CISC 372
Text Analytic IIII
EM & LDA

https://giphy.com/gifs/kid-costume-transformer-R9zXHWAHyTjnq

Last lectures

- Expectation Maximization
  - Mode-based clustering
  - Iterative approach for optimization
  - Domain knowledge integration

Today

- EM
  - An example
- LDA & Gibb Sampling

Coin Flipping

H T H H H H H H T H H T H
H H H H H T H H H T H T H
H H H H H H H H H H T H H
H T H H T H H H H H H T H

Binominal Distribution: $B(n, p)$

https://giphy.com/gifs/Friends-episode-12-friends-tv-the-one-with-embryos-Y4cMaANkENnOxDEPe6

Coin Flipping using Two coins

H T H H H H H H T H H T H
T T T T H H T T T T H T H
H H H H H T H H H T H T H
T T T T T T T H T H T H T
H T T T T T T T T H T H T
T T T T T T H T T T T T H
H H H H H H H H H H T H H
T H T H T T T T T T T T H
H T H H T H H H H H H T H

Flipping coins

- there are two coins
- And each of them has a different chance of observing head/tail

Binominal Distribution: $B(n_1, p_1)$ $B(n_2, p_2)$

https://giphy.com/gifs/Friends-episode-12-friends-tv-the-one-with-embryos-Y4cMaANkENnOxDEPe6

What if we don't know which coin was used in which flip??

H T H H H H H H T H H T H
T T T T H H T T T T H T H
H H H H H T H H H T H T H
T T T T T T T H T H T H T
H T T T T T T T T H T H T
T T T T T T H T T T T T H
H H H H H H H H H H T H H
T H T H T T T T T T T T H
H T H H T H H H H H H T H

Flipping coins

- there are two coins
- And each of them has a different chance of observing head/tail

Binominal Distribution: $B(n_1, p_1)$ $B(n_2, p_2)$

https://giphy.com/gifs/Friends-episode-12-friends-tv-the-one-with-embryos-Y4cMaANkENnOxDEPe6

What if we don't know which coin was used in which flip??

H T H H H H H H T H H T H
T T T T H H T T T T H T H
H H H H H T H H H T H T H
T T T T T T T H T H T H T
H T T T T T T T T H T H T
T T T T T T H T T T T T H
H H H H H H H H H H T H H
T H T H T T T T T T T T H
H T H H T H H H H H H T H

$$x = \{x_1, x_2, \dots x_m\}$$

$$z_i \in \{0,1\}$$

Flipping coins

- there are two coins
- And each of them has a different chance of observing head/tail

Binominal Distribution: $B(n_1, p_1)$ $B(n_2, p_2)$

https://giphy.com/gifs/Friends-episode-12-friends-tv-the-one-with-embryos-Y4cMaANkENnOxDEPe6

What if we don't know which coin was used in which flip??

$$x = \{x_1, x_2, \dots x_m\}$$
$$z_i \in [0,1]$$

H T H H H H H H T H H T H
T T T T H H T T T T H T H
H H H H H T H H H T H T H
T T T T T T T H T H T H T
H T T T T T T T T H T H T
T T T T T T H T T T T T H
H H H H H H H H H H T H H
T H T H T T T T T T T T H
H T H H T H H H H H H T H

EM Algorithm:

1. Randomly assign each flip to a coin (random z)
2. Estimate the parameter for that coin w.r.t. n & p
3. Update the flip assignment Z based on the updated parameters for these two coins
4. Go back to 2, until converge

Flipping coins

- there are two coins
- And each of them has a different chance of observing head/tail

Binominal Distribution: $B(n_1, p_1)$ $B(n_2, p_2)$

https://giphy.com/gifs/Friends-episode-12-friends-tv-the-one-with-embryos-Y4cMaANkENnOxDEPe6

LDA – Document Generation (no labels)

How would you generate (write) a document??

https://www.google.com/url?sa=i&url=https%3A%2F%2Fgiphy.com%2Fexplore%2Fwriting&psig=AOvVaw0o9_6cyMD_wS5Z7ulvHsPF&ust=1583516912871000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNCsiKzyg-gCFQAAAAAdAAAAABAY

LDA – Document Generation

1. What it is about??
   - Pick a set (mixture of topics)

2. Start writing word-by-word
   - Each word is generated one-by-one
   - Which one to use?
   - Depends on the what is the topic to be covered

https://www.google.com/url?sa=i&url=https%3A%2F%2Fgiphy.com%2Fexplore%2Fwriting&psig=AOvVaw0o9_6cyMD_wS5Z7ulvHsPF&ust=1583516912871000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNCsiKzyg-gCFQAAAAAdAAAAABAY

LDA – Document Generation

- i denotes the index of a document

$$i = \{1, \cdots, N_D\}$$

- v denotes the index of a word

$$v = \{1, \cdots, N_w\}$$

- k denotes the index of a topic

$$k = \{1, \cdots, N_k\}$$

LDA – Document Generation

- To generate a document, we need a mixture of topics:

$$\pi_i \sim \mathrm{Dir}(\pi_i \mid \alpha) \qquad i = \{1, \cdots, N_D\}$$

- How about topic? Topic itself is a mixture of words.

$$\mathbf{b}_k \sim \mathrm{Dir}(\mathbf{b}_k \mid \gamma) \qquad k = \{1, \cdots, N_k\}$$

- Given a document, to generate a word, we need to know what topics to be used

$$z_{iv} \sim \mathrm{Cat}(z_{iv} \mid \pi_i) \qquad v = \{1, \cdots, N_w\}$$

- Then given the topic for this word under this document, we can generate the actual word (observation)

$$y_{iv} \sim \mathrm{Cat}(y_{iv} \mid z_{iv} = k, \mathbf{B})$$

LDA – Document Generation

- Learning – Gibbs Sampler
  - Bayesian Inference
  - Iterative approach
  - Update the parameters/assignment for each variable at a time
  - A variants of generalized EM algorithm
  - Hyperparameter: number of topics

  - Nice implementation:
    - Gensim
    - https://radimrehurek.com/gensim/models/ldamodel.html
  - Example:
    - https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html

LDA – After training

- A document is a distribution over topics:

$$\pi_i \sim \mathrm{Dir}(\pi_i \mid \alpha) \qquad i = \{1, \cdots, N_D\}$$

- A topic is a distribution over words

$$\mathbf{b}_k \sim \mathrm{Dir}(\mathbf{b}_k \mid \gamma) \qquad k = \{1, \cdots, N_k\}$$

LDA – Document Generation

- How about topic? Topic itself is a mixture of words.

$$\mathbf{b}_k \sim \mathrm{Dir}(\mathbf{b}_k \mid \gamma) \qquad k = \{1, \cdots, N_k\}$$

Topic 1		Topic 2		Topic 3	
term	**weight**	**term**	**weight**	**term**	**weight**
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

https://databricks.com/blog/2015/03/25/topic-modeling-with-lda-mllib-meets-graphx.html