Topic Models

pLSI (Hofmann, 1999), LDA (Blei, Ng, Jordan, 2002)

Nathan Sutter

Overview

Introduction

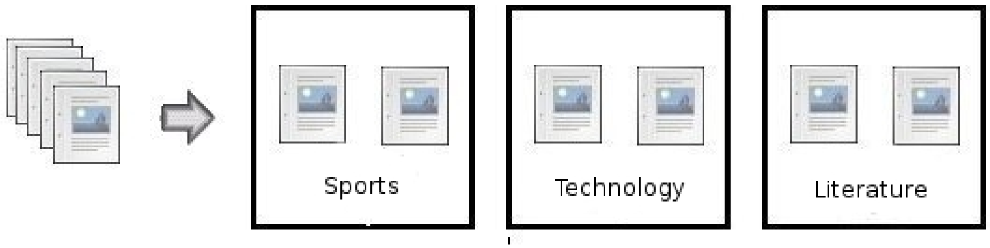
- Topic Modeling
- Review of Relevant Probability Distributions
- LSI
- Evaluating Topic Models

Topic Models

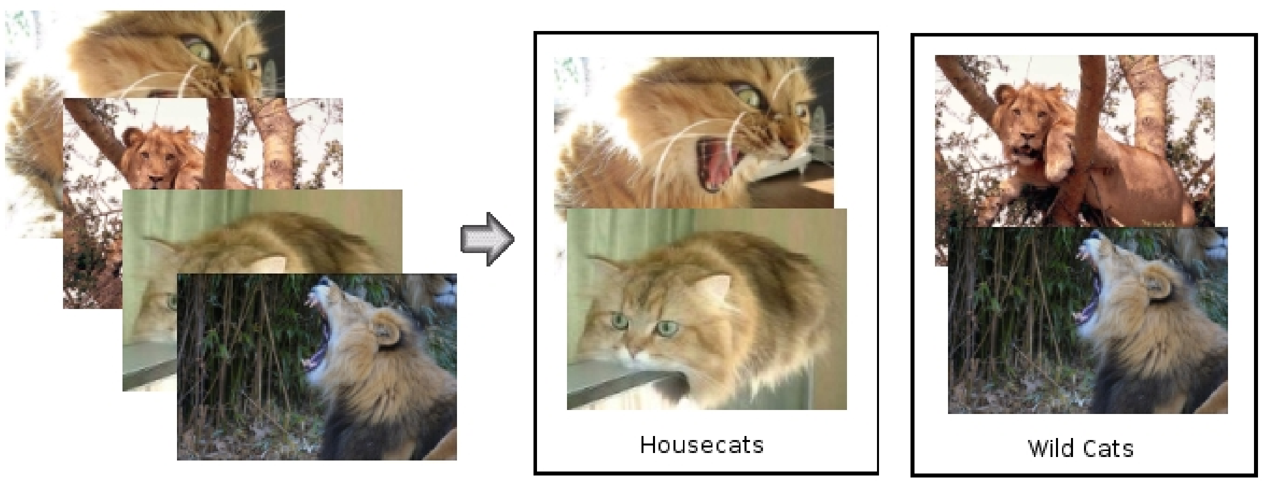
- pLSI
- LDA
- Application to Collaborative Filtering

Topic Modeling

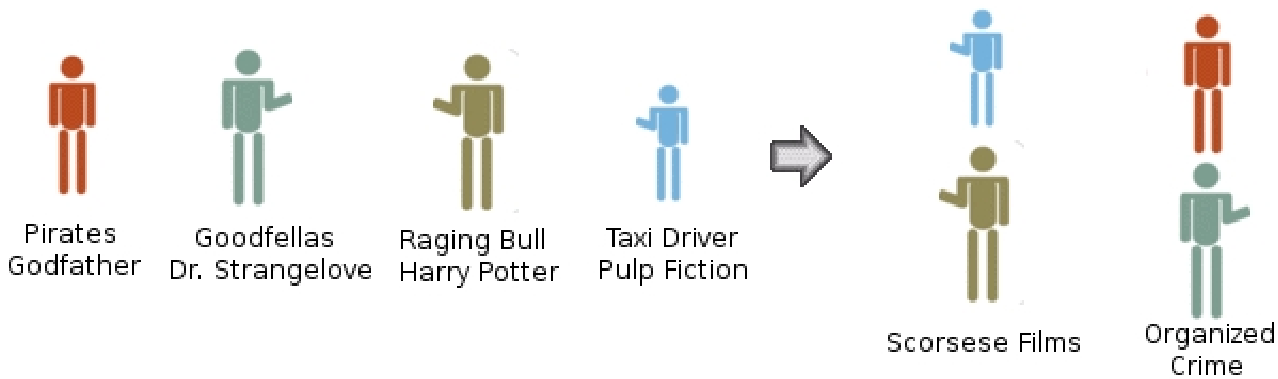
- Given some dataset what topics are persistent in that dataset?
- Can we infer that items within our dataset "belong" to a topic or mixture of topics?
- Example topic groups in datasets :
  - ▷ Documents



  - ▷ Images



  - ▷ Movie preferences per user



- ▷ etc...

- Examples of topic models
  - ▷ SVD/Specific Applications of SVD (like LSI)
  - ▷ pLSI
  - ▷ LDA

- Trivial to apply topic models to collaborative filtering!
  - ▷ Beware of Caviats!

Review of Relevant Probability Distributions

- Multinomial Distribution

Parameters :

n independent events,

$x_1, \ldots, x_k$, where x_i is the number of times event i occurs, $\sum_i x_i = n$.

$p_1, \ldots, p_k$, where p_i is the probabiliy that event i occurs, $\sum_i p_i = 1$..

$$p(x_1, \ldots, x_n) = \frac{n!}{x_1! \ldots x_n!} \prod_i p_i^{x_i}$$

Toy Example :

Have an urn with 7 balls in it, which are either red, green, blue, or yellow. Out of seven draws, what is the probability that a red ball is drawn once, a green ball is drawn 4 times, and a blue ball is drawn 2 times?

Parameters :

7 independent events,

$x_1 = 1, x_2 = 4, x_3 = 2, x_4 = 0$

$p_1 = p_2 = p_3 = p_4 = .25$.

$p(x_1 = 1, x_2 = 4, x_3 = 2) = \frac{7!}{1!4!2!0!}(.25^1)(.25^4)(.25^2)(.25^0)$

Introduction(cont.)

- Dirichlet Distribution

  Returns the probability of multinomial probabilities $\boldsymbol{\theta}$, given counts for each event $\boldsymbol{\alpha}$.

  Parameters :

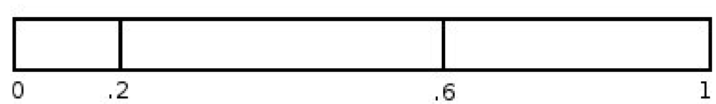
  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, where θ_i is the probabiliy that event $\boldsymbol{i}$ occurs. $\sum_i \theta_i = 1$.

  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$, where $\boldsymbol{\alpha}_i$ is the number of times event $\boldsymbol{i}$ occurs.

$$Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_i \theta_i^{\alpha_i - 1}$$

$$B(\boldsymbol{\alpha}) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}$$

Or put more intuitively :



- ▷ Given a stick of length 1, break stick into k pieces.
- ▷ Allow for variability in the size of pieces.

Introduction(cont.)

Nice transition to topic modeling :

- We have k topics in our dataset, each claiming a part of the stick.
- Interested in how much of the stick each topic claims in our training set, optimally.

Properties :

- Conjugate to the multinomial distribution.
- Finite dimensional sufficient statistics.

Introduction(cont.)

LSI

- Map a document to latent space
  $X = U\Sigma V^T$ — U,V orthonormal, Σ has ordered eigs of X.
  $\hat{\Sigma} = \Sigma(1:n, 1:n) \ni X \approx U\hat{\Sigma}V^T$ — Ignore eigenvalues of dimensions $> n$.
  $V = V^T U \hat{\Sigma}^{-1}$ — Solving for V
  $d = d^T U \hat{\Sigma}^{-1}$ — Individual document mapping in latent space
- Compare that document to other documents
  $cos(\theta) = \frac{d_1 \cdot d_2}{||d_1|| ||d_2||}$

Evaluating Topic Models

- Calculate perplexity on test set, given model parameters learned during training.
- Monotonically Decreasing in the likelihood of the test data
- A good model would assign a high likelihood to held out documents, and thus, low perplexity.

$$perplexity(D_{test}) = -\frac{\sum_m log(p(w_m))}{\sum_m w_m}$$

pLSI (Hoffman, 1999)

- Notation

  Document Data :

  M documents $D = (d_1, ..., d_M)$

  N words per document $d_i = (w_1, ..., w_N)$

  K topics $z_1, ..., z_k$

  Parameters :

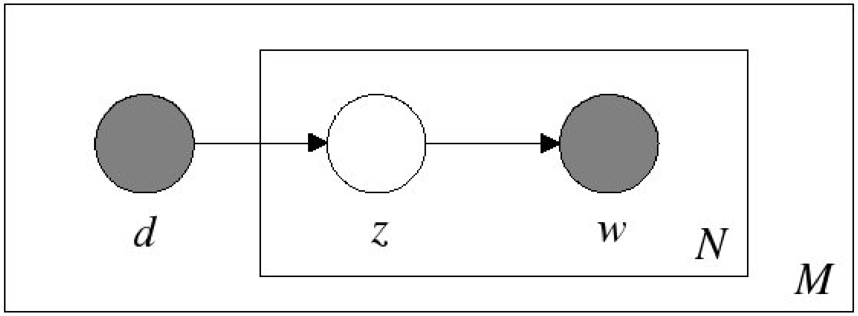
  $P(w|z)$: Probability of a word given a topic.

  $P(d|z)$: Probability of a document given a topic.

  $P(z)$: Probability of a topic.

Topic Models(cont.)

- Model Specification



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

- Relation to LSI

  We can interpret the joint probability $P(d, w)$ as $P = U\Sigma V^T$ such that :

  $U = (P(d_i|z_k))_{i,k}$
  $\Sigma = diag(P(z_k))_k$
  $V = (P(w_j|z_k))_{j,k}$
  $P = U\Sigma V^T$

Topic Models(cont.)

- Fitting the model via an EM algorithm

  E Step:

  $$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}$$

  M Step:

  $$P(w|z) \propto \sum_d n(d,w)P(z|d,w)$$

  $$P(d|z) \propto \sum_w n(d,w)P(z|d,w)$$

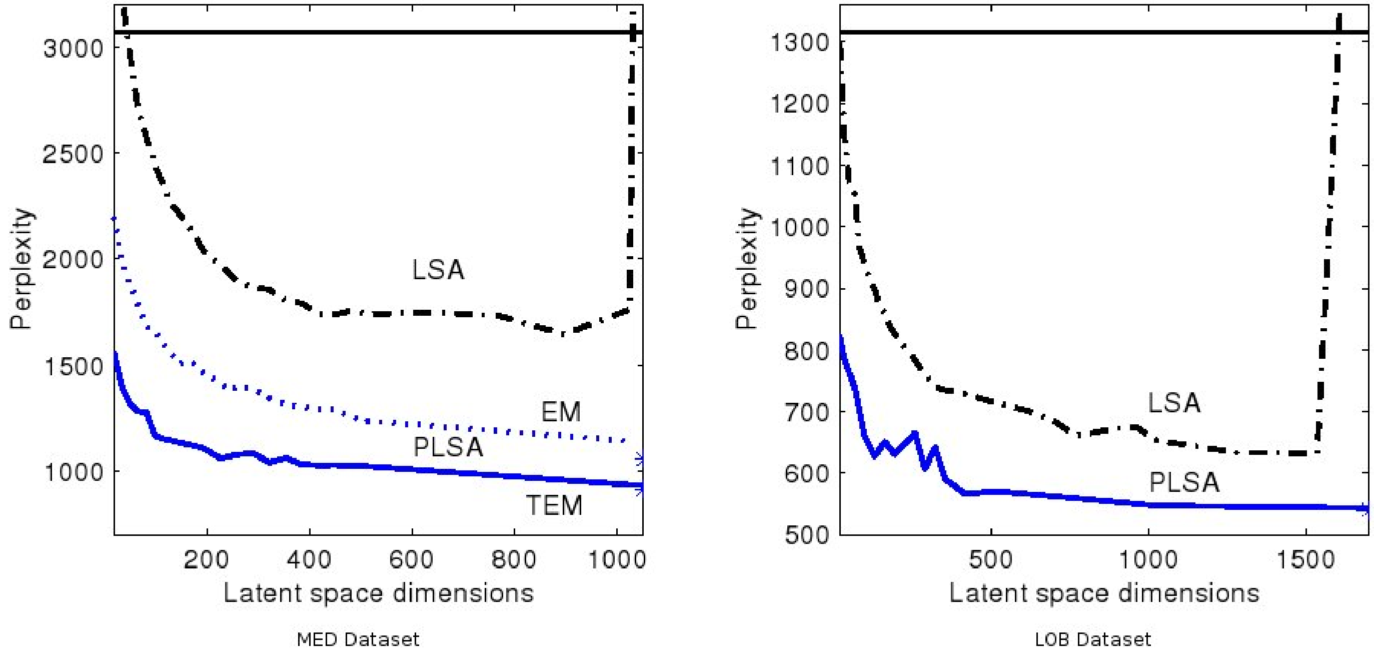
  $$P(z) \propto \sum_{d,w} n(d,w)P(z|d,w)$$

- Model inference on unseen data
  - ▷ Recall we are interested in perplexity as a metric for model performance.
  - ▷ Need to calculate $P(w|\boldsymbol{w}_{obs}) = \sum_z P(w|z)P(z|\boldsymbol{w}_{obs})$.
  - ▷ Fix $P(w|z)$ from training, re-learn $P(z|d,w)$ on test set (called folding-in).

Topic Models(cont.)

- Applications

  pLSI tested on two text datasets, MED and LOB, using bag of words assumption.



- Disadvantages
  - ▷ Assigns zero probability to documents in training set
  - ▷ Folding-in is kind of weird.
  - ▷ Folding-in, as presented by Hoffmann, ignores an intractable normalization constant.
  - ▷ Correct folding-in requires heavy regularization depending on the test data split.
  - ▷ pLSA overfits heavily as $K \rightarrow \infty$.

LDA (Blei, Ng, and Jordan, 2002)

- Notation

  Document Data :

  A vocabulary that is V entries long.

  Words $\boldsymbol{w}$ are V dimensional vectors such that $\boldsymbol{w}^u = 1$, and $\boldsymbol{w}^v = 0$ for $\boldsymbol{u} \neq \boldsymbol{v}$

  A document which is a sequence of N words, $\boldsymbol{w} = \{w_1, w_2, ..., w_N\}$

  A corpus of M documents, $\boldsymbol{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}$

  Parameters :

  $\boldsymbol{\alpha}$: hyperparameter on $\boldsymbol{\theta}$, number of times each topic occurs.

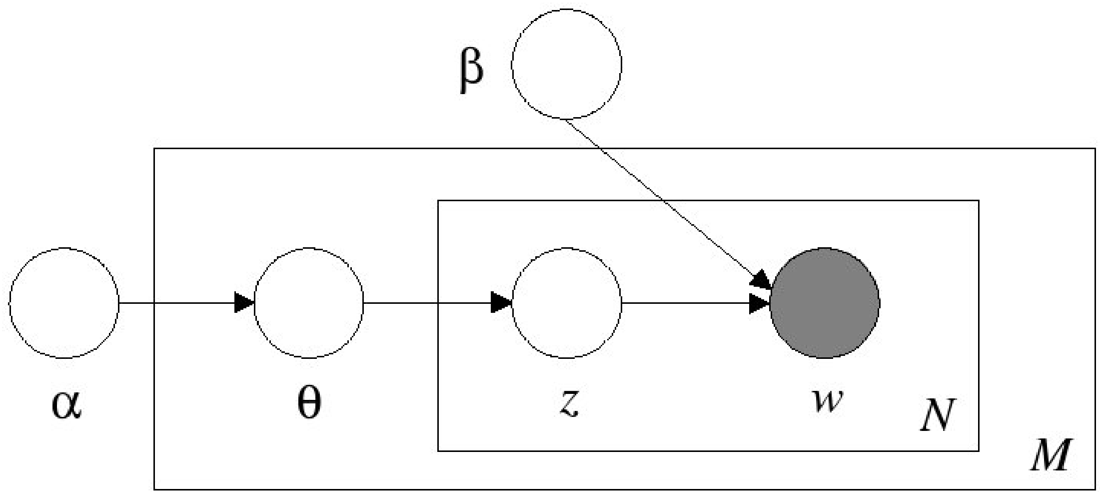
  $\boldsymbol{\beta}$: $\beta_{ij} = p(w_j = 1 | z_i = 1)$

  $\boldsymbol{\theta}$: individual probabilities of each topic occuring.

Topic Models (cont.)

- Model Specification

We imagine a corpus is generated as follows :
For each document $1, \ldots, M$. Choose $N \sim Poisson(\xi)$ Choose $\theta \sim Dirchilet(\alpha)$ For each word in w_n Choose $z_n \sim Multinomial(\theta)$ Choose a word w_n from $p(w_n\|z_n, \beta)$



Full Joint Probability (for one document):

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_N p(z_n|\theta) p(\boldsymbol{w}|z_n, \beta)$$

Topic Models (cont.)

Marginalize out parameters :

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha) \prod_N \sum_z p(z_n|\theta) p(\boldsymbol{w}|z_n, \beta) d\theta$$

We're interested in :

$$\arg\max_{\alpha,\beta} L(p(\boldsymbol{w}|\alpha, \beta))$$

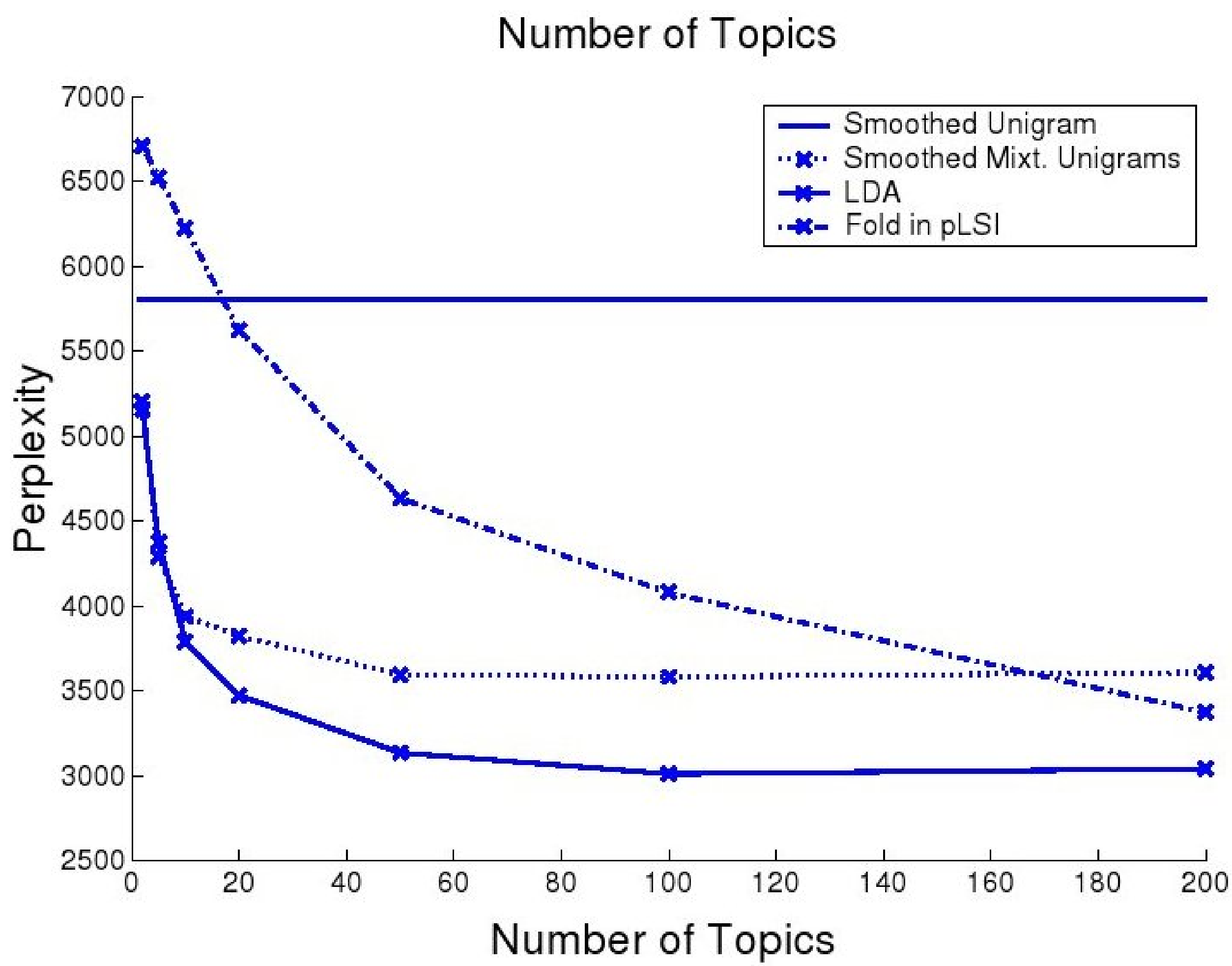
Problem : function we are trying to optimize is intractable for exact inference.

- Learning Model Parameters from data
  - ▷ Use Variational EM to approximate $\boldsymbol{\alpha}, \boldsymbol{\beta}$ (Blei, et al, 2001).
  - ▷ Use a collapsed Gibbs sampler to approximate $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ (Griffiths,Steyvers 2002).

Topic Models (cont.)

- Applications

  LDA tested on text dataset AP, relying on bag of words assumption.



Application to Collaborative Filtering

- Dataset and Framework
  - Using the EachMovie dataset.
  - Each movie rating is converted to either a positive or negative rating.
  - Users are analogous to documents, Movies are analougous to words.
  - Measure Predictive Perplexity.
- pLSA

$$P(w|\boldsymbol{w}_{obs}) = \sum_z P(w|z)P(z|\boldsymbol{w}_{obs})$$

- LDA

$$P(w|\boldsymbol{w}_{obs}) = \int \sum_z P(w|z)P(z|\theta)P(\theta|\boldsymbol{w}_{obs})d\theta$$

Topic Models (cont.)

- Results