Kernel Topic Models

Thore Graepel
Microsoft Research Cambridge

Joint work with Philipp Hennig,
David Stern, and Ralf Herbrich
Overview

• Introduction and Motivation
• The Kernel Topic Model
• The Laplace Bridge: From Hilbert Space to Probabilities and back
• Experimental Results

Talk based on:
Overview

• Introduction and Motivation
• The Kernel Topic Model
• The Laplace Bridge: From Hilbert Space to Probabilities and back
• Experimental Results
Topic Models

- Model documents as bags-of-words
- For each document $d$: Draw probability vector $\pi_d$ over topics from Dirichlet ($\alpha$)
- For each topic $k$: Draw probability vector $\theta_k$ over words from Dirichlet ($\beta_k$)
- For each document $d$ and word slot $i$, draw topic $c_{di}$ from topic distribution $\pi_d$, then word $w_{di}$ from word distribution $\theta_{c_{di}}$
Topic Models and Metadata

• Topic Models are useful to organize text corpora in a meaningful but unsupervised way.

• The classical Latent Dirichlet Allocation disregards context information and metadata such as:
  – Author
  – Time and Place of creation
  – Part of book, journal, proceedings, series etc.
  – People to which document is addressed
  – Web links and citations
Documents in Context

Related Work

• Latent Dirichlet Allocation [Blei, Ng, & Jordan, 2003]
• Dynamic Topic Models [Blei and Lafferty, 2006]
• A correlated topic model of Science [Blei and Lafferty 2007]
• Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. [Mimno and McCallum, 2008]
• Relational Topic Models for document networks [Chang & Blei, 2009]
Overview

• Introduction and Motivation
• The Kernel Topic Model
• The Laplace Bridge: From Hilbert Space to Probabilities and back
• Experimental Results
Topic Model as Matrix Factorization

\[ W \triangleq \Pi \Theta \]

Data
\[ [1; \ldots, D] \rightarrow [0, 1]^K \rightarrow [0, 1]^V \]
Conditional Topic Model

$D \triangleq \sigma$

$[1; \ldots, D] \rightarrow \mathbb{R}^F \rightarrow \mathbb{R}^K \rightarrow [0, 1]^K \rightarrow [0, 1]^V$
Kernel Topic Model

\[ \sigma \triangleq \Phi \]

[1; \ldots, D] \xrightarrow{\mathcal{H}} \mathbb{R}^K \xrightarrow{[0, 1]^K} [0, 1]^V
Latent Dirichlet Analysis

\[ \begin{align*}
\alpha & \quad \pi_d \\
\beta_k & \quad \theta_k & \quad K \\
\pi_d & \quad c_{di} \quad \omega_{di} \\
I_d & \quad D
\end{align*} \]
Kernel Topic Model

\[ \mathcal{GP}(\mu, \Sigma) \rightarrow h \rightarrow y_d \rightarrow \pi_d \rightarrow c_{di} \rightarrow w_{di} \]

- \( h \): document features
- \( y_d \): topic proportions
- \( \pi_d \): topic assignments
- \( c_{di} \): topic assignments
- \( w_{di} \): topic descriptions
- \( \mathcal{GP}(\mu, \Sigma) \): Gaussian Process
- \( \beta_k \): feature weights
- \( \theta_k \): topic descriptions
- \( K \): number of topics
- \( D \): number of documents
- \( I_d \): topic assignments
Kernels/Covariance Functions on Documents

• Kernel over documents based on meta-data
  – Author
  – Time
  – Location
  – Publication

• Kernel over documents based on their link structure
  – Web Link structure
  – Citation structure
  – Social network of authors
Overview

• Introduction and Motivation
• The Kernel Topic Model
• The Laplace Bridge: From Hilbert Space to Probabilities and back
• Experimental Results
Kernel Topic Model: Laplace Bridge
Inference: The Laplace Bridge

• Laplace approximation:
  – Find mode of posterior distribution
  – Fit Gaussian distribution based on mode/curvature
  – Evaluate approximate posterior integrals etc.

• Laplace approximation is basis dependent and MacKay, 1998, suggests using a softmax basis

• Here we use the Laplace approximation in the softmax basis to transform probability messages into unconstrained (Gaussian) messages and back
Laplace approximation: softmax basis

- Dirichlet distribution over probability vectors:

\[
D_\pi(\pi; \alpha) \propto \prod_{i=1}^{K} \pi_i^{\alpha_i - 1} \delta \left( \sum_i \pi_i - 1 \right)
\]

- Transform to new softmax basis

\[
\pi_i(y) = \sigma_i(y) = \frac{\exp(y_i)}{\sum_j \exp(y_j)}
\]

- New parameterized form:

\[
D_y(y|\alpha) \propto \prod_{i=1}^{K} (\pi_i(y))^{\alpha_i} g(1^T y)
\]

- Choose \( g(x) = \exp \left(-\frac{\epsilon}{2} x^2 \right) \), later \( \epsilon \to 0 \)
From Dirichlet to Gaussian (and back)

- Laplace approximation of Dirichlet in form $N(\mathbf{y}; \mu, \Sigma)$
- Mean/Mode is given by
  $$\mu_k = \log \alpha_k - \frac{1}{K} \sum_{\ell=1}^{K} \log \alpha_{\ell}$$
- Approximately diagonal covariance given by
  $$\Sigma_{kk} = \frac{1}{\alpha_k} \left(1 - \frac{2}{K}\right) + \frac{1}{K^2} \sum_{\ell=1}^{K} \frac{1}{\alpha_{\ell}}$$
- And the inverse mapping:
  $$\alpha_k = \frac{1}{\Sigma_{kk}} \left(1 - \frac{2}{K} + \frac{\exp(-\mu_k)}{K^2} \sum_{\ell=1}^{K} \exp(-\mu_{\ell})\right)$$
Beta distributions approximated in 
Softmax basis

Simplex Basis

Softmax Basis
Gaussian “Dirichlet” in Action

• Generation
  – Sample $x \in \mathbb{R}^K \sim N(x; \mu, \Sigma)N(0; 1^T x, \epsilon^2)$
  – Map $\pi := \sigma(x)$
  – Sample data $c$ from $\rho(c = k|\pi) = \pi$

• Inference
  – Use Laplace bridge to obtain Dirichlet belief on $\pi$ from Gaussian prior
  – Update Dirichlet belief on data $c$ (conjugacy, easy!)
  – Reverse Laplace bridge to obtain Gaussian belief on $x$. 
Inference: Laplace Bridge vs MCMC

- Multivariate Gaussian $\boldsymbol{x}$ in $K = 10$ dimensions
- Sample from resulting Discrete distribution
- Infer point estimate of $\boldsymbol{x}$ using MCMC and Laplace
- Estimates are stable and accurate
Overview

• Introduction and Motivation
• The Kernel Topic Model
• The Laplace Bridge: From Hilbert Space to Probabilities and back
• Experimental Results
State of the Union

- Data: State of the Union corpus
  - Continuous time feature “year”, 1790-2011
  - Discrete author feature “president”, 44 values
  - Use $K = 10$ topics

- Dirichlet Multinomial Regression [Mimno et al, 2008]
  - Time features: 100 radial basis functions, width 5 years
  - Author features: 44 mutually exclusive binary features

- Kernel Topic Model
  - Rational Quadratic kernel with width 5 years
  - Additional constant term if authors are not the same
State of the Union: Kernel Topic Model
State of the Union: Linear Model

The chart illustrates the frequency of certain terms over time. The terms are categorized into groups such as "law, secretary, service," "department, act, interest," "citizens, war, commerce," and so on. The x-axis represents years ranging from 1800 to 2000, while the y-axis shows the frequency of the terms, with values ranging from 0 to 1. The chart shows how the frequency of these terms has evolved over time.
State of the Union: Perplexity

Initial perplexity equals size of vocabulary $V = 5,000$
More Perplexity

• Wiki dataset
  – List of probability topics
  – $D = 318$ documents
  – Squared exponential kernel on link distance

• NIPS dataset
  – Globersen et al, 2007
  – NIPS papers and their citation structure

Discontinuities in the learning curve for the KTM correspond to optimization of the hyper parameters
Conclusions

• Make use of document meta-data and link structure to improve topic models
• Use kernel/Gaussian process framework to integrate such context information into topic distribution in LDA
• Laplace bridge is a useful tool to bridge the divide between an unbounded vector space and probability vectors → other applications?