An Introduction to Spectral Learning

Hanxiao Liu

November 8, 2013

Outline

Preliminaries

$$X_1, \cdots, X_n \sim p\left(x;\theta\right),\ \theta = \left(\theta_1, \cdots \theta_m\right)^\top$$

$$\hat{\theta} = \hat{\theta}_n = w\left(X_1, \cdots, X_n\right)$$

- Maximum Likelihood Estimator (MLE)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \mathcal{L}\left(\theta\right)$$

- Bayes Estimator (BE)

$$\hat{\theta} = \mathbb{E}\left(\theta | X\right) = \frac{\int \theta p\left(x|\theta\right) \pi\left(\theta\right) d\theta}{\int p\left(x|\theta\right) \pi\left(\theta\right) d\theta}$$

Preliminaries

QUESTION

What makes a good estimator?

- MLE is consistent
- Both the MLE and BE have asymptotic normality

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \rightsquigarrow N\left(0, \frac{1}{I\left(\theta\right)}\right)$$

under mild (regularity) conditions

Can be computationally expensive

Preliminaries

Example (Gamma distribution)

$$p\left(x_i;\alpha,\theta\right)=\frac{1}{\Gamma\left(\alpha\right)\theta^{\alpha}}x_i^{\alpha-1}\exp\left(-\frac{x_i}{\theta}\right)$$

$$\mathcal{L}\left(\alpha,\theta\right)=\left(\frac{1}{\Gamma\left(\alpha\right)\theta^{\alpha}}\right)^n\left(\prod_{i=1}^{n}x_i\right)^{\alpha-1}\exp\left(-\frac{\sum_{i=1}^{n}x_i}{\theta}\right)$$

MLE is hard to compute due to the existence of $\Gamma\left(\alpha\right)$

Method of Moments

j-th theoretical moment, $j \in [k]$

$$\mu_j(\theta) := \mathbb{E}_\theta\left(X^j\right)$$

j-th sample moment, $j \in [k]$

$$M_j := \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Plug-in and solve the multivariate polynomial equations

$$M_j = \mu_j(\theta) \quad j \in [k]$$

sometimes can be recast as spectral decomposition

Method of Moments

Example (Gamma distribution)

$$p\left(x_i;\alpha,\theta\right)=\frac{1}{\Gamma\left(\alpha\right)\theta^{\alpha}}x_i^{\alpha-1}\exp\left(-\frac{x_i}{\theta}\right)$$

$$\overline{X}=\mathbb{E}\left(X_i\right)=\alpha\theta$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\mathit{Var}\left(X_i\right)=\alpha\theta^2$$

$$\Rightarrow\hat{\theta}=\frac{1}{n\overline{X}}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2,\ \hat{\alpha}=\frac{\overline{X}}{\hat{\theta}}=\frac{n\overline{X}^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}$$

Method of Moments

- lack guarantee about the solution
- high-order sample moments are hard to estimate

To reach a specified accuracy, the required sample size and computational cost is exponential in k (or n)!

Question

Could we recover the true θ from only low-order moments?

Question

Could we lower the sample requirement and computational complexity based on some (hopefully mild) assumptions?

Learning the Topic Models

- Papadimitriou et al. (2000)
  - Non-overlapping separation condition (strong)
- Anandkumar et al. (2012), MoM+SD
  - Full rank assumption (weak)
  - Multinomial Mixture, LDA
- Arora et al. (2012), MoM+NMF+LP
  - Anchor words (mild)
  - LDA, Correlated Topic Model
  - A more practical algorithm proposed in 2013

Learning the Topic Models

Suppose there are n documents, k hidden topics, d features

$$M = [\mu_1|\mu_2|\dots|\mu_k] \in R^{d\times k},\ \mu_j \in \Delta^{d-1}\ \forall j \in [k]$$

$$w = (w_1, \dots, w_k),\ w \in \Delta^{k-1}$$

$$P(h=j) = w_j \quad j \in [k]$$

For the v-th word in a document, $x_v \in \{e_1, \dots e_d\}$

$$P(x_v = e_i | h = j) = \mu_j^i, \quad j \in [k], i \in [d]$$

GOAL: Recover the M using low-order moments

Learning the Topic Models

Construct moment statistics

$$\text{Pairs}_{ij} := P\left(x_1 = e_i, x_2 = e_j\right)$$

$$\text{Triples}_{ij} := P\left(x_1 = e_i, x_2 = e_j, x_3 = e_t\right)$$

$$\text{Pair} = \mathbb{E}[x_1 \otimes x_2] \in R^{d\times d}$$

$$\text{Triples} = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \in R^{d\times d\times d}$$

- Empirical plug-ins i.e. $\hat{\text{Pairs}}$ and $\hat{\text{Triples}}$ could be obtained from data through a straightforward manner
- We want to establish some equivalence between the empirical moments and parameters of interest

Learning the Topic Models

$$\text{Triples}\,(\eta) := \mathbb{E}[x_1 \otimes x_2 \otimes \langle x_3, \eta \rangle] \in R^{d \times d}$$

$$\text{Triples}\,(\eta) : \mathbb{R}^d \to \mathbb{R}^{d \times d}$$

Lemma

$$\text{Pairs} = M \text{diag}\,(w)\, M^\top$$

$$\text{Triples}\,(\eta) = M \left(\text{diag} \left(M^\top \eta \right) \text{diag}\,(w) \right) M^\top$$

The unknown M and w are twisted.

Learning the Topic Models

Assumption (Non-degeneracy)

M has full column rank k

1. Find $U, V \in \mathbb{R}^{d \times k}$ s.t. $\left(U^\top M\right)^{-1}$ and $\left(V^\top M\right)^{-1}$ exist.
2. $\forall \eta \in R^d$, define $B(\eta) \in \mathbb{R}^{k \times k}$

$$B(\eta) := \left(U^\top \mathrm{Triples}(\eta)\, V\right)\left(U^\top \mathrm{Pairs}\, V\right)^{-1}$$

Lemma (Observable Operator)

$$B(\eta) = \left(U^\top M\right) \mathrm{diag}\left(M^\top \eta\right)\left(U^\top M\right)^{-1}$$

Learning the Topic Models

Input: $\hat{\text{Pairs}}$ and $\hat{\text{Triples}}$
Output: topic-word distributions $\hat{M}$
$\hat{U}, \hat{V} \leftarrow$ top k left, right eigenvectors of $\hat{\text{Pairs}}$ [a]
$\eta \leftarrow$ random sample from range$(\hat{U})$
$\left(\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k\right) \leftarrow$ right eigenvectors of $B(\eta)$ [b]
for $j \leftarrow 1$ ***to*** k **do**
  $\hat{\mu}_j \leftarrow \hat{U}\hat{\xi}_j / \langle 1, \hat{U}\hat{\xi}_j \rangle$
end
return $\hat{M} = [\hat{\mu}_1 | \hat{\mu}_2 | \ldots | \hat{\mu}_k]$

---

[a] $\text{Pairs} = M \text{diag}(w) M^\top$
[b] $B(\eta) = \left(U^\top M\right) \text{diag}\left(M^\top \eta\right) \left(U^\top M\right)^{-1}$

Learning the Topic Models

Lemma (Observable Operator)

$$B(\eta) = \left(U^\top M\right) \operatorname{diag}\left(M^\top \eta\right) \left(U^\top M\right)^{-1}$$

We hope $M^\top \eta$ has distinct entries. How to pick η?

$$\eta \leftarrow e_i \Rightarrow M^\top \eta \quad \text{i-}\textit{th} \text{ word's distribution over topics}$$

Prior knowledge required!
Otherwise, $\eta \leftarrow U\theta$, $\theta \sim Uniform(\mathcal{S}^{k-1})$

Learning the Topic Models

- SVD is carried out on $\mathbb{R}^{k \times k}$, $k \ll d$
- Only involves trigram statistics i.e. low-order moments
- Guaranteed to recover the parameters
- Parameters of more complicated models like LDA can be recovered in the same manner

Tensor Decomposition

Recall

$$\text{Pairs} = M \operatorname{diag}(w) M^\top$$

$$\text{Triples}(\eta) = M \left(\operatorname{diag}\left(M^\top \eta\right) \operatorname{diag}(w)\right) M^\top$$

$$\text{Pairs} = \sum_j^k w_j \cdot \mu_j \otimes \mu_j$$

$$\text{Triples} = \sum_j^k w_j \cdot \mu_j \otimes \mu_j \otimes \mu_j$$

Symmetric tensor decomposition? μ_j need to be orthogonal

Tensor Decomposition

Whiten Pairs

$$W := UD^{\frac{1}{2}} \Rightarrow W^\top \mathrm{Pairs}\, W = I$$

$$\mu'_j := \sqrt{w_j}\, W^\top \mu_j$$

We can check that $\mu'_j, j \in [k]$ are orthonormal vectors

Do orthogonal tensor decomposition on

$$\mathrm{Triples}\,(W, W, W) = \sum_{j=1}^{k} w_j \left(W^\top \mu_j\right)^{\otimes 3} = \sum_{j=1}^{k} \frac{1}{\sqrt{w_j}} {\mu'_j}^{\otimes 3}$$

Then recover μ_j from μ'_j

Anchor Words

Drawbacks of previous algorithms

- topics cannot be correlated
- the bound is weak (comparatively speaking)
- empirical runtime performance is not satisfactory

Alternatively assumptions?

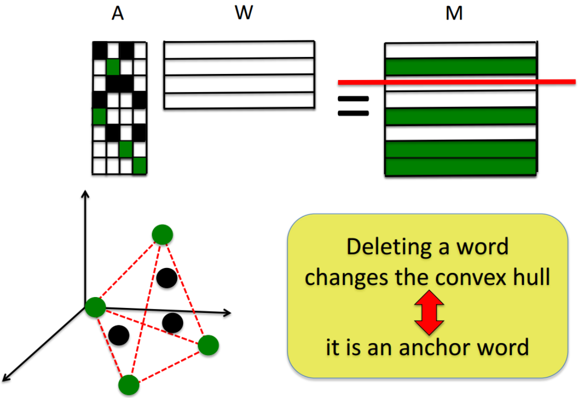
Anchor Words

Definition (p-separable)

M is p-separable if $\forall j$, $\exists i$ s.t. $M_{ij} \geq p$ and $M_{ij'} = 0$ for $j' \neq j$

- Documents do not necessarily contains anchor words
- Two-fold algorithm
  1. Selection: find the anchor word for each topic
  2. Recover: recover M based on anchor words
- Good theoretical guarantees and empirical results

Anchor Words



1

[1]The illustration is taken from Ankur Moitra's slides, http://people.csail.mit.edu/moitra/docs/IASM.pdf

Discussion

Summary

- A brief introduction to MoM
- Learning topic models by spectral decomposition
- Anchor words assumption

Connections with our work?