеченные признаки интерпретируемы
- Повышение устойчивости некоторых методов
  - при линейно-зависимых признаках коэффициенты лин. регрессии не определены

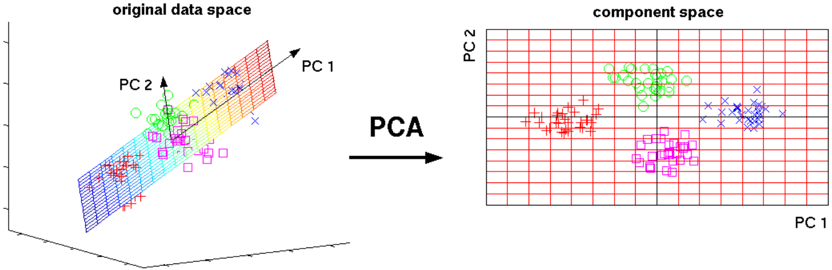
Категоризация методов снижения размерности

- Снижение размерности - с учителем/без учителя, линейное/нелинейное
- Метод главных компонент - линейный метод снижения размерности без учителя.



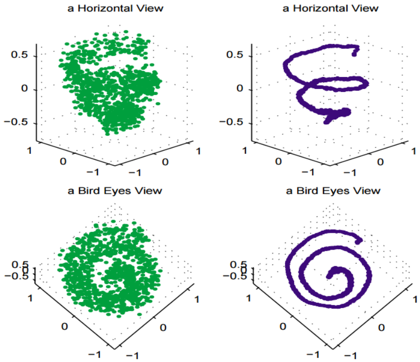
Содержание

Визуализация



Фильтрация данных

Убираем шум из данных[1]:



[1] X. Huo and Jihong Chen (2002). Local linear projection (LLP). First IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS).

Снижение размерности

Задача идентификации человека по лицу:

Для фото *HxW*: *HW* признаков, переобучение.

Главные компоненты (eigenfaces)

Главные компоненты (eigenfaces).

Проекции на гл. компоненты - информативные признаки.

Анализ текстов

- Объекты - текстовые файлы.
- Индикаторные, TF, TF-IDF кодировки приводят в высокому D.
    - вычислительно долгая работа с X и настройкой моделей
- Разреженность данных приводит к проблемам:
    - например, задача поиска:
      "ремонт машины" != "обслуживание автомобилей"

Анализ текстов

- Объекты - текстовые файлы.
- Индикаторные, TF, TF-IDF кодировки приводят в высокому D.
  - вычислительно долгая работа с X и настройкой моделей
- Разреженность данных приводит к проблемам:
  - например, задача поиска:
    "ремонт машины" != "обслуживание автомобилей"
- Снижение размерности PCA позволяет решить эти проблемы.
  - технически-через сокр. сингулярное разложение
  - достаточно 200-300 гл. компонент
  - признаки не центрируются, чтобы не потерять разреженность
  - англ. latent semantic analysis (LSA)

Содержание

Матрица ковариации

- Матрица ковариации

$$\Sigma = \{cov(x^i, x^j)\}_{i,j=1}^{D} = \{\mathbb{E}\{(x^i - \mathbb{E}x^i)(x^j - \mathbb{E}x^j)\}\}$$

- Из определения $\Sigma = \Sigma^T$.
- Свойства симметричных матриц:
  - все СЗ симметричной матрицы вещественные.
  - существует ортонормированный базис из СВ.

Теорема (Спектральное разложение.)

Любая симметричная $\Sigma \in \mathbb{R}^{DxD}$ может быть представлена как

$$\Sigma = A\Lambda A^T$$

где $A \in \mathbb{R}^{DxD}$ - ортогональная матрица, колонки которой $a_1, ...a_D$ - СВ, а $\Lambda = \text{diag}\{\lambda_1, ...\lambda_D\}$ с СЗ Σ на диагонали.

Дисперсия распределения вдоль направления

Для случайной величины $x \in \mathbb{R}^D$, $x \sim F(\mu,\Sigma)$, и $\forall b \in \mathbb{R}^D$:

$$var(b^T x) = \mathbb{E}\left\{\left(b^T x - b^T \mu\right)^2\right\}$$
$$= \mathbb{E}\left\{\left(b^T x - b^T \mu\right)\left(x^T b - \mu^T b\right)\right\}$$
$$= b\mathbb{E}\left\{(x-\mu)(x-\mu)^T\right\} b = b^T \Sigma b$$

Поскольку b - произвольно, то $\Sigma \succeq 0$, т.к. $b^T \Sigma b = var(b^T x) \geq 0$.

- следовательно все СЗ≥ 0, т.к. $0 \leq a_i^T \Sigma a_i = \lambda_i a_i^T a_i = \lambda_i$

Дисперсия распределения вдоль разных направлений

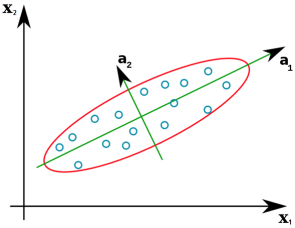
- Для различных $b \in \mathbb{R}^D$, $\|b\| = 1$:

$$b^T x - \text{проекция на ось } b.$$

$$\text{var}(b^T x) = b^T \Sigma b = b^T A \Lambda A^T b =$$

$$= \left(\Lambda^{1/2} A^T b\right)^T \left(\Lambda^{1/2} A^T b\right) = \left\|\Lambda^{1/2} A^T b\right\|^2$$

- Интуиция: $b \to$в базис СВ, координаты масштабируются на $\sqrt{\lambda_1}, ... \sqrt{\lambda_D}$.



Направления максимального разброса

- Упорядочим СВ $a_1, ...a_D$ по убыванию СЗ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$.
- $a_1, ...a_D$ - называются главными компонентами
    - a_1 - направление макс. дисперсии $\text{var}(a_1^T x) = \lambda_1$
    - a_2 - ортогональное a_1 направление макс. дисперсии $\text{var}(a_2^T x) = \lambda_2$
    - a_3 - ортогональное a_1, a_2 направление макс. дисперсии $\text{var}(a_3^T x) = \lambda_3$
    - $\cdots$
- Относительный разброс вдоль осей $a_1, ...a_D$:

$$\frac{\lambda_1}{\lambda_1 + ... + \lambda_D}, ... \frac{\lambda_D}{\lambda_1 + ... + \lambda_D}$$

Оценка разброса распределения

Оценим средний разброс сл. вел. $x \sim F(\mu, \Sigma)$:

- используя инвариантность tr и det к смене базиса

$$\frac{1}{D}(\lambda_1 + ... + \lambda_D) = \frac{1}{D}\operatorname{trace}\Lambda = \frac{1}{D}\operatorname{trace} A\Lambda A^T = \frac{1}{D}\operatorname{trace}\Sigma$$

$$\sqrt[D]{\lambda_1 \cdot ... \cdot \lambda_D} = \sqrt[D]{\det\Lambda} = \sqrt[D]{\det A\Lambda A^T} = \sqrt[D]{\det\Sigma}$$

Метод главных компонент

- Отцентрируем признаки $x_n := x_n - \mu \; \forall n$
- Матрица попарных скалярных произведений признаков:

$$X^T X = N \frac{1}{N} X^T X = N\widehat{\Sigma}$$

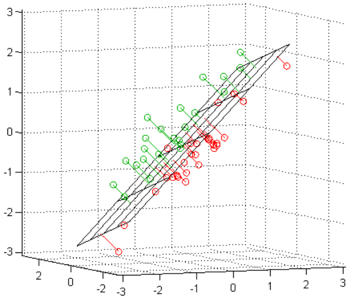
- Главные компоненты: $a_1, ...a_D$ - СВ $\widehat{\Sigma}$, отвечающие СЗ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$.
  - совпадают с СВ $X^T X$, а СЗ - отличаются в N раз при предварительном центрировании признаков.
- В TF-IDF представлениях текстов признаки не центрируются, т.к. потеряем разреженность.
  - в силу разреженности X в любом случае $\mathbb{E}x^i \approx 0$.
- Метод главных компонент: $x \rightarrow$проекции на $a_1, ...a_K$, $K < D$.

Содержание

4 Подпространство наилучшей аппроксимации
- Определение
- Оценка качества аппроксимации
- Проецирование на L_K

Подпространство наилучшей аппроксимации

Метод главный компонент находит подпространство наилучшей аппроксимации:



Первые K главных компонент $a_1, a_2, ...a_K$ - ортонормированный базис этого подпространства.

Проекции, ортогональные дополнения

- Для точки x и подпространства L обозначим:
    - p: проекция x on L
    - h: ортогональное дополнение
    - $x = p + h$, $\langle p, h \rangle = 0$.
- Для обучающей выборки $x_1, x_2, ...x_N$ и подпространства L обозначим:
    - проекции: $p_1, p_2, ...p_N$
    - ортогональные дополнения: $h_1, h_2, ...h_N$.

Подпространство наилучшей аппроксимации

Рассмотрим K-мерное подпространство - линейную оболочку базиса $v_1, v_2, ...v_K$: $L_K = \mathcal{L}(v_1, v_2, ...v_K)$

Определение 1

L_K - подпространство наилучшей аппроксимации для набора точек $x_1, x_2, ...x_N$, если решает задачу

$$\sum_{n=1}^{N} \|h_n\|^2 \to \min_{L:\text{rg}\, L=K}$$

Предложение 1

L_K - подпространство наилучшей аппроксимации для набора точек $x_1, x_2, ...x_N$, если решает задачу[a]*.*

$$\sum_{n=1}^{N} \|p_n\|^2 \to \max_{L:\text{rg}\, L=K}$$

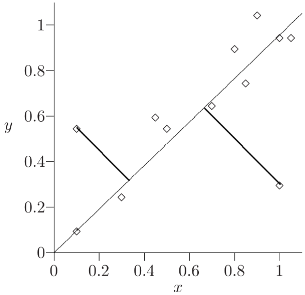
[a] Докажите, используя $\|x\|^2 = \|p\|^2 + \|h\|^2$ для $x = p + h$ и $\langle p, h\rangle = 0$.

Свойства главных компонент

- D главных компонент образуют ортонормированный базис пространства признаков.
- Не инвариантны к сдвигу $x_1, x_2, ...x_D$.
- Не инвариантны к масштабу $x_1, x_2, ...x_D$.
    - рекомендуется центрировать и приводить к одинаковой шкале.
    - не центрируется для текстовых данных:
        - X - разреженная, поэтому уже $\bar{x}_i \approx 0$. Сдвиг сделает X не разреженной.

Пример L_1

- Рассмотрим одномерное подпространство наилучшей аппроксимации L_1:



- В чем отличие от нахождения $y = wx$ в линейной регрессии?

4. Подпространство наилучшей аппроксимации
   - Определение
   - Оценка качества аппроксимации
   - Проецирование на L_K

Оценка качества аппроксимации

- Величина проекции (со знаком) x на a: $\langle x, a\rangle / \|a\|$
- Т.к. $a_1, a_2, ...a_D$ - ОНБ, для любого x

$$x = \langle x, a_1\rangle a_1 + \langle x, a_2\rangle a_2 + ... + \langle x, a_D\rangle a_D$$

Оценка качества аппроксимации

- Величина проекции (со знаком) x на a: $\langle x, a\rangle / \|a\|$
- Т.к. $a_1, a_2, ...a_D$ - ОНБ, для любого x

$$x = \langle x, a_1\rangle a_1 + \langle x, a_2\rangle a_2 + ... + \langle x, a_D\rangle a_D$$

- Пусть p^K - проекция, а h^K- орт. дополнение x на L_K.

$$p^K = \langle x, a_1\rangle a_1 + \langle x, a_2\rangle a_2 + ... + \langle x, a_K\rangle a_K$$

$$h^K = x - p^K = \langle x, a_{K+1}\rangle a_{K+1} + ... + \langle x, a_D\rangle a_D$$

Оценка качества аппроксимации

- Величина проекции (со знаком) x на a: $\langle x, a\rangle / \|a\|$
- Т.к. $a_1, a_2, ... a_D$ - ОНБ, для любого x

$$x = \langle x, a_1\rangle a_1 + \langle x, a_2\rangle a_2 + ... + \langle x, a_D\rangle a_D$$

- Пусть p^K - проекция, а h^K- орт. дополнение x на L_K.

$$p^K = \langle x, a_1\rangle a_1 + \langle x, a_2\rangle a_2 + ... + \langle x, a_K\rangle a_K$$
$$h^K = x - p^K = \langle x, a_{K+1}\rangle a_{K+1} + ... + \langle x, a_D\rangle a_D$$

- Рассчитаем квадраты длин x, p^K, h^K:

$$\|x\|^2 = \langle x, x\rangle = \langle x, a_1\rangle^2 + ... + \langle x, a_D\rangle^2$$
$$\left\|p^K\right\|^2 = \langle p^K, p^K\rangle = \langle x, a_1\rangle^2 + ... + \langle x, a_K\rangle^2$$
$$\left\|h^K\right\|^2 = \langle h^K, h^K\rangle = \langle x, a_{K+1}\rangle^2 + ... + \langle x, a_D\rangle^2$$

Оценка качества аппроксимации

p_n^K, h_n^K - проекция и ортогональное дополнение x_n для L_K.

$$L(K)=\frac{\sum_{n=1}^{N}\left\|h_n^K\right\|^2}{\sum_{n=1}^{N}\left\|x_n\right\|^2},\quad S(K)=\frac{\sum_{n=1}^{N}\left\|p_n^K\right\|^2}{\sum_{n=1}^{N}\left\|x_n\right\|^2},\quad L(K)+S(K)=1$$

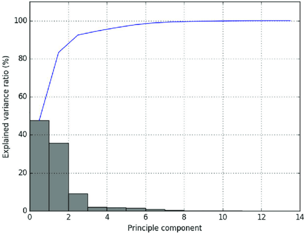
Вклад a_k в описание x: $\langle x, a_k\rangle^2$.
Вклад a_k в описание $x_1, x_2, ...x_N$: $\sum_{n=1}^{N}\langle x_n, a_k\rangle^2$
Относительный вклад (explained variance ratio):

$$E(a_k)=\frac{\sum_{n=1}^{N}\langle x_n, a_k\rangle^2}{\sum_{n=1}^{N}\sum_{d=1}^{D}\langle x_n, a_d\rangle^2}=\frac{\sum_{n=1}^{N}\langle x_n, a_k\rangle^2}{\sum_{n=1}^{N}\left\|x_n\right\|^2}$$

$$E(a_k)\in[0,1];\qquad \sum_{k=1}^{K}E(a_k)=S(K)$$

Выбор числа главных компонент

- Визуализация данных: 2 или 3 компоненты.



- Можно брать a_k, пока $E(a_k)$ не упадет резко вниз.
- Или брать по порогу, например

$$K^* = \arg\min_K E(a_K) < 0.01$$

$$K^* = \arg\min_K \{S(K) > 0.95\} = \arg\min_K \left\{ \sum_{k=1}^{K} E(a_k) > 0.95 \right\}$$

4 Подпространство наилучшей аппроксимации

- Определение
- Оценка качества аппроксимации
- Проецирование на L_K

Расчет p^K по x

$x \to y$ (значения проекций x на $a_1, ...a_D$):

$$y = A^T(x - \mu)$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \;\; A = [a_1|a_2|...|a_D] \in \mathbb{R}^{DxD}$$

Для $A_K = [a_1|a_2|...|a_K] \in \mathbb{R}^{DxK}$, значения проекций на $a_1, ...a_K$:

$$y^K = A_K^T(x - \mu)$$

$x \to p^K$(вектор проекций в исх. базисе):

$$p^K = A\begin{pmatrix} y^K \\ 0 \end{pmatrix} + \mu = A_K y^K + \mu = A_K A_K^T(x - \mu) + \mu$$

Численное нахождение главных компонент

- Определяем вектор средних и станд. отклонений каждого признака:

$$\mu, \sigma \in \mathbb{R}^D$$

- Приводим все признаки к нулевому среднему и единой шкале:

$$x_1, ...x_N \to \frac{x_1 - \mu}{\sigma}, ... \frac{x_N - \mu}{\sigma}$$

- Формируем матрицу объекты-признаки

$$X = [x_1^T; ...x_N^T]^T \in \mathbb{R}^{NxD}$$

- Оцениваем выборочную ковариационную матрицу$\in \mathbb{R}^{DxD}$:

$$\widehat{\Sigma} = \frac{1}{N} X^T X$$

Численное нахождение главных компонент

- По $\widehat{\Sigma}$: находим СЗ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$ и соответствующие СВ $a_1, a_2, ...a_D$.
    - $\widehat{\Sigma} = \widehat{\Sigma}^T$, поэтому существует ОНБ из СВ с вещественными СЗ
    - $\widehat{\Sigma} \succeq 0$, поэтому все СЗ ≥ 0
- $a_1, a_2, ...a_K$ - первые K главных компонент, $k = 1, 2, ...D$.
- Сумма квадратов проекций на a_i:

$$\|Xa_i\|^2 = \sum_{n=1}^{N} \langle x_n, a_i \rangle^2 = \lambda_i$$

- Доля объясненной информации a_i:

$$E(a_i) = \frac{\lambda_i}{\sum_{d=1}^{D} \lambda_d}$$

Содержание

Конструктивное определение главных компонент

- $a_1 = \arg\max_a \|Xa\|^2$, при ограничении $\langle a, a\rangle = 1$
- $a_2 = \arg\max_a \|Xa\|^2$, при ограничениях $\langle a, a\rangle = 1, \langle a, a_1\rangle = 0$
- $a_3 = \arg\max_a \|Xa\|^2$, при ограничениях $\langle a, a\rangle = 1, \langle a, a_1\rangle = 0, \langle a, a_2\rangle = 0$
- $\cdots \qquad \cdots \qquad \cdots$
- $a_D = \arg\max_a \|Xa\|^2$, при ограничениях $\langle a, a\rangle = 1, \langle a, a_1\rangle = 0, ...\langle a, a_{D-1}\rangle = 0$
- $Xa_i = [\langle x_1, a_i\rangle, ...\langle x_N, a_i\rangle]$ - вектор координат (проекций) всех объектов вдоль a_i.
- Квадрат нормы через $\langle\cdot,\cdot\rangle$:

$$\|b\|^2 = b^T b, \qquad \|Xa\|^2 = (Xa)^T (Xa) = a^T X^T Xa$$

Векторные производные некоторых функций[2]

- Рассмотрим $x = [x^1, ...x^D]$ и $f(x) = f(x^1, ...x^D)$. Векторная производная

$$\frac{\partial f(x)}{\partial x} := \begin{pmatrix} \frac{\partial f(x)}{\partial x^1} \\ \frac{\partial f(x)}{\partial x^2} \\ \cdots \\ \frac{\partial f(x)}{\partial x^D} \end{pmatrix}$$

- Для любых $x, b \in \mathbb{R}^D$:

$$\frac{\partial [b^T x]}{\partial x} = b, \quad \frac{\partial [x^T x]}{\partial x} = 2x$$

- Для любых $x \in \mathbb{R}^D$ и симметричной $B \in \mathbb{R}^{DxD}$:

$$\frac{\partial [x^T B x]}{\partial x} = 2Bx$$

[2] Докажите их формулу. Как изменится формула для несимметричной B?

Вычисление 1-й главной компоненты

$$\begin{cases} \|Xa_1\|^2 \to \max_{a_1} \\ \|a_1\| = 1 \end{cases} \tag{1}$$

Лагранжиан оптимизационной задачи (1):

$$L(a_1, \mu) = a_1^T X^T X a_1 - \mu(a_1^T a_1 - 1) \to \text{extr}_{a_1, \mu}$$

$$\frac{\partial L}{\partial a_1} = 2X^T X a_1 - 2\mu a_1 = 0$$

поэтому a_1 - один из СВ матрицы $X^T X$.

Вычисление 1-й главной компоненты

Поскольку мы ищем $\|Xa_1\|^2 \to \max_{a_1}$ и

$$\|Xa_1\|^2 = (Xa_1)^T Xa_1 = a_1^T X^T X a_1 = \lambda a_1^T a_1 = \lambda$$

a_1 должен быть СВ, отвечающим максимальному СЗ λ_1.

Если существует несколько СВ для λ_1, выберем любой единичной нормы.

Вычисление 2-й главной компоненты

$$\begin{cases} \|Xa_2\|^2 \to \max_{a_2} \\ \|a_2\| = 1 \\ a_2^T a_1 = 0 \end{cases} \tag{2}$$

Лагранжиан оптимизационной задачи (2):

$$L(a_2, \mu) = a_2^T X^T X a_2 - \mu(a_2^T a_2 - 1) - \alpha a_1^T a_2 \to \text{extr}_{a_2,\mu,\alpha}$$

$$\frac{\partial L}{\partial a_2} = 2X^T X a_2 - 2\mu a_2 - \alpha a_1 = 0 \tag{3}$$

Вычисление 2-й главной компоненты

Домножая на a_1^T слева, получим:

$$a_1^T \frac{\partial L}{\partial a_1} = 2a_1^T X^T X a_2 - 2\mu a_1^T a_2 - \alpha a_1^T a_1 = 0 \tag{4}$$

$$\text{т.к. } \langle a_2, a_1 \rangle = 0: \quad 2\mu a_1^T a_2 = 0$$

Поскольку $a_1^T X^T X a_2 \in \mathbb{R}$ и a_1 - СВ $X^T X$:

$$a_1^T X^T X a_2 = \left(a_1^T X^T X a_2\right)^T = a_2^T X^T X a_1 = \lambda_1 a_2^T a_1 = 0$$

Следовательно (4) упрощается до $\alpha a_1^T a_1 = \alpha = 0$ и (3) становится

$$X^T X a_2 - \mu a_2 = 0$$

Значит a_2 - тоже СВ $X^T X$.

Вычисление 2-й главной компоненты

Поскольку мы ищем $\|Xa_2\|^2 \to \max_{a_2}$ и

$$\|Xa_2\|^2 = (Xa_2)^T Xa_2 = a_2^T X^T Xa_2 = \lambda a_2^T a_2 = \lambda$$

a_2 должен быть СВ, отвечающим 2-му максимальному СЗ λ_2.

Если существует несколько СВ для λ_1, выберем любой, удовлетворяющий (2).

Вычисление k-й главной компоненты

$$\begin{cases} \|Xa_k\|^2 \to \max_{a_k} \\ \|a_k\| = 1 \\ a_k^T a_1 = \ldots = a_k^T a_{k-1} = 0 \end{cases} \tag{5}$$

Лагранжиан оптимизационной задачи (5):

$$L(a_k, \mu) = a_k^T X^T X a_k - \mu(a_k^T a_k - 1) - \sum_{j=1}^{k-1} \alpha_j a_k^T a_j \to \text{extr}_{a_k, \mu, \alpha_1, \ldots \alpha_{k-1}}$$

$$\frac{\partial L}{\partial a_k} = 2X^T X a_k - 2\mu a_k - \sum_{j=1}^{k-1} \alpha_j a_j = 0 \tag{6}$$

Вычисление k-й главной компоненты

Домножая на a_i^T слева для $i = 1, 2, ...k - 1$ получим:

$$2a_i^T X^T X a_k - 2\mu a_i^T a_k - \alpha_1 a_i^T a_1 - ... - \alpha_{k-1} a_i^T a_{k-1} = 0$$

$$\text{т.к. } \forall i \neq j \ \langle a_i, a_j \rangle = 0: \qquad 2\mu a_i^T a_k = 0, \quad \alpha_j a_i^T a_j = 0 \ \forall i \neq j \tag{7}$$

Поскольку $a_i^T X^T X a_2 \in \mathbb{R}$ и a_i - СВ $X^T X$:

$$a_i^T X^T X a_2 = \left(a_i^T X^T X a_k\right)^T = a_k^T X^T X a_i = \lambda_i a_k^T a_i = 0$$

Следовательно (7) упрощается до $\alpha_i a_i^T a_i = \alpha_i = 0$. Выбирая $i = 1, 2, ...k - 1$, получим $\alpha_1 = \alpha_2 = ... = \alpha_{k-1} = 0$ и (6) становится

$$X^T X a_k - \mu a_k = 0$$

Значит a_k - тоже СВ $X^T X$.

Вычисление k-й главной компоненты

Поскольку мы ищем $\|Xa_k\|^2 \to \max_{a_k}$ и

$$\|Xa_k\|^2 = (Xa_k)^T Xa_k = a_k^T X^T Xa_k = \lambda a_k^T a_k = \lambda$$

a_k должен быть СВ, отвечающим k-му максимальному СЗ λ_k.

Если существует несколько СВ для λ_k, выберем любой, удовлетворяющий (5).

Содержание

$\mathcal{L}(a_1, a_2, ...a_K) = L_K$

Далее все рассматривается в контексте фиксированной выборки X, L_K - подпространство наилучшей аппроксимации ранга K для X.

Теорема 1

Линейная оболочка главных компонент $a_1, a_2, ...a_K$, рассчитанных по X. Тогда

$$\mathcal{L}(a_1, a_2, ...a_K) = L_K \quad \forall K$$

Доказательство: по индукции. Для $K = 1$

$$\begin{cases} \|Xa_1\|^2 \to \max_{a_1} \\ \|a_1\| = 1 \end{cases}$$

$$\|Xa_1\|^2 = \|\langle x_1, a_1\rangle, ...\langle x_N, a_1\rangle\|^2 = \sum_{n=1}^{N} p_n^2 \to \max_{a_1}$$

$\mathcal{L}(a_1, a_2, ...a_K) = L_K$

Предположим, теорема верна для $K-1$. Рассмотрим оптимальное L_K, $\dim L = K$, для которого мы всегда можем выбрать ОНБ $b_1, b_2, ...b_K$ такой, что

$$\begin{cases} \|b_K\| = 1 \\ b_K \perp a_1, b_K \perp a_2, ...b_K \perp a_{K-1} \end{cases} \tag{8}$$

выбирая b_K перпендикулярным проекциям $a_1, a_2, ...a_{K-1}$ на L_K.

$\mathcal{L}(a_1, a_2, ...a_K)$ - подпространство наилучшей аппроксимации

Рассмотрим сумму квадратов проекций:

$$\|Xb_1\|^2 + \|Xb_2\|^2 + ... + \|Xb_{K-1}\|^2 + \|Xb_K\|^2$$

По предположению индукции $L[a_1, a_2, ...a_{K-1}]$ подпространство наилучшей аппроксимации $K-1$ и $L[b_1, ...b_{K-1}]$ - того же ранга, поэтому сумма квадратов проекций не меньше:

$$\|Xb_1\|^2+\|Xb_2\|^2+...+\|Xb_{K-1}\|^2 \leq \|Xa_1\|^2+\|Xa_2\|^2+...+\|Xa_{K-1}\|^2$$

при этом

$$\|Xb_K\|^2 \leq \|Xa_K\|^2$$

т.к. b_K по (8) удовлетворяет (5) а a_K оптимальное решение.

Заключение

- Снижение размерности - преобразование признаков с переходом в пространство меньшей размерности.
- Полезно для повышения точности, интерпретируемости и скорости работы моделей.
- Метод главных компонент - метод линейного снижения размерности без учителя.
- Первые K главных компонент образуют ОНБ подпространства наилучшей аппроксимации.
    - в среднеквадратичном смысле