Матричные разложения и их применения (SVD)

Гайдамашко Даниил

БПМИ161

02.11.2018

О матричных разложениях

Разложение матрицы – ее представление в виде произведения [«хороших»] матриц:

- Примеры «хороших» свойств – диагональность, симметричность, ортогональность;
- Бонусы – эффективное хранение матрицы, оптимизация вычислений, выявление полезной информации в данных;

Определение

Сингулярное разложение матрицы (SVD) – разложение вещественной матрицы $A_{m\times n},\ \mathrm{rk}\,(A) = r$:

$$A_{m\times n} = U_{m\times m} \cdot \Sigma_{m\times n} \cdot V^T_{n\times n}$$

- $U,\ V$ – ортогональные $(U^{-1} = U^T)$;
- Σ – диагональная;
- $\sigma_1 \geqslant \cdots \geqslant \sigma_r > \sigma_{r+1} = \cdots = \sigma_{\min(m,n)} = 0$ – сингулярные значения;
- Усеченное разложение:

$$A_{m\times n} = U_{m\times r} \cdot \Sigma_{r\times r} \cdot V^T_{r\times n}$$

Интуиция из линейной алгебры

1. *Откуда берется?*
   Теорема о сингулярном базисе
   $\Rightarrow \exists U, V$ – ортогональные, т.ч. $U^{-1}AV = \Sigma \Rightarrow A = U\Sigma V^T$;
2. *Распределение ролей?*
   $AA^T = U\Sigma\Sigma^T U^T \Rightarrow AA^T U = U\Sigma\Sigma^T$
   $A^T A = V\Sigma^T\Sigma V^T \Rightarrow A^T AV = V\Sigma^T\Sigma$
   - Квадраты сингулярных значений – собственные значения для AA^T и $A^T A$;
   - Столбцы U – собственные векторы для AA^T, столбцы V – для $A^T A$;
   - Геометрическая интерпретация:

$$\underbrace{A}_{\mathbb{R}^m \to \mathbb{R}^n} = \underbrace{U}_{\text{Поворот}} \times \underbrace{\Sigma}_{\text{Растяжение}} \times \underbrace{V^T}_{\text{Поворот}}$$

Алгоритм нахождения

Наивный алгоритм:

1. Находим спектр $A^T A$;
2. Находим собственные отнормированные векторы (столбцы V);
3. Достраиваем до ортонормированного базиса, если не хватило;
4. Находим $u_i = Av_i/\sigma_i$, т.к. $AV = U\Sigma$;
5. Достраиваем до ортонормированного базиса, если не хватило;

На практике применяются аппроксимирующие алгоритмы.

Псевдообратные матрицы и МНК

Задача наименьших квадратов:

- $Aw = b$, $||Aw - b||_2 \to \min\limits_{w}$;
- Решение: $w = (A^T A)^{-1} A^T b$ – не всегда вычисляется;

Псевдообратная матрица:

- Аксиомы: $AA^+A = A$, $A^+AA^+ = A^+$, $(AA^+)^T = AA^+$, $(A^+A)^T = A^+A$;
- $A^+ = (A^T A)^+ A^T \Rightarrow w = A^+ b$
- Если $A = U\Sigma V^T$, то $A^+ = V^T \Sigma^+ U$

Приближение матрицы по фробениусовой норме

- *Дано*: матрица $A_{m\times n}$, число $k \in \mathbb{N}$;
- *Найти*: матрица $A_k,\ \mathrm{rk}\,(A_k) = k\ :\ ||A - A_k||_F \to \min\limits_{A_k}$;
- *Решение*: если $A = U \cdot \Sigma \cdot V^T$, то зануляем все сингулярные числа, кроме первых k в Σ:

$$A_k = U \cdot \hat{\Sigma}_k \cdot V^T$$

- *Решение*: усеченное разложение ранга k:

$$A_k = U_k \cdot \Sigma_k \cdot V_k^T$$

- *Применение*: сжатие данных, обработка сигналов и др.
- *Интуиция*:

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^T$$

Сжатие изображений (grayscale)

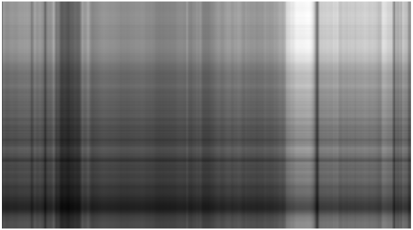


Рис.: $k = 1$

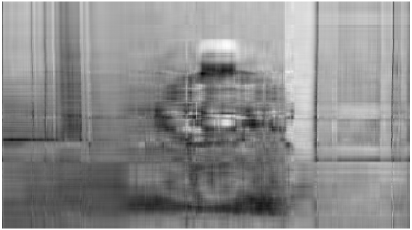


Рис.: $k = 10$

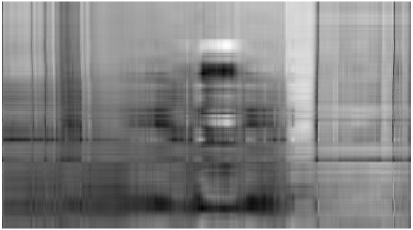


Рис.: $k = 5$

Рис.: $k > 100$

Сжатие изображений (RGB)

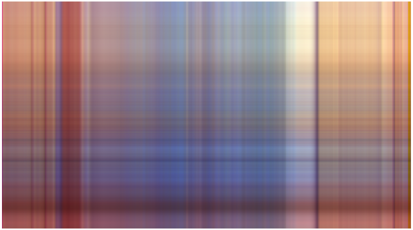


Рис.: $k = 1$

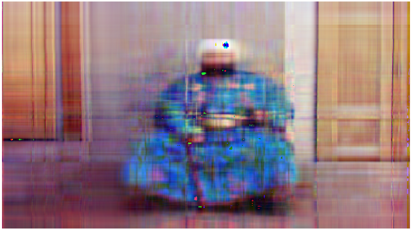


Рис.: $k = 10$

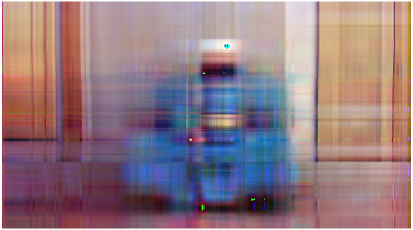


Рис.: $k = 5$

Рис.: $k > 100$

Другой взгляд на матричные разложения

- *Дано*: выборка n объектов с m признаками;
- *Хотим*:найти базис $h_1, \ldots, h_k \in \mathbb{R}^m$, т.ч. $x_i = \sum_{j=1}^{k} w_{ij} h_j$;
- В матричном виде: $X_{n \times m} \approx W_{n \times k} \cdot H_{k \times m}$,
  $||X - WH||_F \to \min$;
- W – новые признаки, H – важные объекты;
- SVD дает точные решения
  $\{U_k \Sigma_k, V_k^T\}, \{U_k \sqrt{\Sigma_k}, \sqrt{\Sigma_k} V_k^T\}, \{U_k, \Sigma_k V_k^T\}$;

Тематическое моделирование

- *Конкретные задачи*: классификация/кластеризация данных;
- *Хотим*: построить коллекцию документов, которая определяет тему каждого документа и слова, образующие данную тему;
- *Как*: ищем взаимосвязь между *текстами* и встречающимися в них *терминами*;

Латентно-семантический анализ (LSA)

- *Задача*: кластеризация данных, классификация текстов;
- *Смысл*: ищем взаимосвязь между *текстами* и встречающимися в них *терминами*;
- *Алгоритм*:
  1. Создаем частотную матрицу терм-текст
  2. Эмпирически подбираем подходящий k;
  3. U_k – матрица термов, V_k – матрица текстов;
  4. Отображаем тексты и термы в $\mathbb{R}^k$, близость характеризуется скалярным произведением;
- Есть PLSA, который больше основан на приемах мат. статистики;

Латентно-семантический анализ (LSA)

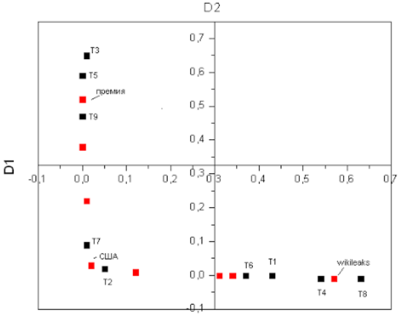
wikileaks	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	-0.64
арестова	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	0.01
великобритан	0.34	-0	0.07	0.41	-0.42	-0.02	0.1	0.17	-0.01
вручен	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.07
нобелевск	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	0.32
основател	0.57	-0.01	0.01	-0.2	0.13	0.16	-0.16	-0.25	0.64
полиц	0.31	-0	0.05	0.07	0.57	-0.6	0.29	0.37	-0
прем	0	0.52	0.07	-0.06	-0.08	-0.15	-0.17	0.02	-0.25
прот	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
стран	0.01	0.22	-0.31	0.39	0.41	0.56	-0.22	0.4	-0
суд	0.12	0.01	-0.38	-0.62	-0.3	0.12	0.21	0.55	-0
сша	0.02	0.03	-0.61	0.13	-0.05	-0.22	0	-0.25	0
церемон	0	0.38	0.03	0.02	0.08	0.31	0.82	-0.29	0

·

3.41	0	0	0	0	0	0	0	0
0	3.3	0	0	0	0	0	0	0
0	0	2.27	0	0	0	0	0	0
0	0	0	1.49	0	0	0	0	0
0	0	0	0	1.19	0	0	0	0
0	0	0	0	0	0.98	0	0	0
0	0	0	0	0	0	0.71	0	0
0	0	0	0	0	0	0	0.43	0
0	0	0	0	0	0	0	0	0

·

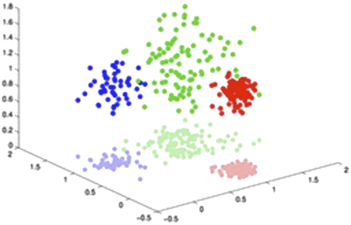
T1	T2	T3	T4	T5	T6	T7	T8	T9
0.43	0.05	0.01	0.54	0	0.37	0.01	0.63	0
-0	0.02	0.65	-0.01	0.59	-0	0.09	-0.01	0.47
0.03	-0.7	-0.04	0.06	0.1	-0.16	-0.67	0.09	0.09
-0.22	-0.24	0.15	0.28	-0.11	-0.68	0.44	0.33	-0.13
0.69	-0.32	0.22	-0.49	-0.12	-0.03	0.27	-0.02	-0.19
-0.27	-0.34	0.44	0.29	-0.13	0.45	0.12	-0.31	-0.45
-0.03	0.3	0.14	-0.17	0.44	-0.15	-0.3	0.24	-0.71
-0.3	0.12	0.4	-0.39	-0.53	0.12	-0.23	0.46	0.13
0.35	0.35	0.35	0.35	-0.35	-0.35	-0.35	-0.35	0

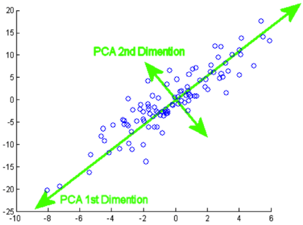


Задача отбора признаков

- *Дано*: много данных-признаков
- *Проблемы*:
  1. Где полезные признаки?
  2. Где шумовые признаки?
  3. Как ускорить модель?
  4. Как сохранить информацию?
- *Пример решения*: понижение размерности.

Задача понижения размерности





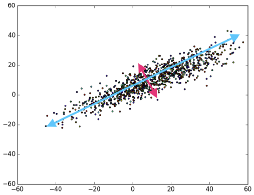
- Иногда можно избавиться от ненужной размерности;
- Иногда нужно объединить признаки в один;

Метод главных компанент (PCA)

- *Было*: $X \in n \times D$ – матрица объект-признак;
- *Хотим*: $Z \in n \times d, d < D$ – новая матрица объект-признак;
- Строим линейные комбинации

$$z_{ij} = \sum_{k=1}^{D} w_{jk} \cdot x_{ik} = \sum_{k=1}^{D} x_{ik} \cdot w_{kj}^{T} \Rightarrow Z = XW^{T}$$

Метод главных компанент (PCA)



- Интерпретация: минимизация расстояния до гиперплоскости;
- Иначе: максимизация дисперсии после ортогональной проекции на гиперплоскость + уменьшение ковариации между разными признаками;
- Искомая система – ортонормированный базис собственных векторов матрицы ковариации $X^T X$;

Метод главных компанент (PCA)

- Если V_k – матрица выбранных k главных компонент, то $Z_k = XV_k$ – проекция данных на эти компоненты;
- Если $X = U\Sigma V^T$, то $Z = XV = U\Sigma V^T V$;
- V – матрица собственных векторов марицы ковариации;
- Выбор k определяется по ошибке:

$$\delta_k^2 = \frac{\lambda_{k+1} + \ldots + \lambda_n}{\lambda_1 + \ldots + \lambda_n}$$

Заключение

- Матричные разложения – способ оптимизировать процесс хранения и обработки данных;
- SVD – универсальный инструмент, применимый для любых матриц;
- Позволяет решать разнообразные задачи:
  1. Решение СЛАУ МНК;
  2. Сжатие данных;
  3. Тематическое моделирование;
  4. Отбор признаков;

Источники

1. Статья про ЛСА – https://habr.com/post/110078/
2. Лекции про SVD и PCA – https://www.coursera.org/lecture/unsupervised-learning/
3. Статьи про SVD и PCA на wikipedia.org
4. Статьи про SVD и PCA на machinelearning.ru
5. etc.