Lecture 24: April 20, 2015
cs 573: Probabilistic Reasoning
Professor Nevatia
Spring 2015

Review

- HW#7 due April 22
- Exam2: April 29, class period, here (except for remote DEN students)
  - Closed book/notes
  - Detailed list of topics in following slides
  - Style similar to Exam1 but likely to be more descriptive
- Previous Lecture
  - Learning from incomplete data
    - Gradient Ascent Approach
    - EM algorithm
- Today's objective
  - EM for clustering
  - Latent Dirichlet Allocation (LDA)

Exam Topics

- Lec 13-25 (both inclusive); Exam1 topics will not be repeated in Exam 2
- Topics:
  - Gaussian Networks: Ch. 7, 14.2
  - Variational Approximation: 11.5 (only as covered in class)
  - MAP inference: 13.1, 13.2 and 13.3
  - Sampling: 12.1, 12.2 and 12.3 (including Metropolis-Hastings algorithm)
  - Temporal Models: 6.2, 15.1, 15.2, 15.3 (as covered in class)
  - Parameter Estimation: 17.1, 17.2 (exclude 17.2.4, 17.2.5), 17.3
  - Partially Observed Data: 19.1, 19.2 (except 19.2.2.5 and 19.2.2.6, 19.2.4)
  - Chapter 20: depends on class-coverage
  - Chapter 18: depends on class-coverage

Practical Issues (EM for HMMs)

- Forward-backward (or the clique-tree) algorithm involves multiplication and summation of probabilities.
  - For a long sequence, we will encounter underflow problems (dynamic range problems)
- We can not just convert to log space (as for the MAP queries) as expressions have both products and summations.
- One solution is to normalize at each step of both the forward and the backward passes
- Baum-Welch algorithm (~1960)
  - Not explicitly formulated as an EM algorithm (EM framework is much newer, ~1977)

Theoretical Analysis of EM

- We have derived formulas for the E-steps
  - Basically the steps of inference
- We have not shown that the M-step maximizes data likelihood
  - Instead, M-step implicitly maximizes an auxiliary function, which is the likelihood of D^+ (augmented set of examples)
  - Can be shown that parameters that maximize the auxiliary function also maximize the original function
  - We skip these derivations
- Problem of local maxima
  - EM algorithm does not guarantee reaching a global maximum
  - Search with different initial parameters
  - Initial parameters based on domain knowledge
  - Simulated annealing and other global optimization methods

EM vs Gradient Ascent

- Both converge to a local optimum
- KF Book states that EM tends to make major improvements in the first few iterations, gradient ascent is more effective in later iterations
  - Combine the two, use EM in the beginning, gradient ascent when near the maximum
- EM seems easier to implement and more commonly used
- Note: we skip the topic of Bayesian learning with missing data

EM for Clustering

- k-means clustering
- Given a set of points (in d-dimensional space), cluster them in k clusters such that the sum of the squares of distances of points from cluster means is minimized
- k-means algorithm assumes an initial assignment of means (may be randomly initialized)
- Based on means, points can be assigned to a cluster
- Given assignments, we can compute means
- Iterate: an EM-like algorithm
- Can be formalized to case where each cluster represents a Gaussian distribution; points are assigned to cluster giving highest probability; task is to estimate parameters of clusters that maximize data observation likelihood.
- This is called “hard-assignment” EM

Bayesian (Soft Assignment) Clustering

- Instead of a hard assignment of points to clusters, we assign a probability of assignment to a cluster, or a "weight vector"
- Thus, we need to compute the parameters of the clusters (means and co-variance matrices for a Gaussian distribution) AND "weight vectors".
- Two variations are possible:
  - Even though, there is a probability of point being assigned to different cluster, only one cluster is chosen
  - A point can be a member of multiple clusters with different weights
- Both variations can be computed by using EM algorithm (but correspond to maximizing different likelihood functions)
- One case given in following slides

Mixture of Gaussians

- Dataset D = $\{x_1, \ldots, x_n\}$, each x_i is a d-dimensional vector
- K mixture components (clusters) $p(\underline{x}|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(\underline{x}|z_k, \theta_k)$
- z_k is the selected cluster, non-zero for only one k; each θ_k is a Gaussian distribution (given by mean μ_k , co-variance Σ_k)
- $\alpha_k = P(z_k)$; probability that z_k is selected (for unknown x_i); add to 1
- Model parameters: $\Theta = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$
- Membership weight of a specific sample:

$$w_{ik} = p(z_{ik} = 1|\underline{x}_i, \Theta) = \frac{p_k(\underline{x}_i|z_k, \theta_k) \cdot \alpha_k}{\sum_{m=1}^{K} p_m(\underline{x}_i|z_m, \theta_m) \cdot \alpha_m}, \quad 1 \le k \le K, \quad 1 \le i \le N.$$

- Given model parameters, we can compute weight vector (E-step)
- Given weight vector, we can compute $\alpha_k = P(z_k)$ and the corresponding distribution parameters (M-step)
- More details in: http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf

MOG: M-step

- Given weight vector, we can compute $\alpha_k = P(z_k)$ and the corresponding distribution parameters (M-step)
- Let $N_k = \Sigma_{i=1,n} w_{ik;}$; number of data points assigned to cluster k

$$\alpha_k^{new} = \frac{N_k}{N}, \quad 1 \le k \le K.$$

$$\underline{\mu}_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot \underline{x}_i \quad 1 \le k \le K.$$

$$\Sigma_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \cdot (\underline{x}_i - \underline{\mu}_k^{new})(\underline{x}_i - \underline{\mu}_k^{new})^t \quad 1 \le k \le K.$$

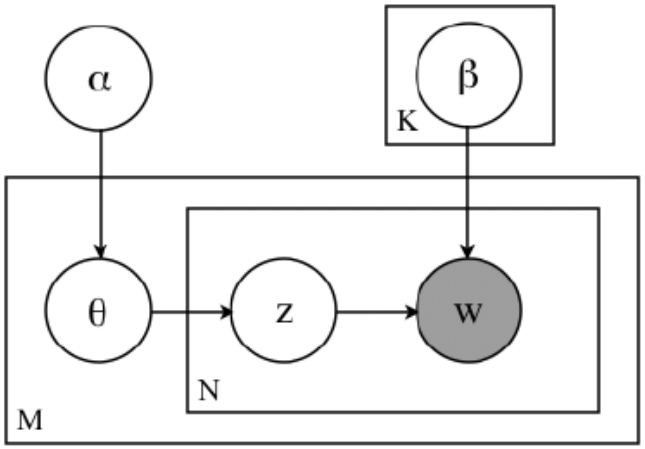
- Above formulas from reference:

http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf

Document Processing

- Need to search relevant documents amongst huge collections
  - Documents on a specific topic
  - Documents similar to a given one
    - Recommend book, find a reviewer….
- Structured, semantic understanding of text remains difficult
- Can simpler analysis still provide useful results?
  - Surprisingly, "bag of words" suffices for many tasks
    - Ignore order of words, scramble text in a document
  - Frequency of words may be indicative of document category
    - Just presence of collections of words may suffice
  - Simple frequency models ("unigrams") and pair-models ("bigrams") have shown to be of high utility
  - tf-idf: term frequency x inverse document frequency
  - Remove stop words, "stem" words (drop suffixes)
    - e.g. wait, waits, waited, waiting..

Latent Dirichlet Allocation Model



Ref: Blei, Ng, Jordan 2003

- Words ("w") in a document are what we can "observe" (1:N)
- Document is a collection (sequence) of words
- Corpus is a collection of documents
- θ is a "topic" distribution; governed by Dirichlet (α); K topics
- z is sampled from θ, it is like an indicator function, specifying a specific topic
- β is a K x v matrix giving probability of words for chosen value of z (v is size of word dictionary)

Generative Analysis

- Distribution given α,β

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

Note z_n is the selected topic, index into distribution given by β

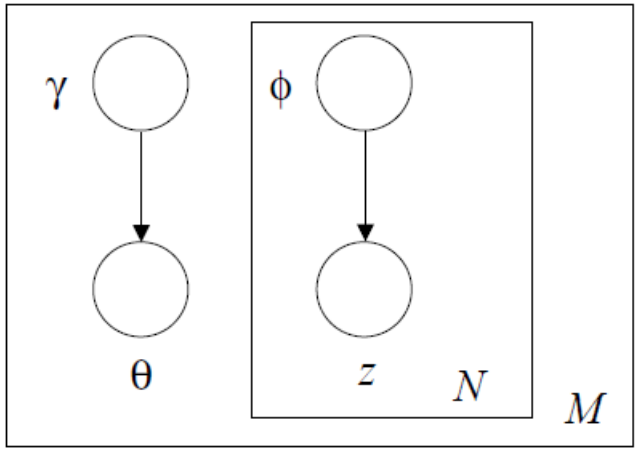
$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta.$$

By marginalizing over z, θ

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d.$$

Sum over the whole document

Model used for Variational Approximation



Inference and Learning

- Choose a k-dimensional topic weight vector, say θ_m from Dirichlet (α)
- Sample a topic (z) given θ_m .
- Sample a word from the distribution given z and β
- Given all the parameters of the model, and a sequence of words in the document, we can infer the distribution of topics in the document (also the likelihood of data)
  - Each category should have a different distribution
- How to infer the model parameters (α, β)
  - Apply EM
    - E-step : compute probability of data given model
    - M-step: Modify model parameters
    - E-step is not tractable, use variational inference
  - Note: LDA is learnt on the whole corpus; K is not learned

An Example from BNJ

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Note: "William Randolph Hearst Foundation" not treated as a single entity in bag of words representation

Applying LDA

- Once LDA parameters are learned, we can infer distribution of topics in a class of documents
- Two documents may be compared based on these distributions
  - e.g. cosine similarity of two vectors
- Topic coefficients can be used as feature vector for classification
  - Large reduction in dimensionality compared to word distributions
- Collaborative filtering
  - Predicting preferences of a user based on partial observations and preferences of other users
- Used for many applications: image segmentation, object recognition, musical analysis…

Next Class

- Read sections 20.1 and 20.2 of the KF book