Topic Models

Special course in Unsupervised Machine Learning
University of Helsinki

Guest Lecture

Ali Faisal
ali.faisal@opuscapita.com

16th May, 2017

Data Scientist
OpusCapita Group Oy
Motivation

- Topic Models has several applications
  - Ranging from NLP to Biology
- Example applications
  - Language modelling
    - text categorization,
    - Innovative search engines
    - speech recognition, etc
  - **Demo**: Evolution of topic in 40 years of *Signs* archive
    - Evolution of game genre
    - Scientific topic evolution PNAS
  - Semantic hierarchies
    etc etc...
Today's journey - Agenda

- Background of Topic models
- Language Models
- Latent Dirichlet Allocation (LDA)
  - Semantic interpretability case
- Demo
- Nonparametric Topic models
  - Information retrieval case
- Demo
Background - Example Topics

- Most often used in analyzing text and image collections
- All data is assumed to be generated from a collection of Topics
  - Topics are sources that generate elements with certain probabilities

Top 5 topics for NIPS conference article collection from 1987-99

<table>
<thead>
<tr>
<th>Strongest Topic</th>
<th>CNP</th>
<th>NS</th>
<th>LT and AA</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning</td>
<td>control</td>
<td>neurons</td>
<td>variables</td>
<td>rules</td>
</tr>
<tr>
<td>network</td>
<td>reinforcement</td>
<td>neuron</td>
<td>kernel</td>
<td>coarse</td>
</tr>
<tr>
<td>model</td>
<td>robot</td>
<td>cortical</td>
<td>markov</td>
<td>instruction</td>
</tr>
<tr>
<td>time</td>
<td>learned</td>
<td>arbors</td>
<td>conditional</td>
<td>fine</td>
</tr>
<tr>
<td>input</td>
<td>policy</td>
<td>dendritic</td>
<td>group</td>
<td>schedule</td>
</tr>
<tr>
<td>neural</td>
<td>tangent</td>
<td>competition</td>
<td>likelihood</td>
<td>instructions</td>
</tr>
<tr>
<td>algorithm</td>
<td>interpolation</td>
<td>cells</td>
<td>face</td>
<td>dec</td>
</tr>
<tr>
<td>data</td>
<td>steps</td>
<td>axonal</td>
<td>database</td>
<td>blocks</td>
</tr>
<tr>
<td>set</td>
<td>initial</td>
<td>cell</td>
<td>generalized</td>
<td>rl</td>
</tr>
<tr>
<td>system</td>
<td>grid</td>
<td>modules</td>
<td>matrix</td>
<td>resolution</td>
</tr>
</tbody>
</table>

Faisal et al, '12
Topic Model

- Extensively used in bag-of-words text data.
- Called Latent Dirichlet Allocation (LDA) or discrete PCA (dPCA)
Background

- Model text as dice rolling:

\[ D = "A \text{ child plays and learns while playing} \]

Generates each word independently

\[ p(D) = p(w_{d,1}, w_{d,2}, \ldots, w_{d,N}) = \prod_n p(w_{d,n}) \]

if, \( p(\text{word}) = \varphi_{\text{word}} \)

\[ p(D|\varphi) = p(A) p(\text{child}) p(\text{play}) p(\text{and}) p(\text{learns}) p(\text{while}) p(\text{play}) \]

\[ = p(A) p(\text{child}) p(\text{play})^2 p(\text{and}) p(\text{learns}) p(\text{while}) \]

\[ = \varphi_A \varphi_{\text{child}} \varphi_{\text{play}}^2 \varphi_{\text{and}} \varphi_{\text{learns}} \varphi_{\text{while}} \]

\[ = \prod_j \varphi_j^{n_j} = \text{Multinomial}(\varphi) \]
Language models

- **Unigram Model**
  
  \[ p(D) = \prod_n p(w_{d,n}) \]

- **Mixture of Unigrams**
  
  For each document \( d \), choose a topic \( k \) (i.e. \( Z^d = k \)) and generate all words of the document from the word distribution of the chosen topic

  \[ p(D) = \sum_z p(z^d) \prod_n p(w_n | z^d) \]
Language models

- Unigram Model

\[ p(D) = \prod_n p(w_{d,n}) \]

One likelihood for the entire text collection

- Mixture of Unigrams

For each document \( d \), choose a topic \( k \) (i.e. \( Z^d = k \)) and generate all words of the document from the word distribution of the chosen topic

\[ p(D) = \sum_z p(z^d) \prod_n p(w_n | z^d) \]

One topic per document
Bayesian Modelling

- All methods we are discussing can be interpreted as performing Max.Lik. (Unigrams, Mixture of Unigrams, PCA, ICA or FA) or Bayesian estimation (LDA) in a probabilistic generative model.
Latent Dirichlet Allocation (LDA)
Generative process

- Draw topic distribution:
  - $\pi_d \sim \text{Dirichlet}(\alpha)$

- Generate n-th word by
  - draw topic index:
    - $z_{d,n} \sim \text{Multinomial}(\pi_d)$
  - Draw word from topic-wise word distribution: $\beta$
    - $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$

- Where $\beta_k$ are probabilities of each word “w” in the k-th topic
  - $\beta_k \sim \text{Dirichlet}(\eta)$
    - $\beta_k$ lives on a simplex
Dirichlet distribution - background

Dirichlet($\alpha$) where $\alpha$ is a multivariate probability distribution over a simplex

PDF: \[
\frac{1}{\Gamma(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \quad \alpha = (\alpha_1, \ldots, \alpha_K)
\]

- Induces uniform Topic distribution when $\alpha_i = 1$
- Induces Sparse Topics when $\alpha < 1$
Latent Dirichlet Allocation (LDA) Generative process

- Draw topic distribution:
  - \( \pi_d \sim \text{Dirichlet}(\alpha) \),

- Generate n-th word by
  - draw topic index:
    - \( z_{d,n} \sim \text{Multinomial}(\pi_d) \)
  - Draw word from topic-wise word distribution: \( \beta \)
    - \( w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}}) \)

- Where where \( \beta_k \) are probabilities of each word “w” in the k-th topic
  - \( \beta_k \sim \text{Dirichlet}(\eta) \)
    \( \beta_k \) lives on a simplex

Dirichlet prior mitigate over fitting, words that do not appear in the training set are still assigned some probability to appear in future documents
Factor Analysis - revisited

\[ X \sim ZW^T \]

Given a matrix \( X \in \mathcal{R}^{N \times D} \) we can factorise it into a product of two smaller matrices \( Z \in \mathcal{R}^{N \times K} \) and \( W \in \mathcal{R}^{D \times K} \).
Demo
Semantic interpretation – case study
Explosion of data in Genomic databases

Keeping research cumulative is a huge challenge for current data-driven science

Growth of EBI ArrayExpress database

How to make maximal use of progressively expanding databases?
Retrieval of Relevant Samples

Annotation-driven query: “leukemia” → “leukemia” → “cancer”

Data-driven query
(leukemia) → “CML” → “Crohn’s disease” → “CLL”
(leukemia) → “CML” → “Crohn’s disease” → “CLL”

Caldas et al, Bioinformatics, 2009
Topic Model

- Extensively used in bag-of-words text data.
- Called Latent Dirichlet Allocation (LDA) or discrete PCA (dPCA)
Components of experiments

Caldas et al, Bioinformatics, 2009
Recipe for retrieval

- A background model for the biology provides $p(exp)$
- The retrieval engine finds experiments that share activated biological processes:

$$\max_{exp_2} p(exp_1 | exp_2) = \int_z p(exp_1 | z)p(z | exp_2)dz$$
NerV 2D Visualizations of the entire collection (color-coded with topics)
Problem: Model selection (How to fix the number of components)

Annealed importance sampler for 1000 iterations over the ArrayExpress database (~7000 samples)
Systematic retrieval evaluation & Comparison

- Retrieval evaluation shows comparable performance with alternatives.
- Here the gold standard is more refined (Experimental factor ontology) and represents relationships between experimental factors.

Discounted Cumulative Gain:
How much an investigator gains when a comparison with particular relevance is found at particular rank in the result list of query.

LDA and REx are our model based approaches

Caldas et al., Bioinformatics, 2012
Nonparametric Topic models

unlike LDA, in nonparametric models we no longer need to pre-specify “k”

Nonparametric find the number of topics “k” automatically from data by utilizing the amazing Dirichlet process (DP) prior and Hierarchical Dirichlet Process (HDP) prior
A Dirichlet Process (DP) is a distribution over distributions which can be seen as an infinite dimensional generalization of the usual Dirichlet distribution

$$G \sim \text{DP}(\alpha_0, G_0)$$

$G$ is a random probability distribution.

$G_0$ is a base measure – a putative mean for $G$.

$\alpha_0$ Is a concentration parameter, controls the amount of variability around $G$.

In a Mixture models (MM) each data item $x_i$ is associated with an underlying factor $\theta_i$ with prior given by $G$:

$$x_i | \theta_i \sim F(\theta) \quad \theta_i | G \sim G$$
Sethuraman's (1994) stick breaking construction shows that samples \( G \sim \text{DP}(\alpha_0, G_0) \) has the form:

\[
G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}
\]

where \( \pi_k \geq 0, \sum_{k=1}^{\infty} \pi_k = 1 \) are random variables depending upon \( \alpha_0 \)

\[
\pi_k = \pi'_k \prod_{j=1 \text{ to } k-1} (1 - \pi'_j)
\]

\[
\pi'_k = \text{Beta}(1, \alpha_0)
\]

Intuitively, consider a stick of length one, at each point we break the stick. The broken part \( \pi'_k \) is taken as the weight of the corresponding atom in DP.

\[
\pi'_1, \pi'_2(1-\pi'_1)
\]
DP Mixture model (cont'd)

\[ \theta_i \sim G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \]

Associate each \( \Phi_k \) with a component

So a DP mixture is a mixture model with potentially infinite no. of components
DP Mixture model Grouped data

Modeling grouped data with a DP MM.

Associate a DP with each group

Each group can learn the appropriate number of components automatically.

There is a problem...
Modeling grouped data with a DP MM.

Associate a DP with each group

Each group can learn the appropriate number of components automatically.

There is a problem...

Each group is modeled independently.

Different groups will never share the same components if $G_0$ is continuous.

Individual atoms are not shared.
The prior on the individual groups can be made discrete by placing a DP on base distribution $G_0$

$$G_0 \sim \text{DP}(\gamma, H)$$

The prior induced on $G_0$ and $G_j$ is called a HDP prior while the model induced on data is a HDP MM.
Stick breaking construction

The factors $\theta_{ji}$ take on values $\Phi_k$ with probability $\pi_{jk}$

This is denoted by indicator variable $z_{ji}$

The DP priors have the form:

\begin{align*}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\Phi_k} \\
G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\Phi_k}
\end{align*}

So The HDP MM. Is simply a mixture model where the mixing weights $\pi_{jk}$ are dependent on each other via $\beta_k$. 
HDP MM vs LDA

HDP MM

LDA
Comparison LDA (cont'd)

Perplexity over held-out set a dataset of ~6000 biology abstracts
Traditional approaches:
Our model vs Multitask HDPLDA
Traditional approaches:
Our model vs Multitask HDPLDA
Low training data: transfer learning

Sharing and strengths of topics are coupled

HDP Multi-task

Sharing and strengths of topics are decoupled

IBP-gamma Multi-task

Faisal et al, Neurocomputing 2013
Humans use earlier knowledge of related tasks to perform new tasks, e.g. knowledge about standing helps walking and running.

Transfer learning transfers knowledge from earlier tasks (data-sets) to a new one and Multi-task learning learn several tasks together from their respective data sets, exploiting their underlying relationships.

Faisal et al, Neurocomputing, 2013
Transfer learning: Advantages

- It is a model for data-sets
- Number of topics can be estimated automatically
- It is capable to model weak topic in multi-task problems
- It is a robust and flexible Bayesian generative model
- Outperforms state of the art HDP topic model, specially when the number of samples is low; scenario central to the multi-task problems
Non-parametric modeling - conclusions

Avoids model selection and thus saves computation time....

The complexity is comparable to parametric model.

Good choice for count data
Information retrieval making
Biology Cumulative
Objectives

To make data-driven biology more cumulative
How to achieve that!!

Efficiently decompose a transcriptomics dataset into earlier datasets.

Retrieve a set of earlier datasets where each explains a certain part of variation in the query dataset.
How to achieve that!!

Efficiently **decompose a transcriptomics dataset** into earlier datasets.

Retrieve **a set of earlier datasets** where each **explains** a certain part of **variation in the query dataset**.
Scalable Supermodels

Definition:

We consider several data sources, $D_i$ and then their collection $D = \{D_1, D_2, ..., D_I\}$. If we compute models for each dataset, $M_i$ then the model for complete data collection is: $M = f(M_i, \theta_i)$.

There are at-least two different distributional assumptions on the whole data.

- Datasets come from a same distribution.
- Datasets come from different distributions.
Model - Trivial

- In the trivial case we assume semi-independent models
- We approximate joint probability of the query dataset by a combination of previously obtained probability distributions

\[
p_D(x, z) = \epsilon_D(x, z) + \hat{p}_D(x, z)
\]

\[
\hat{p}_D(x, z) = \sum_k w_{DP}^k p_k(x, z)
\]
Example Model: Latent Dirichlet Allocation

A generative model for count data e.g. text

\[ P(\Phi) = \prod_t \text{Dir}(\bar{\phi}_t; \beta \bar{n}) \]

\[ P(\Theta) = \prod_d \text{Dir}(\bar{\theta}_d; \alpha \bar{m}) \]

\[ P(\bar{z}^{(d)} | \bar{\theta}_d) = \prod_n \theta_{z_n|d} \]

\[ P(\bar{w}^{(d)} | \bar{z}^{(d)}, \Phi) = \prod_n \phi_{w_n|d} \]
Characteristics of the trivial model

- Simple and Straightforward:
  - We decompose the query model into earlier models using a trivial supermodel; a model for models that reduces to successive Bayesian hierarchical learning if the query decomposition constraints are removed.

- Constraints on the model:
  - The dataset should share a library of latent components;
  - The model is useful when
    - there exists a global library of topics
      - If there is an in-house collection of background datasets we can easily build this.
    - or prior knowledge is sufficient to be used as latent projections or components.
Model - Nontrivial

• Here we assume completely independent component models having different latent spaces
• Generalize and do not assume an existing model for the query

\[
\hat{p}(\{x\}_q|\{M\}) = \prod_i p(\{x_i\}_q|\{M\})
\]

\[
p(\{x\}_q|\{M\}) = \sum_s w_q^s p(\{x\}_q|M^s),
\]

s.t. \( \{w_q^s\}_{s=1:S} \geq 0, \sum_s (w_q^s) = 1 \)

• Compute the posterior probability of approximation weights assuming that our approximation family is correct:

\[
\log p(\{w_q^s\}|\{x\}_q, \{M\}) \propto \log p(\{x\}_q|\{M\}, \{w_q\}) + \log p(\{w_q\})
\]

\[
\propto \sum_i \log \sum_s w_q^s p(\{x_i\}_q|M^s) - \lambda \sum_s I(w_q^s > 0)
\]

• Optimization scheme to estimate \( W \) - two stage convex relaxation to the L-0 or L-1 norm

Faisal et al, PloS ONE, 2014
Are the most cited datasets, most important?

Compare correlation between the importance of each dataset with respect to the no. of times it has been cited.

Characterize the importance by the weighted out degree of a dataset; where the weight is provided by our method.

- Analyze if there are significant and obvious differences in the four corners of the scatter plot using:
  - Impact factor of publication venue
  - H-index of last author
  - Size of the dataset

- Results 1: Upper half of scatter plot:
  - Significantly lower impact factor for the left blue block:
    - Avg IF 6.6532 vs 21.9674 pval = 0.00020.

- Results 2: Lower half of scatter plot:
  - Significantly higher IF and h-index for right red block:
    - Avg IF values 21.93 vs 4.5 (pval: 0.0129)
    - H-index 54.25 vs 21.80 (pval: 0.0053)

Faisal et al, PloS ONE, 2014
Inconsistent annotations in the public databases

- the arrow tails represent original position of datasets based on original records in GEO and EBI ArrayExpress
- the head points to newly corrected positions as suggested by our model.
A generic to making research cumulative

Faisal et al, PloS ONE, 2014
Demo
TAKE HOME.....

• A powerful unsupervised machine learning approach

• Can summarize and interpret a huge collection of documents

• Allows us to study evolution of topics over time

• Allows us to study citation patterns, effectively pointing to outliers
  - e.g. the supermodel can point out datasets who should cite whom, or ones where it does not make sense to cite.
  -

• Nonparametric extensions automatically infer the correct number of topics

• Can handle polysemy and synonymy
How does brain work!

Contact
ali.faisal@opuscapita.com
References

Most results are taken from my articles, available here if article full text is not available then you can get it from me via email.

David Blei's tutorials and lectures are recommended http://www.cs.columbia.edu/~blei/