Non-Negative Matrix Factorization

Andrew An Bian, Francesco Locatello

March 30-31, 2017

Overview

Review

Why non-negativity?

Estimating the pLSA parameters

The pLSA model

EM

NMF Algorithms

Update rules derivation

Review

- Non-negative matrix factorization solves

$$\mathbf{X} \approx \mathbf{U}^{\mathbf{T}} \cdot \mathbf{V}$$

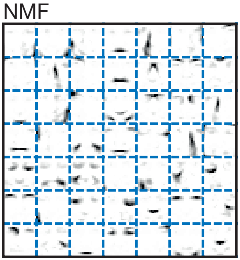
  for matrices $\mathbf{X}$, $\mathbf{U}$ and $\mathbf{V}$ whose entries are **non-negative**.

- More formally, for non-negative $\mathbf{X}$ we minimize the cost function:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^{\mathbf{T}}\mathbf{V}\|_F^2. \\ \text{s.t.} \quad & u_{zi} \in [0,\infty) \ \forall i, z \\ & v_{zj} \in [0,\infty) \ \forall j, z. \end{aligned}$$

Why impose non-negativity constraint?

- In situations where negative values do not make sense.
  - For example, $\# \{\text{occurrences of a word in a document}\} \geq 0$
- More interpretable "decomposition of object into parts", rather than holistic components arising from PCA
  - See **Lee & Seung, Nature (1999)** and **Donoho & Stodden, NIPS (2004)**



Other Applications

- Collaborative filtering

  CF: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs, Gu et al; SDM (2006)

  Learning from Incomplete Ratings Using Non-negative Matrix Factorization; Zhang et al, SDM (2006)

- Sparse coding

  Non-negative Matrix Factorization with Sparseness Constraints; Hoyer, JMLR (2004)

- Compression and face representation

  Two-dimensional non-negative matrix factorization for face representation and recognition; Zhang, Chen, Zhou, ICCV Workshop (2005)

- Image inpainting

  Image inpainting via Weighted Sparse Non-negative Matrix Factorization; Wang and Zhang, IEEE Int Conf Image Processing (2011)

Estimating pLSA Parameters

pLSA: Generative Model

- Topic $z \in \{z_1, \ldots, z_K\}$
- Word $w \in \{w_1, \ldots, w_N\}$
- Document $d \in \{d_1, \ldots, d_M\}$

In order to generate a tuple (w, d):

- Sample a document d according to $P(d)$.
- Sample a topic z according to $P(z|d)$.
- Sample a word w according to $P(w|z)$.
- Assume(!) a factorization:

$$P(w|d) = \sum_z P(w|z)P(z|d).$$

Conditional independence of word and document given topic!

The joint distribution of a document and a word is therefore:

$$P(w, d) = P(w|d)P(d).$$

pLSA: Matrix Factorization View

Normalize $\mathbf{X}$

Normalize the elements of $\mathbf{X}$ so that they can be interpreted as (joint) probabilities:

$$P(w_m, d_n) = \frac{x_{mn}}{\sum_{m',n'} x_{m'n'}}.$$

Matrix Factorization

pLSA can be understood as a matrix factorization of the form

$$\mathbf{X} \approx \mathbf{U^T V}$$

with $\mathbf{U^T} \in \mathbb{R}_+^{M \times K}$ and $\mathbf{V} \in \mathbb{R}_+^{K \times N}$. Additionally we have the constraints:

$$\sum_z^K u_{zi} = 1 \; \forall i \qquad \sum_j^N v_{zj} = 1 \; \forall z.$$

pLSA: Parameter Estimation

- Want to maximize the likelihood of the data under the model.
- Data: the occurrence $\mathbf{X}$.
- The model:

$$P(w|d) = \sum_z P(w|z)P(z|d).$$

- The log-likelihood can be written as

$$\mathcal{L}(\mathbf{U}, \mathbf{V}) = \sum_{i,j} x_{ij} \log p(w_j|d_i) = \sum_{(i,j)\in\mathcal{X}} \log \sum_{z=1}^{K} \underbrace{p(w_j|z)}_{=:\mathbf{v_{zj}}} \underbrace{p(z|d_i)}_{=:\mathbf{u_{zi}}}$$

  - two types of parameters:
  - $u_{zi} \geq 0$ such that $\sum_z u_{zi} = 1$ $(\forall i)$
  - $v_{zj} \geq 0$ such that $\sum_j v_{zj} = 1$ $(\forall z)$

Estimating The Parameters

Parameters Of The pLSA Model

- $P(w|z)$ and $P(z|d)$
- Think of them as probability tables of dimension $M \times K$ and $K \times N$ respectively.

Expectation Maximization

- pLSA: non-convex optimization, many local extrema
- Introduce variational parameters q_{zij}, apply Jensen's inequality

$$\sum_{i,j} x_{ij} \log \sum_{z=1}^{K} q_{zij} \frac{u_{zi} v_{zj}}{q_{zij}} \geq \sum_{i,j} x_{ij} \sum_{z=1}^{k} q_{zij} \left[\log u_{zi} + \log v_{zj} - \log q_{zij}\right]$$

Expectation Maximization

E-Step

$$q_{zij} = \frac{u_{zi} v_{zj}}{\sum_{k=1}^{K} u_{ki} v_{kj}} = \frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^{K} p(w_j|k)p(k|d_i)}$$

- posterior of latent topic variable associated with an occurrence (d_i, w_j).

M-Step

- Solve for optimal parameters

$$u_{zi} = \frac{\sum_j x_{ij} q_{zij}}{\sum_j x_{ij}}, \qquad v_{zj} = \frac{\sum_i x_{ij} q_{zij}}{\sum_{i,l} x_{il} q_{zil}},$$

Updating Rules of the NMF Algorithm

NMF for Quadratic Cost Function

- pLSA: just one instance of a non-negative matrix factorization
- Variation: non-negative data $\mathbf{X}$ with **quadratic** cost function:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top \mathbf{V}\|_F^2. \\ \text{s.t.} \quad & u_{zi}, v_{zj} \geq 0 \quad (\forall i, j, z) \quad (\textbf{non-negativity}) \end{aligned}$$

NMF Algorithm: Quadratic Costs

- **Alternating least squares**
  - objective is convex in $\mathbf{U}$ and $\mathbf{V}$ alone, but not jointly in $(\mathbf{U}, \mathbf{V})$
    $\Rightarrow$ alternate optimization of $\mathbf{U}$ and $\mathbf{V}$, keeping the other fixed
  - **normal equations**: look at single column of $\mathbf{V}$ at a time

$$\left(\mathbf{x}_j - \mathbf{U}^\top \mathbf{v}_j\right)^2 = \|\mathbf{x}_j\|^2 - \mathbf{x}_j^\top \mathbf{U}^\top \mathbf{v}_j - \mathbf{v}_j^\top \mathbf{U} \mathbf{x}_j + \mathbf{v}_j^\top \mathbf{U} \mathbf{U}^\top \mathbf{v}_j$$

optimality condition: $\nabla_{\mathbf{v}_j}(\dots) = 0 \iff \left(\mathbf{U}\mathbf{U}^\top\right) \mathbf{v}_j = \mathbf{U}\mathbf{x}_j$

  - normal equations in matrix notation

$$\left(\mathbf{U}\mathbf{U}^\top\right) \mathbf{V} = \mathbf{U}\mathbf{X}, \quad \text{and} \quad \left(\mathbf{V}\mathbf{V}^\top\right) \mathbf{U} = \mathbf{V}\mathbf{X}^\top$$

    - can be numerically solved in many ways, e.g. with QR-decomposition or via gradient descent methods

NMF Algorithm: Quadratic Cost (cont'd)

- **Projected ALS**
  - need to project in between alternations – **non-negativity**!
  - simply project elementwise by

$$u_{zi} = \max\{0, u_{zi}\}, \quad v_{zj} = \max\{0, v_{zj}\}$$