Context as Filtering

Keywords: Language Modeling, Particle Filters, Change Point Analysis, LDA, Dirichlet Mixtures

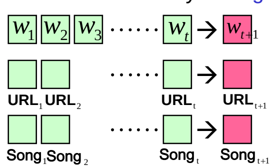
Daichi Mochihashi
ATR, Spoken Language Communication Research Laboratories, Japan
daichi.mochihashi@atr.jp

Yuji Matsumoto
Computational Linguistics Laboratory, NAIST, Japan
matsu@is.naist.jp

Abstract: For a prediction problem for high-dimensional discrete sequences, we propose a solution using online change point analysis by Particle Filters combined with probabilistic text models LDA and DM.

Language Modeling

Prediction from history on High-dimensional discrete data

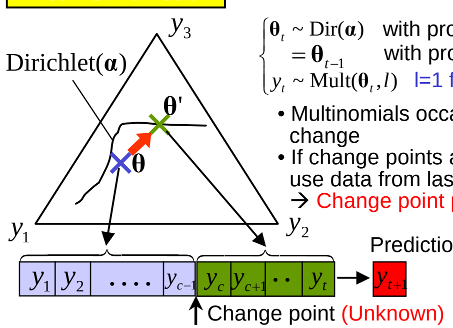


(Speech Recognition, Machine Translation)
(Hyperlink Navigation)
(Collaborative Filtering)
:
This problem is ubiquitous.

Problem: How long context should we use?
– Hidden state of multinomial distributions may differ
– Beyond "Bag of Words" assumption

Aim of this research: Estimate next word from a long history by introducing a state space model in Multinomial space.

Mean Shift Model



$$\begin{cases} \boldsymbol{\theta}_t \sim \mathrm{Dir}(\boldsymbol{\alpha}) & \text{with probability } \rho \\ \quad = \boldsymbol{\theta}_{t-1} & \text{with probability } (1-\rho) \\ y_t \sim \mathrm{Mult}(\boldsymbol{\theta}_t, l) & l=1 \text{ for simplicity} \end{cases}$$

- Multinomials occasionally change
- If change points are known, use data from last change point → Change point problem

Change Point Probabilities

By Bayes' Theorem,

p(change|observed) $\propto$ p(observed|change)p(change)

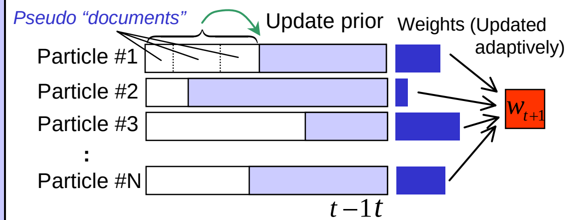
$$= \begin{cases} \text{p(observed|change=1)p(change=1)} \\ \text{p(observed|change=0)p(change=0)} \end{cases}$$

$$= \begin{cases} \text{Prior prediction } p(y_t \mid \boldsymbol{\alpha}) \times \rho \\ \text{Posterior prediction } p(y_t \mid \boldsymbol{\alpha}, y_c \cdots y_{t-1}) \times (1-\rho) \end{cases}$$

$$= \begin{cases} \rho \times \alpha_{y_t} / \sum_y \alpha_y \\ (1-\rho) \times (\alpha_{y_t} + n(y_t)) / \sum_y (\alpha_y + n(y)) \end{cases}$$ $n(y)$: # of y in $y_c \cdots y_{t-1}$

Multinomial Particle Filter

■ Simultaneous Bernoulli trials → Multinomial Particle Filter



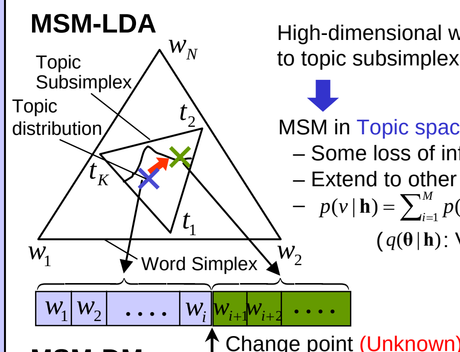
■ Online estimate of ρ: Expectation of Beta posterior

$$\langle \rho_t \rangle = \frac{\alpha + (\#\text{ of change points thus far})}{\alpha + \beta + t - 1}$$ α, β: hyperparameters

■ Problem: Extremely high dimensionality of language (Semantic correlations between words)

LDA, Dirichlet Mixtures (DM) → MSM-LDA, MSM-DM

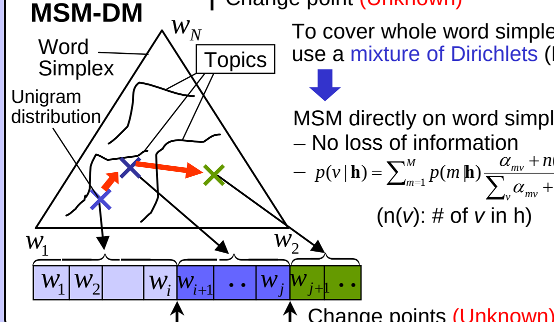
MSM-LDA



High-dimensional word simplex to topic subsimplex (LDA)

MSM in Topic space
– Some loss of information
– Extend to other mixture models
– $p(v \mid \mathbf{h}) = \sum_{i=1}^{M} p(v \mid \theta_i) \langle \theta_i \rangle_{q(\boldsymbol{\theta}|\mathbf{h})}$
($q(\boldsymbol{\theta} \mid \mathbf{h})$: VB-EM on h)

MSM-DM

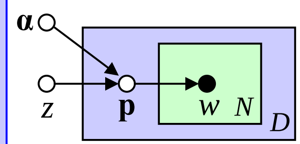


To cover whole word simplex, use a mixture of Dirichlets (DM)

MSM directly on word simplex
– No loss of information
– $p(v \mid \mathbf{h}) = \sum_{m=1}^{M} p(m \mid \mathbf{h}) \dfrac{\alpha_{mv} + n(v)}{\sum_v \alpha_{mv} + n(v)}$
($n(v)$: # of v in h)

Dirichlet Mixtures: Mixture of Polya distributions

(Sjolander et al. 1996; Yamamoto et al. 2005)



- w: Word
- z : Latent variable for each document
- p : Unigram distribution (integrated out)
- $\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_M$: Dirichlet hyperparameters for p

$$p(D \mid \lambda, \alpha_1 \cdots \alpha_M) = \prod_{i=1}^{D} \sum_{m=1}^{M} \lambda_m \frac{\Gamma(\sum_v \alpha_{mv})}{\Gamma(\sum_v \alpha_{mv} + n_{iv})} \prod_v \frac{\Gamma(\alpha_{mv} + n_{iv})}{\Gamma(\alpha_{mv})}$$

n_{iv} : occurrences of word v in document i

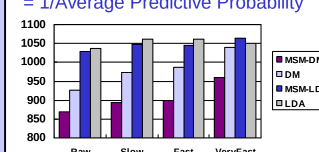
As opposed to LDA:
✓ Unitopic
✓ Can model whole word simplex
✓ Lower document perplexity than LDA
- "Cache" property of Polya distributions (Minka 2000)
✓ Number of parameters equal to LDA ($\boldsymbol{\lambda}, \boldsymbol{\alpha}_1 \ldots \boldsymbol{\alpha}_M$)
✓ Dirichlet Process Extension is now under development

Experiments

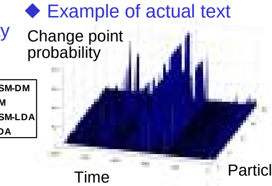
- British National Corpus (wide coverage of topics)
- 11,032,233 words, Lexicon = 52,846 words
- Evaluation texts: 100 documents x 100 sentences
  – Raw: Extract contiguous 100 sentences
  – Slow~VeryFast: Randomly skip to sample 100 sentences (Slow: a little skip, Fast: large, VeryFast: very large)

Experimental Results

◆ Perplexity = 1/Average Predictive Probability



◆ Example of actual text



Summary and Future Directions

✓ Introduced a MSM of natural language with LDA/DM
– Online inference with a Particle Filter
✓ Multiple observations, Gibbs for a whole document (OK)
✓ How to estimate segmentation and LDA/DM parameters simultaneously (without using a slow Gibbs)?