Latent Semantic Indexing

Seminar “Theoretical Topics in Data Science”

Vahe Eminyan

vahe.eminyan@rwth-aachen.de

29.12.2023

Overview

Motivation

- Large datasets, often organized in tabular form, represented as matrices
  - Term-document matrix representing word occurrence in documents
  - Movie-user matrix representing watched movies of users
- Interesting aspects
  - Find documents semantically associated with a query
  - Recommend a new movie to a user

	Doc 1	Doc 2	...	Doc m
Term 1	0	1	...	1
Term 2	1	0	...	1
...	...	...	...	...
Term n	1	0	...	0

$$\longrightarrow \quad \begin{array}{cc} & \text{Documents} \\ \text{Terms} & \begin{pmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \\ & n \times m \end{array}$$

Latent Semantic Indexing

- LSI as an information retrieval method
- Finds the latent (hidden) semantic structure of textual data
- Represent term-document matrix as product of three matrices
- Answer queries with help of these matrices
- Based on singular value decomposition of the matrix

LSI Background

Singular Value Decomposition (SVD) [7]

- Any n by m matrix of rank r can be factored into

$$A_{n\times m} = U_{[n\times r]} D_{[r\times r]} (V_{[m\times r]})^T.$$

- U column-orthonormal matrix: left singular vectors
- V column-orthonormal matrix: right singular vectors
- D diagonal matrix: Singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ in decreasing order
- Vector notation

$$A = UDV^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Singular Value Decomposition (SVD) Example: Matrix A with rank $r = 3$

$$\text{Terms}\quad \overset{\text{Documents}}{\underset{A}{\begin{pmatrix}1&0&1\\1&1&0\\0&1&0\\0&1&1\end{pmatrix}}} = \overset{\text{Term-Topic similarity}}{\underset{U}{\begin{pmatrix}-0.48&-0.79&-0.11\cdot 10^{-14}\\-0.58&0.16&0.71\\\mathbf{-0.34}&\mathbf{0.56}&0.42\cdot 10^{-15}\\-0.56&0.16&-0.71\end{pmatrix}}} \times \overset{\text{Topic ”importance”}}{\underset{D}{\begin{pmatrix}2.1&0&0\\0&1.26&0\\0&0&1\end{pmatrix}}}$$

$$\times \overset{\text{Topic-Document similarity}}{\underset{V^T}{\begin{pmatrix}-0.5&\mathbf{-0.71}&-0.5\\-0.5&\mathbf{0.71}&-0.5\\0.71&0.67\cdot 10^{-15}&-0.711\end{pmatrix}}}$$

LSI Background

Latent Semantic Indexing based on SVD

- LSI considers A_k the rank k approximation of A (I.e. keep only k most relevant topics)

$$A_k = U_k D_k V_k^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

$$\begin{array}{cc} & \text{Documents} \\ \text{Terms} & \begin{pmatrix} 1.0 & 0.01 & 1 \\ 0.51 & 1.01 & 0.51 \\ 0.0 & 1.01 & 0.0 \\ 0.49 & 0.98 & 0.49 \end{pmatrix} \\ & A_k \end{array} = \begin{array}{c} \text{Term-Topic similarity} \\ \begin{pmatrix} -0.48 & -0.79 \\ -0.58 & 0.16 \\ \mathbf{-0.34} & \mathbf{0.56} \\ -0.56 & 0.16 \end{pmatrix} \\ U_k \end{array} \times \begin{array}{c} \text{Topic "importance"} \\ \begin{pmatrix} 2.1 & 0 \\ 0 & 1.26 \end{pmatrix} \\ D_k \end{array} \times \begin{array}{c} \text{Topic-Document similarity} \\ \begin{pmatrix} -0.5 & \mathbf{-0.71} & -0.5 \\ -0.5 & \mathbf{0.71} & -0.5 \end{pmatrix} \\ V_k^T \end{array}$$

- Map a query to k dimensional space with U_k, apply cosine similarity to find similar documents in $D_k V_k^T$

Latent Semantic Indexing based on SVD

Theorem (Eckart and Young [3])

Among all $n \times m$ matrices C of rank at most k, A_k is the one that minimizes $||A - C||_F^2 = \sum_{i,j}(A_{ij} - C_{ij})^2$, where F denotes the Frobenius norm of a matrix.

$$
\begin{array}{cc}
 & \text{Documents} \\
\text{Terms} & \begin{pmatrix} 1.0 & 0.01 & 1 \\ 0.51 & 1.01 & 0.51 \\ 0.0 & 1.01 & 0.0 \\ 0.49 & 0.98 & 0.49 \end{pmatrix} \\
 & A_k
\end{array}
\approx
\begin{array}{cc}
 & \text{Documents} \\
\text{Terms} & \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \\
 & A
\end{array}
$$

Original Paper Overview and Emphasized Aspect

- Strong empirical results of LSI
- Two important aspects
  - Why does LSI find semantically related documents?
  - How can we reduce the computational complexity ?
- Papadimitriou et al. [6] investigated both aspects:
  1. Under certain constraints on the term-document matrix, semantically related documents are mapped to similar vectors
  2. Instead of LSI use LSI by random projection.This reduces the computational time:
     - Map the original term-document matrix into a lower dimensional space
     - Use LSI on the lower dimensional matrix
- We focus on the second aspect

Random Projection for Dimensionality Reduction

Given a matrix $A \in \mathbb{R}^{n \times m}$ and a matrix $R \in \mathbb{R}^{\ell \times n}$. Use matrix R to reduce the dimensionality of matrix A by preserving pairwise distances between any two points:

$$B = \sqrt{\frac{n}{\ell}} \cdot R^T A \in \mathbb{R}^{\ell \times m}$$

Lemma (Johnson and Lindenstrauss [4])

Let $v \in \mathbb{R}^n$ be a unit vector, let H be a random ℓ-dimensional subspace through the origin, and let the random variable X denote the square of the length of the projection of v onto H. Suppose $0 < \epsilon < 0.5$, and $24 \log n < \ell < \sqrt{n}$. Then, $E[X] = \frac{\ell}{n}$, and

$$Pr(|X - \frac{\ell}{n}| > \epsilon \frac{\ell}{n}) < 2\sqrt{\ell} e^{-(\ell-1)\epsilon^2/4}$$

LSI by Random Projection

Two-Step LSI

1. Apply a random projection onto ℓ dimensions on A. ($\ell > k$)

$$B = \sqrt{\frac{n}{\ell}} \cdot \begin{pmatrix} | & | & & | \\ r_1 & r_2 & \cdots & r_\ell \\ | & | & & | \end{pmatrix}^T \cdot A$$

2. Apply rank $O(k)$ LSI

- Improved computational complexity
- With high probability the original matrix A almost as good recovered as by directly using LSI (Formulation and proof of theorem later)

Background and Notation for the Proof

Vector notations of SVD:

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \qquad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \qquad B = \sum_{i=1}^{\ell} \lambda_i a_i b_i^T, \qquad B_{2k} = A \sum_{i=1}^{2k} b_i b_i^T.$$

- A: original term-document matrix
- A_k: rank k approximation of A
- B: matrix after randomly projecting and scaling A
- B_{2k}: rank $2k$ approximation of A

LSI by Random Projection

Background and Notation for the Proof

Lemma (3)

Let ϵ be an arbitrary positive constant. If $\ell \geq c((\log n)/\epsilon^2)$ for a sufficiently large constant c then, for $p = 1, \dots, \ell$

$$\lambda_p^2 \geq \frac{1}{k}\left[(1-\epsilon)\sum_{i=1}^{k}\sigma_i^2 - \sum_{j=1}^{p-1}\lambda_j^2\right].$$

Corollary (4)

$$\sum_{p=1}^{2k}\lambda_p^2 \geq (1-\epsilon)\|A_k\|_F^2.$$

LSI by Random Projection

Background and Notation for the Proof

Lemma (5)

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2 .$$

Theorem (Parsevals identity [2])

Let $b_1, \ldots, b_n$ be an orthonormal basis for a space S. Then for each $s \in S$, $|s|^2 = \sum_{i=1}^{n} (s b_i)^2$.

LSI by Random Projection

Main Theorem

Theorem (Papadimitriou et al. [6])

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Informally, the theorem states that the original matrix A after applying random projection and then LSI is with high probability almost as good recovered as by using one-step LSI on the original matrix.

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \le \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof

We have

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \qquad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \qquad B = \sum_{i=1}^{\ell} \lambda_i a_i b_i^T, \qquad B_{2k} = A \sum_{i=1}^{2k} b_i b_i^T.$$

$b_1, \dots, b_n$ Are orthonormal vectors spanning the row space of A and B_{2k}.
Hence using the Parseval's identity we can write:

$$\|A - B_{2k}\|_F^2 = \sum_{i=1}^{n} |(A - B_{2k})b_i|^2. \tag{1}$$

For $i = 1, \dots, 2k$, because $b_i^T b_i = 1$, we have

$$(A - B_{2k})b_i = Ab_i - Ab_i = 0, \tag{2}$$

and for $i = 2k + 1, \dots, n$, because $b_j^T b_i = 0$, we have

$$(A - B_{2k})b_i = Ab_i. \tag{3}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

Now we continue from the equation

$$\|A - B_{2k}\|_F^2 = \sum_{i=1}^{n} |(A - B_{2k})b_i|^2 \tag{4}$$

$$= \sum_{i=2k+1}^{n} |Ab_i|^2 \tag{5}$$

$$= \sum_{i=1}^{n} |Ab_i|^2 - \sum_{i=1}^{2k} |Ab_i|^2 \tag{6}$$

$$\overset{\text{Parseval's id.}}{=} \|A\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2 \tag{7}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \le \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

On the other hand, we have

$$\|A - A_k\|_F^2 \overset{\text{Lemma 5}}{=} \sum_{i=k+1}^{n} \sigma_i^2 \tag{8}$$

$$\overset{\text{Frob. norm [5]}}{=} \|A\|_F^2 - \|A_k\|_F^2. \tag{9}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

Now we consider

$$\|A - B_{2k}\|_F^2 - \|A - A_k\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2 - (\|A\|_F^2 - \|A_k\|_F^2) \tag{10}$$

$$= \|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2, \tag{11}$$

that is equivalent to

$$\|A - B_{2k}\|_F^2 = \|A - A_k\|_F^2 + (\|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2) \tag{12}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

For the next step, we show

$$(1+\epsilon)\sum_{i=1}^{2k}|Ab_i|^2 \geq \sum_{i=1}^{2k}\lambda_i^2. \tag{13}$$

We write

$$\sum_{i=1}^{2k}\lambda_i^2 \overset{|Bb_i|=\lambda_i}{=} \sum_{i=1}^{2k}|Bb_i|^2 \tag{14}$$

$$\overset{\text{sbst. B}}{=} \sum_{i=1}^{2k}\left|\sqrt{\frac{n}{\ell}}R^T(Ab_i)\right|^2 \tag{15}$$

$$= \sum_{i=1}^{2k}\frac{n}{\ell}\left|R^T(Ab_i)\right|^2 \tag{16}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \le \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

Now from the Johnson-Lindenstrauss lemma [4] for very large $\ell \in \Omega((\log n)/\epsilon^2)$ we have for each i

$$\frac{n}{\ell}|R^T(Ab_i)|^2 \le (1+\epsilon)|Ab_i|^2 \tag{17}$$

with high probability .

Hence with a high probability

$$(1+\epsilon)\sum_{i=1}^{2k}|Ab_i|^2 \ge \sum_{i=1}^{2k}\lambda_i^2. \tag{18}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

Now we have

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq \frac{1}{(1+\epsilon)} \sum_{i=1}^{2k} \lambda_i^2 \tag{19}$$

$$\overset{\text{Cor. 4}}{\geq} \frac{(1-\epsilon)}{(1+\epsilon)} \|A_k\|_F^2 \tag{20}$$

$$\geq (1-2\epsilon)\|A_k\|_F^2 \tag{21}$$

I.e.

$$\sum_{i=1}^{2k} |Ab_i|^2 \geq (1-2\epsilon)\|A_k\|_F^2 \tag{22}$$

LSI by Random Projection

Theorem

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2$$

where $\epsilon \in (0, 0.5)$

Proof (continued)

Remember the Equation (12):

$$\|A - B_{2k}\|_F^2 = \|A - A_k\|_F^2 + (\|A_k\|_F^2 - \sum_{i=1}^{2k} |Ab_i|^2)$$

Now we substitute the result of Equation (22) in equation (12):

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + \|A_k\|_F^2 - (1 - 2\epsilon)\|A_k\|_F^2 \tag{23}$$

$$\iff \|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A_k\|_F^2 \tag{24}$$

Due to the formulation of Frobenius norm as in Lemma 5, we have $\|A\|_F^2 \geq \|A_k\|_F^2$.
Hence

$$\|A - B_{2k}\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A\|_F^2. \tag{25}$$

□

LSI by Random Projection

Comparison of Computational Time

Given the term-document matrix $A \in \mathbb{R}^{n \times m}$.

Time complexity of one-step LSI:

- LSI computation: $O(mnc)$ if A is sparse with about c nonzero entries per column

Time complexity of LSI by random projection :

- Random projection to ℓ dimensions: $O(mc\ell)$
- LSI computation: $O(m\ell^2)$
- Together: $O(mc\ell + m\ell^2) = O(m(c\ell + \ell^2))$, with $\ell \in O(\frac{\log n}{\epsilon^2})$
- Hence we get a time complexity: $O(m(\log^2 n + c \log n))$

$O(m(\log^2 n + c \log n))$ better than $O(mnc)$

Summary and Newer Approaches

- Latent semantic analysis: SVD-based technique for information retrieval
- Papadimitriou et al. analysed two important aspects [6]
  - Why does LSI find semantically related documents?
  - How can we reduce the computational time ? (Our main focus)
- LSI by random projection: reduction of computational complexity, while preventing the expressiveness of original matrix with high probability.
- There are newer techniques based on neural networks [8, 1]

References

William L. Hamilton, Rex Ying, and Jure Leskovec.
Representation learning on graphs: Methods and applications.
IEEE Data Eng. Bull., 40(3):52–74, 2017.
URL: `http://sites.computer.org/debull/A17sept/p52.pdf`.

Leslie Hogben, editor.
Handbook of Linear Algebra.
Chapman and Hall/CRC, 2nd edition, 2013.
`https://doi.org/10.1201/b16113`.

C. Reinsch J. H. Wilkinson.
Handbook for Automatic Computation.
Springer Berlin, Heidelberg, volume ii: linear algebra edition, 1971.

William B Johnson.
Extensions of lipshitz mapping into hilbert space.
In *Conference modern analysis and probability, 1984*, pages 189–206, 1984.

Changxue Ma, Y. Kamp, and L.F. Willems.
A frobenius norm approach to glottal closure detection from the speech signal.
IEEE Transactions on Speech and Audio Processing, 2(2):258–265, 1994.
`doi:10.1109/89.279274`.

Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala.
Latent semantic indexing: A probabilistic analysis.
Journal of Computer and System Sciences, 61(2):217–235, 2000.
URL: `https://www.sciencedirect.com/science/article/pii/S0022000000917112`, `doi:10.1006/jcss.2000.1711`.

Gilbert Strang.
Linear Algebra and Its Applications.
Cengage Learning, 4th edition edition, 2005.

References

Liang Yao, Chengsheng Mao, and Yuan Luo.
Graph convolutional networks for text classification.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.