Sequence Learning

Le Song

Machine Learning II: Advanced Topics
CSE 8803ML, Spring 2012
**Topic Modeling in Document Analysis**

**Document Collections**

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

<table>
<thead>
<tr>
<th></th>
<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>
<td></td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
<td></td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
<td></td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
<td></td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
<td></td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
<td></td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
<td></td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
<td></td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
<td></td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
<td></td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
<td></td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
<td></td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
<td></td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
<td></td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
<td></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]
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)
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
Approximate the posterior distribution $P$ of latent variable by a simpler distribution $Q(\Phi^*)$ in a parametric family

$$\Phi^* = \text{argmin } KL(Q(\Theta) \mid \mid P), \text{ where } KL = \text{Kullback-Leibler}$$

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
Sequence Learning

- Hidden Markov Models
  - Speech recognition
  - Genome sequence modeling
  - ...

- Kalman Filter (Gaussian noise models and linear models)
  - Tracking
  - Weather forecasting
  - ...

- Conditional random fields (discriminative model)
  - Natural language processing
  - Image processing
  - ...
More complex sequence models

- **Discrete** $Y$ and $X$
  - **Mixture model**
    - e.g., mixture of multinomials
  - **HMM**
    - (for discrete sequential data, e.g., text)

- **Continuous** $Y$ and $X$
  - **Mixture model**
    - e.g., mixture of Gaussians
  - **HMM**
    - (for continuous sequential data, e.g., speech signal)

- **State space model**

- **Factorial HMM**

- **Switching SSM**
Hidden Markov Models

- **Hidden States** $Y_t$
- **Transition Model** $P(Y_{t+1}|Y_t)$
- **Observation Model** $P(X_t|Y_t)$

- **Observation** $X_t$
Three Problems in Hidden Markov Models

- Given an observation sequence and a model, how to compute the probability of the observed sequence?

- Given an observation sequence and a model, how do we choose the hidden states that best explain the observations?

- How do we learn the model given the observation sequence?
Hidden Markov Model

- Directed graphical model

- The joint distribution factorizes according to parent child relation

\[ P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = P(Y_1) \prod_{t=1}^{n} P(X_t | Y_t) \prod_{t=2}^{n} P(Y_t | Y_{t-1}) \]
HMM Problem I: sequence probability

- Given an observation sequence and a model, how to compute the probability of the observed sequence?

Inference problem in graphical models: compute the marginal distribution of observation by summing out the hidden variables

\[
P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) = \sum_{Y_1, \ldots, Y_n} P(X_1, \ldots, X_n, Y_1, \ldots, Y_n)
\]

\[
= \sum_{Y_1, \ldots, Y_n} P(Y_1) \prod_{t=1}^{n} P(X_t | Y_t) \prod_{t=2}^{n} P(Y_t | Y_{t-1})
\]
HMM Problem I: sequence probability

Nice localization in computation

\[ Z = \sum_e \sum_d \sum_c \sum_b \sum_a f(a) f(b|a) f(c|b) f(d|c) f(E|d) \]
\[ Z = \sum_e \sum_d f(E|d) (\sum_c f(d|c) (\sum_b f(c|b) (\sum_a f(b|a) f(a)))) \]
HMM Problem II: best hidden states

- Given an observation sequence and a model, how to compute the hidden state with highest probability? (decoding)

- Inference problem in graphical models: compute the maximum a posterior assignment (MAP) for hidden states given the observations

\[
\arg\max_{Y_1, \ldots, Y_n} P(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \\
= \arg\max_{Y_1, \ldots, Y_n} P(Y_1) \prod_{t=1}^{n} P(X_t | Y_t) \prod_{t=2}^{n} P(Y_t | Y_{t-1})
\]
The general problem solved by the junction tree algorithm is the sum-of-product problem:

\[ P(X_i) \propto \sum_{V \setminus X_i} f(X_1, \ldots, X_n) = \sum_{V \setminus X_i} \prod_j \Psi(D_j) \]

The property used by the junction tree algorithm is the distributivity of \( \times \) over \( + \):

For HMM, it is distributivity of max over product. Or if you take log of the probability, it is distributivity of max over sum.

You just need to keep track of the hidden sequence that achieve the maximum score.
HMM problem II: find model parameters

- Given just an observation sequence, find the model parameters that maximize the likelihood of the data

- Suppose the set of parameters are \( \Theta \)
  
  \[
  \argmax_{\Theta} \log P(X_1, \ldots, X_n)
  \]
  
  \[
  \argmax_{\Theta} \log \sum_{Y_1, \ldots, Y_n} P(Y_1) \prod_{t=1}^{n} P(X_t | Y_t) \prod_{t=2}^{n} P(Y_t | Y_{t-1})
  \]
EM algorithm

EM: Expectation-maximization for finding $\theta$
- $l(\theta; D) = \log \sum_z p(x, z|\theta) = \log \sum_z p(x_i | z, \theta_2) P(z|\theta_1)$

Iterate between E-step and M-step until convergence
- Expectation step (E-step)
  - $f(\theta) = E_{q(z)}[\log p(x, z|\theta)]$, where $q(z) = P(z|x, \theta^t)$
- Maximization step (M-step)
  - $\theta^{t+1} = arg\max_\theta f(\theta)$

Also called Baum-Welch algorithm
More complex sequence models

Mixture model
e.g., mixture of multinomials

HMM
(for discrete sequential data, e.g., text)

Factorial HMM

Switching SSM

Mixture model
e.g., mixture of Gaussians

HMM
(for continuous sequential data, e.g., speech signal)

State space model
Kalman Filter

- Also called state space models, or linear dynamical system
  
  \[ X_t = AY_t + W_t \]
  \[ Y_t = CY_{t-1} + V_t \]
  \[ W_t \sim N(0; Q), V_t \sim N(0; R) \]
  \[ x_0 \sim N(0; \Sigma) \]

- HMM with special transition and observation models
  
  \[ P(Y_t | Y_{t-1}) \sim N(Y_t - CY_{t-1}; R) \]
  \[ P(X_t | Y_t) \sim N(X_t - AY_t; Q) \]
  \[ P(X_0) \sim N(0; \Sigma) \]

*In general can be nonlinear system* \( F(Y_t) \) and \( G(Y_{t-1}) \)
Inference problem in Kalman filter

- Filtering: given an observation sequence $x_1 \ldots x_t$, estimate $y_t$

- Inference problem in graphical models: estimate the marginal probability:
  - $P(y_t | x_1 \ldots x_t)$
  - By summing out all value of $y_1, \ldots, y_{t-1}$

- Forward message passing

![Graphical model diagram](image-url)
Inference problem in Kalman filter

- Smoothing: given an observation sequence $x_1 \ldots x_t$, estimate $y_1, \ldots, y_t$

- Inference problem in graphical models: estimate the marginal probability:
  - For all $t$, $P(y_t|x_1 \ldots x_t)$
  - By summing out all internal variables other than $y_t$

- Forward and backward message passing
Learning Kalman Filters

- HMM with special transition and observation models
  - $P(Y_t | Y_{t-1}) \sim N(Y_t - CY_{t-1}; R)$
  - $P(X_t | Y_t) \sim N(X_t - AY_t; Q)$
  - $P(X_0) \sim N(0; \Sigma)$

- Usually, people have chosen some dynamics for $A$ and $C$, just estimate the noise model for this case
  - New position = old position + $\Delta$ x velocity + noise
  - Observation = true position + noise

- If no true position is provided, use EM algorithm for learning
Tracking and Smoothing with Kalman Filters
Parameter Learning for Markov Random Fields

\[
P(X_1, ..., X_k | \theta) = \frac{1}{Z(\theta)} \exp(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \theta_i X_i)
\]

\[
= \frac{1}{Z(\theta)} \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i)
\]

\[
Z(\theta) = \sum_x \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i)
\]

\[
l(\theta, D) = \log(\prod_{l=1}^N \frac{1}{Z(\theta)} \prod_{ij} \exp(\theta_{ij} x_i^l x_j^l) \prod_i \exp(\theta_i x_i^l))
\]

\[
= \sum_i^N \left( \sum_{ij} \log(\exp(\theta_{ij} x_i^l x_j^l)) + \sum_i \log(\exp(\theta_i x_i^l)) - \log Z(\theta) \right)
\]

\[
= \sum_i^N \left( \sum_{ij} \theta_{ij} x_i^l x_j^l + \sum_i \theta_i x_i^l - \log Z(\theta) \right)
\]

can be other feature function \( f(x_i) \)

Term \( \log Z(\theta) \) does not decompose!
Derivatives of log likelihood

\[ l(\theta, D) = \frac{1}{N} \sum_{l}^{N} \left( \sum_{ij} \theta_{ij} x_{i}^{l} x_{j}^{l} + \sum_{i} \theta_{i} x_{i}^{l} - \log Z(\theta) \right) \]

\[ \frac{\partial l(\theta, D)}{\partial \theta_{ij}} = \frac{1}{N} \sum_{l}^{N} \sum_{ij} x_{i}^{l} x_{j}^{l} - \frac{\partial \log Z(\theta)}{\partial \theta_{ij}} \]

\[ \frac{\partial l(\theta, D)}{\partial \theta_{ij}} = \frac{1}{N} \sum_{l}^{N} \sum_{ij} x_{i}^{l} x_{j}^{l} - \frac{1}{Z(\theta)} \frac{\partial Z(\theta)}{\partial \theta_{ij}} \]

\[ = \frac{1}{N} \sum_{l}^{N} x_{i}^{l} x_{j}^{l} - \frac{1}{Z(\theta)} \sum_{x} \Pi_{ij} \exp(\theta_{ij} X_{i} X_{j}) \Pi_{i} \exp(\theta_{i} X_{i}) \ X_{i} X_{j} \]

A convex problem
Can find global optimum

need to do inference
Moment matching condition

\[ \frac{\partial l(\theta, D)}{\partial \theta_{ij}} = \frac{1}{N} \sum_{l}^{N} x_i^l x_j^l - \frac{1}{Z(\theta)} \sum_{X} \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i) \ X_i X_j \]

\[ \widehat{P}(X_i, X_j) = \frac{1}{N} \sum_{l=1}^{N} \delta(X_i, x_i^l) \delta(X_j, x_j^l) \]

\[ \text{Moment matching: } \frac{\partial l(\theta, D)}{\partial \theta_{ij}} = E_{\widehat{P}(X_i, X_j)}[X_i X_j] - E_{P(X|\theta)}[X_i X_j] \]
Optimize MLE for undirected models

- \( \max_{\theta} l(\theta, D) \) is a convex optimization problem.

- Can be solve by many methods, such as gradient descent, conjugate gradient.

- Initialize model parameters \( \theta \)
- Loop until convergence
  - Compute \( \frac{\partial l(\theta, D)}{\partial \theta_{ij}} = E_{\tilde{P}(X_i,X_j)}[X_iX_j] - E_P(X|\theta)[X_iX_j] \)
  - Update \( \theta_{ij} \leftarrow \theta_{ij} - \eta \frac{\partial l(\theta, D)}{\partial \theta_{ij}} \)

- Or use the gradient equation for fixed point iteration: iterative proportional fitting
Conditional random Fields

- Focus on conditional distribution
  - \( P(X_1, \ldots, X_n | Y_1, \ldots, Y_n, Y) \)

- Do not explicitly model dependence between \( Y_1, \ldots, Y_n, Y \)

- Only model relation between \( X - Y \) and \( X - Y \)

\[
P(X_1, X_2, X_3, X_4 | Y_1, Y_2, Y_3, Y_4, Y) \]

\[
= \frac{1}{Z(Y_1,Y_2,Y_3,Y_4,Y)} \Psi(Y_1, Y_2, X_1, X_2, Y) \Psi(Y_2, Y_3, X_2, X_3, Y)
\]

\[
Z(Y_1, Y_2, Y_3, Y_4, Y) = \sum_{Y_1Y_2Y_3} \Psi(Y_1, Y_2, X_1, X_2, X) \Psi(Y_2, Y_3, X_2, X_3, X)
\]
Parameter Learning for Conditional Random Fields

\[ P(X_1, \ldots, X_k | Y, \theta) = \frac{1}{Z(Y, \theta)} \exp\left( \sum_{ij} \theta_{ij} X_i X_j Y + \sum_i \theta_i X_i Y \right) \]

\[ = \frac{1}{Z(Y, \theta)} \prod_{ij} \exp(\theta_{ij} X_i X_j Y) \prod_i \exp(\theta_i X_i Y) \]

\[ Z(Y, \theta) = \sum_X \prod_{ij} \exp(\theta_{ij} X_i X_j Y) \prod_i \exp(\theta_i X_i Y) \]

Maximize long conditional likelihood

\[ cl(\theta, D) = \log(\prod_{l=1}^{N} \frac{1}{Z(y^l, \theta)} \prod_{ij} \exp(\theta_{ij} x_i^l x_j^l y^l) \prod_i \exp(\theta_i x_i^l y^l)) \]

\[ = \sum_{l=1}^{N} \left( \sum_{ij} \log(\exp(\theta_{ij} x_i^l x_j^l y^l)) + \sum_i \log(\exp(\theta_i x_i^l y^l)) - \log Z(y^l, \theta) \right) \]

\[ = \sum_{l=1}^{N} \left( \sum_{ij} \theta_{ij} x_i^l x_j^l y^l + \sum_i \theta_i x_i^l y^l - \log Z(y^l, \theta) \right) \]

\[ \text{can be other feature function } f(x_i) \]

\[ \text{Term } \log Z(y^l, \theta) \text{ does not decompose!} \]
Derivatives of log likelihood

\[ cl(\theta, D) = \frac{1}{N} \sum_i^N \left( \sum_{ij} \theta_{ij} x_i^l x_j^l y^l + \sum_i \theta_i x_i^l y^l - \log Z(y^l, \theta) \right) \]

\[ \frac{\partial cl(\theta, D)}{\partial \theta_{ij}} = \frac{1}{N} \sum_i^N \sum_{ij} x_i^l x_j^l y^l - \frac{\partial \log Z(y^l, \theta)}{\partial \theta_{ij}} \]

\[ = \frac{1}{N} \sum_i^N \sum_{ij} x_i^l x_j^l y^l - \frac{1}{Z(y^l, \theta)} \frac{\partial Z(y^l, \theta)}{\partial \theta_{ij}} \]

\[ = \frac{1}{N} \sum_i^N x_i^l x_j^l y^l - \frac{1}{Z(y^l, \theta)} \sum_x \prod_{ij} \exp(\theta_{ij} X_i X_j y^l) \prod_i \exp(\theta_i X_i y^l) \ X_i X_j y^l \]

\[ \text{need to do inference for each } y^l!!! \]
Moment matching condition

\[
\frac{\partial \text{cl}(\theta, D)}{\partial \theta_{ij}} = \\
\frac{1}{N} \sum_{l}^{N} x_i^l x_j^l y^l - \frac{1}{Z(y^l, \theta)} \sum_{x} \Pi_{ij} \exp(\theta_{ij} x_i x_j y^l) \Pi_i \exp(\theta_i x_i y^l) x_i x_j y^l
\]

- empirical covariance matrix
- covariance matrix from model

\[
\hat{P}(X_i, X_j, Y) = \frac{1}{N} \sum_{l=1}^{N} \delta(X_i, x_i^l) \delta(X_j, x_j^l) \delta(Y, y^l)
\]

\[
\hat{P}(Y) = \frac{1}{N} \sum_{l=1}^{N} \delta(Y, y^l)
\]

Moment matching:

\[
\frac{\partial \text{cl}(\theta, D)}{\partial \theta_{ij}} = E_{\hat{P}(X_i, X_j, Y)}[X_i X_j Y] - E_{P(X|Y, \theta)} \hat{P}(Y)[X_i X_j Y]
\]
Optimize MLE for undirected models

- \( \max_{\theta} c_l(\theta, D) \) is a convex optimization problem.

- Can be solved by many methods, such as gradient descent, conjugate gradient.

- Initialize model parameters \( \theta \)
- Loop until convergence

- Compute \( \frac{\partial c_l(\theta, D)}{\partial \theta_{ij}} = E_{\tilde{P}(X_i, X_j, Y)}[X_i X_j Y] - E_{P(X|Y, \theta)\tilde{P}(Y)}[X_i X_j Y] \)

- Update \( \theta_{ij} \leftarrow \theta_{ij} - \eta \frac{\partial c_l(\theta, D)}{\partial \theta_{ij}} \)