10-708: Probabilistic Graphical Models 10-708, Spring 2016

13 : Mean Field Approximation and Topic Models

Lecturer: Eric P. Xing *Scribes: Shichao Yang, Mengtian Li, Haoqi Fan*

1 Mean field

1.1 Recall Loopy Belief Propagation

For a distribution $p(X|\theta)$ associated with a complex graph especially with loops, it is intractable to compute the approximation distribution $q(X)$ (from KL-divergence view) directly and marginal (or conditional) probability of arbitrary random variables. So instead, the variational methods optimize over approximation objective:

$$q^* = \arg min_{q \in M} F_{Betha}(p, q)$$

However, optimization $F_{Betha}(p, q)$ with respect to $q(X)$ is still difficult. So we don't explicit optimize $q(X)$, but optimize b in the following form:

$$b = \{b_{i,j} = \tau(x_i, x_j), b_i = \tau(x_i)\}$$

namely a set of beliefs of singleton and edges. The constraints on b is the local consistence:

$$M_o = \{\tau \geq 0 | \sum_{x_i} \tau(x_i) = 1, \sum_{x_i} \tau(x_i, x_j) = \tau(x_j)\}$$

M_o is an over-approximation of original feasible set M, namely $M_o \supseteq M$. The LBP can be viewed as a fixed point iteration procedure that want to optimize $F(b)$. LBP often not converge to the correct solution, although empirically it often performs well.

1.2 Mean Field Introduction

The core idea behind MF is to optimize the posterior distribution $q(x_H)$ in the space of tractable families to make it easier to compute. For example, we can approximate $q(x_H)$ as:

$$q(z_1, ..., z_m) = \prod_{j==1}^{m} q(z_j)$$

This assumes the complete factorization of the distribution over individual latent variables, which is often referred as "naive mean field". From graphical view, we remove all the edges between variables. For a more general settings, we can also assume the variational distribution factorizes into R group: $z_{G_1}, ..., z_{G_R}$ which is referred to as "generalized mean field": [picture]

$$q(z_1, ..., z_m) = \prod_{r==1}^{R} q(z_{G_r})$$

So there are different ways of subgraph representation that could be used to approximate the true distribution. The problem thus has changed to:

$$q^* = \arg min_{q \in T}(< E >_q - H_q)$$

We are optimizing the exact objective H_q but on a tightened feasible set $T \subseteq Q$.

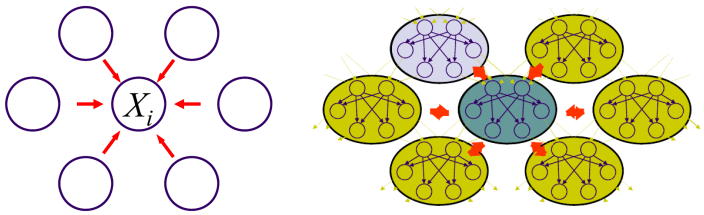
In the next lecture, we will introduce a unified point of view based on the variational principle, here we briefly conclude: Mean field method is non-convex inner bound with exact form of entropy. Loopy belief propagation is polyhedral outer bound with non-convex Bethe approximation.

1.3 Naive Mean Field

We approximate $p(x)$ by fully factorized $q(X) = \prod_i q_i(X_i)$. For the Boltzmann distribution $p(X) = exp\{E_{i<j} q_{ij} X_i X_j + q_{io} X_i\}/Z$. The mean field update equation is that:

$$q_i(X_i) = \{\theta_{io} X_i + \sum_{j \in N_i} \theta_{ij} X_i < X_j >_{q_j} + A_i\}$$

where $< X_j >_{q_j}$ resembles a "message" sent from node j to i. $< X_j >_{q_j} : j \in N_i$ forms the "mean field" applied to X_i from its neighborhood shown in Figure 1(a).



(a) (b)

Figure 1: (a) Naive mean field update. (b) Generalized mean field update.

1.4 Generalized Mean Field

We can also apply more general forms of the mean field approximations, i.e. clusters of disjoint latent variables are independent, while the dependencies of latent variables in each clusters are preserved shown in Figure 1(b).

Figure 2 is the naive mean field for Ising model while Figure 3 shows the original Ising model and, generalized mean field with 2×2 clusters and generalized mean field with 4×4 clusters.

The generalized mean filed theorem shows that the optimum GMF approximation to the cluster marginal is isomorphic to the cluster posterior of the original distribution given internal evidence and its generalized mean fields:

$$q^*(X_{H,C_i}) = p(X_{H,C_i} | X_{E,C_i}, < X_{H,MB_i} >_{q_{j \neq i}})$$

where $< X_{H,MB_i} >_{q_{j \neq i}}$ is the neighbour cluster. The convergence theorem shows that GMF is guaranteed to converge to a local optimum and provides a lower bound for the likelihood of evidence.

The inference accuracy and computation comparison between NMF and GMF, BP is shown in Figure 4. We can see that mean filed especially larger GMF has the lowest error while also at much high computational cost. GMF with grid 2×2 is better than BP in both accuracy and speed.

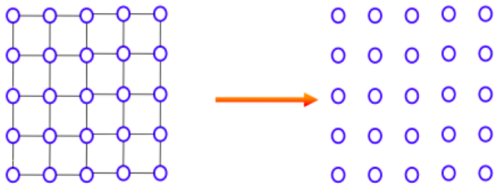


Figure 2: Naive mean field for Ising models.

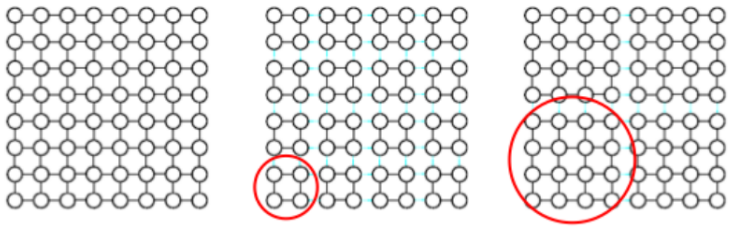


Figure 3: General mean field for Ising models

1.5 Factorial HMMs

Lets take a look at another example: factorial HMM. We can make naive mean field approximations such that all variables are assumed to be independent in posterior. We can also make generalized mean field approximations such that each disjoint cluster contains one hidden Markov chain, two hidden Markov chains or three hidden Markov chains. The original factorial HMM and a generalized mean field approximation based on clusters with two hidden Markov chains are shown in Figure 5. Figure 5 shows the singleton marginal error and CPU time for naive mean field, generalized mean field with clusters of different number of Markov chains and exact inference BP algorithm. From the Singleton marginal error histogram, we can see that as expected, naive mean field drops all edges in the posterior, thus has the highest error, while generalized mean field with clusters of more chains generally behave better in terms of singleton marginal error. From the CPU time histogram, we can see that the exact inference method takes the most time, while mean field approximations generally take less time. From the above example, we can draw the conclusion that mean field approximation makes inference tractable or cheap compared with exact inference methods, at the cost of additional independency assumptions, thus leads to a bias in the inference result.

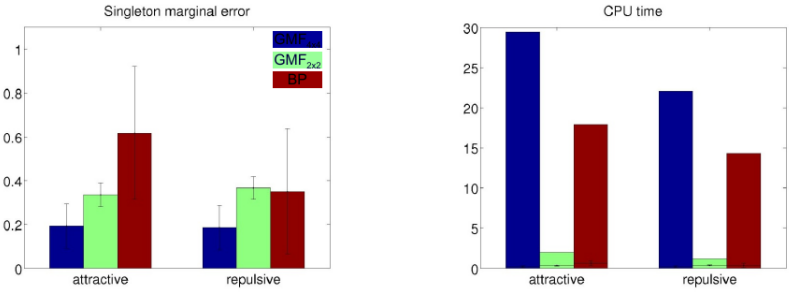


Figure 4: Inference comparison on Ising models.

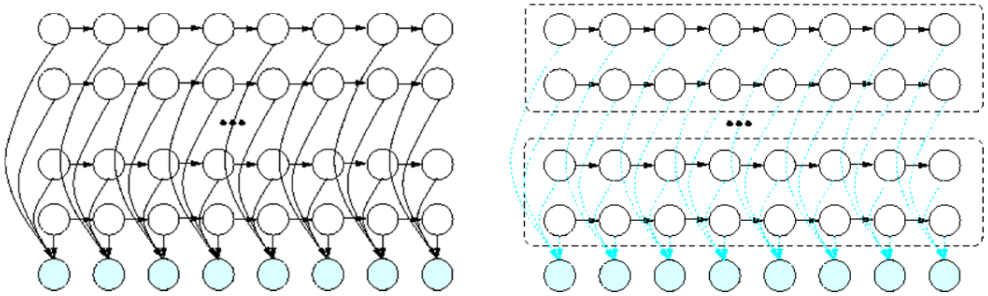


Figure 5: Mean Field Approximation for HMM.

2 Topic Model

2.1 General framework

The general framework for solving problems with graphical models can be decomposed into these major steps:

- Task: embedding, classification, clustering, topic extraction.
- Data representation: input and output, data types (continuous, binary, counts)
- Model: belief networks, Markov random fields, regression, support vector machine
- Inference: exact inference, MCMC, variational
- Learning: MLE, MCLE, max-margin
- Evaluation: visualization, human interpretability, perplexity, predictive accuracy

We should better consider one step at a time.

2.2 Task and Data Representation

The motivation underlying the probabilistic topic models is the incapability of human processing a huge number of text documents (e.g., search, browse, or measure similarity). We need new computational tools

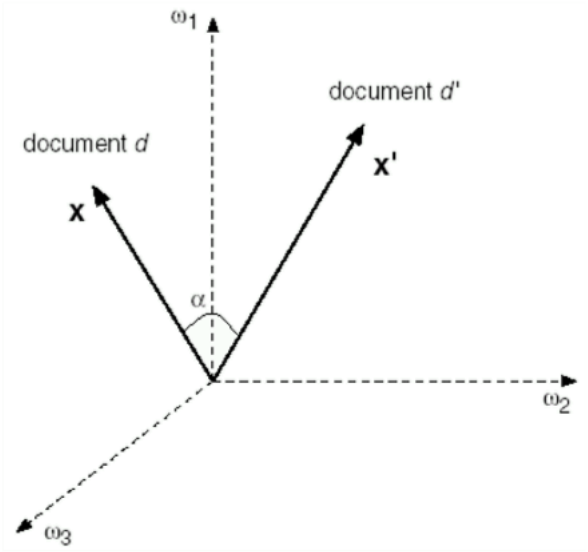


Figure 6: Document Embedding.

to help organize, search and understand these vast amounts of information. To this end, the probabilistic topic models to help to organize the documents according to the topics in order for us to perform a variety of tasks with large number of documents.

This task can be done by finding a mapping from each document to a vector space, i.e. document embedding (Fig 6). Formally, it reads $D \to R^d$, where D is the spaces of documents and Rd is an Euclidean space. Document embedding enable us to compare the similarity of two documents, classify contents, group documents into clusters, distill semantics and perspectives, etc.

One common representation is bag of words (Fig 7). Bag of words is an orderless high-dimensional sparse representation, where each document is represented by the frequency of words over a fixed vocabulary.

There are a few drawbacks of the bag of words representation. It is not very informative. It is not very efficient for text processing tasks, e.g., search, document classification, or similarity measure. It is also not effective for browsing.

2.3 The Big Picture of Topic Models

The big picture is presented in (Fig 8). The blue crosses denote topics while the red crosses denotes documents. Each topic is modeled by a distribution of words, thus it can be viewed as a point in a word simplex. Similarly, each document is modeled by a distribution of words, thus it can be viewed as a point in the topic simplex.

2.4 Latent Dirichlet Allocation

Below is the scheme to generate a document from a topic model. If the prior is a Dirichlet distribution, the model is called Latent Dirichlet Allocation (LDA).

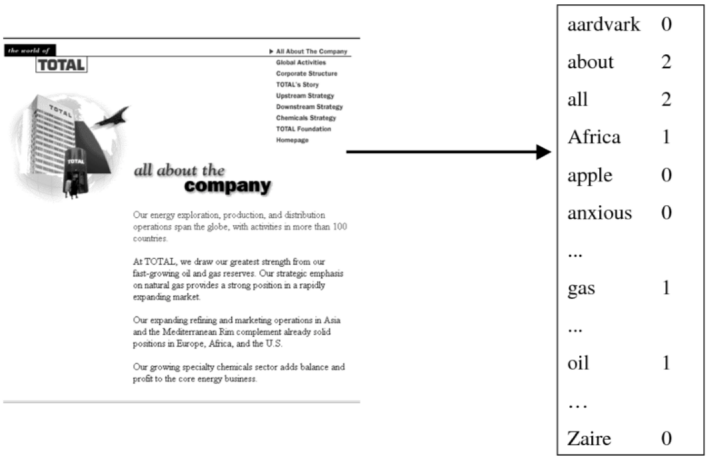


Figure 7: The Bag of Words Representation of a Document

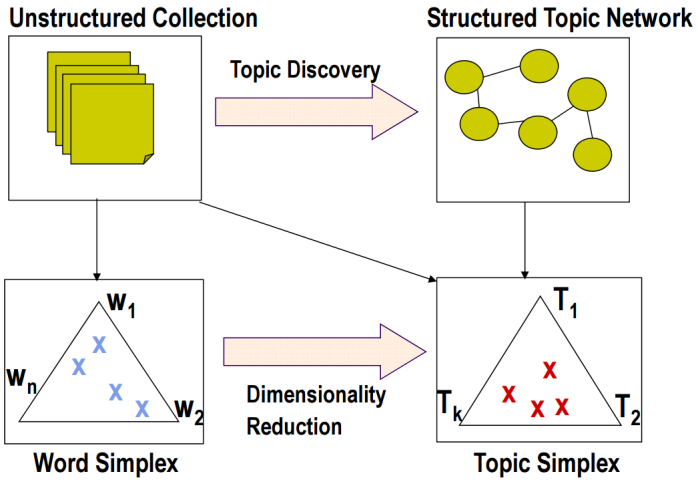


Figure 8: The Big Picture of Probabilistic Topic Models

Draw θ from the prior;
for *each word* n **do**
  Draw z_n from $multinomial(\theta)$;
  Draw $w_n|z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$;
end

2.5 Inference for LDA

The posterior inference is the fundamental problem to reveal the mysteries behind the LDA. The joint distribution of the network could be easily computed by chain rule. According to the Fig. 9

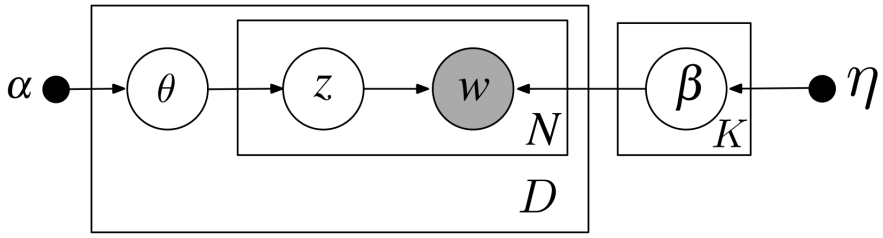


Figure 9: The Latent Dirichlet allocation (LDA) Model.

$$P(\beta, \theta, z, w) = \prod_{k=1}^{K} P(\beta_k|\zeta) \prod_{d=1}^{D} P(\theta_d; \alpha) \prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(w_{d,n}|z_{d,n}, \beta)$$

In order to utilize the LDA model, there some conditional possibility we may be need in order to calculate the entire posterior possibility. For example, in order to know $p(theta_n|D)$ as well as $p(z_{n,m}|D)$. Also, in order to calculate $p(\theta_n|D) = \frac{p(\theta_n, D)}{p(D)}$, we might need to integrate over θ, β:

$$\frac{\sum_{\{z_{n,m}\}} \int \left(\prod_n \left(\prod_m p(w_{n,m}|\beta_{z_n}) p(z_{n,m}|\theta_n) \right) p(\theta_n|\alpha) \right) p(\beta|\eta) d\theta_{-i} d\beta}{p(D)}$$

Where Then, we could get an acceptable result by approximate inference. There are several inference

$$p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left(\prod_n \left(\prod_m p(x_{n,m}|\beta_{z_n}) p(z_{n,m}|\theta_n) \right) p(\theta_n|\alpha) \right) p(\beta|\eta) d\theta_1 \cdots d\theta_N d\beta$$

algorithm for the posterior possibility.

- Mean field approximation
- Expectation propagation
- Variational 2nd-order Taylor approximation
- Markov Chain Monte Carlo (Gibbs sampling) predictive accuracy

Since in the lecture, it is actually running our of time when introducing inference approach, so barely detail of inference is revealed. We are going to cover some important ones to in order to reflect the main points of LDA.

2.6 Mean-Field Approximation

When we doing the mean field approximation, the variational distribution over the latent variables factorizes could be asserted as:

$$q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{d,n})$$

By the formulation, it is obvious to see that the variational approximation q over β, θ, d are independent.

Using the mean field, we could optimize the lower bound of the real posterior as the following form, given

$$q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{d,n})$$

$$L(q(\beta, \theta, z) = E_q \log p(w, \beta, \theta, z) + H(q(\beta, \theta, z))$$

Then it is easy to derive the coordinate ascent algorithm by

$$L(q(\beta_i, \theta_i, z_i) = \int q(\beta_i, \theta_i, z_i) E_{q-i} \log p(w, \beta, \theta, z) d\beta_i \theta_i z_i + H(q(\beta, \theta, z))$$

In order to avoid the confusion we should clearify that E_{q-i} is the expectation over all other latent variables except for the j then variable. Then we could directly get the update rule as:

$$q(\theta_d|\alpha) \propto \exp\left(\sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{dk}\right)$$

$$q(z_{dn}|\theta_d) = \exp\left(\sum_{k=1}^{K} 1_{[z_{dn}=k]} \log \theta_{dk}\right)$$

$$q(\theta_d) = \exp\left(\sum_{k=1}^{K} \left(\sum_{n=1}^{N} q(z_{dn} = k)\right) \log \theta_{dk}\right)$$

We could have a better review of the entire algorithm by going through the following algorithm.

initialize varientional topics $q(\beta_k)$;
while *Lower bound* $L(q)$ *not converge* **do**
  for *for each document* $d \in \{1, 2, 3..., D\}$ **do**
    Initialize varientional topic assignment $q(z_{dn})$;
    while *Change of* $q(\theta)$ *is not small enough* **do**
      Update varientional topic proportions $q(\theta_d)$;
      Update varientional topic assignments $q(z_{dn})$;
    end
    Update varientional topics $q(\beta_k)$;
  end
end

Algorithm 1: Coordinate ascent algorithm for LDA

By revealing the algorithm, we could see there is 3 loops here. If we have millions of documents here, it is going to be very slow.

2.7 Variational Inference for LDA

The general idea of variational inference is minimizing the KL divergence between the variational distribution $q(\theta, z|\gamma, \phi)$ and the true posterior distribution is $q(\theta, z|w, \alpha, \beta)$, and the γ and ϕ are variational parameters for q. Worth to notice that $q(\theta|\gamma)$ follows a Dirichlet distribution. And the distribution is parameterized by γ. And this is a clique that the $q(z_n|\phi_n)$ follows the multinomial distribution which is parameterizing by ϕ_n. So we could have the KL divergence as We omit some derivation and draw the conclusion here as

$$\begin{aligned} KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) &= E_q(\theta, z|\gamma, \phi) \log \frac{q(\theta, z|\gamma, \phi)}{p(\theta, z|w, \alpha, \beta))} \\ &= E_q \log q(\theta, z|\gamma, \phi) - E_q p(\theta, z|w, \alpha, \beta) \\ &= E_q \log q(\theta, z|\gamma, \phi) - E_q p(\theta, z, w|\alpha, \beta) + E_q p(w|\alpha, \beta) \end{aligned}$$

$$L(\gamma, \phi : \alpha, \beta) = E_q[\log p(\theta|\alpha)] + E_q[\log(z|\theta)] + E_q[\log p(w|z, \beta)] - E_q[\log q(\theta)] - E_q[\log(q(z))]$$

$L(\gamma, \phi; \alpha, \beta)$ could be optimized by Variational EM algorithm which maximizes the lower bound with respect to the variational parameters γ and ϕ in E step. Then maximazes the lower bound with respect to the model parameters for fixed values of the variational parameters in M step.

2.8 Gibbs Sampling for LDA

We actually going to cover the Gibbs sampling in Lecture 16, but in order to make the note self-contain, we are going to have a brief introduce about the Gibbs sampling based approximate inference on LDA as a special case Markov-chain Monte Carlo. The latent variables in the graphical model are sampled iteratively given the rest based on the conditional distribution.

Using $\mathbf{z}$ denote the concatenation of z for all words and z_{-n} is the topics for all words except w_n. Given all variables, the conditional probability of w_n being assigned to topic k is

$$\begin{aligned} p(z_i = k|\mathbf{z_{-n}}, \mathbf{w}, \alpha, \eta) &\propto \frac{p(\mathbf{w}, \mathbf{z}|\alpha, \eta)}{p(\mathbf{w}, \mathbf{z_{-n}}|\alpha, \eta)} = \frac{p(\mathbf{w}|\mathbf{z}, \eta)}{p(\mathbf{w}|\mathbf{z_{-n}}, \eta)} \frac{p(\mathbf{z}|\alpha)}{p(\mathbf{z_{-n}}|\alpha)} \\ &= \int p(\mathbf{z}|\theta) p(\theta|\alpha) d\theta \propto \frac{n_k^w + \eta_k}{\sum_{k'} n_{k'}^w + \eta_k} (n_d^k + \alpha_k) \end{aligned}$$

Where n_k^w denotes the counting of appearence of word w in topic k. n_d^k denotes the counting number of words assigning to topic k in document d.

2.9 Conclusion

In this section, we reviewed the essential background of Latent Dirichlet Allocation. Then we introduced the general frameworks. After that we reviewed three standard inference method for LDA. Worth to mention that the topic model are one of the most activate models in the cutting edge research, especially in multi-media and unsupervised data mining areas. This could be proved easily by reviewing Fig. 10.

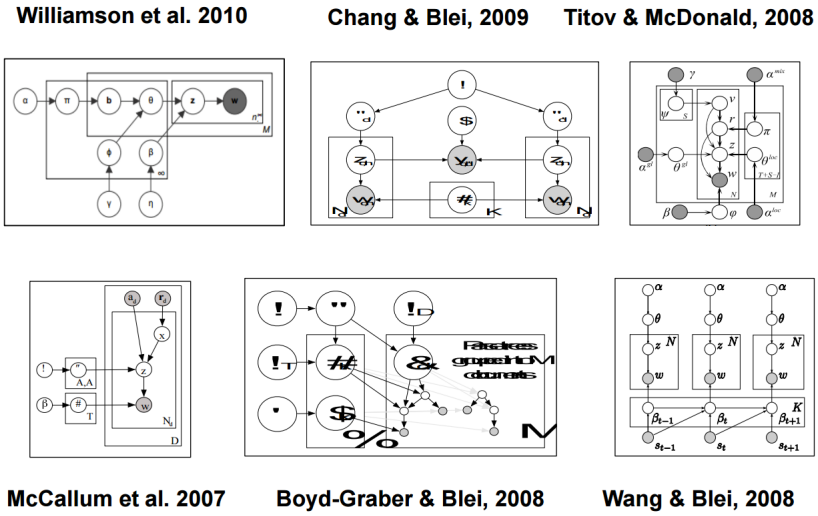


Figure 10: Topic Model zoo.