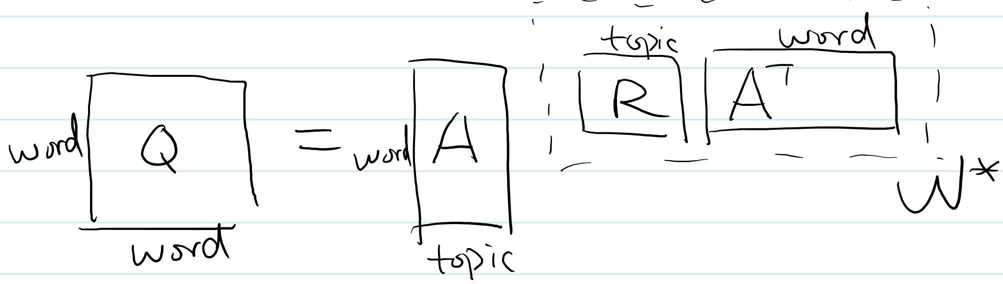
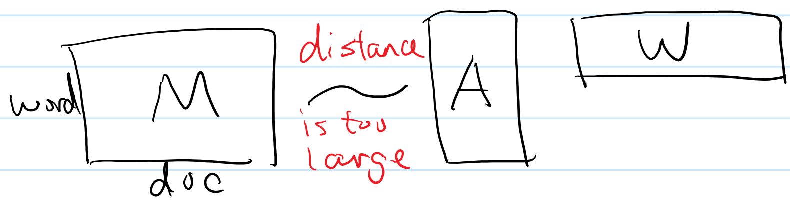
Lecture 4 Topic Modeling in Practice

Monday, September 5, 2016 9:24 PM

- Recap: Using NMF for topic modeling
  - estimate Q: $Q_{i,j} = \Pr[z_1 = i, z_2 = j]$
  - NMF: $\bar{Q} = \bar{A}\,\bar{W^*}$ $(W^* = RA^T)$
  - Fix normalization in $\bar{A}$.





- Problem 1. how to estimate Q?

Lemma: Let u be the word count vector for a length N document, then $\mathbb{E}[uu^T - \mathrm{diag}(u)] = N(N-1)Q$

Proof. $u = \sum_{i=1}^{N} z_i \leftarrow$ i-th word

we know $\mathbb{E}[z_i | w] = Aw \leftarrow$ mixture of topics

$\mathbb{E}[z_1 z_2^T] = ARA^T = Q$ (see last lecture)

but $\mathbb{E}[z_i z_j^T] = \mathbb{E}[z_1 z_2^T]$ (symmetry)

So $\mathbb{E}[uu^T] = \mathbb{E}[(\sum_{i=1}^{N} z_i)(\sum_{j=1}^{N} z_j)^T]$

$= \sum_{i \neq j} \mathbb{E}[z_i z_j^T] + \sum_{i=1}^{N} \mathbb{E}[z_i z_i^T]$

$$= \sum_{i \neq j} E[z_i z_j] + \sum_{i=1} E[z_i c_i]$$

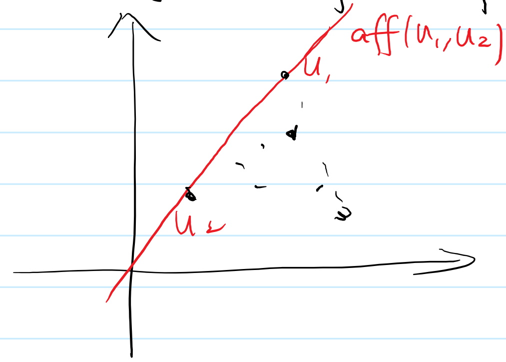
diag(u)

$$E[uu^T - \mathrm{diag}(u)] = N(N-1)Q \qquad \square$$

Why is this better? use more samples, get better estimate of Q.

- Problem 2: algorithm for NMF is slow.

  Fast algorithm: repeatedly look for the farthest point.



Given: $v_1, v_2, \cdots, v_n$

Find: vertices $u_1, u_2, \cdots, u_K$

u_1 = vector with largest norm

u_2 = vector farthest away from u_1

for $i = 3$ to n

  u_i = vector farthest to $\mathrm{aff}(u_1, u_2, \cdots, u_{i-1})$

$$\mathrm{aff}(u_1, u_2, \cdots, u_K) = \left\{ u \,\middle|\, u = \sum_{i=1}^{K} c_i u_i,\ \sum_{i=1}^{K} c_i = 1 \right\}$$

recall $\mathrm{conv}(u_1, u_2, \cdots, u_K) = \mathrm{cone}(u_1, \cdots, u_K) \cap \mathrm{aff}(u_1, \cdots, u_K)$

- Inference Problem

  Given A, document u, how to find w?

  - Observations:
    - cannot hope to get exact w
    - in practice, try to approximate posterior dist.

      prior $\Pr[w] \propto \prod_{i=1}^{K} w_i^{\alpha_i - 1}$

      $w \mid A, u$
    - can also try to find a matrix B such that

      $BA = I$,

and use $\hat{w} = \frac{Bu}{|u|_1}$

$$\mathbb{E}[\hat{w}] = \frac{B\,\mathbb{E}[u]}{N} = \frac{B \cdot N \cdot Aw}{N}$$

$$= BAw$$

$$= w$$

Problem: how to control variance?

See http://128.84.21.199/abs/1605.08491