Tensor Methods in Machine Learning

Evgeny Kurbatsky

June 3, 2016
Rank 1 matrix

All rows and columns linearly dependent

\[ M = ab^\top = a \otimes b \]

\[ M[i, j] = a[i]b[j] \]
Low rank matrix

\[ M = a_1 \otimes b_1 + a_2 \otimes b_2 \]
Rank 1 tensor

\[ T = a \otimes b \otimes c \]

\[ T[i, j, k] = a[i]b[j]c[k] \]
Low rank tensor

\[ T \approx \sum_i a_i \otimes b_i \otimes c_i \]
Low rank tensor decomposition problem

- Have unique solution when \( a_i, b_i, c_i \) linearly independent
- Hard to solve
- No guarantees for arbitrary tensor
- rank > tensor size
We have \( k \)-topics for documents.

Topics probabilities: \( \Pr[h = i] = w_i \)

Every word has independent probability assuming topic

\( x_t \) – indicator vector for \( t \)-th word in document
\( x_t = e_j \) if \( t \)-th word in document is \( j \)

\( \mu_i \) – word probabilities for topic \( i \)

\[ \mathbb{E}[x_t|h = i] = \mu_i \]
Topic model

\[ P[h = i] = w_i \]
\[ \mathbb{E}[x_t | h = i] = \mu_i \]
\[ \mathbb{E}[x_t] = \sum_{i=1}^{k} w_i \mu_i \]

\[ M_3 = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \]
\[ M_3 = \sum_{i=1}^{k} \mathbb{E}[x_1 \otimes x_2 \otimes x_3 | h = i] P[h = i] \]
\[ M_3 = \sum_{i=1}^{k} \mathbb{E}[x_1 | h = i] \otimes \mathbb{E}[x_2 | h = i] \otimes \mathbb{E}[x_3 | h = i] P[h = i] \]
\[ M_3 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \]
Multiview model

$h$ – has categorical distribution

$x_1, x_2, x_3$ – indicator variable (categorical distribution)

$a_i, b_i, c_i$ – vector of probabilities

\[
\mathbb{E}[x_1|h=i] = a_i \\
\mathbb{E}[x_2|h=i] = b_i \\
\mathbb{E}[x_3|h=i] = c_i
\]
Multiview model

\[ E[x_1|h = i] = a_i, \ E[x_2|h = i] = b_i, \ E[x_3|h = i] = c_i \]

\[ E[x_1 \otimes x_2 \otimes x_3] = \]
\[ = \sum_{i=1}^{k} E[x_1 \otimes x_2 \otimes x_3|h = i]P[h = i] = \]
\[ = \sum_{i=1}^{k} E[x_1|h = i] \otimes E[x_2|h = i] \otimes E[x_3|h = i]P[h = i] \]

\[ E[x_1 \otimes x_2 \otimes x_3] = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i \]
HMM as multiview model

$$
\Pr[x_1|h] = \sum_i \Pr[x_1|h_1] \Pr[h_2|h_1] \Pr[h_1]/\Pr[h_2]
$$

$$
\Pr[x_2|h_2] = \Pr[x_2|h_2]
$$

$$
\Pr[x_3|h_2] = \sum_i \Pr[x_3|h_3] \Pr[h_3|h_2]
$$
Find some statistics expressed as low rank tensor
Decompose tensor
...
Profit!
Resent results

- Beating the Perils of Non-Convexity: Guaranteed Training of Neural Networks using Tensor Methods
- Provable Tensor Methods for Learning Mixtures of Generalized Linear Models
- Learning Mixed Membership Community Models in Social Tagging Networks through Tensor Methods
- Training Input-Output Recurrent Neural Networks through Spectral Methods
<table>
<thead>
<tr>
<th>Tensor</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method of moments</td>
<td>Max likelihood</td>
</tr>
<tr>
<td>Not need all data</td>
<td>Need all data</td>
</tr>
<tr>
<td>Give optimal solution for some data</td>
<td>Give some solution for any data</td>
</tr>
<tr>
<td>For specific problems</td>
<td>Generic</td>
</tr>
</tbody>
</table>
Questions?
Tensor Decompositions for Learning Latent Variable Models
http://arxiv.org/abs/1210.7559
https://www.youtube.com/watch?v=NB5NHWzTRSY

A Spectral Algorithm for Learning Hidden Markov Models
http://arxiv.org/abs/0811.4413

Tensor Methods in Machine Learning
http://www.offconvex.org/2015/12/17/tensor-decompositions/

Tensor Decompositions and their Applications
https://www.youtube.com/watch?v=HcIN27_WqPU
We now consider a mixture of $k$ Gaussian distributions with spherical covariances.

\[ P[h = i] = w_i \]
\[ x = \mu_i + z, \quad z \sim N(0, \sigma I) \]
Gaussian mixture

Minimal eigenvalue of $E[x \otimes x] - E[x] \otimes E[x]$ is $\sigma^2$

$$\bar{\mu} = E[x] = \sum_i E[x|h = i]P[h = i]$$

$$E[x \otimes x] - E[x] \otimes E[x] = E[(x - E[x]) \otimes (x - E[x])] = E[(x - \bar{\mu}) \otimes (x - \bar{\mu})]$$

$$x = \mu_i + z, z \sim N(0, \sigma I)$$

$$E[(x - \bar{\mu}) \otimes (x - \bar{\mu})] = \sum_i w_i(E[(\mu_i + z - \bar{\mu}) \otimes (\mu_i + z - \bar{\mu})]) =$$

$$= \sum_i w_i((\mu_i - \bar{\mu}) \otimes (\mu_i - \bar{\mu}) + \sigma I) =$$

$$= \sum_i w_i((\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top) + \sigma I$$
Gaussian mixture

\[ \mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x] = \sum_i w_i((\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top) + \sigma I \]

\[ \tilde{z}^\top \sum_i w_i((\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top)z = \sum_i w_i(\tilde{z}^\top(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top z) = \]

\[ = \sum_i w_i((\bar{z}(\mu_i - \bar{\mu}))(\mu_i - \bar{\mu})z) \geq 0 \]

\[ \sum_i w_i((\mu_i - \bar{\mu})^\top(\mu_i - \bar{\mu})) - \text{positive semidefinite} \]

\[ \sum_i w_i(\mu_i - \bar{\mu}) = 0 \]

\[ \sum_i w_i((\mu_i - \bar{\mu})^\top(\mu_i - \bar{\mu})) \text{ rank } r < k \]

Minimal eigenvalue of \( \mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x] \) is \( \sigma^2 \)
Gaussian mixture

\[ M_2 = \mathbb{E}[x \otimes x] - \sigma^2 I \]

\[ M_2 = \sum_{i=1}^{k} w_i \otimes \mu_i \otimes \mu_i \]

\[ M_3 = \mathbb{E}[x \otimes x \otimes x] - \sigma^2 \sum_{i=1}^{k} (\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]) \]

\[ M_3 = \sum_{i=1}^{k} w_i \mu_i \otimes \mu_i \otimes \mu_i \]
Jenrich’s Algorithm

\[ T = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i \]

\[ T[\bullet, \bullet, u] = \sum_{j=1}^{n} T[\bullet, \bullet, i] u_j = \sum_{j=1}^{n} \sum_{i=1}^{k} w_i a_i \otimes b_i (c_{ij} u_j) = \]

\[ = \sum_{i=1}^{k} w_i a_i \otimes b_i (\sum_{j=1}^{n} c_{ij} u_j) = \sum_{i=1}^{k} w_i (u^T c_i) (a_i b_i^T) = \]

\[ = AD_u B^T \]
Jenrich’s Algorithm

\[ T = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i \]

\[ T_u = T[\bullet, \bullet, u] = AD_u B^\top \]

\[ T_v = T[\bullet, \bullet, v] = AD_v B^\top \]

\[ T_u T_v^+ = AD_u B^\top (AD_v B^\top)^+ = AD_u D_v^{-1} A^+ \]

\[ T_u^\top = BD_u A^\top \]

\[ T_v^\top = BD_v A^\top \]

\[ (T_u^\top)(T_v^\top)^+ = BD_u A^\top (BD_v A^\top)^+ = BD_u D_v^{-1} B^+ \]
Jenrich’s Algorithm

\[ T = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i \]

1. Pick two random vectors \( u, v \)
2. Compute \( T_u = \sum_{i}^{n} u_i \cdot T[\bullet, \bullet, i] = \sum_{i=1}^{k} w_i (u^\top c_i) a_i b_i^\top \)
3. Compute \( T_v = \sum_{i}^{n} v_i \cdot T[\bullet, \bullet, i] = \sum_{i=1}^{k} w_i (v^\top c_i) a_i b_i^\top \)
4. \( a_i \) are eigenvectors \( T_u(T_v)^+, b_i \) are eigenvectors \( (T_v^T)^+(T_u^T) \)
Another view on HMM

\( O \) – emission matrix, \( T \) – transition matrix, \( \pi \) – prior probabilities

\[
A_x = T_{\text{diag}}(O_{x,1}, \ldots, O_{x,m})
\]

\[
Pr[x_1, \ldots, x_t] = 1_m^\top A_{x_t} \ldots A_{x_2} A_{x_1} \pi
\]
Another view on HMM

\[
[P_1]_i = \Pr[x_1 = i] \\
[P_{2,1}]_{ij} = \Pr[x_2 = i, x_1 = j] \\
[P_{3,x,1}]_{ij} = \Pr[x_3 = i, x_2 = x, x_1 = j]
\]

\[
[P_1] = O\pi \\
[P_{2,1}] = OTdiag(\pi)O^T \\
[P_{3,x,1}] = OA_x Tdiag(\pi)O^T
\]
Observable HMM Representation

\[ b_1 = (U^T O) \pi \]
\[ b_\infty^T = 1_m (U^T O)^{-1} \]
\[ B_x = (U^T O) A_x (U^T O)^{-1} \]
\[ \Pr[x_1, ..., x_t] = b_\infty^T B_{x_t} ... B_{x_2} B_{x_1} b_1 = 1_m^T A_{x_t} ... A_{x_2} A_{x_1} \]