Topic modeling

Jae Yeon Kim

12 April, 2019

Review

- Text data is high-dimensional data (N $<$ P)
- 1000 common English words for 30-word tweets: 1000^{30} similar to N of atoms in the universe (Gentzknow, Kelly, and Taddy)

Bag of Words

- Easy and useful
- Ignored grammatical structures and rich interactions among words
- Words $\rightarrow$ Meaning

Grimmer and Stewart (2013)

1. All models are wrong but some are useful. (Think of DGP)
2. All models augment humans, not replace them. (Why domain knowledge matters.)
3. No globally best method. (Needs experience.)
4. Validate, validate, and validate.

We're going to learn ...

- Dictionary-based
- Unsupervised (attributes are latent): topic modeling
- Supervised (attributes are observed): text classification

Topic modeling

- Good at capturing **approximated** topics (= issues, themes)
- Co-occurence of words (clustering)
- Words lie on a lower dimensional space (dimension reduction)

Clustering

- Co-occurence of words
- Giving more contexts to how words are used

Dimension reduction

- Topics as **distributions** of words
- Documentions as **distributions** of topics
- What distributions?
  - Probability
  - Multinominal (e.g., Latent Dirichlet Distribution)