Lecture 3 Topic Models

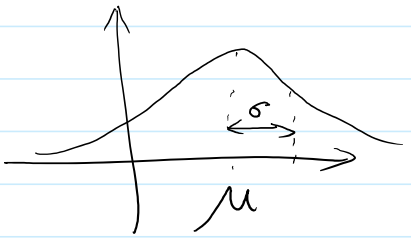
Friday, September 2, 2016 4:23 PM

Topic Models

Goal: uncover thematic structure

Approach: Bayesian, probabilistic model.

- Probabilistic Model
  - Def. (informal) A prob. model is a class of distributions $D(\theta)$ $(\theta \in (-))$, θ is called the parameter.
    $D(\theta)$ is a distribution.
  - Def. parameter estimation: Given samples from $D(\theta)$, estimate θ
  - Example: Gaussian distribution $N(\mu, \sigma^2)$
  - parameter estimation: easy (sample mean)



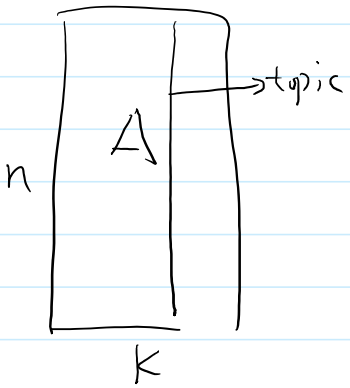
Note: cannot get exact μ
only hope for $\frac{1}{\sqrt{\#\text{samples}}}$

- Topic models
  - plan: Define a topic model (hard!)
    (distribution of documents, parameters specify topics)
    Estimate the parameters.
  - Simplifying assumptions. (for the model)
    - bag of words: ordering does not matter.
      Consequence: doc $\Leftrightarrow$ frequency of words

(distribution)

- topic structure
  topic := distribution over words
  document = (linear) mixture of topics.

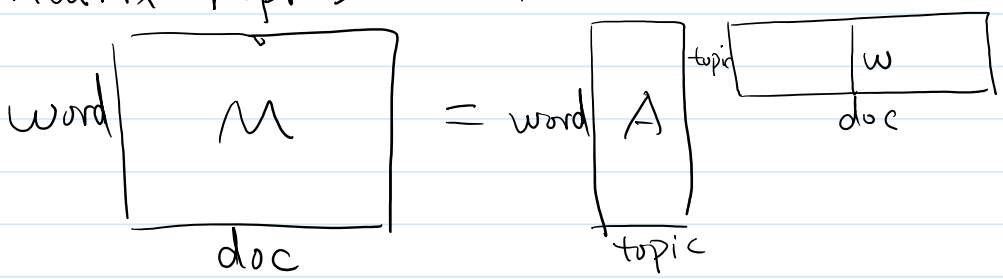
- "Algorithm" for generating documents
  - Given: $A \in \mathbb{R}^{n \times k}$ "word-topic" matrix



  - pick $w \sim D_{topic}$
    $(w \in \mathbb{R}^k, w_i \geq 0, \sum w_i = 1)$
  - pick length N
  - for $i = 1$ to N
    - pick topic $t_i \sim w$
    - pick word $z_i \sim A_{:,t_i}$

independent / bag-of-words

- matrix representation



- different choices for D_{topic}
  - pure documents:
    $w = e_i$ with probability α_i
  - Latent Dirichlet Allocation (Blei, Ng, Jordan '04)

$$Pr[w] \propto \prod_{i=1}^{k} w_i^{\alpha_i - 1} \qquad \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$$

Property: w is sparse (when $\alpha_i \ll 1$)
$(\alpha_0 = \sum_{i=1}^{k} \alpha_i$ is roughly # large entries)

  - Correlated Topic Model (Blei, Lafferty), ...

- Using NMF for topic models
  - separability
    - Recall: $\forall i = 1, \ldots, k$, there is a row r_i such that $A_{r_i, i} > 0$, $\forall j \neq i \; A_{r_i, j} = 0$.
    - For every topic, there is a word ("anchor-word") that only appears in this topic. (also: has reasonable prob.)

$$\boxed{M^T} = \boxed{W^T} \boxed{A^T}$$

For every topic, there is a pure document.

(Q: why do we use "anchor-words" instead of "pure doc"?)

  - sampling noise

Example: $\left(\frac{1}{10000}, \ldots, \frac{1}{10000}\right)$, 100 samples $\Rightarrow$ vector with ≤ 100 nonzero entry

ℓ_1 distance ≈ 2!

too large to be called "perturbation"

- idea: reducing noise with more documents

- word-word correlation matrix

$$Q_{i,j} = \Pr[z_1 = i, z_2 = j]$$

Claim: $Q = ARA^T$, $R = \mathbb{E}_{w \sim D_{topic}}[ww^T]$,

Proof: z_1, z_2 independent conditioned on w

$$Q = \mathbb{E}[z_1 z_2^T]$$

(abuse notation: z_i = index of ith word = indicator vector e_{z_i})

$$= \mathbb{E}_w[\mathbb{E}[z_1 z_2^T | w]]$$

$$= \mathbb{E}_w[\mathbb{E}[z_1 | w] \mathbb{E}[z_2^T | w]]$$

$$= \mathbb{E}_w[A_{w,\cdot} w w^T A^T] = ARA^T$$

□

$= \mathbb{E}_w[\mathbb{E}[c_1|w]\mathbb{E}[c_2|w]]$

$= \mathbb{E}_w[AwwA^T] = A\underbrace{R}_{W^*}A^T$ □

words in vocabulary fixed, # doc → infinity

⟹ can estimate Q

(recall: why not "pure doc"? hard to reduce noise)

- High level Alg.

scaling may not be correct

① estimate Q matrix

② apply separable NMF to get $Q = \tilde{A}\tilde{W}^*$

③ fix the scaling to find A.

$$\bar{Q}_{i,j} = \frac{Q_{i,j}}{\|Q_{i,:}\|_1} = \Pr[z_2 = j \mid z_1 = i]$$

$$\Pr[z_2 = j \mid z_1 = i] = \sum_{l=1}^{k} \Pr[z_2 = j \mid t_1 = l]\Pr[t_1 = l \mid z_1 = i]$$

(convex combination)

$$\boxed{\bar{Q}} = \boxed{\bar{A}}\ \boxed{\bar{W}^*}$$

$\bar{W}^*$: $\Pr[z_2 = j \mid t_1 = l]$

$\bar{A}$: $\Pr[t_1 = l \mid z_1 = i]$

r_l: anchor word for topic l, then

$$\Pr[z_2 = j \mid z_1 = r_l] = \Pr[z_2 = j \mid t_1 = l]$$

NMF: $\bar{A}_{i,l} = \Pr[t_1 = l \mid z_1 = i]$

want: $A_{i,l} = \Pr[z_1 = i \mid t_1 = l]$