Correlated Topic Models

Richard (Fangjian) Guo[1] Yezhou Huang[1]

[1] Department of Computer Science,
Duke University

Jan 22, 2014

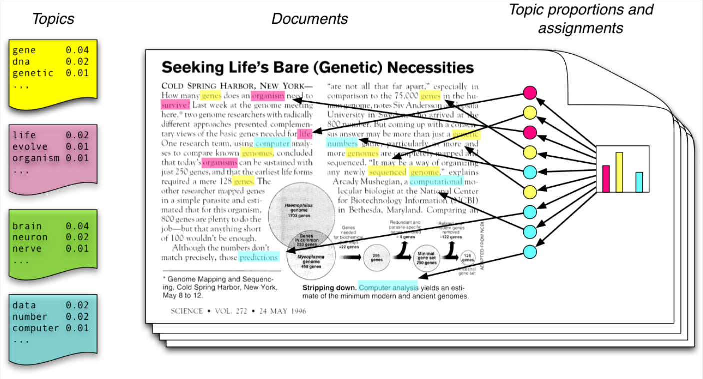
Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Task of Topic Models


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

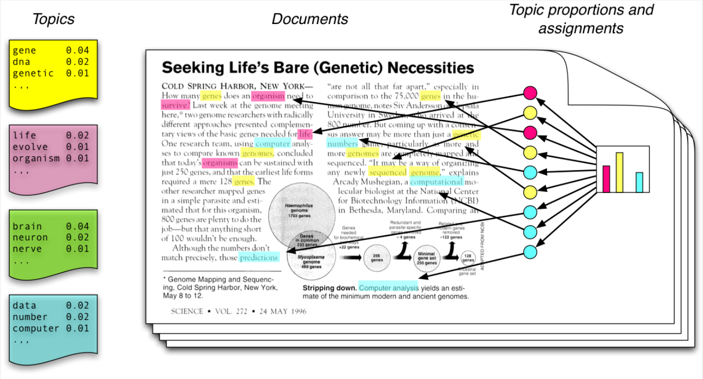
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure : The thematic information discovered by LDA (from Blei 2012).

Task: discovering and annotating large archives of documents with thematic information.

Intuitions


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

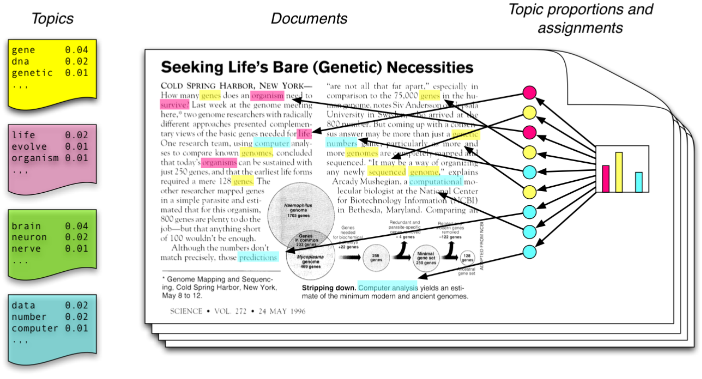
SCIENCE • VOL. 272 • 24 MAY 1996

Figure : "Documents exhibit multiple topics" and are generated by "an imaginary random process" (from Blei 2012).

Outline

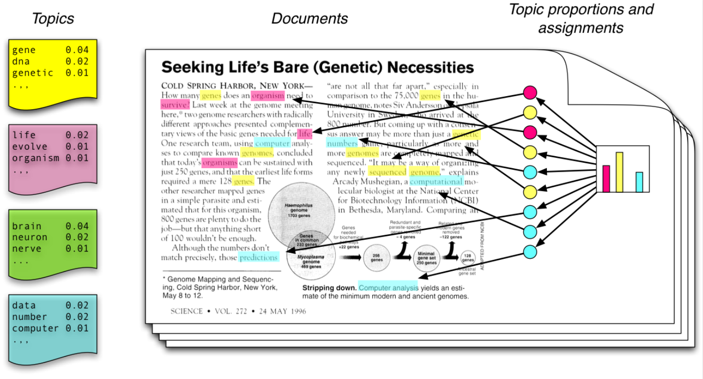
1. From LDA to CTM
   - Intuitions behind LDA
   - Generative Process
   - Relaxing An Assumption of LDA to Obtain CTM
   - Relaxing Some Other Assumptions of LDA
2. Inference
   - EM scheme
   - Variational EM
   - The CTM case
3. Supplementary
   - Review of Probability Distributions
   - Model Comparison with Predictive Perplexity
   - Examples of Language Generation

Description of Generative Process


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Topic: a distribution over a fixed vocabulary; different topics have different distributions over the same vocabulary.
Assumption: topics are specified before any data has been generated; in other words, the probabilistic model, which we will see, assumes that the topics are generated first, before the documents.

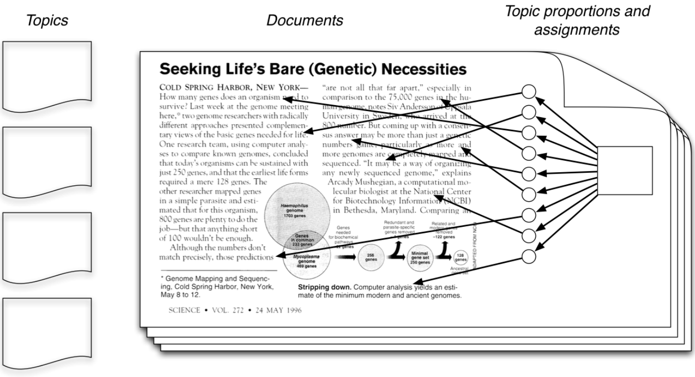
Description of Generative Process


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

For each document, we generate the words in a two-stage process.

- Randomly choose a distribution over the topics (the histogram at right).
- For each word in the document,
  - Randomly choose a topic assignment (the colored coins) from the distribution over topics selected in step #1.
  - Randomly choose a word from the corresponding topic (distribution over the vocabulary).

Goal with Generative Process



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways
256 genes
Redundant and parasite-specific genes removed
Minimal gene set 250 genes
Related and modern genes removed
−122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

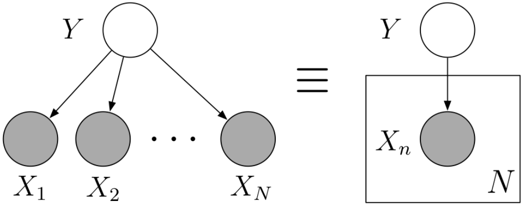
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure
- We represent the Generative Process as a Probabilistic Graphical Model

slide from Blei MLSS '09 talk

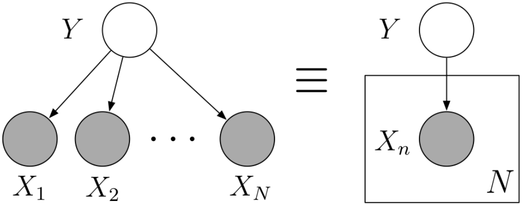
Graphical models (Aside)



- Nodes are random variables
- Shaded nodes are observed variables
- Edges denote possible dependence
- Plates denote replicated structure (sharing the same set of attributes and same probabilistic model)

slide from Blei MLSS '09 talk

Graphical models (Aside)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- This graph corresponds to

$$p(y, x_1, \ldots, x_n) = p(y) \prod_{n=1}^{N} p(x_n | y)$$

slide from Blei MLSS '09 talk

LDA as a Graphical Model

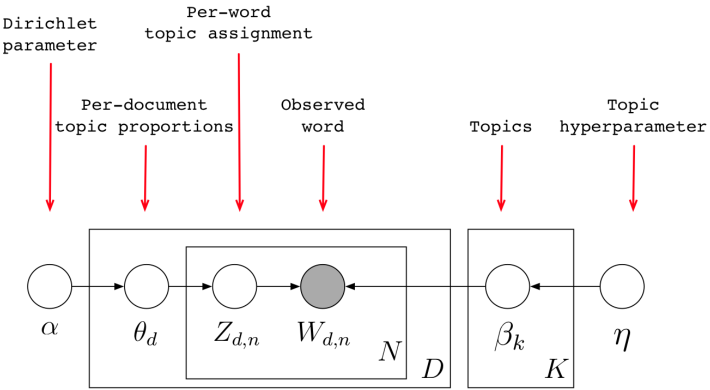


Figure : from Blei MLSS '09 talk

LDA as a Graphical Model

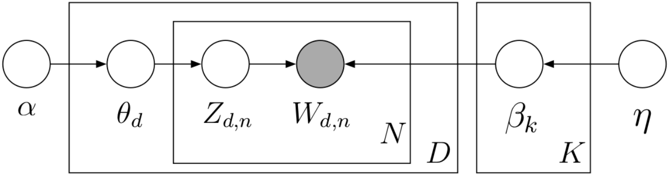


Figure : from Blei MLSS '09 talk

LDA is a mixed-membership model:

- Each group (document) is modeled with a mixture (of topics)
- The mixture components (topics) are shared across all the groups (documents)
- The mixture proportions (topic proportions) are vary from group (document) to group (document)

Refer to Blei, Mixed Membership Models, Princeton COS597C 2011 Fall

LDA as a Graphical Model

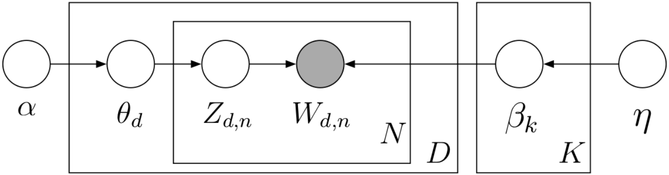


Figure : from Blei MLSS '09 talk

Observed variables are words of the documents.

- Each document d is a group of words $w_{d,1:N}$.
- Each word $w_{d,n}$ is a multinomial (for better understanding, multinoulli or categorical) value among V words.

LDA as a Graphical Model

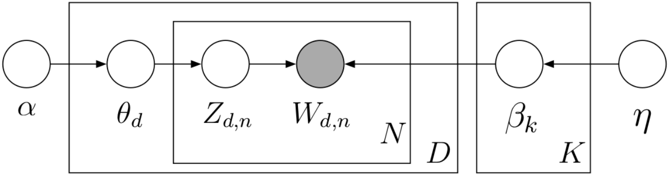


Figure : from Blei MLSS '09 talk

Hidden variables are the topic structure.

- Topics $\beta_{1:K}$:
  - Each β_k is proportions of words in the vocabulary, a point on the $V-1$ simplex.
  - $\beta_{1:K}$ are the multinomial (multinoulli or categorical) parameters for any word $w_{d,n}$.

LDA as a Graphical Model

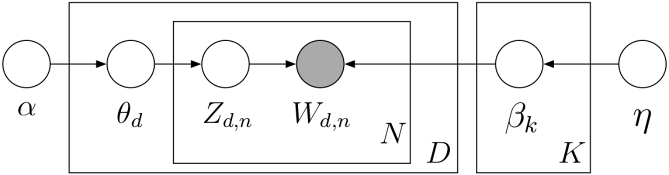


Figure : from Blei MLSS '09 talk

Hidden variables are the topic structure.

- Topic proportions $\theta_{1:D}$:
  - θ_d is the topic proportions for the dth document, a point on the $K-1$ simplex.
  - $\theta_{d,k}$ is the topic proportion for topic k in document d.
- Topic assignments $z_{1:D,1:N}$:
  - z_d is the topic assignments for the dth document.
  - $z_{d,n}$ is the topic assignment for the nth word in document d: a multinomial (multinoulli or categorical) indicator for the word, namely, the k value.

LDA as a Graphical Model

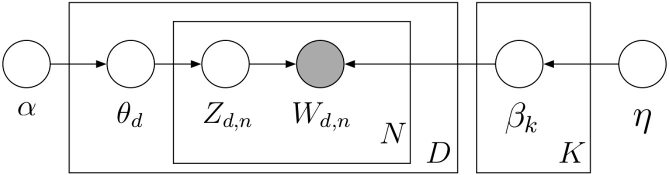


Figure : from Blei MLSS '09 talk

Generative Process with Probabilistic Model:From LDA to CTM

- Draw $\beta_k \sim \text{Dir}_V(\eta)$, for $k \in \{1, \ldots, K\}$
- For each document d, $d \in \{1, \ldots, D\}$,
    - Draw $\theta_d \sim \text{Dir}_K(\alpha)$
    - For each word n in document d, where $n \in \{1, \ldots, N\}$
        - Draw $z_{d,n} \sim \text{Cat}(\theta_d)$
        - Draw $w_{d,n} \sim \text{Cat}(\beta_{z_{d,n}})$

LDA as a Graphical Model

Joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}) \\ = \prod_{k=1}^{K} p(\beta_k) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right)$$

Posterior of the hidden variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N})}{P(w_{1:D,1:N})}$$

- The denominator is the marginal probability.
- It is hard to be computed by summing the joint distribution over every possible instances of hidden topic structure.
- There are exponentially large number of possible instances hidden topic structure (with just $z_{d,n}$, we have billions of hidden variables).

LDA as a Graphical Model

Approximate posterior inference algorithms:

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)

For comparison, see Mukherjee and Blei (2009) and Asuncion et al. (2009).

Richard will introduce the application of Mean field variational method in Correlated Topic Models.

slide from Blei MLSS '09 talk

Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Respective correlation Assumptions of LDA and CTM

LDA Assumption: No (very weak) correlation among topics.

- Truth: a document about **fossil fuels** is more likely to also be about **geology** than about **genetics**.
- Solution: the correlated topic model (Blei and Lafferty, 2005, 2007), and pachinko allocation machine (Li et al., 2010)

Correlated Topic Model: Given a K-vector μ and a $K \times K$ covariance matrix Σ:

- Draw $\eta_d | \mu, \Sigma \sim \mathrm{N}(\mu, \Sigma)$
- $\theta_d = f(\eta_d) = \frac{\exp(\eta_d)}{\sum_{k=1}^{K} \exp(\eta_{d,k})}$

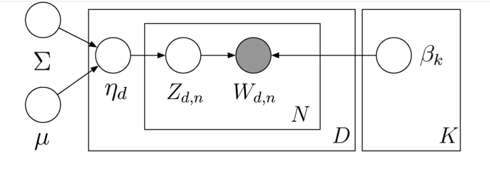


Figure : Graphical model representation of the correlated topic model (from Blei and Lafferty, 2005, 2007).

Dirichlet

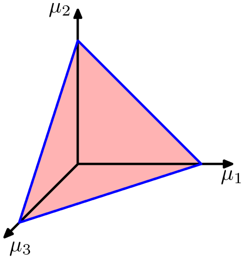


Figure : from Bishop, Pattern Recognition and Machine Learning, 2006

- The **Dirichlet** is a distribution on a simplex (a bounded manifold), positive vectors that sum to 1.
- It assumes that components are nearly independent (only very weak correlation from the constraint that they sum to 1).

Dirichlet

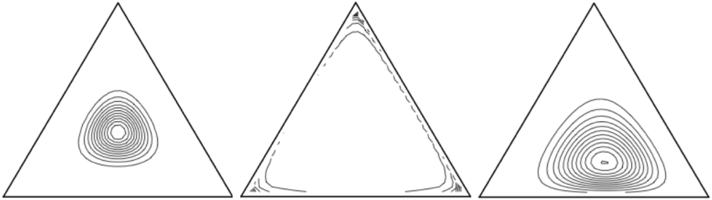


Figure : from Blei MLSS '12 talk.

- The **Dirichlet** is a distribution on a simplex (a bounded manifold), positive vectors that sum to 1.
- It assumes that components are nearly independent (only very weak correlation from the constraint that they sum to 1).

Logistic Normal

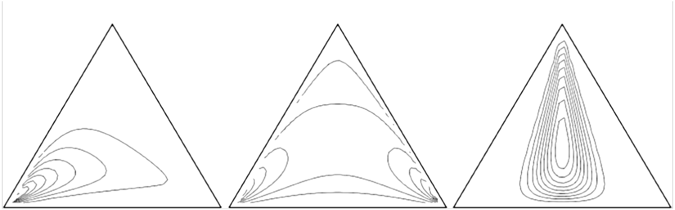


Figure : from Blei and Lafferty, 2005, 2007.

- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The Figure shown above is an example densities of the logistic normal on the 2-simplex.
- From left: diagonal covariance and nonzero-mean, negative correlation between topics 1 and 2, positive correlation between topics 1 and 2.

Outline

1. From LDA to CTM
   - Intuitions behind LDA
   - Generative Process
   - Relaxing An Assumption of LDA to Obtain CTM
   - Relaxing Some Other Assumptions of LDA
2. Inference
   - EM scheme
   - Variational EM
   - The CTM case
3. Supplementary
   - Review of Probability Distributions
   - Model Comparison with Predictive Perplexity
   - Examples of Language Generation

Relaxing Some Other Assumptions for Better Models

Assumption: "Bag of words" (in NLP).

- It assumes that the order of the words in the document does not matter.
- It is about exchangeability.
    - De Finetti's theorem: If a collection of random variables are exchangeable, then their joint can be written as a Bayesian model.
    - $p(x_1, x_2, \ldots, x_n) = \int p(\theta) \prod_{i=1}^{n} p(x_i|\theta) d\theta$
- For uncovering the thematic information, it is reasonable, though unrealistic:
    - imagine shuffling the words of the article in 1st Figure
    - we will still be able to realize the article is related to genetics.
- For some other tasks, like language generation, this assumption is not appropriate:
    - The topics generate words should be conditional on the previous word.
    - Solution: a topic model that switches between LDA and a standard HMM (Griffiths et al., 2005).
    - Parameter space is significantly larger, but language modeling performance is improved.

Refer to Blei, 2012

Relaxing Some Other Assumptions for Better Models

Assumption: Order of documents does not matter.

- It may be unrealistic when analyzing long-running collections.
- In such collections we might assume that the topics change over time.
  - Solution: the dynamic topic model (Blei and Lafferty, 2006).
  - A topic is now a sequence of distributions over words.
  - We can find and track how a topic has changed over time.

Assumption: The number of topics is assumed known and fixed.

- Solution: Bayesian nonparametric topic model (HDP, Teh et al., 2006).
- Recall the DP we have seen in first lecture.
- The number of topics is determined by the data during posterior inference.
- New documents can exhibit previously unseen topics.
- Have been extended to hierarchies of topics (Blei et al., 2003 & 2010):
  - Find a tree of topics;
  - Move from more general to more concrete;
  - A particular structure of the model is inferred from the data.
  - We will see it in next lecture.

Refer to Blei, 2012

Relaxing Some Other Assumptions for Better Models

Assumption: every word is likely in any topic.

- Truth: "wrench" will be particularly unlikely in a topic about cats.
- Solution: the spherical topic model allows words to be unlikely in a topic (Reisinger et al., 2010)

Assumption: The count of a word in a document does not influence the probability that the word appears in the document.

- Truth: Suppose that there is a natural "sport" topic in a corpus, with the words "rugby" and "hockey" being equally common overall. Within a document, though, one appearance of "rugby" makes a second appearance of "rugby" more likely than a first appearance of "hockey." (Doyle and Elkan, 2009).
- Recall the Zipf's law (the probability of occurrence of a word follows a power law) we have seen in last lecture about Two-Stage Language Models (Pitman-Yor CRP).
- Solution: "bursty" topic models (Doyle and Elkan, 2009).

Refer to Blei, 2012

Some Other Models

- Sparse topic models enforce further structure in the topic distributions (Wang and Blei, 2009).
- Author-topic model makes a topic be drawn from a author-specific topic proportions.
- Relational topic model assumes that each document is modeled as in LDA and that the links between documents depend on the distance between their topic proportions.

Refer to Blei, 2012

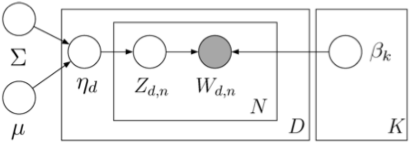
Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Inference objective: MLE estimates



Excerpted from Blei et al's Correlated Topic Models

Objective: the MLE estimates for model parameters $\Sigma, \mu, \beta_{1:K}$ by maximizing

$$P(\{W_{1,n}, W_{2,n}, \cdots, W_{D,n}\} \mid \Sigma, \mu, \beta_{1:K}) = \prod_{d=1}^{D} P(W_{d,n} \mid \Sigma, \mu, \beta_{1:K}).$$

Difficulty: the observed-data likelihood is available only by marginalizing out the latent variables $\eta_d, \{Z_{d,n}\}$.

EM algorithm

The EM algorithm is a framework for getting MLE estimates (extendable for MAP estimates) on models with **latent variables**.
Considering a model with observed variable X and latent variable Z,

$$P(X \mid \theta) = \sum_Z P(X, Z \mid \theta).$$

Introducing a distribution $q(Z)$ as an **approximation** for the **posterior** of the latent variable. For **any** q, we have the following equality due to $P(X, Z \mid \theta) = P(X \mid \theta) P(Z \mid X, \theta)$.

$$\log P(X \mid \theta) = L(q, \theta) + KL(q \| p),$$

where

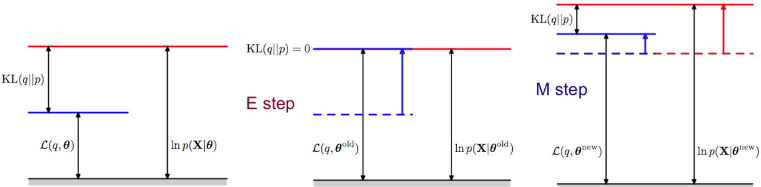
$$L(q, \theta) = \sum_Z q(Z) \log \left[\frac{P(X, Z \mid \theta)}{q(Z)} \right],$$

$$KL(q \| p) = - \sum_Z q(Z) \log \left[\frac{P(Z \mid X, \theta)}{q(Z)} \right] \geq 0.$$

Recall that $KL(q \| p) = 0$ only when $p(Z \mid X, \theta) = q(Z)$.

EM algorithm

Two-step iteration scheme for maximizing $\log P(X \mid \theta) = L(q, \theta) + KL(q\|p)$.


$\mathrm{KL}(q\|p)$
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$
$\mathrm{KL}(q\|p) = 0$
E step
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$
$\mathrm{KL}(q\|p)$
M step
$\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$
$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{new}})$

Excerpted from Bishop's PRML Book

E step: Keeping the model parameters θ^{old} fixed, maximizing $L(q, \theta^{old})$ by setting

$$q \leftarrow P(Z \mid X, \theta^{old}),$$

so that the KL term vanishes.
M step: Keeping q fixed, maximizing $L(q, \theta)$ by optimizing θ

$$\theta^{new} \leftarrow \operatorname*{argmax}_{\theta} L(q, \theta),$$

after which both L and KL will increase.

Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Variational EM

- **Difficulty with EM**: Setting $q \leftarrow P(Z \mid X, \theta^{old})$ requires **solving the posterior for the latent Z analytically**, which is usually **impossible without conjugacy**.
- **Variational EM**:
  In the **E step**, set q to an **approximated posterior** of the latent variables so that the KL-divergence between q and the true posterior is **as small as possible**.
  This approximation is done by **parametrizing q within a family of functions** and then

  $$q \leftarrow q(Z \mid \phi),$$

  where we seek

  $$\phi \leftarrow \operatorname*{argmin}_{\phi} KL(q(Z \mid \phi) \| p) = \operatorname*{argmax}_{\phi} L(q(Z \mid \phi), \theta).$$

  That is, we construct **a lower bound** on $\log p(X \mid \theta)$ that is as high as possible by choosing q from **a parametric family**.
  The **M step** remains the same as before

  $$\theta^{new} \leftarrow \operatorname*{argmax}_{\theta} L(q(Z \mid \phi), \theta).$$

Variational EM

1. What parametric family are looking into?
   Mean-field approximation: q is factorized into

$$q(Z \mid \phi) = q(Z_1 \mid \phi_1)q(Z_2 \mid \phi_2) \cdots q(Z_r \mid \phi_r), \quad Z \in \mathbb{R}^r$$

.

2. What is $L(q(Z \mid \phi), \theta)$?

$$\begin{aligned} L(q(Z \mid \phi), \theta) &= \sum_Z q(Z \mid \phi) \log \left[\frac{P(X, Z \mid \theta)}{q(Z \mid \phi)} \right] \\ &= \sum_Z q(Z\phi) \log P(X, Z \mid \theta) - \sum_Z q(Z \mid \phi) \log q(Z \mid \phi) \\ &= \mathbb{E}_{q(\cdot \mid \phi)} \log P(X, Z \mid \theta) + H_q(\phi). \end{aligned}$$

3. Can we maximize it analytically?
   1. If yes, that would require **a closed-form solution** to $\frac{\partial L}{\partial \phi} = 0$.
   2. If not, what can we do?

Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

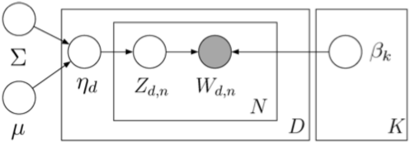
3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Variational EM for Correlated Topic Models

In CTM, considering **only one document** (it suffices because likelihood is factorized into documents), we have

$$
\begin{aligned}
&\log P(w_{1:N} \mid \mu, \Sigma, \beta) \\
&= \mathbb{E}_{q(\cdot|\phi)} \log P(w_{1:N}, \eta, z_n \mid \theta) + H_q(\phi) + KL(q(\eta, z_n \mid \phi) \| P(\eta, z_n \mid w_{1:n}, \mu, \Sigma, \beta)) \\
&\geq \mathbb{E}_{q(\cdot|\phi)} \log P(w_{1:N}, \eta, z_n \mid \theta) + H_q(\phi) \quad \text{(KL dropped)} \\
&\geq \mathbb{E}_q \log \left[\prod_{n=1}^{N} P(w_n \mid \beta, z_n) \right] + \mathbb{E}_q \log \left[\prod_{n=1}^{N} P(z_n \mid \eta) \right] + \mathbb{E}_q \log N_k(\eta \mid \mu, \Sigma) + H_q(\phi)
\end{aligned}
$$

due to the form of the likelihood



Excerpted from Blei et al's Correlated Topic Models

All we need is the L. We can forget the KL.

Variational EM for Correlated Topic Models

We use mean-field approximation and set $q(\cdot)$ to the **parametric form** of

$$q(\eta_{1:K}, z_{1:N} \mid \lambda, \nu^2, \phi) = \prod_{i=1}^{K} N(\eta_i \mid \lambda_i, \nu_i^2) \prod_{n=1}^{N} \text{Category}(z_n \mid \phi_n),$$

where the variational parameters are $\lambda \in \mathbb{R}^K, \nu^2 \in \mathbb{R}_+^K, \phi \in \Delta_K^N$.
The $q(\cdot)$ for $\eta_{1:K}$ is a product of **univariate** Gaussians.

- *"Since the variational parameters are fit using a* **single** *observed document* $w_{1:N}$*, there is* **no advantage** *in introducing a* **non-diagonal** *variational covariance matrix."*
- As long as we are *happy* with this, we can go on to do the math, term by term.
- Okay: $\mathbb{E}_q \log \left[\prod_{n=1}^N P(w_n \mid \beta, z_n)\right]$
- Straightforward: $H_q(\lambda, \nu^2, \phi)$
- Somewhat *trace-tricky*: $\mathbb{E}_q \log N_k(\eta \mid \mu, \Sigma)$

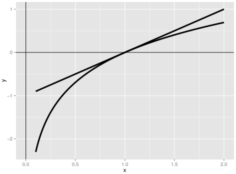
Variational EM for Correlated Topic Models

Finally, we are stuck by this one.

$$\begin{aligned}\mathbb{E}_q \log \prod_{n=1}^{N} P(z_n \mid \eta) &= \mathbb{E}_q \sum_{n=1}^{N} \log \left[\frac{\exp(\eta_{z_n})}{\sum_{k=1}^{K} \eta_k} \right] \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \lambda_k \phi_{n,k} - N \mathbb{E}_{\mathcal{N}(\eta \mid \lambda, \nu^2)} \log \left[\sum_{k=1}^{K} \exp(\eta_k) \right].\end{aligned}$$

To pull the intractable sum out of the log, we use the approximation

$$\log x \approx x - 1, \quad \text{with } x \text{ around } 1$$



$\log \left[\sum_{k=1}^{K} \exp(\eta_k) \right] = \log \left[\zeta \zeta^{-1} \sum_{k=1}^{K} \exp(\eta_k) \right] \geq$
$\log \zeta + \zeta^{-1} \sum_{k=1}^{K} \exp(\eta_k) - 1.$
Now we can get closed-form expectation.
Note: This is guaranteed to be a **lower-bound**!

Variational EM for Correlated Topic Models

- **E-step**: Now with *an analytical lower bound* $L_d(\lambda_d, \nu_d^2, \phi_d, \zeta_d; \mu, \Sigma, \beta)$ for each document, we can maximize it w.r.t. *variational parameters* $(\lambda_d, \nu_d^2, \phi_d, \zeta_d)$ either by closed-form solution or gradient-based optimization.
- **M-step**: Maximize sum of L_d over *all documents* w.r.t. *model parameters* (μ, Σ, β). Surprisingly, they have *simple solutions*, i.e. those formed by *expected sufficient statistics*. This is an advantage of *the exponential family*.

$$L(\mu, \Sigma, \beta \mid \boldsymbol{w}_{1:D}) \geq \sum_{d=1}^{D} \mathbb{E}_{q_d} \log P(\eta_d, z_d, w_d \mid \mu, \Sigma, \beta) + \sum_{d=1}^{D} H(q_d)$$

Summary

- **Using logistic Normal to introduce covariance structure among categorical proportions.**
- **Using variational EM to get MLE/MAP estimates for models with latent variables, yet without full conjugacy.**
- **Inside the E step, the variational inference constructs a non-tight lower bound on the observed data likelihood.**
- **The lower bound should be sought as high as possible by seeking an approximated posterior distribution for the latent variables within some parametric family.**

Outline

1. From LDA to CTM
   - Intuitions behind LDA
   - Generative Process
   - Relaxing An Assumption of LDA to Obtain CTM
   - Relaxing Some Other Assumptions of LDA
2. Inference
   - EM scheme
   - Variational EM
   - The CTM case
3. Supplementary
   - Review of Probability Distributions
   - Model Comparison with Predictive Perplexity
   - Examples of Language Generation

Review of Probability Distributions

Bernoulli: $\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$, where $x \in 0, 1$ and $0 \leq \mu \leq 1$.

Binomial: $\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$.

Beta: $\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$, where $0 \leq \mu \leq 1$.

Conjugacy: Given a Beta prior, and a Binomial (Bernoulli as a special case) likelihood, we have the posterior as

$$\begin{aligned} p(\mu|m,N,a,b) &\sim \mu^{m+a-1}(1-\mu)^{N-m+b-1} \\ &= \frac{\Gamma(N+a+b)}{\Gamma(m+a)(N-m+b)}\mu^{m+a-1}(1-\mu)^{N-m+b-1} \\ &= \text{Beta}(\mu|m+a, N-m+b). \end{aligned}$$

Review of Probability Distributions

Multinoulli (Categorical): Cat$(\mathbf{x}|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$, where $x_k \in 0, 1$, $\sum_{k=1}^{K} x_k = 1$, $0 \leq \mu_k \leq 1$, and $\sum_{k=1}^{K} \mu_k = 1$.

Multinomial: Mult$(\mathbf{m}|N, \mu) = \binom{N}{m_1 m_2 \cdots_K} \prod_{k=1}^{K} \mu_k^{m_k}$.

Dirichlet: Dir$(\mu|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_1)...\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$, where $0 \leq \mu \leq 1$.

Conjugacy: Given a Dirichlet prior, and a Multinomial (Multinoulli or Categorical as a special case) likelihood, we have the posterior as

$$
\begin{aligned}
p(\mu|\mathbf{m}, N, \alpha) &= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k + N)}{\Gamma(\alpha_1 + m_1) \ldots \Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1} \\
&= \text{Dir}(\mathbf{m}|\alpha + \mathbf{m}).
\end{aligned}
$$

Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Performance Comparison with Predictive Perplexity

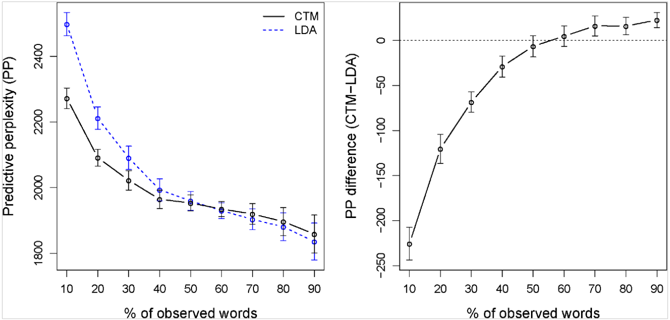
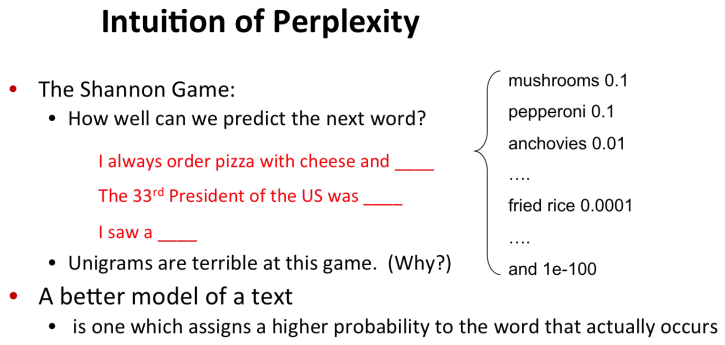


Figure : (Left) The 10-fold cross-validated predictive perplexity for partially observed held-out documents from the 1960 Science corpus ($K = 50$). Lower numbers indicate more predictive power from the CTM. (Right) The mean difference in predictive perplexity. Numbers less than zero indicate better prediction from the CTM.

Figure from Blei and Lafferty, 2007.

Perplexity



Intuition of Perplexity

- The Shannon Game:
  - How well can we predict the next word?

    I always order pizza with cheese and ____

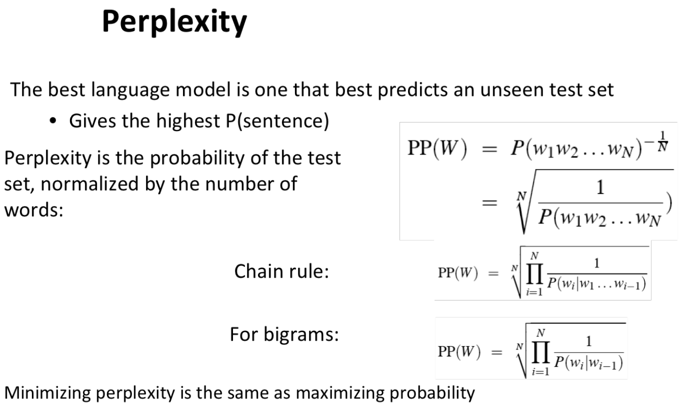
    The 33rd President of the US was ____

    I saw a ____
  - Unigrams are terrible at this game. (Why?)
- A better model of a text
  - is one which assigns a higher probability to the word that actually occurs

mushrooms 0.1
pepperoni 0.1
anchovies 0.01
….
fried rice 0.0001
….
and 1e-100

Slide from Jurafsky and Manning, NLP course, 2012

Perplexity



Perplexity

The best language model is one that best predicts an unseen test set

- Gives the highest P(sentence)

Perplexity is the probability of the test set, normalized by the number of words:

$$
\begin{aligned}
\mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\
&= \sqrt[N]{\frac{1}{P(w_1 w_2 \ldots w_N}\Big)}
\end{aligned}
$$

Chain rule:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1 \ldots w_{i-1})}}$$

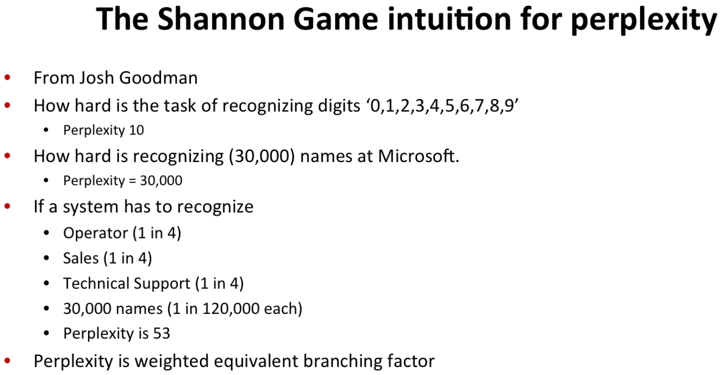
For bigrams:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

Minimizing perplexity is the same as maximizing probability

Slide from Jurafsky and Manning, NLP course, 2012

Perplexity

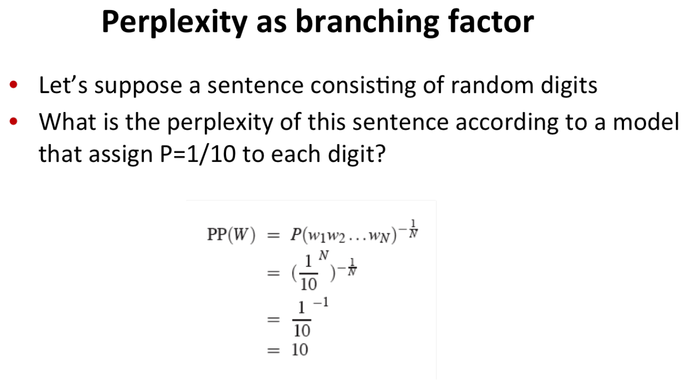


The Shannon Game intuition for perplexity

- From Josh Goodman
- How hard is the task of recognizing digits '0,1,2,3,4,5,6,7,8,9'
  - Perplexity 10
- How hard is recognizing (30,000) names at Microsoft.
  - Perplexity = 30,000
- If a system has to recognize
  - Operator (1 in 4)
  - Sales (1 in 4)
  - Technical Support (1 in 4)
  - 30,000 names (1 in 120,000 each)
  - Perplexity is 53
- Perplexity is weighted equivalent branching factor

Slide from Jurafsky and Manning, NLP course, 2012

Perplexity



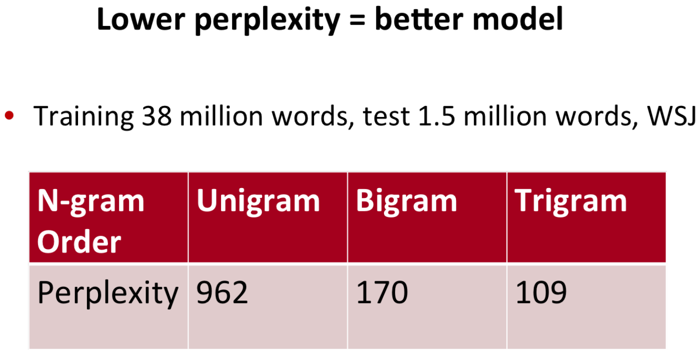
Perplexity as branching factor

- Let's suppose a sentence consisting of random digits
- What is the perplexity of this sentence according to a model that assign P=1/10 to each digit?

$$\begin{aligned} \mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\ &= (\frac{1}{10}^{N})^{-\frac{1}{N}} \\ &= \frac{1}{10}^{-1} \\ &= 10 \end{aligned}$$

Slide from Jurafsky and Manning, NLP course, 2012

Perplexity



Lower perplexity = better model

- Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Slide from Jurafsky and Manning, NLP course, 2012

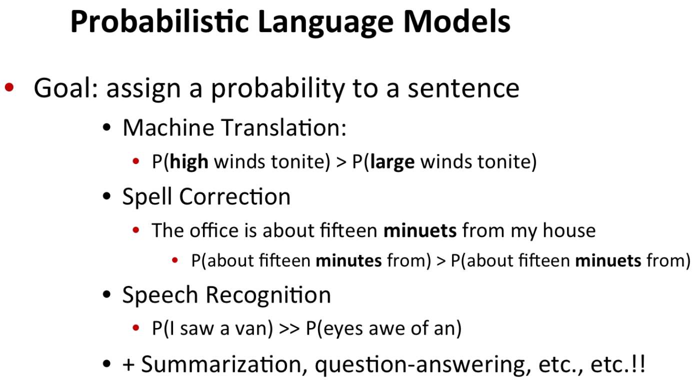
Outline

1 From LDA to CTM
- Intuitions behind LDA
- Generative Process
- Relaxing An Assumption of LDA to Obtain CTM
- Relaxing Some Other Assumptions of LDA

2 Inference
- EM scheme
- Variational EM
- The CTM case

3 Supplementary
- Review of Probability Distributions
- Model Comparison with Predictive Perplexity
- Examples of Language Generation

Language Generation



Probabilistic Language Models

- Goal: assign a probability to a sentence
    - Machine Translation:
        - P(**high** winds tonite) > P(**large** winds tonite)
    - Spell Correction
        - The office is about fifteen **minuets** from my house
            - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
    - Speech Recognition
        - P(I saw a van) >> P(eyes awe of an)
    - + Summarization, question-answering, etc., etc.!!

Slide from Jurafsky and Manning, NLP course, 2012