

Уменьшение эффективной размерности пространства

Алексей Романенко , alexromsput@gmail.com,
Анастасия Зухба, a__l@mail.ru

май 2018

Задачи обучения без учителя

Задача кластеризации:

X^ℓ — признаковое описание объектов Y — отсутствует
 $f_j: X \rightarrow D_j, j = 1, \dots, n$ — признаки объектов (features).

Типы признаков:

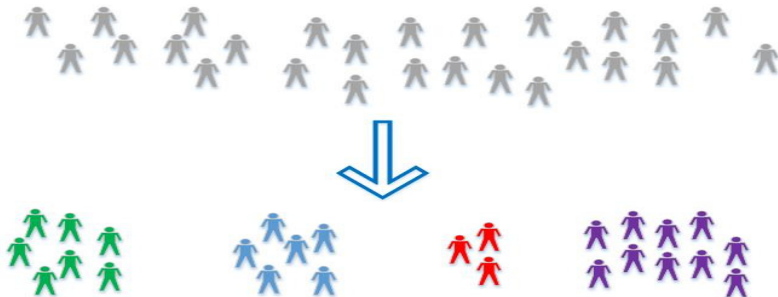
- $D_j = \{0, 1\}$ — бинарный признак f_j ;
- $|D_j| < \infty$ — номинальный признак f_j ;
- $|D_j| < \infty, D_j$ упорядочено — порядковый признак f_j ;
- $D_j = \mathbb{R}$ — количественный признак f_j .

Вектор $(f_1(x), \dots, f_n(x))$ — признаковое описание объекта x .

Матрица «объекты–признаки» (features data)

$$F = \left\| f_j(x_i) \right\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

Задачи обучения без учителя

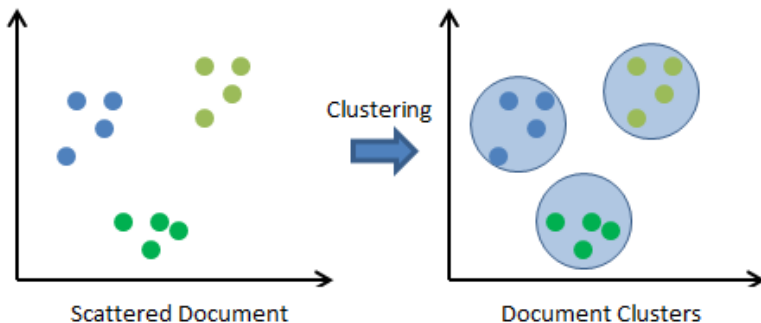


Типы задач без учителя:

- Кластеризация
- Поиск правил ассоциации
- Сокращение размерности
- Визуализация данных

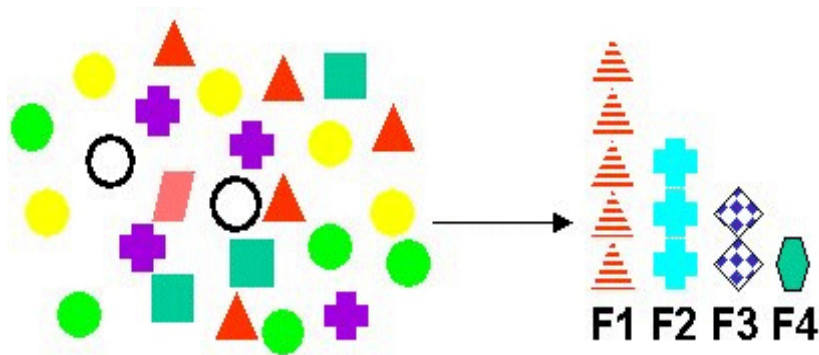
Примеры задач

Кластеризация



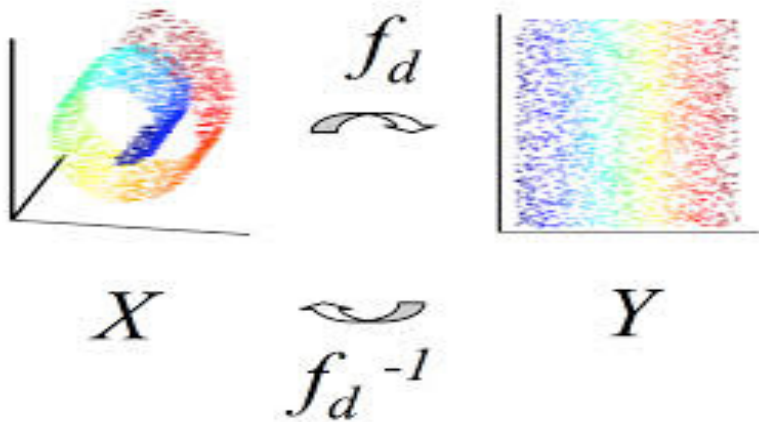
Примеры задач

Feature Extraction



Примеры задач

Снижение размерности



Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Мультиколлинеарность



Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Проклятие размерности



Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Неинтерпретируемость



Многомерная линейная регрессия

$f_1(x), \dots, f_n(x)$ — числовые признаки;

Модель многомерной линейной регрессии:

$$f(x, \alpha) = \sum_{j=1}^n \alpha_j f_j(x), \quad \alpha \in \mathbb{R}^n.$$

Матричные обозначения:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y_{\ell \times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \alpha_{n \times 1} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}.$$

Функционал квадрата ошибки:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2 = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}.$$

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^T(F\alpha - y) = 0,$$

откуда следует *нормальная система* задачи МНК:

$$F^T F \alpha = F^T y,$$

где $F^T F_{n \times n}$ — ковариационная матрица набора признаков f_1, \dots, f_n .

Решение системы: $\alpha^* = (F^T F)^{-1} F^T y = F^+ y$.

Значение функционала: $Q(\alpha^*) = \|P_F y - y\|^2$,

где $P_F = F F^+ = F(F^T F)^{-1} F^T$ — проекционная матрица.

Сингулярное разложение

Произвольная $\ell \times n$ -матрица представима в виде
сингулярного разложения (singular value decomposition, SVD):

$$F = VDU^T.$$

Основные свойства сингулярного разложения:

- 1 $\ell \times n$ -матрица $V = (v_1, \dots, v_n)$ ортогональна, $V^T V = I_n$, столбцы v_j — собственные векторы матрицы FF^T ;
- 2 $n \times n$ -матрица $U = (u_1, \dots, u_n)$ ортогональна, $U^T U = I_n$, столбцы u_j — собственные векторы матрицы $F^T F$;
- 3 $n \times n$ -матрица D диагональна, $D = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$, $\lambda_j \geq 0$ — собственные значения матриц $F^T F$ и FF^T .

Решение МНК через сингулярное разложение

Псевдообратная F^+ , вектор МНК-решения α^* ,
МНК-аппроксимация целевого вектора $F\alpha^*$:

$$F^+ = (UDV^T VDU^T)^{-1}UDV^T = UD^{-1}V^T = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j v_j^T;$$

$$\alpha^* = F^+ y = UD^{-1}V^T y = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y);$$

$$F\alpha^* = P_F y = (VDU^T)UD^{-1}V^T y = VV^T y = \sum_{j=1}^n v_j (v_j^T y);$$

$$\|\alpha^*\|^2 = \|D^{-1}V^T y\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^T y)^2.$$

Проблема мультиколлинеарности

Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- мультиколлинеарность влечёт *переобучение*.

Проблема мультиколлинеарности

Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

Проблема мультиколлинеарности

Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \rightarrow \min$;

Проблема мультиколлинеарности

Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \rightarrow \min$;
- Отбор признаков: $f_1, \dots, f_n \rightarrow f_{j_1}, \dots, f_{j_m}, \quad m \ll n$.

Проблема мультиколлинеарности

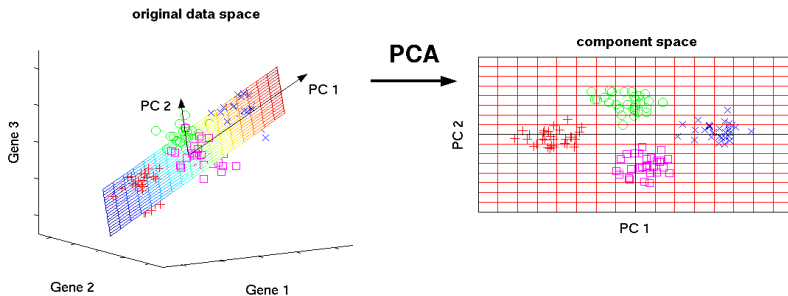
Если имеются $\lambda_j \rightarrow 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо:
 $\|\alpha\| \rightarrow \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»:
 $Q(\alpha^*) = \|F\alpha^* - y\|^2 \rightarrow 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \rightarrow \min$;
- Отбор признаков: $f_1, \dots, f_n \rightarrow f_{j_1}, \dots, f_{j_m}, \quad m \ll n$.
- Преобразование признаков: $f_1, \dots, f_n \rightarrow g_1, \dots, g_m, \quad m \ll n$;

Метод главных компонент: постановка задачи



PCA visalisation:

<http://setosa.io/ev/principal-component-analysis/>

Метод главных компонент: постановка задачи

$f_1(x), \dots, f_n(x)$ — исходные числовые признаки;

$g_1(x), \dots, g_m(x)$ — новые числовые признаки, $m \leq n$;

Требование: старые признаки должны линейно
восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^m g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке x_1, \dots, x_ℓ :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 \rightarrow \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad G_{\ell \times m} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$U_{n \times m} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^T \overset{\text{ХОТИМ}}{\approx} F.$$

Найти: и новые признаки G , и преобразование U :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\|^2 \rightarrow \min_{G, U},$$

Основная теорема метода главных компонент

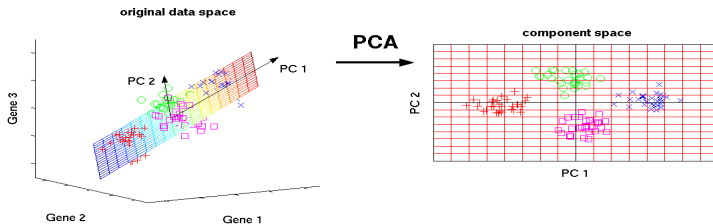
Теорема

Если $m \leq \text{rk } F$, то минимум $\|GU^T - F\|^2$ достигается, когда столбцы U — это с.в. матрицы $F^T F$, соответствующие m максимальным с.з. $\lambda_1, \dots, \lambda_m$, а матрица $G = FU$.

При этом:

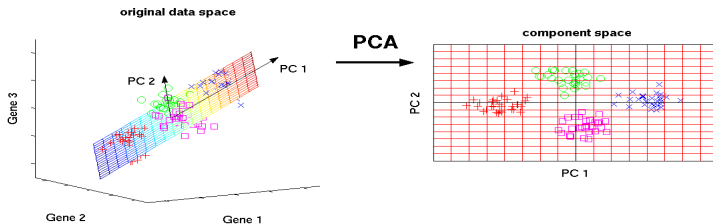
- ① матрица U ортонормирована: $U^T U = I_m$;
- ② матрица G ортогональна: $G^T G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$;
- ③ $U\Lambda = F^T F U$; $G\Lambda = FF^T G$;
- ④ $\|GU^T - F\|^2 = \|F\|^2 - \text{tr } \Lambda = \sum_{j=m+1}^n \lambda_j$.

Три интерпретации метода PCA



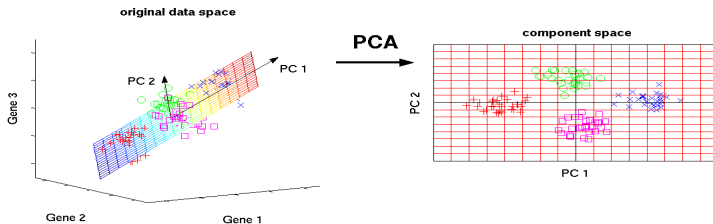
- 1 найти подпространства меньшей размерности, в ортогональной проекции на которые разброс данных (то есть среднеквадратичное отклонение от среднего значения) максимален;

Три интерпретации метода PCA



- 1 ...
- 2 найти подпространства меньшей размерности, в ортогональной проекции на которые среднееквadraticное расстояние между точками максимально;

Три интерпретации метода PCA



1 ...

2 ...

3 построить такое ортогональное преобразование координат, в результате которого корреляции между отдельными координатами обратятся в нуль.

Связь с сингулярным разложением

Если взять $m = n$, то:

- 1 $\|GU^T - F\|^2 = 0$;
- 2 представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$:
$$F = GU^T = V\sqrt{\Lambda}U^T; \quad U^T U = I_m; \quad V^T V = I_m.$$
- 3 линейное преобразование U работает в обе стороны:

$$F = GU^T; \quad G = FU.$$

Поскольку новые признаки некоррелированы ($G^T G = \Lambda$), преобразование U называется *декоррелирующим* (или преобразованием Карунена–Лоэва).

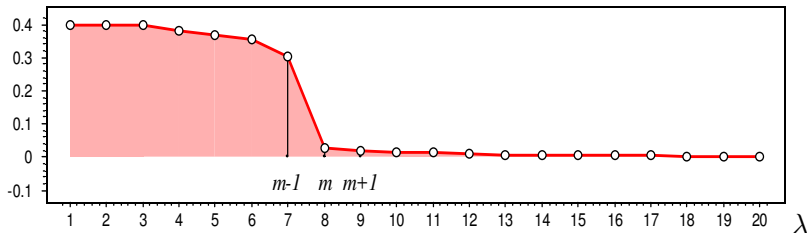
Эффективная размерность выборки

Упорядочим с.з. $F^T F$ по убыванию: $\lambda_1 \geq \dots \geq \lambda_n \geq 0$.

Эффективная размерность выборки — это наименьшее целое m , при котором

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon.$$

Критерий «крутого склона»: находим m : $E_{m-1} \gg E_m$:



Решение задачи НК в новых признаках

Заменяем F на её приближение GU^T :

$$\|G \underbrace{U^T \alpha}_{\beta} - y\|^2 = \|G\beta - y\|^2 \rightarrow \min_{\beta}.$$

Связь нового и старого вектора коэффициентов:

$$\alpha = U\beta; \quad \beta = U^T \alpha.$$

Решение задачи наименьших квадратов относительно β
(единственное отличие — m слагаемых вместо n):

$$\beta^* = D^{-1} V^T y = \sum_{j=1}^m \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y);$$

$$G\beta^* = VV^T y = \sum_{j=1}^m v_j (v_j^T y);$$

PCA, SVD

- PCA помогает избавиться от "лишних" размерностей
- PCA строит оптимальное линейное преобразование
- PCA применим для больших размерностей
- существуют ядерные варианты PCA (KernelPCA)
- если требуется сложное нелинейное преобразование, PCA не работает
- не подходит для визуализации на двухмерной плоскости при большой размерности исходных пространств

t-SNE. Шаг первый.

x_1, \dots, x_N – точки (data points) в исходном пространстве \mathbb{R}^D .

z_1, \dots, z_N – точки (map points) в пространстве \mathbb{R}^2 .

Построение плотностей в изначальном пространстве:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Это выражение показывает, насколько точка x_j близка к x_i , при гауссовом распределении вокруг x_i с заданной дисперсией σ_i^2 .

Дисперсия различна для каждой точки.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

t-SNE. Шаг второй.

Построение приближения в пространстве меньшей размерности:

$$q_{ij} = \frac{(1 + \|z_i - z_j\|^2)^{-1}}{\sum_{k \neq m} (1 + \|z_k - z_m\|^2)^{-1}}$$

Минимизируется расстояния Кульбака-Лейблера:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

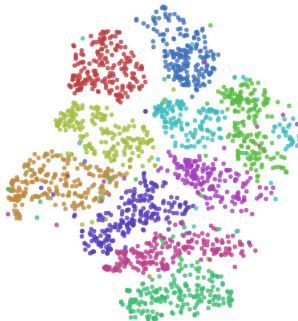
Градиент может быть вычислен аналитически:

$$\frac{\partial KL(P||Q)}{\partial z_i} = 4 \sum_j (p_{ij} - q_{ij}) g(\|x_i - x_j\|) u_{ij}$$

где $g(z) = \frac{z}{1+z^2}$; u_{ij} – единичный вектор от z_j к z_i .

tSNE for MNIST

MNIST with tSNE



A t-SNE plot of MNIST

Demo see here

<http://colah.github.io/posts/2014-10-Visualizing-MNIST/>

Multi-dimensional Scaling

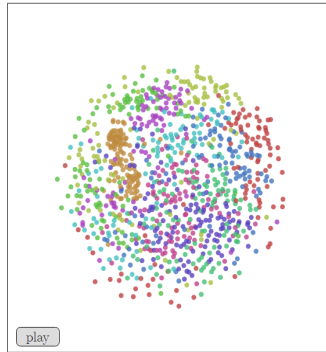
Многомерное масштабирование (Multidimensional scaling, MDS) пытается моделировать сходство данных как расстояния в геометрических пространствах.

MDS ищет для данных представление низкой размерности максимально учитывая расстояния в изначальном многомерном пространстве.

- 1 **Метрический вариант:** приближает или сохраняет расстояния.
- 2 **Неметрический вариант:** приближает или сохраняет порядок расстояний.

MDS for MNIST

MNIST with MDS



Visualizing MNIST with MDS

See demo here

<http://colah.github.io/posts/2014-10-Visualizing-MNIST/>

Заклучение по tSNE

PRO tSNE:

- отлично визуализируют (2D, 3D) данные
- используются для предобработки данных, анализа структуры данных see here
<http://distill.pub/2016/misread-tsne/>
-

Cons tSNE:

- стохастичность (PCA детерминирован)
- интерпретируемость данных
- не применим (сложно применим) для новых данных
- не даёт ответ об оптимальной размерности пространства