FYS: AI in Healthcare

Unsupervised Learning Part 2

John Lalor

October 16, 2018

Admin

- Assignment 4 graded
- Assignment 5 follow-up
  - `https://www.umass.edu/peoplefinder/`
- Source control follow-up
- Midterm posted

Student developer packs

Github: `https://education.github.com/pack`

Bitbucket: `https://bitbucket.org/product/education`

Midterm

Due: Friday, November 9 at 5:00pm

Instructions posted on Piazza

Midterm

Due: Friday, November 9 at 5:00pm

Instructions posted on Piazza

Course tasks

1. AIHC in the media
2. ML Understanding
3. Personal health technology
4. Medical decision making

Midterm

Due: Friday, November 9 at 5:00pm

Instructions posted on Piazza

Course tasks

1. AIHC in the media
2. ML Understanding
3. Personal health technology
4. Medical decision making

UMass tasks

1. Academic advising
2. Meet your professors
3. Time management skills
4. Building good study habits

Dimensionality Reduction

Today

- Dimensionality Reduction
  - Principal Components Analysis
  - Autoencoders
  - Word embeddings

PCA

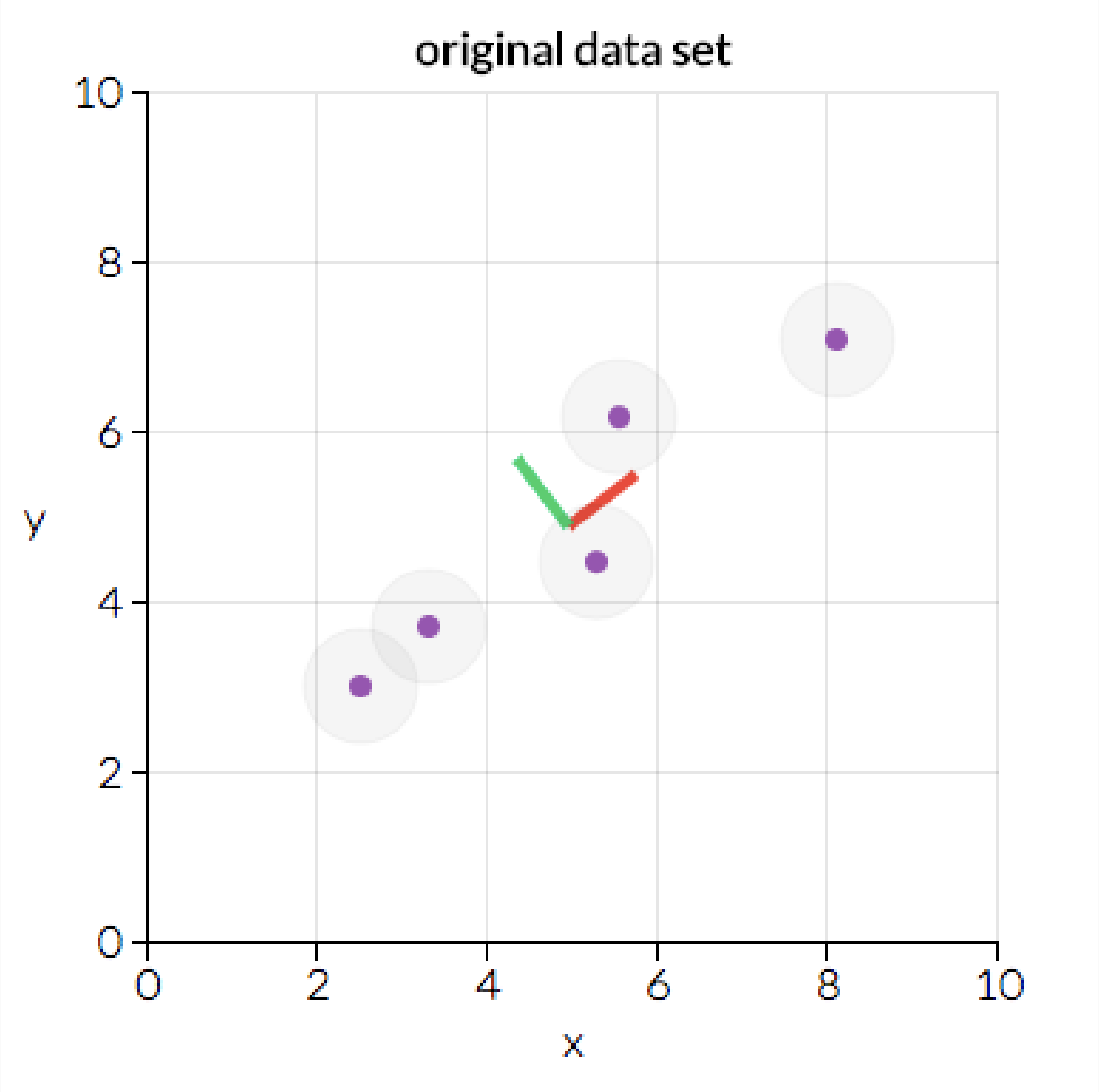
Principal Components Analysis

- Transformation of data to align with important directions
- Reduce dimensionality by dropping least important directions
- Start with most important direction, add features until you hit a threshold

---

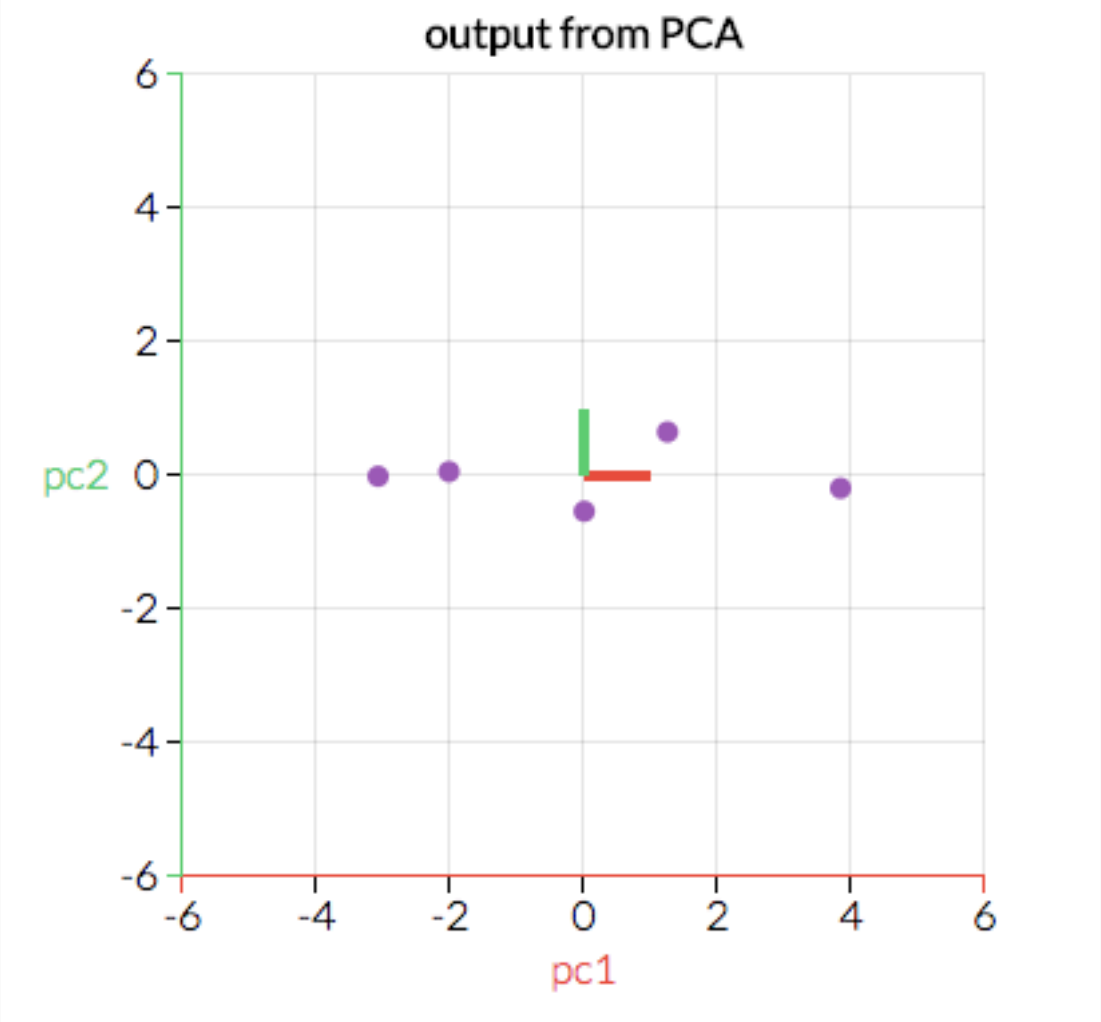
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Principal Components Analysis



https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Principal Components Analysis



https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

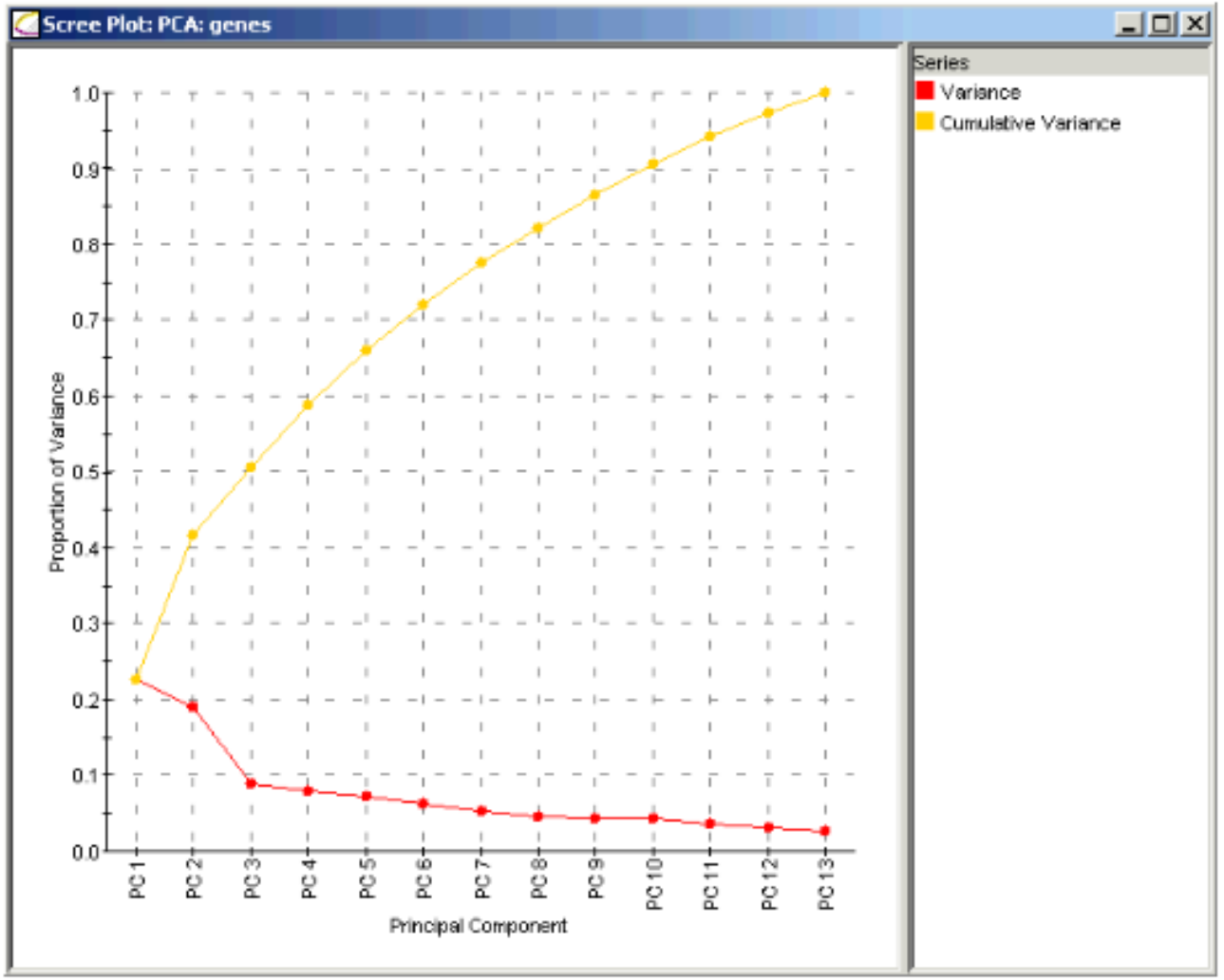
PCA Algorithm

1. Start with your data matrix X
2. Zero-mean and (maybe) normalize the columns: $X \rightarrow Z$
3. Decompose $Z^\top Z$ into PDP^{-1}
   - P: matrix of eigenvectors
   - D: diagonal matrix of eigenvalues
4. sort P based on the eigenvalues in D: $P \rightarrow P^*$
5. $Z^* = ZP^*$
   - Z^* columns of Z^* are independent of each other
6. Keep the desired number of features from Z^*

---

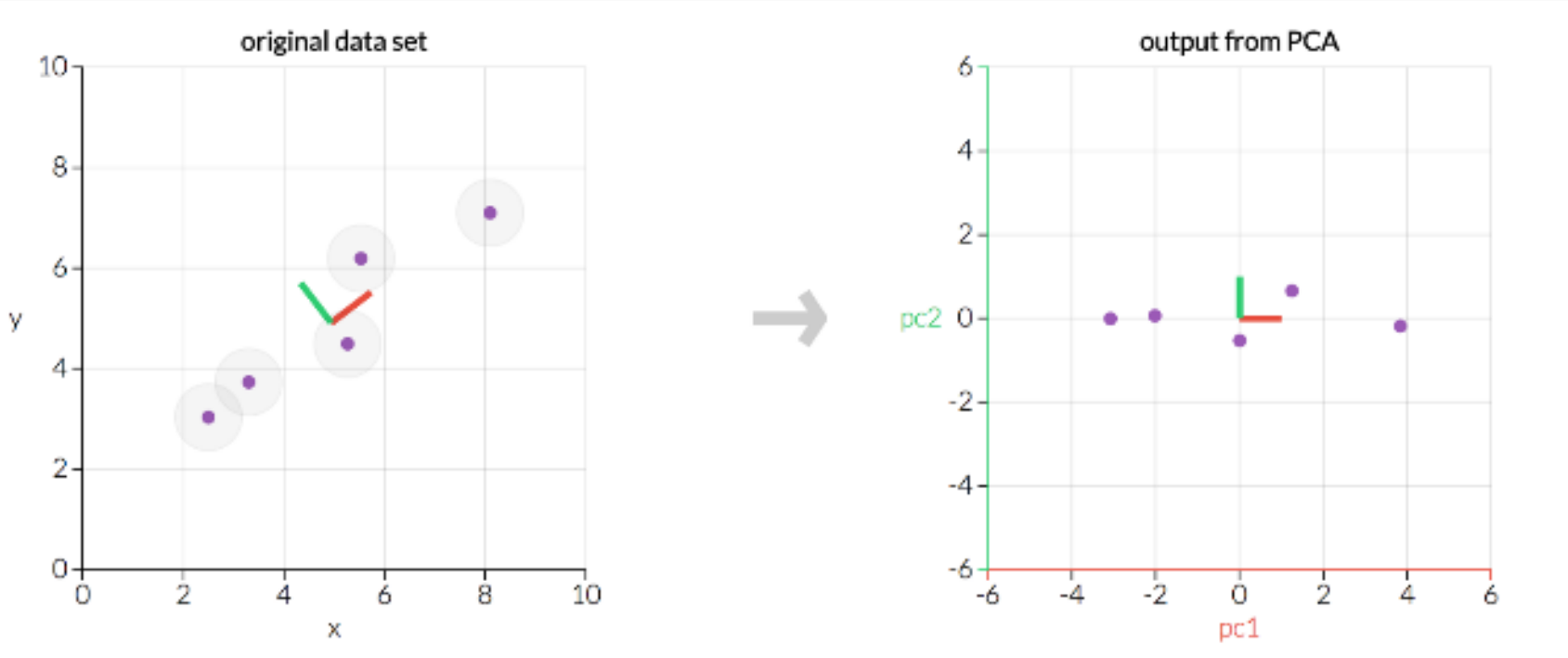
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Choosing Components



https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

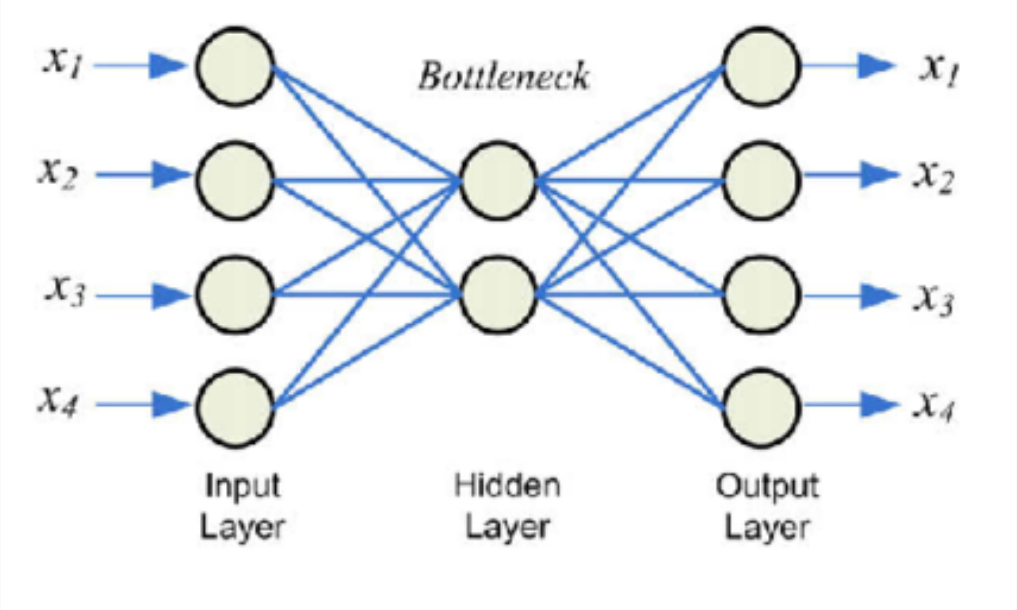
Choosing Components



https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

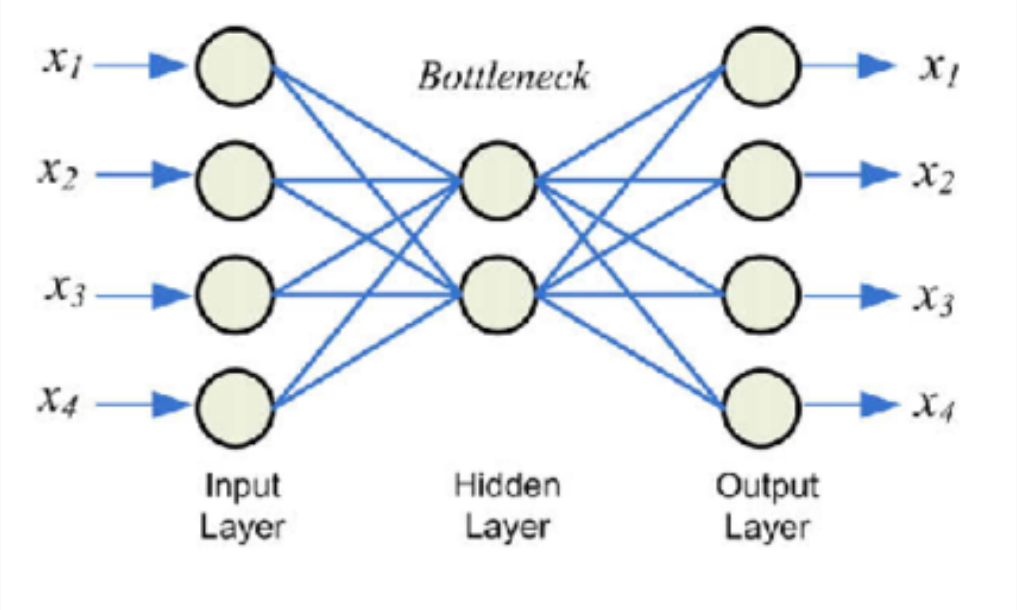
Autoencoders

Autoencoders



Nelwamondo, Fulufhelo V., Dan Golding, and Tshilidzi Marwala. "A dynamic programming approach to missing data estimation using neural networks." Information Sciences 237 (2013): 49-58.

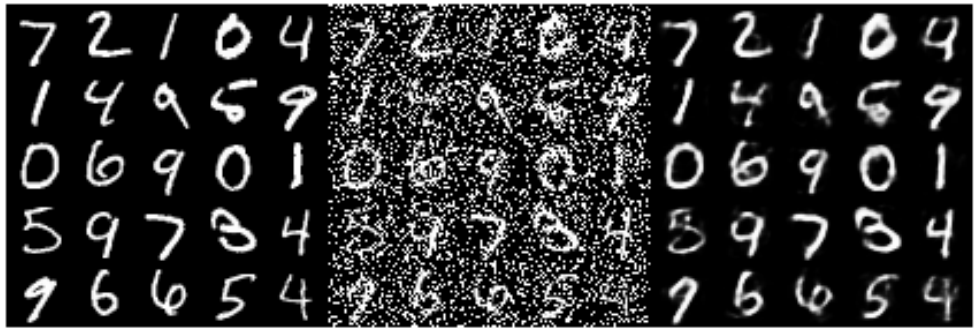
Autoencoders



$$z = \sigma(W_e x + b_e)$$
$$x' = \sigma(W_d z + b_d)$$
$$L(x, x') = \sum (x - x')^2$$

Nelwamondo, Fulufhelo V., Dan Golding, and Tshilidzi Marwala. "A dynamic programming approach to missing data estimation using neural networks." Information Sciences 237 (2013): 49-58.

Denoising Autoencoders



$$z = \sigma(W_e \tilde{x} + b_e)$$

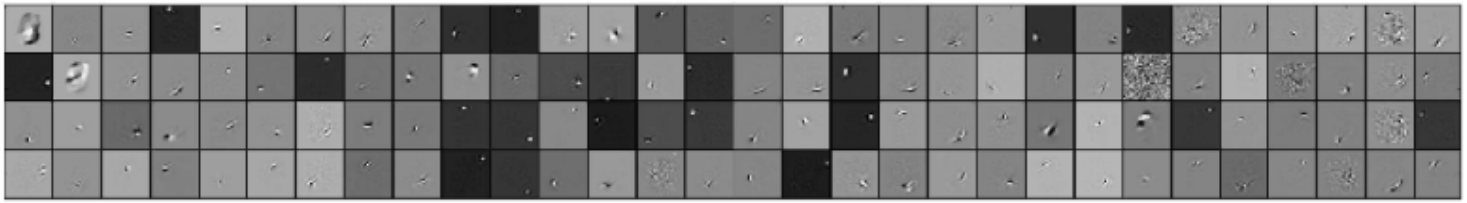
$$x' = \sigma(W_d z + b_d)$$

$$L(x, x') = \sum (x - x')^2$$

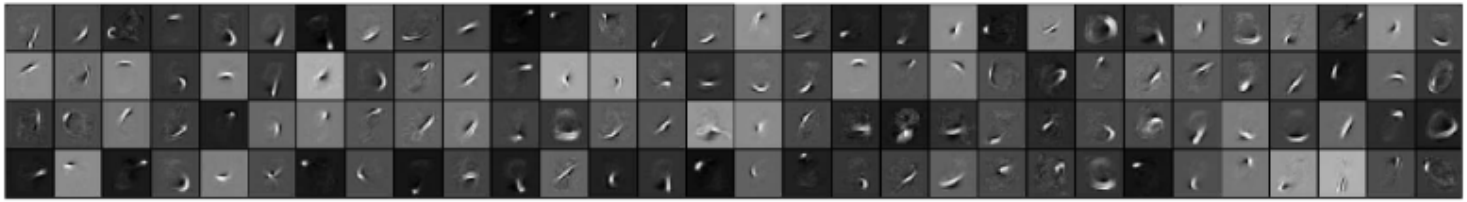
---

http://www.opendeep.org/v0.0.5/docs/tutorial-your-first-model

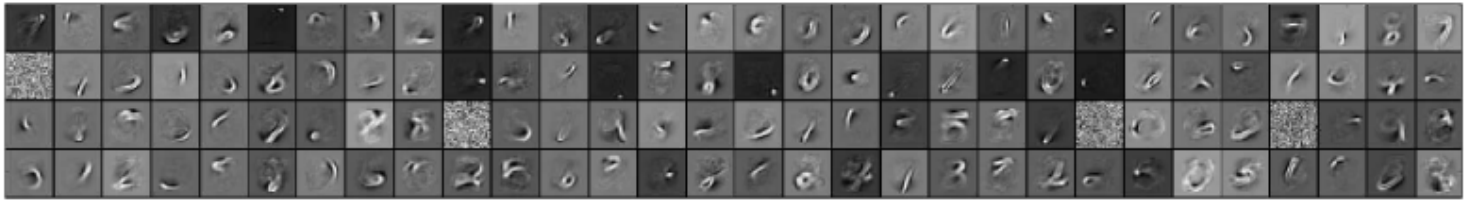
Sparse Autoencoders



(a) $k = 70$



(b) $k = 40$



(c) $k = 25$

(d) $k = 10$

Makhzani, Alireza, and Brendan Frey. "K-sparse autoencoders." ICLR 2014.

Word Embeddings

Before Word Embeddings: Bag of Words

D1: "Patient has a fever, otherwise not sick"

D2: "Patient very sick, not a fever though"

	patient	fever	very	sick	not	otherwise
D1	1	1	0	1	1	1
D2	1	1	1	1	1	0

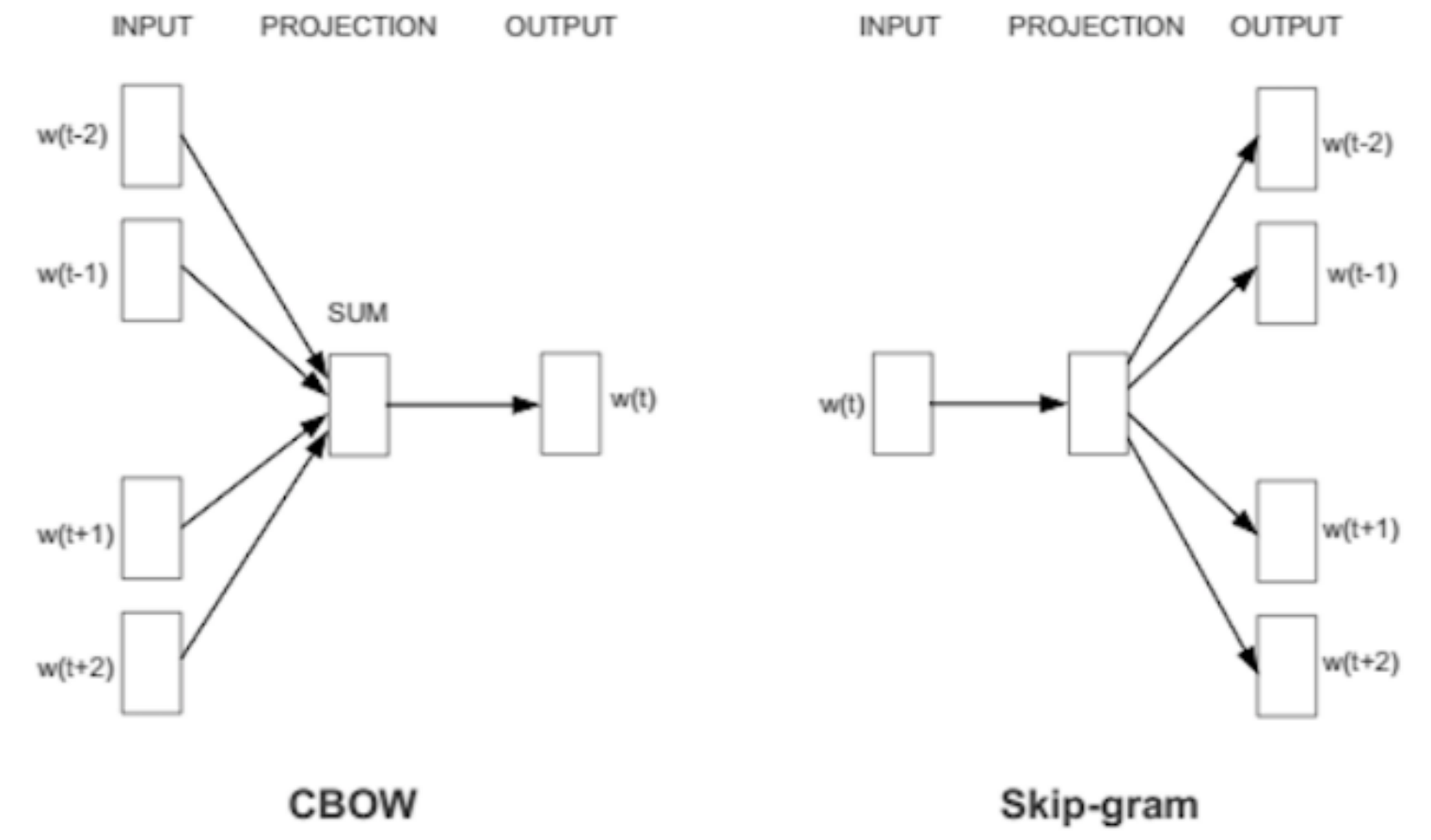
Before Word Embeddings: Bag of Words

D1: "Patient has a fever, otherwise not sick"

D2: "Patient very sick, not a fever though"

	patient	fever	very	sick	not	otherwise	...
D1	1	1	0	1	1	1	0
D2	1	1	1	1	1	0	0

Word Embeddings: Word2Vec



Tomas Mikolov et al. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.

Medical Word Embeddings

TOP 10 SIMILAR WORDS TO 'DIABETES' USING WORD2VEC MODELS

PubMed	MedHelp	Wikipedia
T2DM	diabeties	chemotherapy
prediabetes	diabetis	asthma
mellitus	lupus	schizophrenia
T1DM	Diabetes	hypertension
T2D	RA	radiotherapy
IDDM	diabetese	neonatal
DM2T	anemia	diabetic
DMT2	diabetic	infertility
DM2	diabites	malaria
T1D	hypoglycemia	prognosis

Huang, Jian, Keyang Xu, and VG Vinod Vydiswaran. "Analyzing Multiple Medical Corpora Using Word Embedding." Healthcare Informatics (ICHI), 2016 IEEE International Conference on. IEEE, 2016.

Medical Word Embeddings

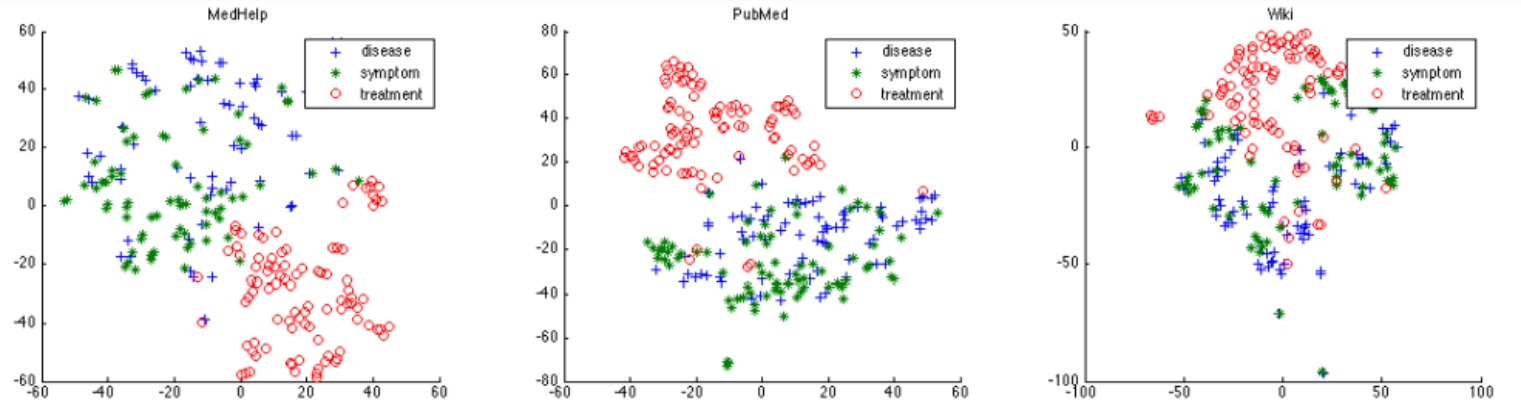


Fig. 2. t-SNE projection of Ranker word embedding

Huang, Jian, Keyang Xu, and VG Vinod Vydiswaran. "Analyzing Multiple Medical Corpora Using Word Embedding." Healthcare Informatics (ICHI), 2016 IEEE International Conference on. IEEE, 2016.

Activity: Dimensionality Reduction

Student Features

- Major
- Dorm/Residence
- Favorite TV show
- Age
- Hobby
- Favorite course at UMass
- Number of siblings
- Programming language
- Home state/region
- Best dining hall meal