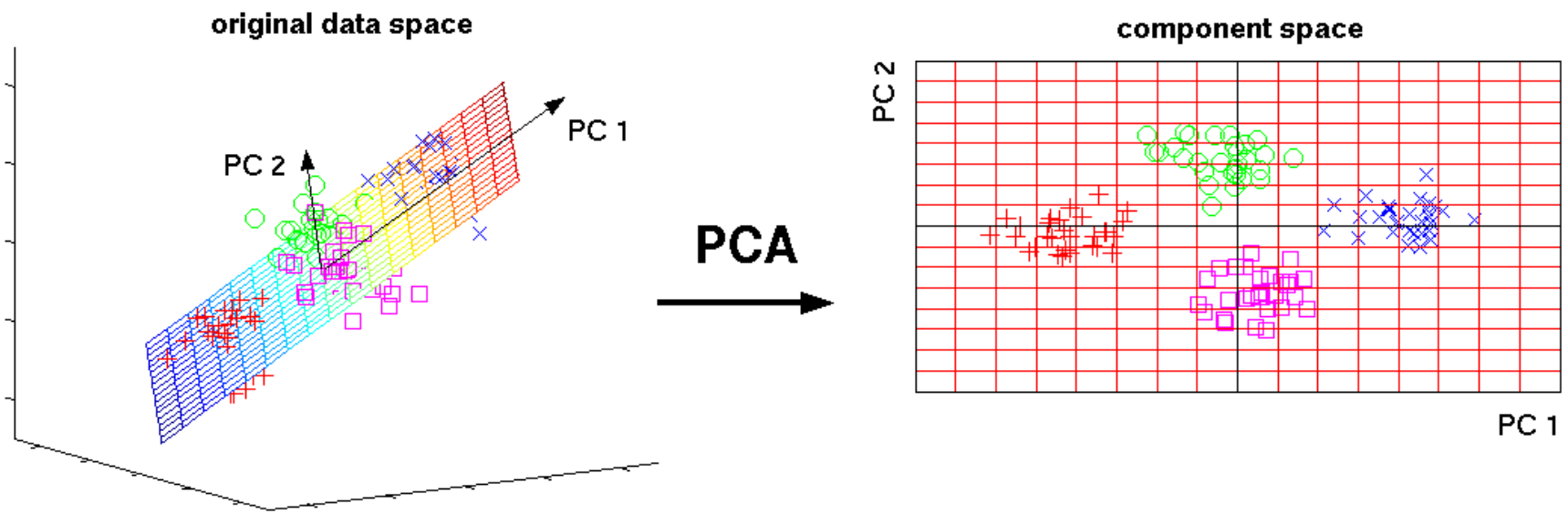
Метод главных компонент

Виктор Китов

v.v.kitov@yandex.ru



Содержание

Скалярное произведение

- Определяем $\langle a, b\rangle = a^T b$
- $\|a\| = \sqrt{\langle a, a\rangle}$
- Величина проекции (со знаком) x на a: $\langle x, a\rangle / \|a\|$
- Величина проекции (без знака) x на a: $|\langle x, a\rangle| / \|a\|$
- K-мерная плоскость L_K м. быть представлена как линейная оболочка ортонормированного базиса пространства (ОНБ) $a_1, a_2, ...a_K$:

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

Собственные вектора и собственные значения

- Если для матрицы $A \in \mathbb{R}^{DxD}$ найдется $\lambda \in \mathbb{R}$ и $v \in \mathbb{R}^D$ такой, что $Av = \lambda v$, то
  - v - собственный вектор (СВ) A
  - λ - собственное значение (СЗ) A, отвечающее собственному вектору v.
- $\exists v \neq 0 : \; Av = \lambda v \Leftrightarrow (A - \lambda I)\, v = 0 \Leftrightarrow det(A - \lambda I) = 0$. Таким образом, все собственные значения удовлетворяют $det(A - \lambda I) = 0$, которое
  - является полиномом порядка D
  - имеет D решений (возможно, повторяющиеся, могут быть комплексными)

Симметричные матрицы

- Матрица $A \in \mathbb{R}^{D \times D}$ называется симметричной, если $A^T = A$.
- Свойства:
  - Все собственные значения симметричной матрицы вещественные.
  - Собственные вектора, соответствующие различным λ, ортогональны друг другу.
  - если λ-повторяющийся корень $\det(A - \lambda I) = 0$ m раз, то существуют m ортогональных СВ, соответствующих СЗ λ.
  - $\forall A \in \mathbb{R}^{D \times D}, A = A^T$ существует ортонормированный базис из СВ A.

Спектральное разложение

Теорема 1 (Спектральное разложение.)

Любая симметричная $A \in \mathbb{R}^{D \times D}$ *может быть представлена как*

$$A = P \Lambda P^T$$

где $P \in \mathbb{R}^{D \times D}$ *- ортогональная матрица, колонки которой* $p_1, ...p_D$ *- СВ* A*, а* $\Lambda = \text{diag}\{\lambda_1, ...\lambda_D\}$ *с СЗ* A *на диагонали.*

Интерпретация: трансформация Ax симметричной матрицей A эквивалентна

1. переводу x в ортонормированный базис СВ A
2. масштабированию координат пропорционально $\lambda_1, ...\lambda_D$.
3. возврату в исходный базис.

Неотрицательная определенность $A \succeq 0$

Определение

Симметричная матрица $A \in \mathbb{R}^{DxD}$ называется неотрицательно определенной ($A \succeq 0$), если

$$\forall x \in \mathbb{R}^D : \langle x, Ax \rangle = x^T Ax \geq 0$$

- Являются ли следующие матрицы неотрицательно определенными: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$?

Теорема

Симметричная матрица A неотрицательно определена <=> все её СЗ ≥ 0.

Оценка разброса распределения

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

$$\begin{aligned} var(\alpha^T x) &= \mathbb{E}\left\{\left(\alpha^T x - \alpha^T \mu\right)^2\right\} \\ &= \mathbb{E}\left\{\left(\alpha^T x - \alpha^T \mu\right)\left(x^T \alpha - \mu^T \alpha\right)\right\} \\ &= \alpha \mathbb{E}\left\{(x-\mu)(x-\mu)^T\right\}\alpha = \alpha^T \Sigma \alpha \end{aligned}$$

- Т.к. $\forall \alpha\ \alpha^T \Sigma \alpha = var(\alpha^T x) \geq 0$, то $\Sigma \succeq 0$.
- Выборочная ковариационная матрица $\widehat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T = \frac{1}{N}X^T X \succeq 0$ $(X = [x_1^T, ...x_N^T]^T \in \mathbb{R}^{N \times D})$ т.к.

$$\alpha X^T X \alpha = (X\alpha)^T (X\alpha) = \|X\alpha\|^2 \geq 0$$

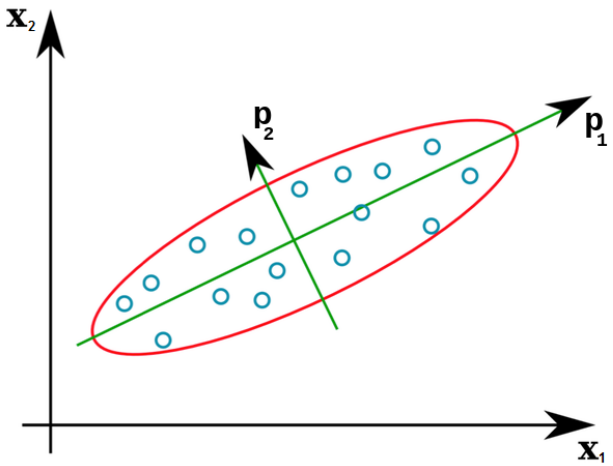
Оценка разброса распределения

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

$$\text{var}(\alpha^T x) = \alpha^T \Sigma \alpha = \alpha^T P \Lambda P^T \alpha =$$

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

- $\alpha \to$в базис СВ, координаты масштабируются $\sqrt{\lambda_1}, ...\sqrt{\lambda_D}$.



- Направления, отвечающие максимальному изменению данных - СВ Σ, отвечающие максимальным СЗ.

Оценка разброса распределения

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

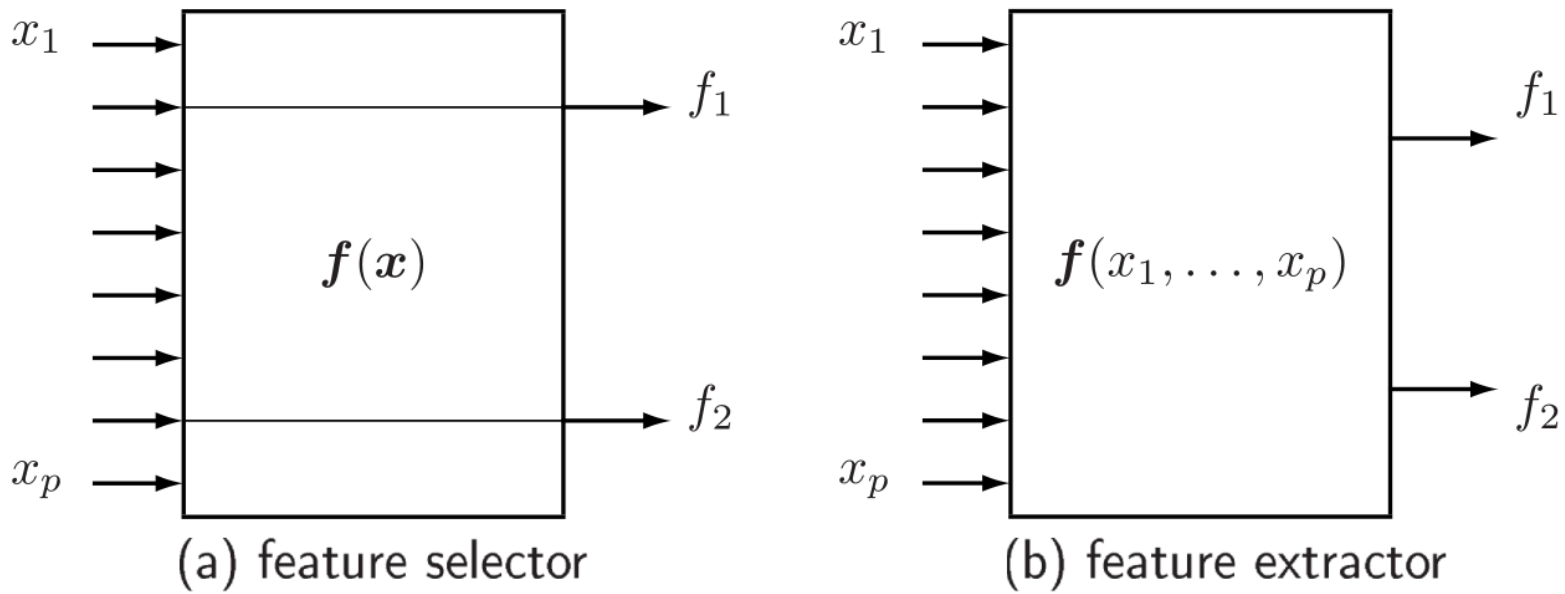
- используя инвариантность tr и det к смене базиса

$$\frac{1}{D}(\lambda_1 + ... + \lambda_D) = \frac{1}{D}\text{trace}\,\Lambda = \frac{1}{D}\text{trace}\,P\Lambda P^T = \frac{1}{D}\text{trace}\,\Sigma$$

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

Содержание

Задача снижения размерности



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

Использование откликов:

- снижение размерности с учителем (по X, Y)
- снижение размерности без учителя (по X)

Преобразование признаков:

- линейное
- нелинейное

Метод главных компонент - линейный метод снижения размерности без учителя.

Содержание

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].

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

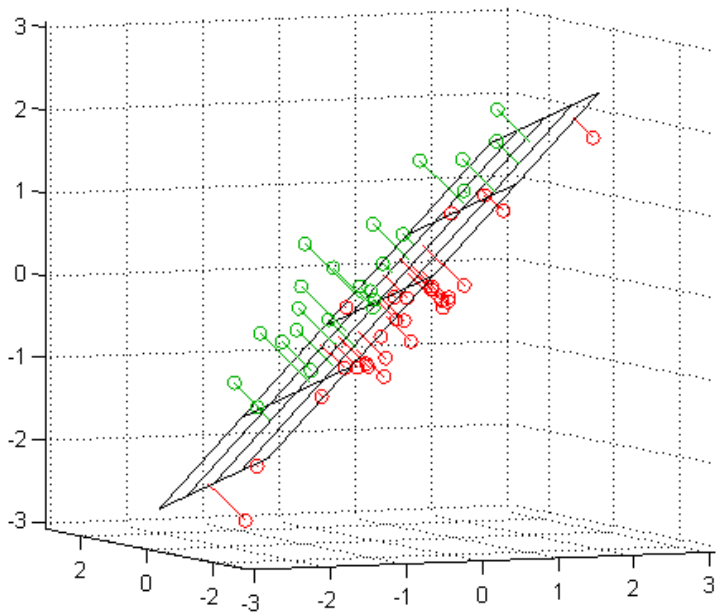
[a] Докажите, используя $\|x\|^2 = \|p\|^2 + \|h\|^2$ для $x = p + h$ и $\langle p, h \rangle = 0$.

Главные компоненты (principal components)

- 1ая главная компонента a_1 : $L_1 = \mathcal{L}(a_1),\ \|a_1\| = 1$
- 2ая главная компонента
  a_2 : $L_2 = \mathcal{L}(a_1, a_2),\ \|a_2\| = 1, \langle a_1, a_2\rangle = 0$
- D-я главная компонента a_D : $L_3 = \mathcal{L}(a_1, a_2, ...a_D),\ \|a_D\| = 1, \langle a_1, a_i\rangle = 0,\ i = 1, 2, ...D-1$
- Метод главных компонент (principal component analysis): нахождение разложения в первых K гл. компонентах для всех объектов:

$$x = \alpha_1 a_1 + \alpha_2 a_2 + ... + \alpha_K a_K$$

Главные компоненты



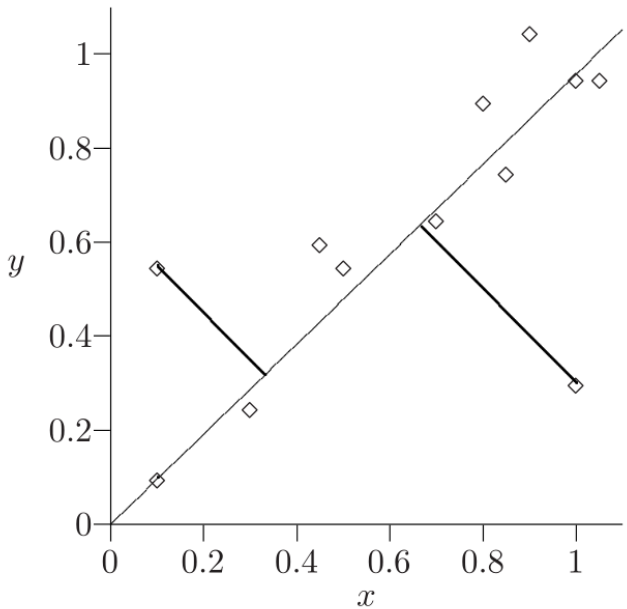
- На практике главные компоненты находятся из сингулярного разложения матрицы X.

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 делается для текстовых данных:
    - X - разреженная, поэтому уже $\bar{x}_i \approx 0$. Сдвиг сделает X не разреженной.
    - если признаки - индикаторы встречаемости или частоты слов, они уже в единой шкале $[0, 1]$.

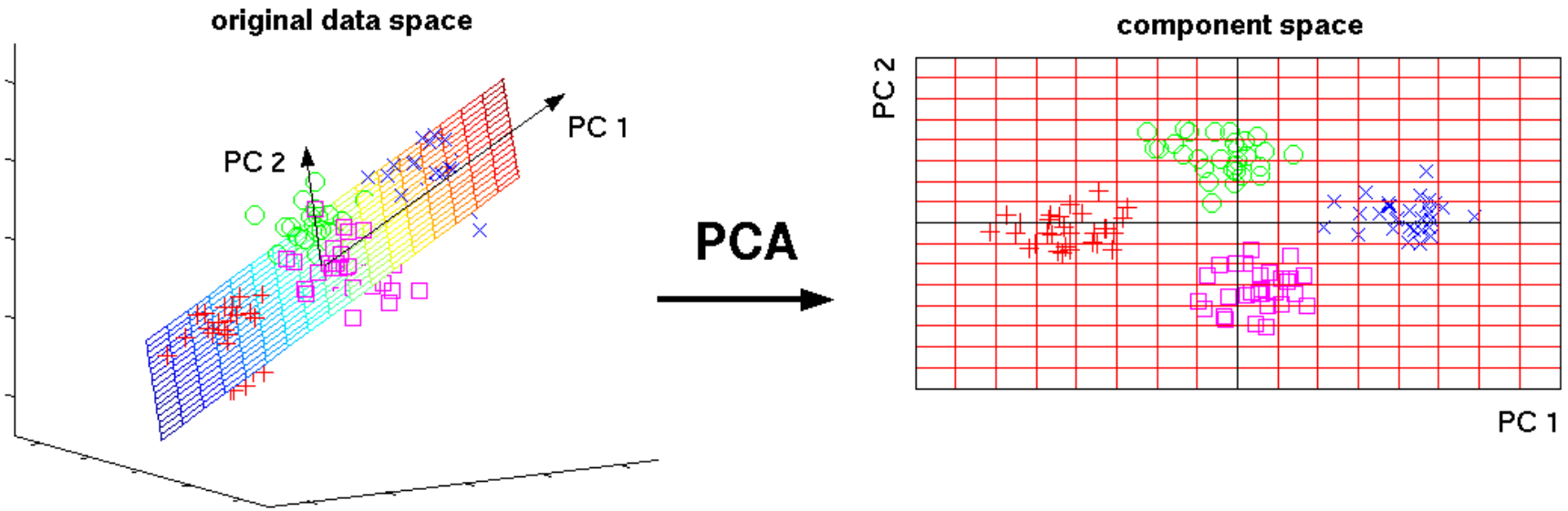
Пример L_1

- Рассмотрим одномерное подпространство наилучшей аппроксимации L_1:



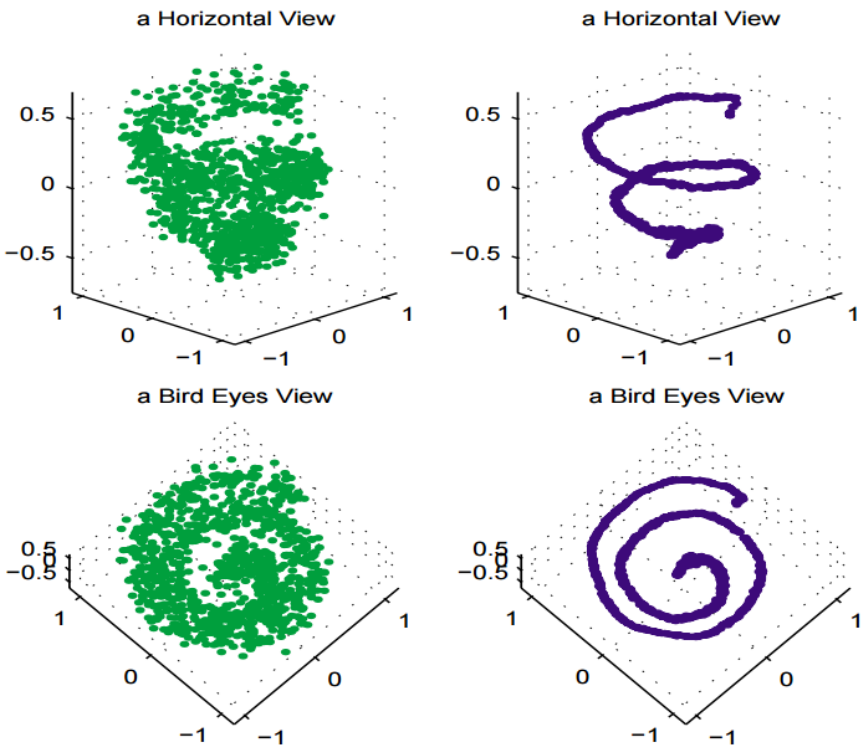
- В чем отличие от нахождения $y = wx$ в линейной регрессии?

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)

Оценка качества аппроксимации

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

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

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.

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

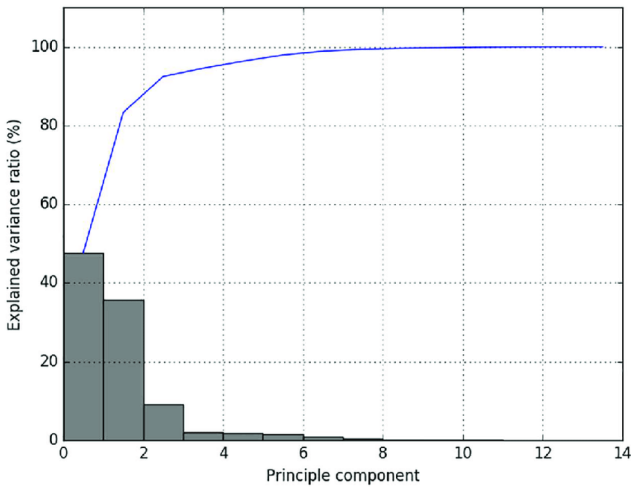
Вклад a_k в описание x: $\langle x, a_k\rangle^2$.
Вклад a_k в описание $x_1, x_2, ...x_N$: $\sum_{n=1}^N \langle x_n, a_k\rangle^2$
Относительный вклад (explained variance ratio):

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

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

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

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

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

Расчет p^K по x

Если y - вектор проекций x на $a_1, ...a_D$, то

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

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

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

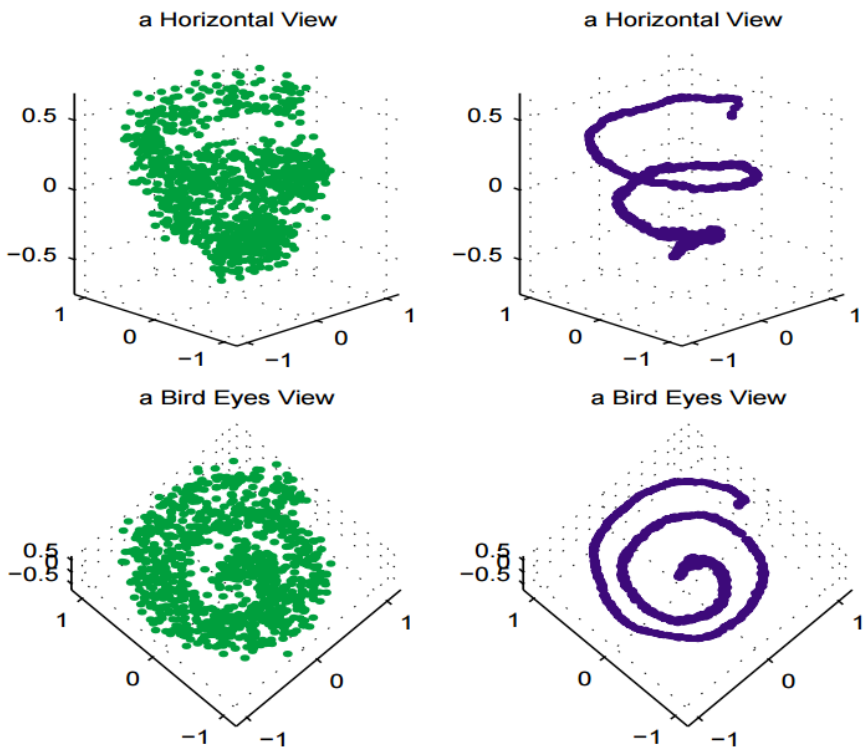
$$y^K = A_K^T(x - \mu)$$

Проекция p^K для x на L_K:

$$p^K = A\begin{pmatrix} y^K \\ 0 \end{pmatrix} + \mu = A_K y^K + \mu$$

$$p^K = A_K A_K^T(x - \mu) + \mu, \quad \text{rg}\left[A_K A_K^T\right] = \text{rg}\,[A_K] = K\ \forall A_K$$

Метод локальной линейной проекции[2]



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

Метод локальной линейной проекции

Метод локальной линейной проекции удаляет шум в данных за счет замены $x_n \to p_n$ (на гиперплоскость, локально описывающую данные).

ВХОД:

K-локальная размерность данных

M-число ближайших соседей

для каждого x_i в X:

найти M ближайших соседей x_i.

определить $a_1, ...a_K$ по центрированным ближайшим соседям.

составить μ и A_K

$x_i \to p_i = A_K A_K^T (x - \mu) + \mu$

ВЫХОД:

данные, очищенные от шума $p_1, p_2, ...p_K$.

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 $\langle a, a\rangle = 1, \langle a, a_1\rangle = 0, \langle a, a_2\rangle = 0$
- ... ...
- $a_D = \arg\max_a \|Xa\|^2$, 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[3]

- 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 $B \in \mathbb{R}^{DxD}$:

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

[3] 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

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

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:

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

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

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 $\|Xa_1\|^2 \to \max_{a_1}$ 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

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

Лагранжиан оптимизационной задачи (5):

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

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

Вычисление k-й главной компоненты

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

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

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

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 2

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 уменьшенное признаковое пространство.
- Полезно для повышения точности, интерпретируемости и скорости работы моделей.
- Метод главных компонент - метод линейного снижения размерности без учителя.
  - центрируем признаки и приводим их к единой шкале
  - вычисляем выборочную ковариационную матрицу $\widehat{\Sigma} = \frac{1}{N} X^T X$
  - определяем СВ $a_1, a_2, ...a_D$, отвечающие СЗ $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$
  - $\mathcal{L}(a_1, a_2, ...a_K)$ - подпространство наилучшей аппроксимации ранга K:

$$\|h_1\|^2 + ... + \|h_N\|^2 \to \min_{b_1, ...b_k}$$

  - $x = \alpha_1 a_1 + ... + \alpha_D a_D$: $\left(x^1, ...x^D\right) \to \left(\alpha^1, ...\alpha^K\right)$