Лекция 7
Снижение размерностей

Габдуллин Р.А., Макаренко В.А.

МГУ им. М.В. Ломоносова

21 февраля 2021

Снижение размерности данных

Цель:

- Построить меньшее количество признаков на основе исходных, сохранив при этом максимум информации.

Для чего это нужно:

- Ускоряем процесс обучения/инференса.
- Используем меньше памяти для хранения признаков.
- Боремся с проклятием размерности.
- Боремся с мультиколлинеарностью.
- Можем визуализировать данные.

Метод главных компонент (Principal component analysis)

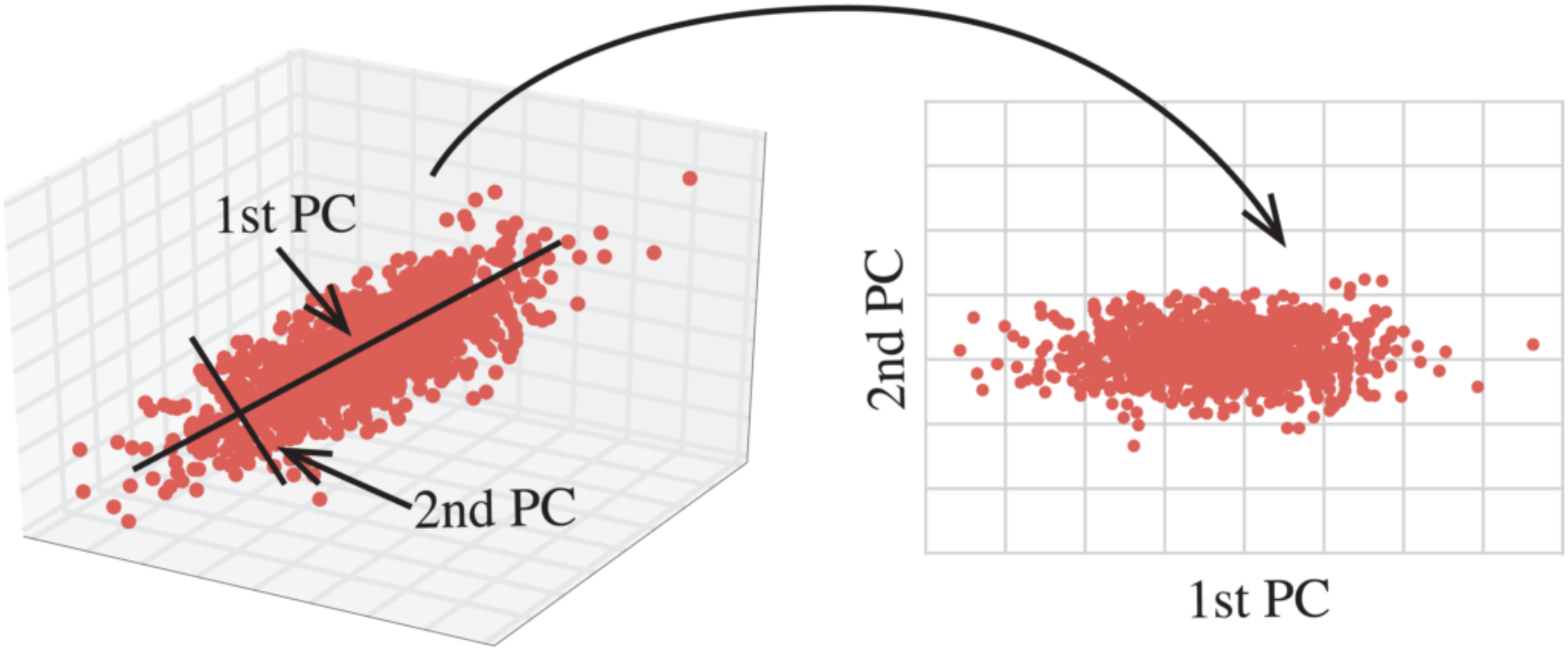


Рис.: Источник: medium.com

Метод главных компонент. Постановка задачи

Исходное признаковое описание объекта $x \in X$:

$$f_1(x), f_2(x), \ldots, f_n(x).$$

Новое признаковое описание объекта $x \in X$:

$$g_1(x), g_2(x), \ldots, g_m(x).$$

Старое представление (естественный базис):

$$f(x) = f(x) I_{n \times n}.$$

Новое представление (новый базис):

$$f(x) \approx g(x) U^T.$$

Требования:

- $m \leqslant n$.
- $\hat{f}_j(x) = \sum\limits_{k=1}^{m} g_k(x) u_{j,k}, \quad 1 \leqslant j \leqslant n, \quad x \in X.$
- $\sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{n} (f_j(x_i) - \hat{f}_j(x_i))^2 \to \min\limits_{\{g_k(x_i)\},\{u_{j,k}\}}$

Примеры. Главные компоненты Olivetti faces

Метод главных компонент. Постановка задачи

Матрицы признаков (старая и новая):

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell\times m}{G} = \begin{pmatrix} g_1(x_1) & \ldots & g_m(x_1) \\ \ldots & \ldots & \ldots \\ g_1(x_\ell) & \ldots & g_m(x_\ell) \end{pmatrix},$$

Матрица линейного преобразования:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{1,1} & \ldots & u_{1,m} \\ \ldots & \ldots & \ldots \\ u_{n,1} & \ldots & u_{n,m} \end{pmatrix}, \quad \hat{F} = GU^T.$$

Цель:

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

Основная теорема метода главных компонент

Теорема

Если $m \leqslant \operatorname{rank} F$, *то минимум* $\|F - GU^T\|^2$ *достигается, когда столбцы* U *– это собственные векторы матрицы* $F^T F$, *соответствующие* m *максимальным собственным значениям* $\lambda_1, \ldots, \lambda_m$, *а матрица* $G = FU$.

При этом:

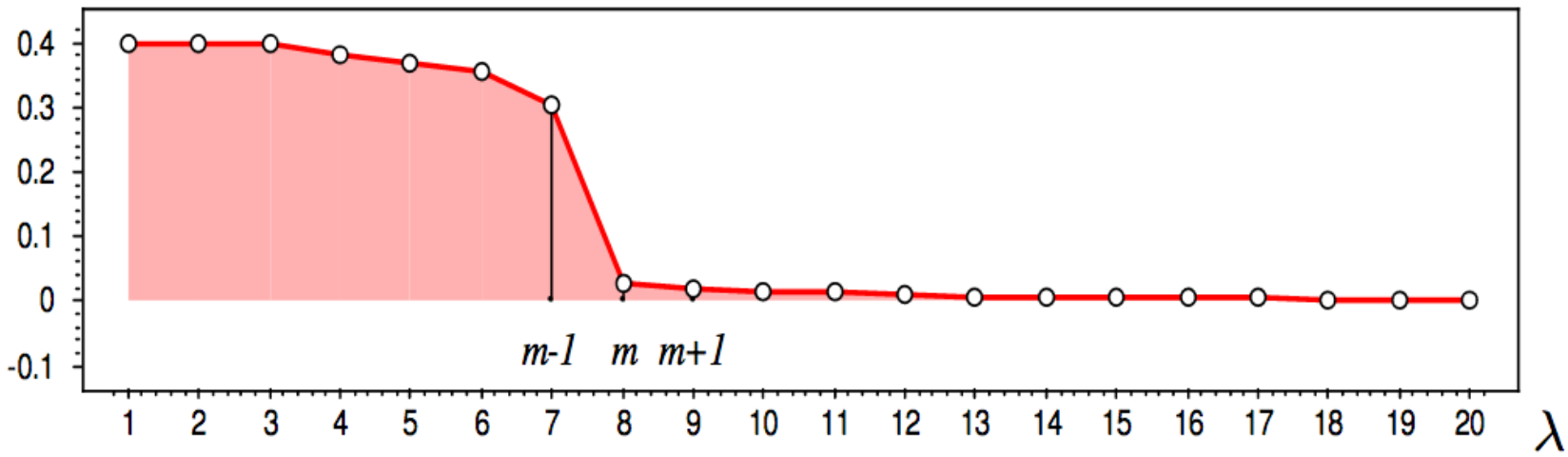
- *Матрица* U *ортогональна:* $U^T U = I_m$.
- *Матрица* G *такова, что* $G^T G = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$.
- $U\Lambda = F^T FU, \quad G\Lambda = FF^T G$.
- $\|GU^T - F\|^2 = \|F\|^2 - \operatorname{tr}\Lambda = \sum\limits_{j=m+1}^{n} \lambda_j$.

Выбор количества новых признаков

Упорядочиваем с.з. F^TF по убыванию: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$.

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \ldots + \lambda_n}{\lambda_1 + \lambda_2 + \ldots + \lambda_n}.$$

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



Сингулярное разложение (SVD)

Любая вещественная матрица $\underset{n \times d}{A}$ разложима следующим образом:

$$A = U \Sigma V^T,$$

где $\underset{n \times n}{U}$, $\underset{d \times d}{V}$ – ортогональные матрицы, $\underset{n \times d}{\Sigma}$ – диагональная матрица.

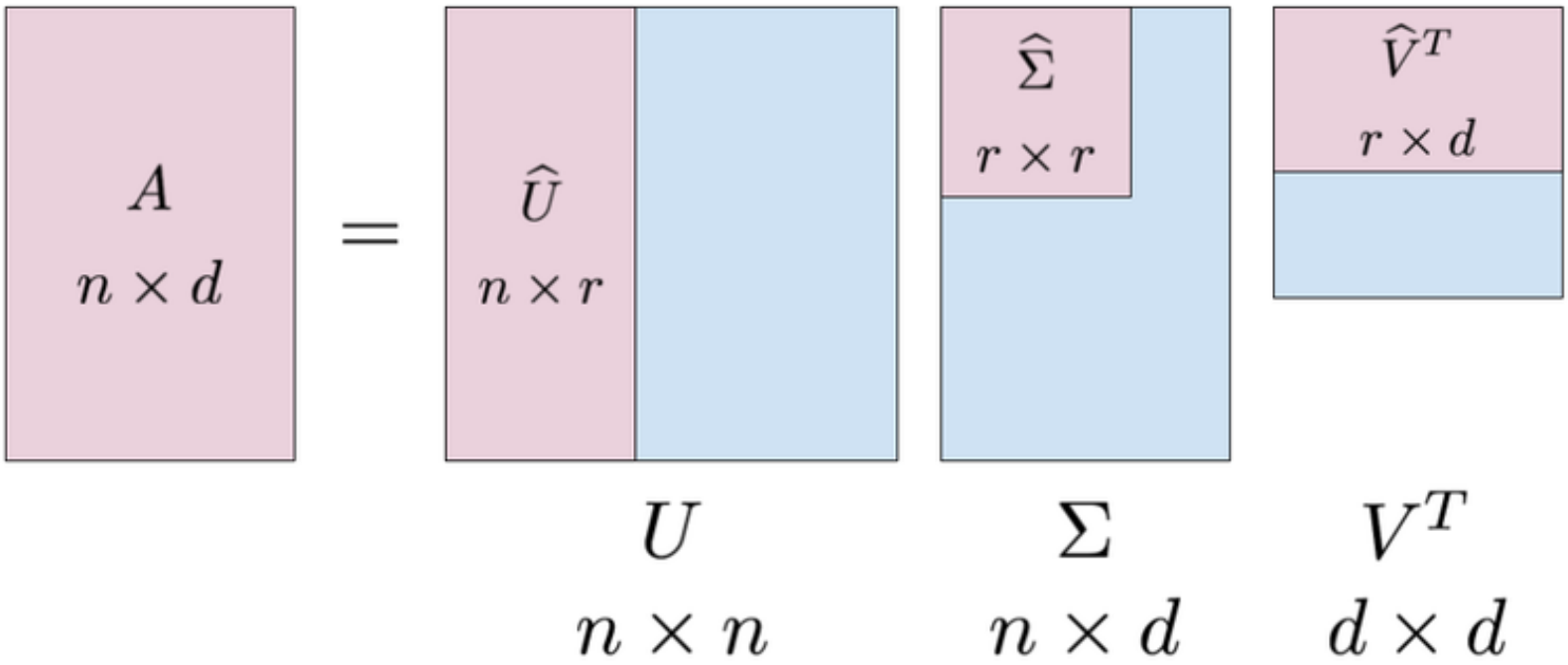


Рис.: Источник: intoli.com

PCA и SVD

Сингулярное разложение:

$$F = VDU^T.$$

Если взять $m = n$, то

- $\|GU^T - F\| = 0$.
- Представление $\hat{F} = GU^T = F$ точное и совпадает с сингулярным разложением при $G = VD, \Lambda = D^2$:

$$F = GU^T = V\sqrt{\Lambda}U^T; \quad U^TU = I_m; \quad V^TV = I_m$$

- Линейное преобразование U работает в обе стороны:

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

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

Нелинейные методы снижения размерности

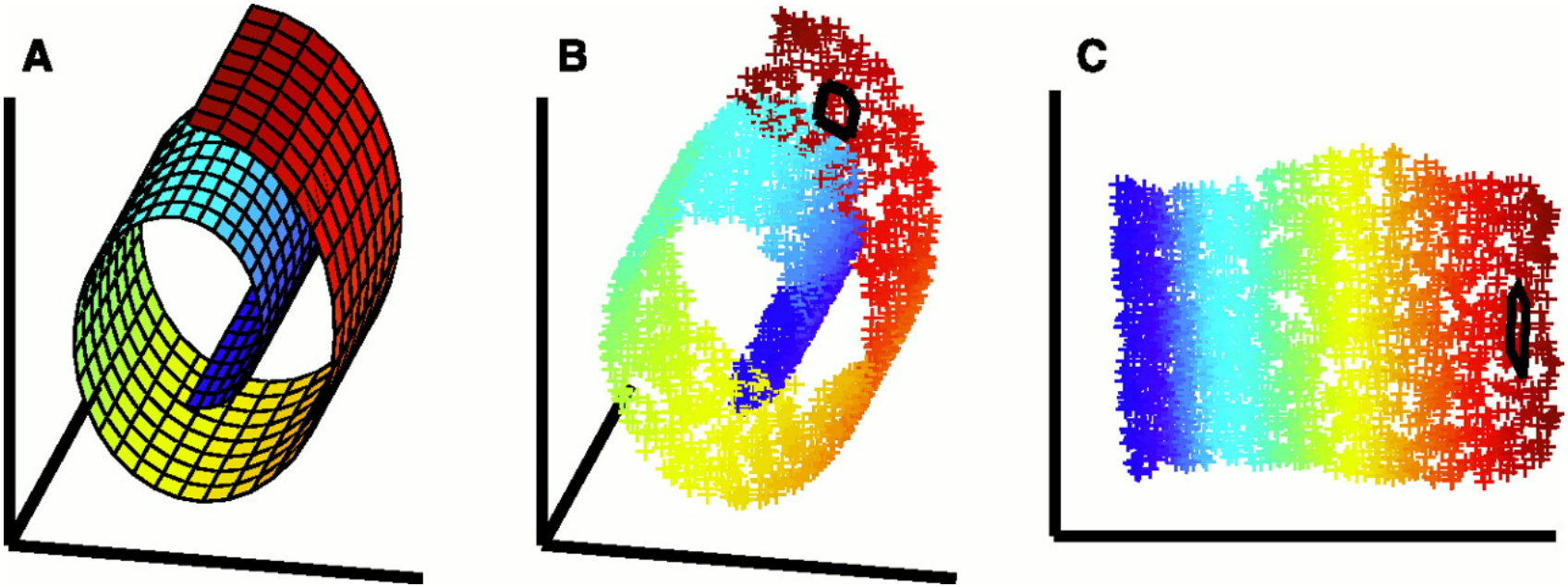


Рис.: Источник: science.sciencemag.org

Multidimensional scaling (MDS)

Цель: при снижении размерности сохранить попарные расстояния $d_{i,j}$ между объектами

Исходное признаковое описание объектов:

$$x_1, x_2, \ldots, x_\ell.$$

Новое признаковое описание объектов:

$$\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_\ell.$$

Задача оптимизации:

$$\sum_{i<j}^{\ell} (\|\widetilde{x}_i - \widetilde{x}_j\| - d_{i,j})^2 \to \min.$$

t-distributed stochastic neighbor embedding (t-SNE)

Условная вероятность соседства двух точек в исходном признаковом пространстве:

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

Вероятность в исходном пространстве признаков:

$$p_{i,j} = \frac{p_{i|j} + p_{j|i}}{2\ell}.$$

В новом пространстве:

$$q_{i,j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum\limits_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

Задача оптимизации:

$$\sum_i \mathrm{KL}(P_i||Q_i) = \sum_i \sum_{j \neq i} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j}}\right) \to \min_{y_1,\ldots,y_\ell}.$$