Structure Learning

Le Song

Machine Learning II: Advanced Topics
CSE 8803ML, Spring 2012
Where are we now?

- **Graphical model representation**
  - Bayesian networks (directed graphical models)
  - Markov networks (undirected graphical models)
  - Conditional independence statements + factorization of joint distribution

- **Inference in graphical models**
  - Variable elimination, message passing on trees and junction trees
  - Sampling (rejection, importance and Gibbs sampling)

- **Learning graphical model parameters (given structure)**
  - Maximum likelihood estimation (just counts for discrete BN)
  - Bayesian learning (posterior)
  - EM for models with latent variables
Structure Learning

- The goal: given set of independent samples (assignments of random variables), find the best (the most likely) graphical model structure

\[(A,F,S,N,H) = (T,F,F,T,F)\]
\[(A,F,S,N,H) = (T,F,T,T,F)\]
\[\ldots\]
\[(A,F,S,N,H) = (F,T,T,T)\]
Mutual information

\[ I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \]

KL-divergence between joint distribution \( P(X_i, X_j) \) and product of marginal distributions \( P(X_i)P(X_j) \)

\[ KL(P(X_i, X_j) || P(X_i)P(X_j)) \]

A “distance” away from independence

- \( X_i \) and \( X_j \) are independent if and only if \( I(X_i, X_j) = 0 \)

Given \( M \) iid data point \( D = \{x^l\} \), \( \hat{P}(x_i, x_j) = \frac{\#(x_i, x_j)}{M} \)
Decomposable score and Equivalent trees

\[ l(D, G) = \log \hat{p}(D | G) = M \sum_{i} \hat{I}(x_i, x_{\pi_i}) - M \sum_{i} \hat{H}(x_i) \]

Score decompose according to edges in the tree!

\[ l(D, G) = \log \hat{p}(D | G) = M \sum_{(i,j) \in T} \hat{I}(x_i, x_j) - M \sum_{i} \hat{H}(x_i) \]
Chow-liu algorithm

\[ T^* = \arg \max_T M \sum_{(i,j) \in T} \hat{I}(x_i, x_j) - M \sum_i \hat{H}(x_i) \]

Chow-liu algorithm

- For each pair of variables \( X_i, X_j \), compute their empirical mutual information \( \hat{I}(x_i, x_j) \)
- Now you have a complete graph connecting variable nodes, with edge weight equal to \( \hat{I}(x_i, x_j) \)
- Run maximum spanning tree algorithm
Structural Learning for general graphs

- Theorem: The problem of learning a Bayesian Network structure with at most $d$ parent is NP-hard for any fixed $d \geq 2$

- Most structure learning approaches use heuristics
  - Exploit score decomposition
  - Two heuristic that exploit decomposition in different ways
    - Greedy search through space of node-orders
    - Local search of graph structure
Search through DAG with d parents

- Know order (K2 algorithm)
  - Given a total ordering of the nodes $X_1 < X_2 < \ldots < X_n$ and want to find a DAG consistent with this that maximize the score
  - The choice of parents of $X_i$ from $\{X_1, \ldots, X_{i-1}\}$ is independent of the choice for $X_j$ (since we obey the ordering, we can not create a cycle
  - Hence we can pick the best set of parents for each node independently
  - For $X_i$, we need to search $\binom{i-1}{d}$ subsets of size up to $d$ for the set that maximizes the score, use greedy algorithm for this

- What if order is unknown?
  - Search in the space of orderings, then conditioned on it and pick the best DAG using K2
Learn Bayesian Networks using local search

- Start from tree and do local refinement

(A,F,S,N,H) = (T,F,F,T,F)
(A,F,S,N,H) = (T,F,T,T,F)
...
(A,F,S,N,H) = (F,T,T,T,T)

Chow-liu Tree

Local search

Possible moves:
- Only if acyclic
  - Add edge
  - Delete edge
  - Invert edge

Select using some score (likelihood etc.)
Exploit score decomposition in local search

- Add edge and delete edge
  - Only change one term in the score

- Reverse edge
  - Change two terms in the score

- Simplest search algorithm
  - Greed hill climbing

![Diagram](attachment:image.png)
Local maxima

- Greedy hill climbing will stop when it reaches a local maximum or a plateau (a set of neighboring networks that have the same score)
  - Unfortunately, plateaus are common, since equivalence classes form contiguous regions of search space and such classes can be very large.

Partial solutions:
- Random restarts
- TABU search (prevent algorithm from undoing an operation applied in the last few steps, thereby forcing it to explore new terrain)
- Data perturbation
- Simulated annealing:
A Gaussian distribution can be represented by a fully connected graph with pairwise edge potentials over continuous variable nodes.

The overall exponential form is:

\[ P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i,j \in E}(X_i - \mu_i)\Sigma_{ij}^{-1}(X_j - \mu_j)\right) \]
\[ \propto \exp(- (X - \mu)^T \Sigma^{-1} (X - \mu)) \]

Also known as Gaussian graphical models (GGM)
Sparse precision vs. sparse covariance in GGM

\[ \Sigma^{-1} = \begin{bmatrix}
1 & 6 & 0 & 0 & 0 \\
6 & 2 & 7 & 0 & 0 \\
0 & 7 & 3 & 8 & 0 \\
0 & 0 & 8 & 4 & 9 \\
0 & 0 & 0 & 9 & 5 \\
\end{bmatrix} \]

\[ \Sigma = \begin{bmatrix}
0.10 & 0.15 & -0.13 & -0.08 & 0.15 \\
0.15 & -0.03 & 0.02 & 0.01 & -0.03 \\
-0.13 & 0.02 & 0.10 & 0.07 & -0.12 \\
-0.08 & 0.01 & 0.07 & -0.04 & 0.07 \\
0.15 & -0.03 & -0.12 & 0.07 & 0.08 \\
\end{bmatrix} \]

\[ \Sigma_{15}^{-1} = 0 \iff X_1 \perp X_5 \mid \text{TheRest} \]

\[ X_1 \perp X_5 \iff \Sigma_{15} = 0 \]
Structure learning for Gaussian GM

- Structure: the zero patterns in the inverse covariance matrix

- Key idea:
  - maximize the log-likelihood of the data (fit data)
  - and inducing sparsity in the inverse covariance matrix (regularization)

- Log-likelihood of the data

\[
\begin{align*}
l(D, \Sigma) &\propto \log \prod_i |\Sigma^{-1}|^{\frac{1}{2}} \exp \left( - \frac{X_i^T \Sigma^{-1} X_i}{2} \right) \\
&\propto \sum_i \log |\Sigma^{-1}| - X_i^T \Sigma^{-1} X_i \\
&\propto \sum_i \log |\Sigma^{-1}| - tr \Sigma^{-1} X_i X_i^T \\
&\propto n \log |\Sigma^{-1}| - tr \Sigma^{-1} S
\end{align*}
\]
Graphical Lasso

- Maximize Gaussian log-likelihood with l1 regularization on the inverse covariance matrix
  \[ \max_{\Theta} n \log |\Theta| - tr \Theta S - \lambda \|\Theta\|_1 \]
  \[ \Theta = \Sigma^{-1} \]
  is positive semidefinite
  \[ \|\Theta\|_1 \]
  denotes the sum of absolute values of the elements of the matrix

- This is a convex optimization problem and can be solved by many optimization algorithm
  - Coordinate descent: Graphical lasso (http://www.stat.stanford.edu/~tibs/ftp/graph.pdf)
  - Interior point methods (Yuan & Lin 2007)
Topic Modeling and Latent Dirichlet Allocation
Topic Modeling in Document Analysis

*Document Collections*

**Topic Discovery:** Discover Sets of Frequently Co-occurring Words

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

**Topics Attribution:** Attribute Words to a Particular Topic

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.

[Blei et al. 03]
Multi-scale Topics over Time

Time is divided into multiple scale of different lengths.

Topics can be learned for each time span at different scale

[Nallapati et al. 07]
Hierarchical Topic Modeling

A hierarchy of coarse to fine grained topics

The hierarchy is discovered automatically from the data

[Blei et al. 10]
Basic Model: Latent Dirichlet Allocation

- A generative model for text documents
  - For each document, choose a topic distribution $\Theta$
  - For each word $w$ in the document
    - Choose a specific topic $z$ for the word
    - Choose the word $w$ from the topic (a distribution over vocabulary)

Simplified version without the prior distribution $P(\theta | \alpha)$ corresponds to probabilistic latent semantic indexing (pLSI)

[Blei et al. 03]
Learning Problem in Topic Modeling

- Automatically discover the topics from document collections
- Given: a collection of $M$ documents
- Basic learning problems: discover the following latent structures
  - $k$ topics
  - Per-document topic distribution
  - Per-word topic assignment
- Application of topic modeling
  - Visualization and interpretation
  - Use topic distribution as document features for classification or clustering
  - Predict the occurrence of words in new documents (collaborative filtering)
Advanced Model: Topic Sharing

- Several document collections can share topics and have their own topics.
- Topics are shared by having a common base measure in the Dirichlet process.
- Gibbs sampling or variational inference is used for learning.

[Teh et al. 06]
Advanced Model: Hierarchical Topics

- Generative model for documents with a hierarchy of topics
  - For each topic in the hierarchy draw a topic $\beta$
  - For each document
    - Draw a path $c$ in the hierarchy from nested Chinese restaurant process
    - Draw a distribution $\theta$ over the levels in the tree
    - For each word $w$
      - Choose the level of the $z$ in path $c$
      - Choose a word $w$ from topic $\beta$ corresponding to node $z$ in path $c$
Gibbs Sampling for Learning Topic Models

- Use samples from the posterior distribution of the latent topic assignment

\[ P(z_1, \ldots, z_n \mid w_1, \ldots, w_n) \]

as an approximation to the distribution

- Gibbs sampling
  - Iterate until convergence
    - Sample \( z_1 \) from \( P(z_1 \mid z_2, z_3, \ldots, z_n, w_1, \ldots, w_n) \)
    - Sample \( z_2 \) from \( P(z_2 \mid z_1, z_2, z_3, \ldots, z_n, w_1, \ldots, w_n) \)
    - ...
    - Sample \( z_n \) from \( P(z_n \mid z_1, z_2, \ldots, z_{n-1}, w_1, \ldots, w_n) \)
  - The final sample \((z_1, \ldots, z_n)\) is used for topic learning

- Gibbs sampling can take very long time to converge
Variational Inference for Learning Topic Models

- Approximate the posterior distribution $P$ of latent variable by a simpler distribution $Q(\Phi^*)$ in a parametric family
  \[ \Phi^* = \arg\min_{\Phi} KL(Q(\Phi) \mid \mid P), \text{ where } KL = \text{Kullback-Leibler} \]
- The approximated posterior is then used in the E-step of the EM algorithm
- Only find a local minima

Variational approximation of mixture of 2 Gaussians by a single Gaussian

Each figure showing a different local optimum
Pros and Cons of Graphical Model Approach

- Very flexible in modeling rich structure in the data (e.g., hierarchical topic modeling)

- Learning algorithms are not scalable
  - Sampling approaches are typically very slow
  - Variational inference methods find only local optima, require approximation in each round of iterations making it difficult to understand the behavior of algorithm

- Graphical model formalisms to topic models involve hyper-parameters making them difficult to tune and deploy in practice
The objective of probabilistic latent semantic indexing

- Maximize the log-likelihood of the observed words
  \( (\theta^*, \beta^*) = \arg\max_{\theta, \beta} \prod_n \log p(w_n | \theta, \beta) \)

- Equivalent to minimize the KL-divergence between the empirical distribution \( \hat{p}(w) = \frac{1}{N} \sum_n \delta(w_n, w) \) and the model:
  \( (\theta^*, \beta^*) = \arg\min_{\theta, \beta} \sum_w \hat{p}(w) \log \frac{\hat{p}(w)}{p(w_n | \theta, \beta)} \)
Divergence minimization

- Word frequency-document matrix $A$ from data

- Word probability-document matrix $\tilde{A}$ from model

- pLSI minimizes column-wise KL-divergence between $A$ and $\tilde{A}$
Projects

- It has to be relatively sophisticated models which you need to explicitly take into account the dependency of variables or parts of the models.

- It can be new models on a new dataset, or improved algorithm for existing models.

- The project will be judged by novelty and elegance in several criteria:
  - Novelty of the application
  - Novelty of the method
  - Significance of the results
  - Project report and presentation
Example Projects

- KDD Cup 2012: User Modeling based on Microblog Data and Search Click Data (http://kdd.org/kdd2012/kddcup.shtml)

- Project 1: Social Network Mining on Microblogs

  The released data represents a sampled snapshot of the users' preferences for various items - the recommendation to users and follow-relation history. In addition, items are tied together within a hierarchy. That is, each person, organization or group belongs to specific categories, and a category belongs to higher-level categories. In the competition, both users and items (person, organizations and groups) are represented as anonymous numbers that are made meaningless, so that no identifying information is revealed. The data consists of 10 million users and 50,000 items, with over 300 million recommendation records and about three million social-networking "following" actions. Items are linked together within a defined hierarchy, and the privacy-protected user information is very rich as well. The data has timestamps on user activities.