Binary Factorizations in Data Mining

Kick-off meeting

27 October 2016

Today’s Agenda

- Short intro to binary factorizations
- Check attendance
- Goals of the seminar
- Organization of the seminar
- Grading & guidelines

First Things First

- A block seminar
  - Preliminary work + one (or two) day(s) of presentations
- 7 ECTS credits
- Meeting all DLs and attending all talks is **mandatory** for passing the seminar
- Attending this kick-off meeting is mandatory

Short intro

Matrix multiplication

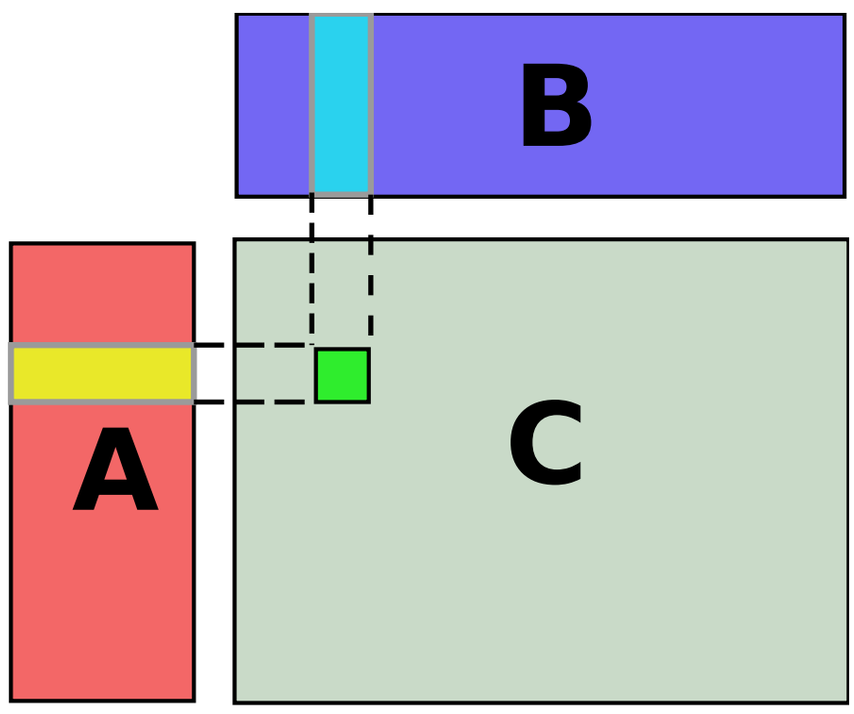
- The product of two matrices, ***A*** and ***B***, is defined element-wise as

$$(\boldsymbol{AB})_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j}$$

  - The number of columns in ***A*** and number of rows in ***B*** must agree
    - inner dimension

Intuition for Matrix Multiplication

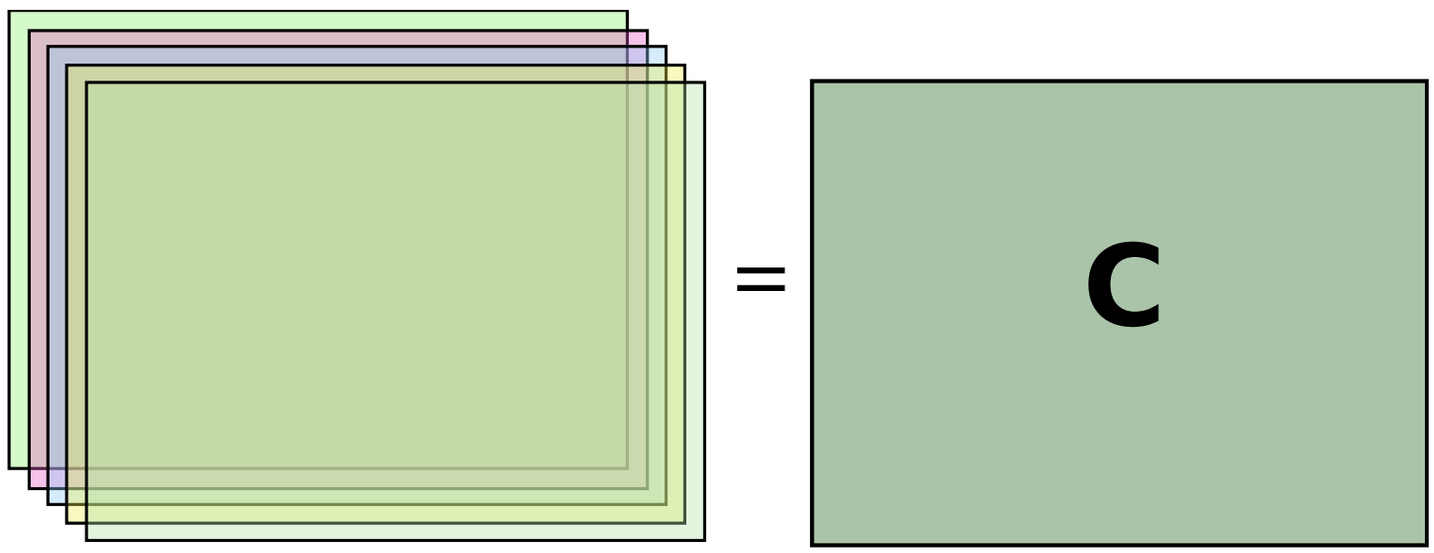
- Element $(\boldsymbol{AB})_{ij}$ is the inner product of row i of $\boldsymbol{A}$ and column j of $\boldsymbol{B}$



$$\boldsymbol{c}_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j}$$

Intuition for Matrix Multiplication

- Matrix $\boldsymbol{AB}$ is a sum of k matrices $\boldsymbol{a}_l\boldsymbol{b}_l^T$ obtained by multiplying the l-th column of $\boldsymbol{A}$ with the l-th row of $\boldsymbol{B}$



$$\boldsymbol{C} = \sum_{\ell=1}^{k} \boldsymbol{a}_\ell \boldsymbol{b}_\ell^T$$

Matrix decompositions

- A **decomposition** of matrix ***A*** expresses it as a product of two (or more) **factor matrices**
  - $\boldsymbol{A} = \boldsymbol{BC}$
- Every matrix has decomposition $\boldsymbol{A} = \boldsymbol{AI}$ (or $\boldsymbol{A} = \boldsymbol{IA}$ if $n < m$)
- The size of the decomposition is the inner dimension of the product

Matrices in data mining

$$\begin{array}{l} \\ \text{Anna} \\ \text{Bob} \\ \text{Charlie} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Bread} & \text{Butter} & \text{Beer} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

Customer transactions

$$\begin{array}{l} \\ \text{Book 1} \\ \text{Book 2} \\ \text{Book 3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Data} & \text{Matrix} & \text{Mining} \end{array} \\ \begin{pmatrix} 5 & 0 & 3 \\ 0 & 0 & 7 \\ 4 & 6 & 5 \end{pmatrix} \end{array}$$

Document-term matrix

$$\begin{array}{l} \\ \text{Alice} \\ \text{Bob} \\ \text{Charlie} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Avatar} & \text{The Matrix} & \text{Up} \end{array} \\ \begin{pmatrix} & 4 & 2 \\ 3 & 2 & \\ 5 & & 3 \end{pmatrix} \end{array}$$

Incomplete rating matrix

$$\begin{array}{l} \\ \text{Saarbrücken} \\ \text{Helsinki} \\ \text{Cape Town} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Jan} & \text{Jun} & \text{Sep} \end{array} \\ \begin{pmatrix} 1 & 11 & 10 \\ 6.5 & 10.9 & 8.7 \\ 15.7 & 7.8 & 8.7 \end{pmatrix} \end{array}$$

Cities and monthly temperatures

Matrix decompositions in data mining

- A common goal in data mining is to find regularities (or patterns) in the data
  - Often, to summarize the data
- A *matrix decomposition* presents the data as a sum of “simple” elements, i.e. patterns
  - but there’s also other uses... *stay tuned!*

Text mining and pLSA

- Consider a document–term matrix ***A***
  - a_{ij} is the number of times term j appears in document i

Can we find these topics automatically?

Environmet

$$\boldsymbol{A} = \begin{array}{c} \\ \text{doc 1} \\ \text{doc 2} \\ \text{doc 3} \\ \text{doc 4} \\ \text{doc 5} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{air} & \text{water} & \text{pollution} & \text{democrat} & \text{republican} \end{array} \\ \begin{pmatrix} 3 & 2 & 8 & 0 & 0 \\ 1 & 4 & 12 & 0 & 0 \\ 0 & 0 & 0 & 10 & 11 \\ 0 & 0 & 0 & 8 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

Politics

pLSA example

air wat pol dem rep

0.04	0.03	0.12	0	0
0.01	0.06	0.17	0	0
0	0	0	0.14	0.16
0	0	0	0.12	0.07
0.01	0.01	0.01	0.01	0.01

A

Here, A is normalized

0.39	0
0.52	0
0	0.58
0	0.36
0.09	0.06

W

How strong the topic is in the document?

0.48	0
0	0.52

Σ

Overall frequency

air wat pol dem rep

0.15	0.21	0.64	0	0
0	0	0	0.53	0.47

H

How strong the word is in the topic?

Binary factorizations

- Often we deal with **binary** matrices
  - presence/absence data, graph adjacency matrices, market basket data, ...
- Such matrices are commonly better decomposed into binary factor matrices
  - Or one matrix binary, other numerical
  - Interpretability, sparsity, speed, ...
- Also the algebra might change to Boolean

Binary factorizations

long-haired	✔ 1	✔ 1	✘ 0
well-known	✔ 1	✔ 1	✔ 1
male	✘ 0	✔ 1	✔ 1

Example factorization

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{matrix} A & B & C \end{matrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \Psi \begin{matrix} A & B & C \end{matrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Papers in this seminar…

- Factorize binary (or ternary) matrices
- Have at least one factor matrix discrete/ binary
- Either present new binary factorisation methods, or apply the binary factorisations to other problems
  - Or both

Head Count

Goals

- To learn how to read and understand recent research literature
- To learn how to write a concise report of a research article
- To learn how to present research
- To boldly read what no one (at this seminar) has read before
- To keep young people out of streets

Workflow

1. You read the paper (+ other papers)
2. You write a draft report and send it to me
3. I comment your report
4. You improve your report and prepare your presentation, which you also send to me
5. I comment your presentation
6. You improve your presentation and send me the final report
7. I distribute the reports to everybody
8. You read others’ reports
9. You present your work and follow and discuss others’ presentations

Schedule

Day	Topic
27 October	Kick-off
27 November	Report draft DL
9 December	Slides draft DL
Early January	Report DL
January/February	Seminar

Selecting the dates

- The seminar takes two full days (approx. 9:00–16:00)
  - The days have to be consecutive
- I’ve created a doodle where you can indicate which days would work with you
  - http://doodle.com/poll/bgst3hwns83vdwqn
  - Fill in the doodle **by Wednesday, 2 November**
  - Use *no* if you have hard constraints and *if-need-be* if you have soft constraints
    - The more *no* answers you have, the more likely I’m to violate them

Odds and ends

- Remember to register to HISPOS
  - You can de-register for three weeks from now, but not after that
- The papers are now behind a username and password
  - see whiteboard or contact the lecturer

Grading Overview

- Report (3–5 pages):
  - Correctness, connections, criticism, style
- Slides & presentation (20 min):
  - Delivery, clearness, presentation skills
- Discussion (5–10 min):
  - Participation, correctness, connections

Report

- 3–5 pages
- **In your own words**
  - No verbatim copy
  - Explain the main ideas of the paper
    - Research questions
    - Proposed solutions and their evaluation
- Write to somebody who hasn’t read the paper
- Provide also extra connections & criticism towards the approach

More on report

- Length is not a hard constraint
  - As long as needed, but no longer
  - Figures, math, choice of margins, etc. effect the final length
- You must cite your sources using the standard academic practices
  - Everything that’s not your own must be cited
- I only accept reports in PDF format
- More information is send later in the semester

Presentation

- 20 minutes (+ 5–10 minutes of discussion)
- Explain the main ideas of the paper
  - What, why, and how
- Build connections and provide criticism where appropriate
- Target to audience that knows CS and basics of DM and ML
  - Some audience knows more on your topic, others less; cater for both groups

Discussion

- After you've followed a presentation, discuss
  - You know the topic a bit: you've read the report
  - Ask if something was left unclear
  - Tell if you know more on something
  - After technical part, give constructive feedback on the presentation itself
  - Actually discussing is **mandatory**

Lecture on giving presentations?

- If there’s demand, I might give one lecture on giving presentations
  - No soft-skills seminar, just few tidbits on how to give a presentation and how to prepare slides in this seminar
- Schedule: November