COS 597C: New Directions in Theoretical Machine Learning	Fall 2017

Lecture 9: 12 October 2017

Lecturer: Sanjeev Arora | *Scribe: Suqi Liu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

9.1 Tensor Decomposition

Let T be a tensor of order 3 with each entry

$$T_{ijk} = \Pr\{i, j, k \text{ appear in some document}\}.$$

If there are n words in the vocabulary, it takes $\mathcal{O}(n^3)$ time to set up T.

Here we restate the model we are interested in. Each of the k topics is identified with a distribution over words, represented by n-dimensional vectors

$$\begin{bmatrix} | \\ A_1 \\ | \end{bmatrix} \cdots \begin{bmatrix} | \\ A_k \\ | \end{bmatrix}.$$

Each document is generated by picking its topic proportions from a distribution, which can also be viewed as a vector x in k-dimensions, where the value in coordinate i represents the proportion of topic i present in the document. Finally, each word is independently sampled according to the distribution represented by $\sum x_i A_i$, where $\sum x_i = 1$.

This formulation is very general and includes most widely used probabilistic topic models. When the vector $\bar{x}$ is sampled from Dirichlet distribution Dir, it becomes the *Latent Dirichlet Allocation* (LDA) model.

9.2 The Method of Moments

Let us describe the approach in the context of topic modeling, working with second order moments. Let M be a $n \times n$ matrix, the entry of which

$$M_{ij} = \Pr\{i, j \text{ are first two words in the document}\}.$$

denotes the probability that the first and second words in a randomly generated document are word i and j respectively.

Claim: $M = A\mathbb{E}[xx^\top]A^\top$.

Proof.

$$\begin{aligned} M_{ij} &= \mathbb{E}[p_i p_j] \\ &= \mathbb{E}[(A^{(i)}x)(A^{(j)}x)] \\ &= A^{(i)}\mathbb{E}[xx^\top]{A^{(j)}}^\top. \end{aligned}$$

□

9.3 Nonnegative Matrix Factorization (NMF) [Lee, Seung '99]

In the *Nonnegative Matrix Factorization* (NMF) problem we are given an $n \times m$ nonnegative matrix M and an integer $r > 0$. Our goal is to express M as AB where A and B are nonnegative matrices of size $n \times r$ and $r \times m$ respectively. In some applications, it makes sense to ask instead for the product AB to approximate M – i.e. (approximately) minimize $\|M - AB\|_F$ where $\|\|_F$ denotes the Frobenius norm; we refer to this as *Approximate NMF*.

Trivial heuristic in this case is Alternating Minimization.

- Fix A, find best B.
- Fix B, find best A.
- Repeat.

Issues:

(i) If the columns of A are not linearly independent then Radons Lemma implies that this expression can be far from unique.

(ii) The NMF problem is NP-hard when r is large.

(iii) [AGKM '12] Fixed parameter hard, require n^r time assuming complexity assumptions. There is also a matching n^r algorithm.

9.4 The Anchor Word Algorithm

"Anchor words" are specialized words that are specific to a single topic. The condition of separability requires that each topic contains at least one (unknown) anchor word. That is, $\forall$ topics A_i, $\exists$ a word j that appears only in that topic, "anchur word for topic i".

$$\begin{matrix} A_1 & A_2 & & A_k \\ \begin{bmatrix} * \\ \\ \\ \\ \vdots \end{bmatrix} & \begin{bmatrix} \\ * \\ \\ \\ \vdots \end{bmatrix} & \ddots & \begin{bmatrix} \\ \\ \\ * \\ \vdots \end{bmatrix} \end{matrix}$$

Let $\overline{M}$ be the row normalized version of M, i.e. each row of $\overline{M}$ sums up to 1. It follows that

$$\overline{M}_{ij} = \Pr\{\text{2nd word is } j \text{ given that first word was } i\}$$

Claim: All rows of $\overline{M}$ are convex combinations of rows corresponding to anchor words.

$$\overline{M} = \left(\overline{A}\right)(B)$$

where $\overline{A}$ is row normalized.

$$\overset{\overline{M}}{\begin{bmatrix} \\ \underline{\quad\quad i \quad\quad} \\ \\ \end{bmatrix}} = \overset{\overline{A}}{\begin{bmatrix} \\ \underline{\quad\quad i \quad\quad} \\ \\ \end{bmatrix}} \begin{bmatrix} & B & \\ & & \\ & & \end{bmatrix}$$

Let $B_1, \ldots, B_k$ denote anchor rows. All other rows can be written as $\sum \lambda_i B_i, \sum_i \lambda_i = 1$, which is in the simplex determined by anchor rows.

The anchor word algorithm

Alg. 1

Take a row. Try to write it as convex combination of other rows. If not possible, declare it as one of the anchor rows (i.e. corresponding word i as an anchor word).

Alg. 2

For $i = 1, \ldots, k$, find row furthest from subspace spanned by first i rows you've identified.

9.5 Pointwise Mutual Information (PMI)

Diagnose which disease(s) a patient may have by observing the symptoms he/she exhibits. Suppose there are n symptoms, denoted by s_i and m diseases, which is latent variable denoted by d_j.

$$\Pr\{s_i \text{ absent}\} = 1 - \exp(-w^{(i)} \cdot d)$$

Can you infer $\overline{w}$ given patient symptom data?

$$PMI(x, y) = \lg \frac{P(xy)}{P(x)P(y)} \qquad \text{"NOISY OR"}$$

$$PMI_{ij} = PMI(1 - s_i, 1 - s_j) = \sum_i w^{(i)} {w^{(i)}}^\top + \rho \sum_i w^{(i)} \otimes w^{(i)}$$

9.6 Robust Jennrich (Guest lecture by Tengyu Ma)

Given $T = \sum_{i=1}^d a_i \otimes b_i \otimes c_i + E$
$a_i, b_i, c_i \in \mathbb{R}^d$
a_i's are orthogonal
b_i's are orthogonal
c_i's are orthogonal
Goals: to recover $\{(a_i, b_i, c_i)\}$

Jennrich ($E = 0$)

$$\begin{aligned} M &= (g \otimes I \otimes I)T \\ &= \left(\sum_{i=1}^{d} g_i T_{ijk} \right)_{\substack{j=1,\dots,d \\ k=1,\dots,d}} \\ &= \sum_{i=1}^{d} (g^\top a_i) b_i c_i^\top \\ &= \begin{bmatrix} b_1 & \dots & b_d \end{bmatrix} \begin{bmatrix} g^\top a_1 & & \\ & \ddots & \\ & & g^\top a_d \end{bmatrix} \begin{bmatrix} c_1^\top \\ \vdots \\ c_d^\top \end{bmatrix} \end{aligned}$$

$((A \otimes B)(C \otimes D)) = AC \otimes BD)$

Robust Jennrich

$S = \emptyset$
For $s = 1$ to $O(d^{1+\delta} \log d)$

$g \sim N(0, I_{d \times d})$

$M = (g^\top \otimes I \otimes I)T$

$v, w =$ left and right top s.v. of M

$u = (I \otimes v^\top \otimes w^\top)T$

check if (u, v, w) are good by $\sum u_i v_j w_k T_{ijk} \geq 1 - \epsilon$

add $(u, v, w) \in S$ if good

$$M = \underbrace{\sum \langle g, a_i \rangle b_i c_i^\top}_{\overline{M}} + \underbrace{(g \otimes I \otimes I)E}_{E'}$$

w.p. $\frac{1}{d^{1-\delta}}$ $\langle g, a_i \rangle$ is the largest
$\langle g, a_i \rangle \geq \underbrace{\left(\max_{j \neq i} \langle g, a_j \rangle \right)}_{\approx \sqrt{\log d}} * (1 + \delta)$
$(\langle g, a_1 \rangle, \dots, \langle g, a_d \rangle$ i.i.d. normal)
eigengap in $\overline{M}$ is $\geq \delta \sqrt{\log d}$
$\Rightarrow \|$Top l.s.v. of $M - b_1\| \leq \frac{\|E'\|_{sp}}{\delta \sqrt{\log d}}$ (Wedin's)

$$\begin{aligned} \|E\|_{\{1\}\{2,3\}} &= \|E \text{ viewed as } d \times d^2\|_{sp} \\ &= \max_{\substack{u \in \mathbb{R}^d \\ v \in \mathbb{R}^{d \times d}}} \sum_{i,jk} u_i v_{jk} T_{ijk} \end{aligned}$$

Lem (Ma Shi Steurer)
With high probability

$$\|(g \otimes I \otimes I)T\|_{sp} \leq \sqrt{\log d} \max\{\|E\|_{\{2,3\}\{1\}}, \|E\|_{\{1,3\}\{2\}}\}$$