On a Guided Nonnegative Matrix Factorization

Joshua Vendrow, Jamie Haddock, Elizaveta Rebrova, Deanna Needell

Department of Mathematics, UCLA

Background

- **Nonnegative Matrix Factorization (NMF)** is an approach typically applied in unsupervised tasks such as dimensionality-reduction, latent topic modeling, and clustering.
- Given nonnegative data matrix $X \in_{\geq 0}^{m \times n}$ and a user-defined target dimension $k \in \mathbb{N}$, NMF seeks nonnegative factor matrices $A \in_{\geq 0}^{m \times k}$, and $S \in_{\geq 0}^{k \times n}$ such that $X \approx AS$, formulated as

$$\arg\min_{A \geq 0, S \geq 0} \|X - AS\|_F^2. \tag{1}$$

- **Semi-supermised Nonnegative Matrix Factorization (SSNMF)** jointly factorizes a data matrix $X \in_{\geq 0}^{m \times n}$ and a supervision information matrix $Y \in_{\geq 0}^{c \times n}$, formulated as

$$\arg\min_{A,S,B \geq 0} \underbrace{\|X - AS\|_F^2}_{\text{Reconstruction Error}} + \lambda \underbrace{\|Y - BS\|_F^2}_{\text{Classification Error}}. \tag{2}$$

Method

- We denote a seed topic as a vector $v = (v_1, \ldots, v_m)$ (where m denotes the vocabulary size), where $v_i = 0$ if the ith word in the vocabulary is not in the seed topic and some positive weight otherwise.
- Let the data matrix $X \in^{m \times n}$ have examples along the columns and features along the rows and suppose we have seed topics $v^{(1)}, v^{(2)}, \ldots, v^{(c)} \in R^m$.
- Let the *seed matrix* be

$$Y = [v^{(1)}, v^{(2)}, \ldots, v^{(c)}] \in_{\geq 0}^{m \times c} \tag{3}$$

- **Guided NMF** is formulated as

$$\min_{A \geq 0, S \geq 0, B \geq 0} \|X - AS\|_F^2 + \lambda \|Y - AB\|_F^2. \tag{4}$$

- Figure 1 shows a visualization of Guided NMF with seed word *space*.

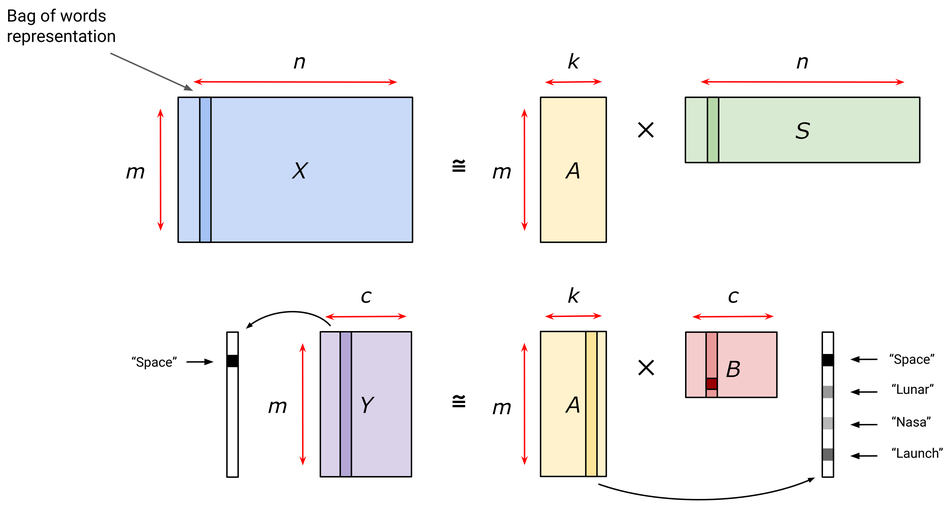

Figure 1: Visualization of Guided NMF

Conclusions

We propose an NMF-based model, Guided NMF, which incorporates seed topic supervision to guide learned topics towards meaningful and coherent sets of features. Our initial experiments illustrate the promise of this model in text-based topic modeling applications.

Contact Information

jvendrow@math.ucla.edu
www.joshvendrow.com
github.com/jvendrow

Results

- The **20 Newsgroups dataset** is a collection of approximately 20,000 text documents containing the text of messages from 20 different newsgroups on the distributed discussion system Usenet
- We use Guided NMF to encourage topics to form around specific newsgroups (e.g. space, medicine, religion).

Rank 4 Guided NMF on the 20 Newsgroups dataset with seed word *pitch*, *medical*, and *space*.

Topic 1	Topic 2	Topic 3	Topic 4
pitch	*medical*	*space*	people
expected	tests	nasa	know
curveball	disease	shuttle	think
stiffness	diseases	launch	time
loosen	prejudices	sci	use
shoulder	services	lunar	new
shea	graduates	orbit	see
rotation	health	earth	say
game	patients	station	us
giants	available	mission	god

Rank 4 Guided NMF on the 20 Newsgroups dataset with seed words *motorcycle*, *sale*, and *religion*.

Topic 1	Topic 2	Topic 3	Topic 4
motorcycle	*sale*	*religion*	people
bike	offer	christian	know
dod	condition	judaism	think
wheelie	shipping	freedom	time
shaft	asking	christians	use
bikes	includes	islam	new
rider	mb	compulsion	space
riding	excellent	avi	see
scene	price	life	say
ski	best	gunpoint	us

- The **Twitter political data set** is a data set of tweets sent by political candidates during the 2016 election season.
- We use Guided NMF to encourage topics to form around *issues* rather than *candidates*.

Rank 8 NMF on the Twitter political data set.

Topic 1	Topic 2	Topic 3	Topic 4
thank	govpencein	gopdebate	tedcruz
trump2016	indiana	imwithhuck	cruz
maga[1]	indiana_edc	jeb	cruzcrew
great	state	tonight	ted
america	jobs	president	choosecruz
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
kasich	hillary	randpaul	fitn
john	trump	iowa	new
johnkasich	people	iacaucus	hampshire
ohio	donald	caucus	johnkasich
gov	president	tonight	nh

[1]Here "maga" abbreviates "makeamericagreatagain."

Rank 8 Guided NMF on Twitter political data set with seed words *economy* and *obamacare*.

Topic 1	Topic 2	Topic 3	Topic 4
economy	*obamacare*	govpencein	gopdebate
jobs	fullrepeal	indiana	kasich
tax	repeal	indiana_edc	randpaul
plan	replace	state	john
create	fight	jobs	tonight
Topic 5	**Topic 6**	**Topic 7**	**Topic 8**
tedcruz	hillary	johnkasich	people
thank	trump	new	need
cruz	donald	fitn	must
cruzcrew	clinton	kasich	berniesanders
ted	president	hampshire	country

Ablation and Comparison

- We explore the impact of adding additional seedwords and varying the rank of the factorization, and compare to SeededLDA
- Guided NMF consistently has an AUC above 0.8 for all rank and number of seedword choices.
- For few seed words and/or only a small rank, Guided NMF significantly out-performs Seeded LDA.

AUC scores for 20 Newsgroups dataset on documents from space class.

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.83**	**0.88**	**0.88**	**0.87**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.86**	**0.87**	0.88	0.87
	Seeded LDA	0.37	0.5	**0.91**	**0.89**
10	Guided NMF	**0.88**	0.89	0.89	0.89
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**

AUC scores for 20 Newsgroups dataset on documents from baseball class.

Rank	Method	# Seed words 1	2	4	8
4	Guided NMF	**0.89**	**0.9**	**0.9**	**0.9**
	Seeded LDA	0.31	0.42	0.74	0.86
6	Guided NMF	**0.9**	**0.9**	0.9	**0.9**
	Seeded LDA	0.37	0.5	**0.91**	0.89
10	Guided NMF	**0.87**	0.9	0.9	0.9
	Seeded LDA	0.45	**0.95**	**0.95**	**0.95**