

Матричные разложения и их применения в анализе данных

Руслан Хайдуров, Анастасия Иовлева

16 октября 2017

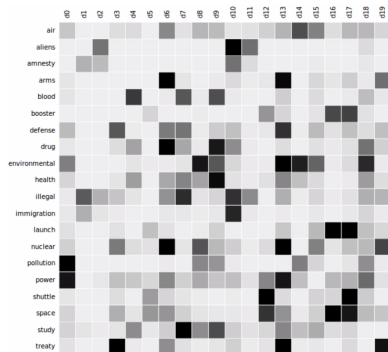
Содержание

- 1 Latent semantic analysis (LSA)
 - Что это такое?
 - При чем тут матричные разложения?
 - Преимущества и недостатки
 - Probabilistic latent semantic analysis
 - Применение: тематическое моделирование
- 2 Non-negative matrix factorization (NMF)
 - Non-negative matrix factorization
 - Где используется?
 - Проблемы
 - Как решать?
 - Функции потерь
 - Блочнo-покоординатная оптимизация
 - Инициализация
 - Пример применения

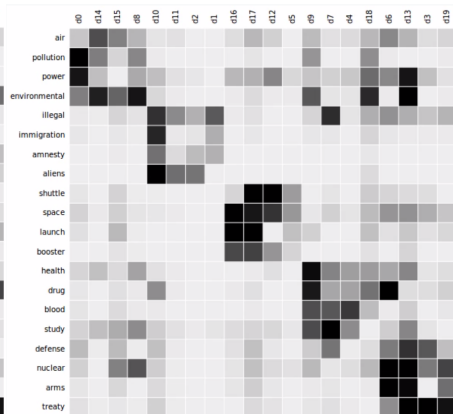
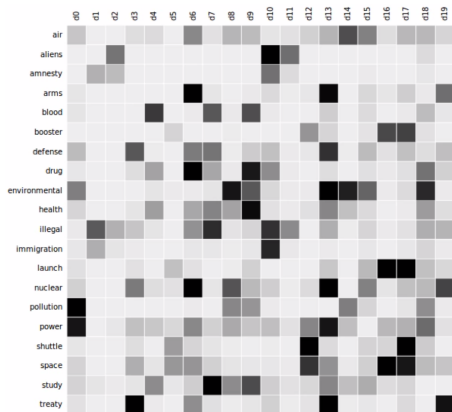
Latent semantic analysis

Выявление латентных зависимостей
внутри множества документов

- сравнить документы
- сравнить слова
- отношения между словами
(например, синонимы)



Пример



Матричные разложения

Используем SVD:

$$X \approx USV^T$$

$X \in \text{Mat}(m, n)$ — исходная document-term матрица,
 $U \in \text{Mat}_{m \times t}$, $V \in \text{Mat}_{n \times t}$ — ортогональные матрицы,
 S — диагональная матрица.

Столбцы U и V соотносятся с темами текстов. Теперь мы можем, например, сравнить похожесть i -го и j -го слова, сравнив i -ую и j -ую строчки V .

Преимущества

- Уменьшение размерности
- Может использоваться без обучения (кластеризация)
- Кластеризует документы почти как человек
- Уменьшает количество синонимов, различает омонимы
- Независимость от языка
- Устойчив к шуму (опечатки, ошибки типографии...)

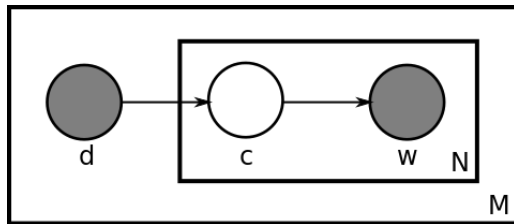
Недостатки

- Вероятностная модель метода не соответствует реальности (распределение Пуассона).
- Неясная интерпретация с точки зрения естественного языка

$$\begin{aligned} \{(\text{car}), (\text{truck}), (\text{flower})\} &\rightarrow \{(x \cdot \text{car} + y \cdot \text{truck}), (\text{flower})\} \\ \{(\text{car}), (\text{bottle}), (\text{flower})\} &\rightarrow \{(x \cdot \text{car} + y \cdot \text{bottle}), (\text{flower})\} \end{aligned}$$

- Не отличает разные значения одного слова
- Зависит от SVD

Probabilistic latent semantic analysis



Использует мультиномиальное распределение:

$$P(w \mid d) = P(d) \sum_c P(c \mid d) P(w \mid c).$$

Обучается с помощью EM-алгоритма

Тематическое моделирование

Темы

Документы

Пропорции и состав
тем в документе

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organism** can be sustained with just 250 **genes**, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a geneticist at the University in Sweden. He arrived at the 800 number. But coming up with a consensus answer may be more than just a **science** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains

Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

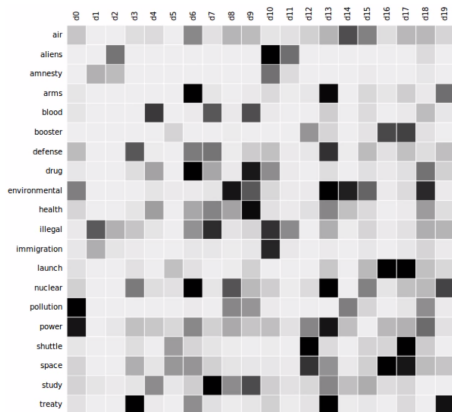


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Тематическое моделирование



Зачем?

- Тематический поиск
- Ранжировать по сходству с заданным фрагментом
- Классификация текстов, правила каталогизации новых документов
- Суммаризация и аннотация коллекций документов
- Как темы изменялись со временем (если есть время создания документа)
- Определить тематику авторов, журналов, конференций и т.д. (если они известны)
- Разделить документ на тематически однородные фрагменты

Где используется?

- Анализ коллекций научных статей, новостных потоков
- Рубрикация коллекций текстов, изображений, видео, музыки
- Задачи биоинформатики (например, аннотация генома)
- ...и многое другое

Содержание

- 1 Latent semantic analysis (LSA)
 - Что это такое?
 - При чем тут матричные разложения?
 - Преимущества и недостатки
 - Probabilistic latent semantic analysis
 - Применение: тематическое моделирование
- 2 Non-negative matrix factorization (NMF)
 - Non-negative matrix factorization
 - Где используется?
 - Проблемы
 - Как решать?
 - Функции потерь
 - Блочнo-покоординатная оптимизация
 - Инициализация
 - Пример применения

Non-negative matrix factorization

$$\begin{bmatrix} W \\ \times \end{bmatrix} \begin{bmatrix} H \\ \approx \end{bmatrix} \begin{bmatrix} V \end{bmatrix}$$

W и H — матрицы с неотрицательными элементами.

$W \in \text{Mat}_{n \times r}$, $H \in \text{Mat}_{r \times m}$, $V \in \text{Mat}_{n \times m}$.

Зачем?

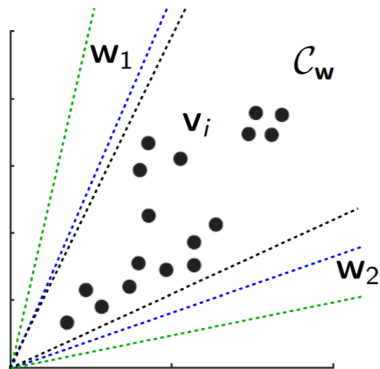
- Понижение размерности.
- Неотрицательные элементы удобно интерпретировать.
- Автоматически кластеризует данные: столбцы W — центроиды кластеров, H — индикаторы классов (если $H_{kj} > 0$, то j -ый столбец V принадлежит классу k).

Проблемы

- Решения может не существовать
- Или их бесконечно много:

$$V = WH = WB B^{-1} H.$$

- Требование на неотрицательность усложняет вычисления
- Чувствителен к начальному приближению



Как решать?

Фактически это задача оптимизации:

$$(W^*, H^*) = \operatorname{argmin} D(V, WH).$$

Функция потерь:

$$D(X, \hat{X}) = \sum_{i=1}^n \sum_{j=1}^m d(x_{ij}, \hat{x}_{ij}); \quad d(x, \hat{x}) \geq 0, \quad d(x, \hat{x}) = 0 \leftrightarrow x = \hat{x}.$$

Функции потерь

Функция потерь — замаскированное правдоподобие. Поэтому в разных областях применения (биоинформатика, тематическое моделирование, анализ аудиторизацией и т.д.) используются разные функции потерь.

Несколько примеров:

	$d(x, \hat{x})$
Норма Фробениуса	$(x - \hat{x})^2$
Обобщенная дивергенция Кульбака-Лейблера	$x \ln \frac{x}{\hat{x}} - x + \hat{x}$
Дивергенция Итакура-Саито	$\ln \frac{\hat{x}}{x} + \frac{x}{\hat{x}} - 1$

Блочно-покоординатная оптимизация

Проблема: обычно функционал ошибки невыпуклый по совокупности элементов — нельзя использовать методы поиска глобального минимума.

Решение: Блочно-покоординатная оптимизация. Фиксируем одну из матриц, обновляем вторую матрицу. Повторяем, пока не сойдемся.

$$\begin{aligned}H &= f(V, W, H_{\text{prev}}), \\W &= g(V, W_{\text{prev}}, H), \\f(V, W, H) &= g^T(V^T, W^T, H^T).\end{aligned}$$

Инициализация

От выбора начального приближения зависит, в какой локальный минимум мы попадем.

Возможные варианты генерации начальных значений матриц W и H :

- Случайные матрицы
- Кластеризация грубым методом
- SVD

$$\underbrace{X(:,j)}_{\substack{j\text{th facial image} \\ \text{[Image of a man's face]}}} \approx \sum_{k=1}^r \underbrace{W(:,k)}_{\substack{\text{facial features} \\ \text{[Grid of 20 feature images]}}} \underbrace{H(k,j)}_{\substack{\text{importance of features} \\ \text{in } j\text{th image} \\ \text{[Heatmap of feature importance]}}} = \underbrace{WH(:,j)}_{\substack{\text{approximation} \\ \text{of } j\text{th image} \\ \text{[Approximate image of the man's face]}}} .$$