Tensor Methods in Machine Learning

Evgeny Kurbatsky

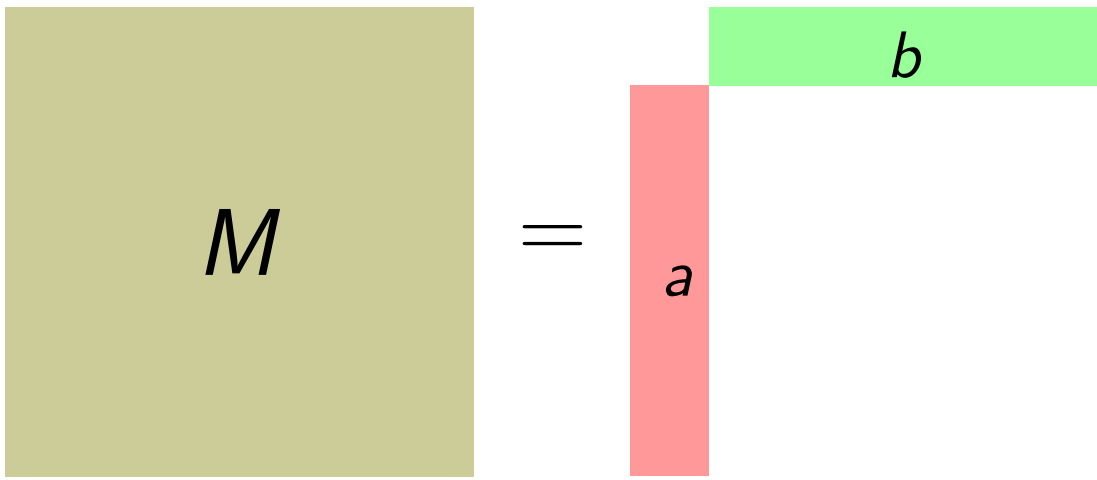
June 3, 2016

Rank 1 matrix

All rows and columns lineary dependent

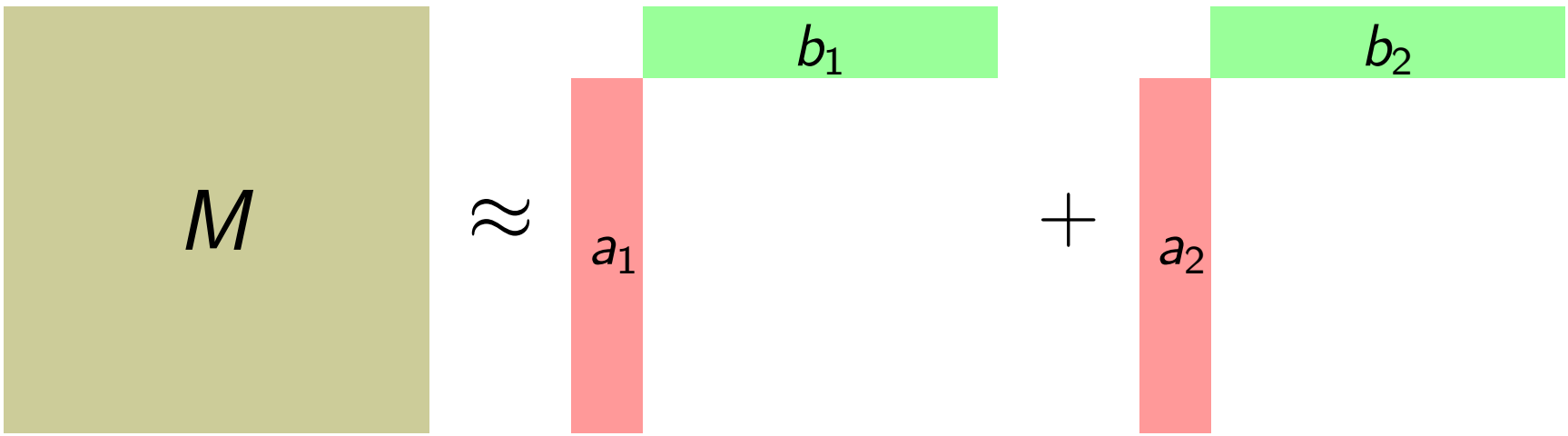
$$M = ab^\top = a \otimes b$$

$$M[i,j] = a[i]b[j]$$

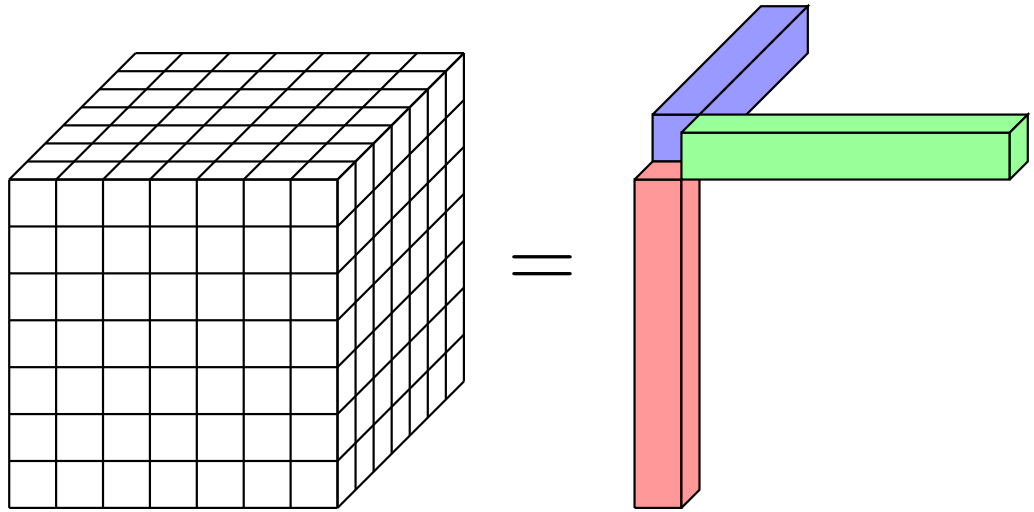


Low rank matrix

$$M = a_1 \otimes b_1 + a_2 \otimes b_2$$



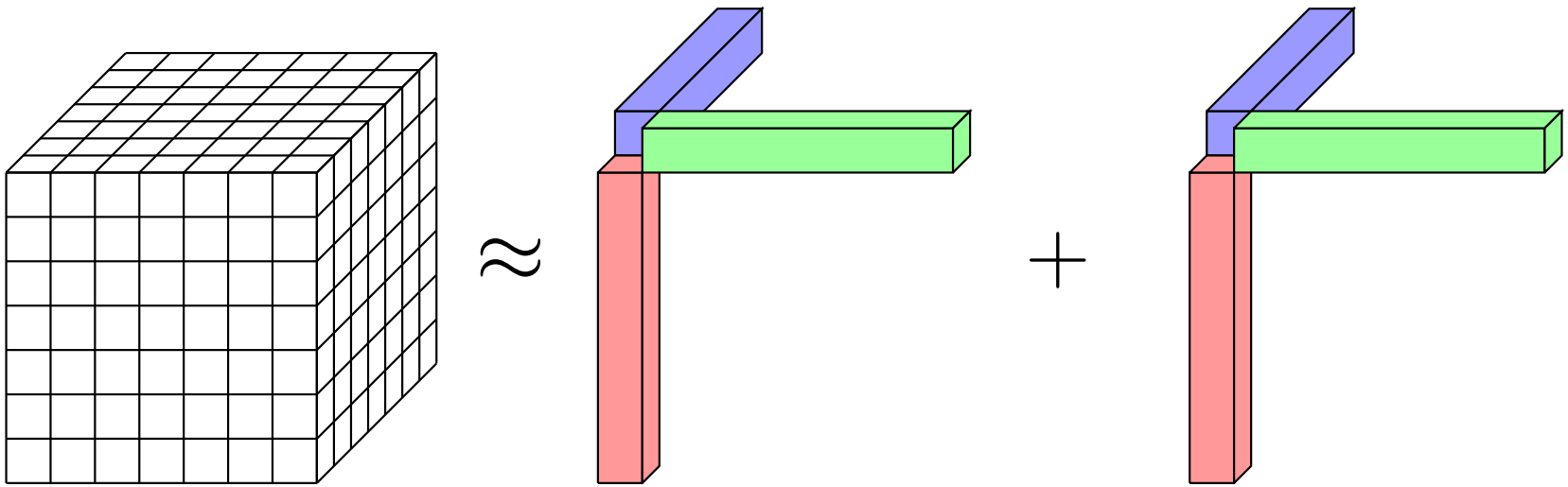
Rank 1 tesnor



$$T = a \otimes b \otimes c$$

$$T[i, j, k] = a[i]b[j]c[k]$$

Low rank tensor



$$T \approx \sum_i a_i \otimes b_i \otimes c_i$$

Low rank tensor depomposition problem

- Have uniqe solution when a_i, b_i, c_i lineary independent
- Hard to solve
- No garanties for arbitrary tensor
- rank $>$ tensor size

Topic model

- We have k-topics for documets.
- Topics probabilites: $\Pr[h = i] = w_i$
- 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
- μ_i – word probabilities for topic i

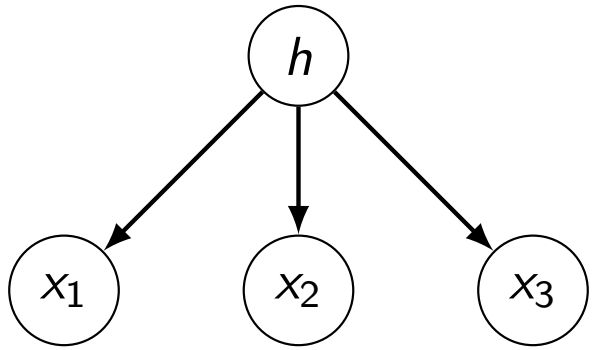
$$\mathbb{E}[x_t | h = i] = \mu_i$$

Topic model

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

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

Multiview model



h – has categorical distribution

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

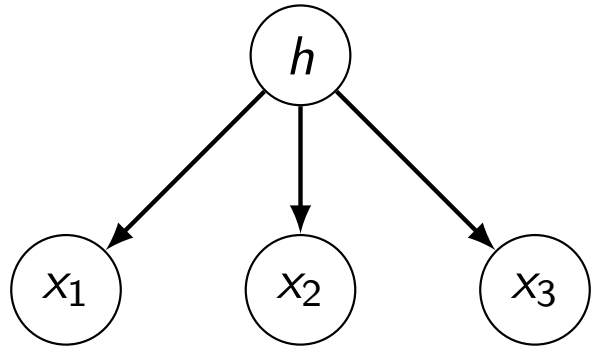
a_i, b_i, c_i – vector of probabilities

$$\mathbb{E}[x_1|h=i] = a_i$$

$$\mathbb{E}[x_2|h=i] = b_i$$

$$\mathbb{E}[x_3|h=i] = c_i$$

Multiview model



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

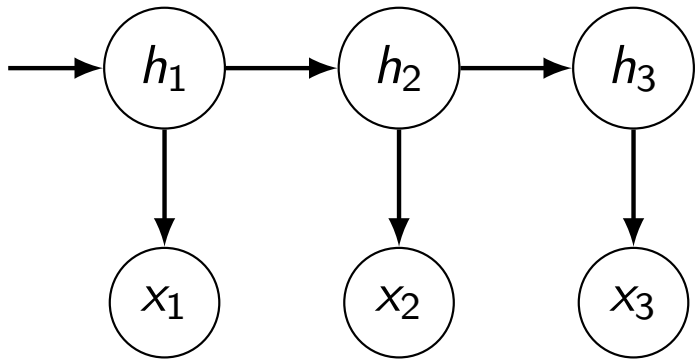
$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] =$$

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

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

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i$$

HMM as multiview model



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

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

$$\Pr[x_3|h_2] = \sum_i \Pr[x_3|h_3]\Pr[h_3|h_2]$$

Scheme

- 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

Tensor vs EM

Tensor	EM
Method of moments	Max likelihood
Not need all data	Need all data
Give optimal solution for some data	Give some solution for any data
For specific problemms	Generic

Questions?

Links

- 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
  http://people.csail.mit.edu/moitra/docs/Tensors.pdf
  https://www.youtube.com/watch?v=HcIN27_WqPU

Gaussian mixture

We now consider a mixture of k Gaussian distributions with spherical covariances.

$$\mathrm{P}[h = i] = w_i$$

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

Gaussian mixture

Minimal eigenvalue of $\mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x]$ is σ^2

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

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

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

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

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

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

Gaussian mixture

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

$$\bar{z}^\top \sum_i w_i((\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top)z = \sum_i w_i(\bar{z}^\top(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top z) =$$

$$= \sum_i w_i((\overline{z(\mu_i - \bar{\mu})})(\mu_i - \bar{\mu})z) \geq 0$$

$\sum_i w_i((\mu_i - \bar{\mu})^\top(\mu_i - \bar{\mu}))$ – positive semidefinite

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

$\sum_i w_i((\mu_i - \bar{\mu})^\top(\mu_i - \bar{\mu}))$ rank $r < k$
Minimal eigenvalue of $\mathbb{E}[x \otimes x] - \mathbb{E}[x] \otimes \mathbb{E}[x]$ is σ^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

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

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

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

$$= A D_u B^\top$$

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 $T_u = \sum_i^n u_i T[\bullet, \bullet, i] = \sum_{i=1}^{k} w_i (u^\top c_i) a_i b_i^\top$
3. Compute $T_v = \sum_i^n v_i T[\bullet, \bullet, i] = \sum_{i=1}^{k} w_i (v^\top c_i) a_i b_i^\top$
4. a_i are eigenvectors $T_u (T_v)^+$, b_i are eigenvectors $(T_v^\top)^+ (T_u^\top)$

Another view on HMM

O – emission matrix, T – transition matrix, π – prior probabilities

$$A_x = T\mathit{diag}(O_{x,1}, ..., O_{x,m})$$

$$\Pr[x_1, ..., x_t] = \mathbf{1}_m^\top A_{x_t} ... A_{x_2} A_{x_1} \pi$$

Another view on HMM

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

$$
\begin{aligned}
[P_1] &= O\pi \\
[P_{2,1}] &= OTdiag(\pi)O^\top \\
[P_{3,x,1}] &= OA_x Tdiag(\pi)O^\top
\end{aligned}
$$

Observable HMM Representation

$$
\begin{aligned}
b_1 &= (U^\top O)\pi \\
b_\infty^\top &= \mathbf{1}_m^\top (U^\top O)^{-1} \\
B_x &= (U^\top O) A_x (U^\top O)^{-1} \\
\Pr[x_1, ..., x_t] &= b_\infty^\top B_{x_t} ... B_{x_2} B_{x_1} b_1 = \mathbf{1}_m^\top A_{x_t} ... A_{x_2} A_{x_1}
\end{aligned}
$$