Hierarchical Dirichlet Process

How measure on measures measures measure on measures

Nanxin Chen

March 16, 2017

Outline

Outline

Different groups in reality

Example

Population Group: African, Asian, European, ...

Medical research:

- binary markers(SNPs)
- haplotypes
- genotype(pair of haplotypes)

Different groups in reality

Example

Population Group: African, Asian, European, ...

Medical research:

- binary markers(SNPs)
- haplotypes
- genotype(pair of haplotypes)

Different groups in reality

Example

Population Group: African, Asian, European, ...

Medical research:

- binary markers(SNPs)
- haplotypes
- genotype(pair of haplotypes)

Different groups in reality

Example

Population Group: African, Asian, European, ...

Medical research:

- binary markers(SNPs)
- haplotypes
- genotype(pair of haplotypes)

Different groups in reality

Example

Population Group: African, Asian, European, ...

Medical research:

- binary markers(SNPs)
- haplotypes
- genotype(pair of haplotypes)

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Different groups in reality

Example

Documents: University, Sports, ...

Information retrieval:

- words
- topics
- documents
- corpora

Dirichlet Process

$DP(\alpha_0, G_0)$:

- α_0: scaling parameter
- G_0: base probability measure
- If $G \sim DP(\alpha_0, G_0)$, then

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

with probability 1.

Dirichlet Process

$DP(\alpha_0, G_0)$:

- α_0: scaling parameter
- G_0: base probability measure
- If $G \sim DP(\alpha_0, G_0)$, then

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

with probability 1.

Dirichlet Process

$DP(\alpha_0, G_0)$:

- α_0: scaling parameter
- G_0: base probability measure
- If $G \sim DP(\alpha_0, G_0)$, then

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

with probability 1.

Dirichlet Process

$DP(\alpha_0, G_0)$:

- α_0: scaling parameter
- G_0: base probability measure
- If $G \sim DP(\alpha_0, G_0)$, then

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

with probability 1.

Hierachical Dirichlet Process

One of the dependent dirichlet process.

Dependent Dirichlet Process (DDP)

Nonparametric approaches to linking multiple DPs

The stick-breaking parameters β_k and ϕ_k becoms general stochastic processes.

Group of Dirichlet Process

$G_j \sim DP(\alpha_{0j}, G_{0j})$ In order to link them, consider

- $G_j \sim DP(\alpha_0, G_0(\tau))$
- G_0 in discrete parametric family

Group of Dirichlet Process

$G_j \sim DP(\alpha_{0j}, G_{0j})$ In order to link them, consider

- $G_j \sim DP(\alpha_0, G_0(\tau))$
- G_0 in discrete parametric family

Group of Dirichlet Process

$G_j \sim DP(\alpha_{0j}, G_{0j})$ In order to link them, consider

- $G_j \sim DP(\alpha_0, G_0(\tau))$
- G_0 in discrete parametric family

Solution with Hierarchical DP

$$G_0|\gamma, H \sim DP(\gamma, H) \tag{1}$$

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{2}$$

β_k becomes β_{jk} in group j and fixed k they are dependent.

Outline

Prerequisite

Exchangeable:

- Observation in each group $x_{j0}, x_{j1}, x_{j2}, \ldots$
- Groups $\mathbf{x_0}, \mathbf{x_1}, \ldots$

Prerequisite

Exchangeable:

- Observation in each group $x_{j0}, x_{j1}, x_{j2}, \dots$
- Groups $\mathbf{x_0}, \mathbf{x_1}, \dots$

Prerequisite

Exchangeable:

- Observation in each group $x_{j0}, x_{j1}, x_{j2}, \ldots$
- Groups $\mathbf{x_0}, \mathbf{x_1}, \ldots$

Notations

- factor θ_{ji}: components associated with the observation x_{ji}
- $F(\theta_{ji})$: distribution of x_{ji} given θ_{ji}

Example

Equation

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{3}$$

can be written as

$$\theta_{ji}|G_j \sim G_j \ \forall\text{group j}\ \forall\text{observation i}$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \ \forall\text{group j}\ \forall\text{observation i}$$

Notations

- factor θ_{ji}: components associated with the observation x_{ji}
- $F(\theta_{ji})$: distribution of x_{ji} given θ_{ji}

Example

Equation

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{3}$$

can be written as

$$\theta_{ji}|G_j \sim G_j \forall \text{group j} \forall \text{observation i}$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \forall \text{group j} \forall \text{observation i}$$

Notations

- factor θ_{ji}: components associated with the observation x_{ji}
- $F(\theta_{ji})$: distribution of x_{ji} given θ_{ji}

Example

Equation

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{3}$$

can be written as

$$\theta_{ji}|G_j \sim G_j \, \forall \text{group j} \, \forall \text{observation i}$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \, \forall \text{group j} \, \forall \text{observation i}$$

Notations

- factor θ_{ji}: components associated with the observation x_{ji}
- $F(\theta_{ji})$: distribution of x_{ji} given θ_{ji}

Example

Equation

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \tag{3}$$

can be written as

$$\theta_{ji}|G_j \sim G_j \,\forall \text{group j} \,\forall \text{observation i}$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}) \,\forall \text{group j} \,\forall \text{observation i}$$

Hierarchical Dirichlet Process

$$G_0|\gamma, H \sim DP(\gamma, H) \quad (4)$$

while H is the baseline probability measure, γ is the concentration parameter

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \quad (5)$$

while α_0 is another concentration parameter and can be group-dependent.

Outline

Stick-Breaking Construction

Stick-breaking construction for DP

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$$

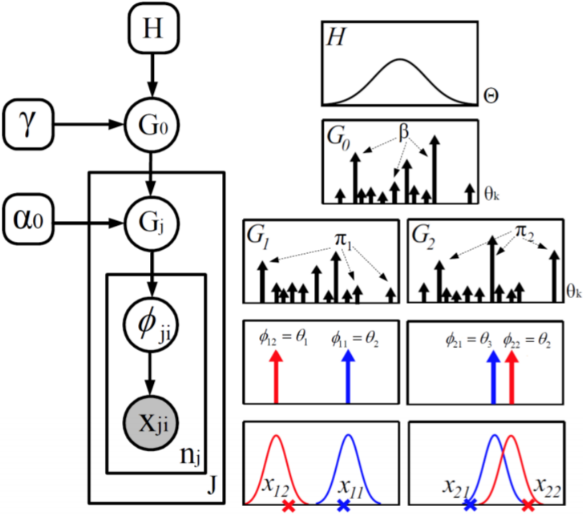
where $\phi_k \sim H$ independently and $\beta \sim GEM(\gamma)$

Support of each group

Because G_0 has support at points ϕ, all G_j has support at these points as well:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

Example


H
γ
G0
α0
Gj
φ ji
Xji
nj
J
H
Θ
G0
β
θk
G1
π1
G2
π2
θk
φ12 = θ1
φ11 = θ2
φ21 = θ3
φ22 = θ2
x12
x11
x21
x22

Outline

3 Two Construction
- Stick-Breaking Construction
- Chinese Restaurant Franchis

Connection between Integer Partition and Measurable Partition

Measurable Partition

$(A_1, \ldots, A_r)$ is a measurable partition.

Integer Partition

$(K_1, \ldots, K_r)$ is a finite partition of postive integers.

The connection is built by checking ϕ_k whether in partition A_l individually.

Relationship between π and β

$$(G_j(A_1), \dots G_j(A_r)) \sim Dir(\alpha_0 G_0(A_1), \dots \alpha_0 G_0(A_r))$$

can be simplified as

$$(\sum_{k \in K_1} \pi_{jk}, \dots, \sum_{k \in K_r} \pi_{jk}) \sim Dir(\alpha_0 \sum_{k \in K_1} \beta_k, \dots, \alpha_0 \sum_{k \in K_r} \beta_k)$$

Relationship between β and π_j

Integer partition $(\{1,...,k-1\},\{k\},\{k+1,k+2,...\})$ coresponding to

$$(\sum_{l=1}^{k-1}\pi_{jl},\pi_{jk},\sum_{l=k+1}^{\infty}\pi_{jl}) \sim Dir(\alpha_0\sum_{l=1}^{k-1}\beta_l,\alpha_0\beta_k,\alpha_0,\sum_{l=k+1}^{\infty}\beta_l) \quad (6)$$

Removing first element we get

$$\frac{1}{1-\sum_{l=1}^{k-1}\pi_{jl}}(\pi_{jk},\sum_{l=k+1}^{\infty}\pi_{jl}) \sim Dir(\alpha_0\beta_k,\alpha_0\sum_{l=k+1}^{\infty}\beta_l) \quad (7)$$

With replacement we get

$$\pi_{jk}^{'} \sim beta(\alpha_0\beta_k,\alpha_0(1-\sum_{l=1}^{k}\beta_l)) \quad (8)$$

Outline

3 Two Construction

- Stick-Breaking Construction
- Chinese Restaurant Franchis

Chinese Restaurant Franchise

Imagine a chain of restaurants,

- Menu are shared among all restaurants
- One dish is order by the first customer in each table

Chinese Restaurant Franchise

Imagine a chain of restaurants,

- Menu are shared among all restaurants
- One dish is order by the first customer in each table

Chinese Restaurant Franchise

Imagine a chain of restaurants,

- Menu are shared among all restaurants
- One dish is order by the first customer in each table

Chinese Restaurant Franchise

Obtain samples θ_{ji}, two steps:

- decide to follow some table else or create new table:
  - follow some existing table:

$$\sum_{\text{all table t in restaurant j}} \frac{\text{number of customers in table t}}{i-1+\alpha_0} \delta_{\text{dish in t}} \tag{9}$$

  - create new table:

$$\frac{\alpha_0}{i-1+\alpha_0} G_0 \tag{10}$$

Chinese Restaurant Franchise

Obtain samples θ_{ji}, two steps:

- decide to follow some table else or create new table:
  - follow some existing table:

$$\sum_{\text{all table t in restaurant j}} \frac{\text{number of customers in table t}}{i-1+\alpha_0} \delta_{\text{dish in t}} \tag{9}$$

  - create new table:

$$\frac{\alpha_0}{i-1+\alpha_0} G_0 \tag{10}$$

Chinese Restaurant Franchise

Obtain samples θ_{ji}, two steps:

- decide to follow some table else or create new table:
  - follow some existing table:

$$\sum_{\text{all table t in restaurant j}} \frac{\text{number of customers in table t}}{i-1+\alpha_0}\delta_{\text{dish in t}} \tag{9}$$

  - create new table:

$$\frac{\alpha_0}{i-1+\alpha_0}G_0 \tag{10}$$

Chinese Restaurant Franchise

Obtain samples θ_{ji}, two steps:

- decide to follow some table else or create new table:
  - follow some existing table:

$$\sum_{\text{all table t in restaurant j}} \frac{\text{number of customers in table t}}{i-1+\alpha_0}\delta_{\text{dish in t}} \tag{9}$$

  - create new table:

$$\frac{\alpha_0}{i-1+\alpha_0}G_0 \tag{10}$$

If new table is needed:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} \delta_{\text{dish k}} \tag{11}$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} H \tag{12}$$

If new table is needed:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} \delta_{\text{dish k}} \tag{11}$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} H \tag{12}$$

If new table is needed:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} \delta_{\text{dish k}} \tag{11}$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} H \tag{12}$$

Outline

Outline

4 Inference
- Posterior Sampling in the Chinese Restaurant Franchise
- Posterior Sampling with an Augmented Representation

Posterior Sampling in the Chinese Restaurant Franchise

Let z_{ji} be the component associated with observation x_{ji}.
Conditional Density of x_{ji} under component k given all other data is

$$f_k^{-x_{ji}}(x_{ji}) = \frac{\int f(x_{ji}|\phi_k)\prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\phi_k)h(\phi_k)d\phi(k)}{\int \prod_{j'i' \neq ji, z_{j'i'}=k} f(x_{j'i'}|\phi_k)h(\phi_k)d\phi(k)} \tag{13}$$

Posterior Sampling in the Chinese Restaurant Franchise

Sample table t_{ji} for customer i in restaurant j:

- if t is previously used, then it is proportional to number of customers in this table(*prior*) times the likelihood $f_{k_{jt}}^{-x_{ji}}(x_{ji})$
- if t is some new table, consider whether we need to sample a new component

Posterior Sampling in the Chinese Restaurant Franchise

Sample table t_{ji} for customer i in restaurant j:

- if t is previously used, then it is proportional to number of customers in this table(*prior*) times the likelihood $f_{k_{jt}}^{-x_{ji}}(x_{ji})$
- if t is some new table, consider whether we need to sample a new component

Posterior Sampling in the Chinese Restaurant Franchise

Sample table t_{ji} for customer i in restaurant j:

- if t is previously used, then it is proportional to number of customers in this table(*prior*) times the likelihood $f_{k_{jt}}^{-x_{ji}}(x_{ji})$
- if t is some new table, consider whether we need to sample a new component

Posterior Sampling in the Chinese Restaurant Franchise

If $t = t^{new}$, similar the equation in CRF:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} f_k^{x_{-ji}}(x_{ji}) \tag{14}$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} \int f(x_{ji}|\phi)h(\phi)d\phi \tag{15}$$

Posterior Sampling in the Chinese Restaurant Franchise

If $t = t^{new}$, similar the equation in CRF:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} f_k^{x_{-ji}}(x_{ji}) \tag{14}$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} \int f(x_{ji}|\phi)h(\phi)d\phi \tag{15}$$

Posterior Sampling in the Chinese Restaurant Franchise

If $t = t^{new}$, similar the equation in CRF:

- serve some existing dish:

$$\sum_{\text{dish k}} \frac{\text{number of tables with dish k}}{\text{number of tables} + \gamma} f_k^{x_{-ji}}(x_{ji}) \qquad (14)$$

- sample some new dish

$$\frac{\gamma}{\text{number of tables} + \gamma} \int f(x_{ji}|\phi)h(\phi)d\phi \qquad (15)$$

Posterior Sampling in the Chinese Restaurant Franchise

Similarly, in order to sample the component for table t:

- serve some existing dish with prob proportional to number of tables with dish k $* f_k^{-x_{ji}}(x_{ji})$
- sample some new dish with prob proportional to $\gamma * f_{k^new}^{-x_{ji}}(x_{ji})$

Posterior Sampling in the Chinese Restaurant Franchise

Similarly, in order to sample the component for table t:

- serve some existing dish with prob proportional to $\text{number of tables with dish k} * f_k^{-x_{ji}}(x_{ji})$
- sample some new dish with prob proportional to $\gamma * f_{k^new}^{-x_{ji}}(x_{ji})$

Posterior Sampling in the Chinese Restaurant Franchise

Similarly, in order to sample the component for table t:

- serve some existing dish with prob proportional to $\text{number of tables with dish k} * f_k^{-x_{ji}}(x_{ji})$
- sample some new dish with prob proportional to $\gamma * f_{k^new}^{-x_{ji}}(x_{ji})$

Outline

4 Inference
- Posterior Sampling in the Chinese Restaurant Franchise
- Posterior Sampling with an Augmented Representation

Posterior Sampling with an Augmented Representation

One drawback of previous approach is that the sampling for different groups is coupled due to the reason that we integrated out G_0.
Another approach is to instantiated and sample G_0, so calculation for different groups can be factorized.

Posterior Sampling with an Augmented Representation

First givens a posterior sample (t, k), we can get the posterior of G_0.
Noticed that $G_0 \sim DP(\gamma, H)$.
Because components for each table is drawn from G_0, conditioning on k_{jt}'s, G_0 is distributed as

$$DP(\gamma + \text{number of tables}, (\gamma H + \sum_{dish k} \text{number of tables with k} * \delta_{\phi_k}))$$

Posterior Sampling with an Augmented Representation

An construction of G_0:

- $G_u \sim DP(\gamma, H)$
- $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
- $p(\phi_k|\mathbf{t}, \mathbf{k}) \propto h(\phi_k) \prod_{x_{ji} \text{ with dish k}} f(x_{ji}|\phi_k)$
- $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$

Posterior Sampling with an Augmented Representation

An construction of G_0:

- $G_u \sim DP(\gamma, H)$
- $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
- $p(\phi_k|\mathbf{t}, \mathbf{k}) \propto h(\phi_k) \prod_{x_{ji} \text{ with dish k}} f(x_{ji}|\phi_k)$
- $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$

Posterior Sampling with an Augmented Representation

An construction of G_0:

- $G_u \sim DP(\gamma, H)$
- $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
- $p(\phi_k | \mathbf{t}, \mathbf{k}) \propto h(\phi_k) \prod_{x_{ji} \text{ with dish k}} f(x_{ji}|\phi_k)$
- $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$

Posterior Sampling with an Augmented Representation

An construction of G_0:

- $G_u \sim DP(\gamma, H)$
- $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
- $p(\phi_k|\mathbf{t}, \mathbf{k}) \propto h(\phi_k) \prod_{x_{ji} \text{ with dish k}} f(x_{ji}|\phi_k)$
- $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$

Posterior Sampling with an Augmented Representation

An construction of G_0:

- $G_u \sim DP(\gamma, H)$
- $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
- $p(\phi_k|\mathbf{t}, \mathbf{k}) \propto h(\phi_k) \prod_{x_{ji} \text{ with dish k}} f(x_{ji}|\phi_k)$
- $G_0 = \sum_{k=1}^{K} \beta_k \delta_{\phi_k} + \beta_u G_u$

Posterior Sampling with an Augmented Representation

Sample β: $\beta = (\beta_1, ..., \beta_k, \beta_u) \sim Dir(m_1, ..., m_k, \gamma)$
Sample (t, k): similar with CRF, and count m_k replaced with β_k and γ replaced with β_u.

Outline

Document Modeling

A document is considered as "bag of words", which means exchangeability assumptions for the words in document. Typical parametric approach is to use *latent Dirichlet allocation (LDA)*.

LDA

In more detail, LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you

- Decide on the number of words N the document will have (say, according to a Poisson distribution).
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).

Example

Assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

LDA

In more detail, LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you

- Decide on the number of words N the document will have (say, according to a Poisson distribution).
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).

Example

Assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

LDA

In more detail, LDA represents documents as mixtures of topics that spit out words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you

- Decide on the number of words N the document will have (say, according to a Poisson distribution).
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).

Example

Assuming that we have the two food and cute animal topics above, you might choose the document to consist of 1/3 food and 2/3 cute animals.

LDA

Generate each word w_i in the document by:

- First picking a topic (according to the multinomial distribution that you sampled above;

Example

You might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability)

- Using the topic to generate the word itself (according to the topic's multinomial distribution)

Example

If we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on

LDA

Generate each word w_i in the document by:

- First picking a topic (according to the multinomial distribution that you sampled above;

Example

You might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability)

- Using the topic to generate the word itself (according to the topic's multinomial distribution)

Example

If we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on

LDA

Generate each word w_i in the document by:

- First picking a topic (according to the multinomial distribution that you sampled above;

Example

You might pick the food topic with 1/3 probability and the cute animals topic with 2/3 probability)

- Using the topic to generate the word itself (according to the topic's multinomial distribution)

Example

If we selected the food topic, we might generate the word "broccoli" with 30% probability, "bananas" with 15% probability, and so on

Document Modeling with HDP

- H is a measure on multinomial probability vectors
- G_0 is sampled and provides a countably infinite collection of multinomial probability vectors, which is corresponding to all available topics for the corpus
- For each document G_j is sampled and represents the subset of topics used in document
- Sample word by sample multinomial vector θ_{ji} first and then sample words x_{ji} with probabilities θ_{ji}

Document Modeling with HDP

- H is a measure on multinomial probability vectors
- G_0 is sampled and provides a countably infinite collection of multinomial probability vectors, which is corresponding to all available topics for the corpus
- For each document G_j is sampled and represents the subset of topics used in document
- Sample word by sample multinomial vector θ_{ji} first and then sample words x_{ji} with probabilities θ_{ji}

Document Modeling with HDP

- H is a measure on multinomial probability vectors
- G_0 is sampled and provides a countably infinite collection of multinomial probability vectors, which is corresponding to all available topics for the corpus
- For each document G_j is sampled and represents the subset of topics used in document
- Sample word by sample multinomial vector θ_{ji} first and then sample words x_{ji} with probabilities θ_{ji}

Document Modeling with HDP

- H is a measure on multinomial probability vectors
- G_0 is sampled and provides a countably infinite collection of multinomial probability vectors, which is corresponding to all available topics for the corpus
- For each document G_j is sampled and represents the subset of topics used in document
- Sample word by sample multinomial vector θ_{ji} first and then sample words x_{ji} with probabilities θ_{ji}

Document Modeling with HDP

- H is a measure on multinomial probability vectors
- G_0 is sampled and provides a countably infinite collection of multinomial probability vectors, which is corresponding to all available topics for the corpus
- For each document G_j is sampled and represents the subset of topics used in document
- Sample word by sample multinomial vector θ_{ji} first and then sample words x_{ji} with probabilities θ_{ji}

Experiment Details

A symmetric Dirichlet distribution with parameters of .5 for the prior H over topic distributions is used.

$$\gamma \sim gamma(1, .1)$$

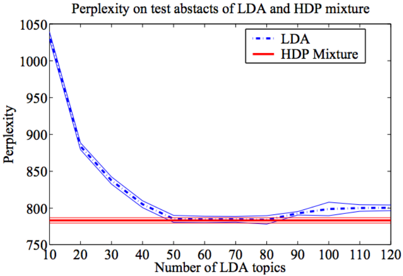
$$\alpha_0 \sim gamma(1, 1)$$

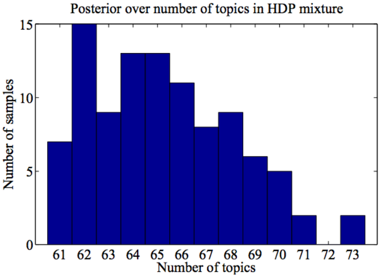
Posterior samples were obtained using the Chinese restaurant franchise sampling scheme.
10-fold cross-validation is used and the evaluation metric is perplexity :

Perplexity

$$exp(-\frac{1}{I} logp(w_1, ..., w_I | \text{training corpus}))$$

Experiment Results





Extension to multiple corpora

The documents that we used for these experiments consist of articles from the proceedings of the Neural Information Processing Systems (NIPS) conference for the years 1988-1999.

The NIPS conference deals with a range of topics covering both human and machine intelligence.

Articles are separated into nine sections:

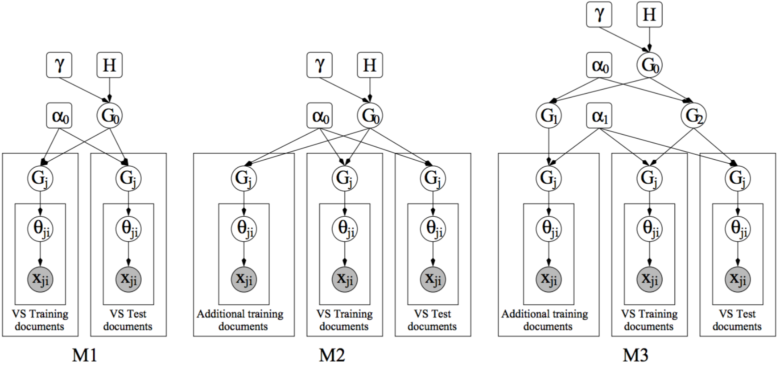
- algorithms and architectures (AA)
- applications (AP)
- cognitive science (CS)
- control and navigation (CN)
- implementations (IM)
- learning theory (LT)
- neuroscience (NS)
- signal processing (SP)

Extension to multiple corpora

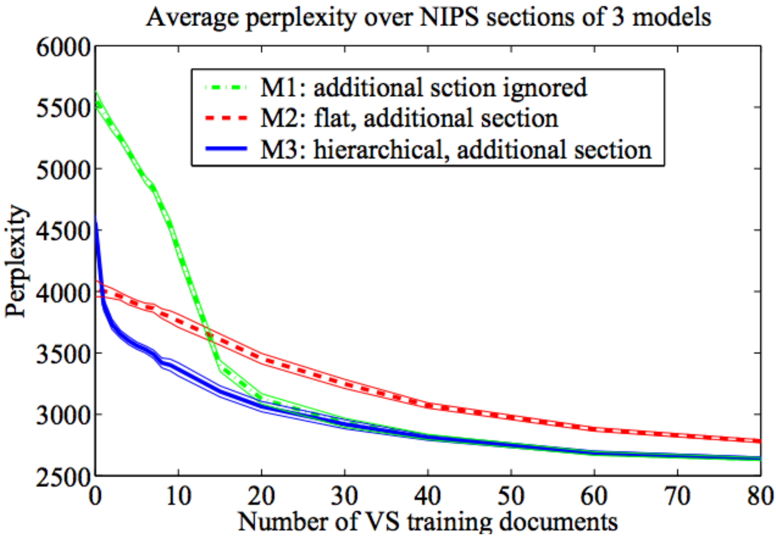
These sections are treated as "corpora," and are interested in the pattern of sharing of topics among these corpora.

Given a set of articles from a single NIPS section that we wish to model (the VS section in the experiments that we report below), we wish to know whether it is of value (in terms of prediction performance) to include articles from other NIPS sections.

Extension to multiple corpora



Extension to multiple corpora



Questions ?

Nanxin Chen
bobchennan@gmail.com