Graphical Model & Gibbs Sampling
Hung-yi Lee
Structured Learning

Problem 1: Evaluation
- What does $F(x,y)$ look like? $F(x, y) = w \cdot \phi(x, y)$

Problem 2: Inference
- How to solve the “arg max” problem

\[ y = \operatorname{arg\,max}_{y \in Y} F(x, y) \]

Problem 3: Training
- Given training data, how to find $F(x,y)$

Structured SVM, etc.

We also know how to involve hidden information.
Difficulties

**Difficulty 1. Evaluation**

\[ F(x, y) = w \cdot \phi(x, y) \]

Hard to figure out? Hard to interpret the meaning?

**Difficulty 2. Inference**

We can use Viterbi algorithm to deal with sequence labeling. How about other cases?
Graphical Model

A language which describes the evaluation function
Graphical Model

Define and describe your evaluation function $F(x, y)$ by a graph.

There are three kinds of graphical model.

- *Factor graph*, *Markov Random Field* (MRF) and *Bayesian Network* (BN).
- Only *factor graph* and *MRF* will be briefly mentioned today.
Decompose $F(x,y)$

- $F(x,y)$ is originally a **global** function
  - Define over the whole $x$ and $y$
- Based on graphical model, $F(x,y)$ is the composition of some **local** functions
  - $x$ and $y$ are decomposed into smaller components
  - Each local function defines on only a few related components in $x$ and $y$
  - Which components are related $\rightarrow$ defined by Graphical model
Decomposable x and y

- x and y are decomposed into smaller components

**POS Tagging**

x: John saw the saw.

y: PN V D N

x: x₁ x₂ x₃ x₄

y: y₁ y₂ y₃ y₄

{word}

{tags}
Each factor influences some components.
Each factor corresponds to a local function.

\[ F(x, y) = f_a(x_1, y_1) + f_b(x_2, y_1, y_2) + f_c(y_2) \]

Larger value means more compatible.

You only have to define the factors.
The local functions of the factors are learned from data.
Factor Graph - Example

- **Image De-noising**

Each pixel is one component

Noisy image $x$

Clean image $y$

http://cs.stanford.edu/people/karpathy/visml/ising_example.html
Factor Graph - Example

Noisy and clean images are related

- **a**: the values of $x_i$ and $y_i$
  - The colors in the clean image is smooth.
- **b**: the values of the neighboring $y_i$

The weights can be learned from data.

**Factor:**

\[ f_a(x_i, y_i) = \begin{cases} 1 & x_i = y_i \\ -1 & x_i \neq y_i \end{cases} \]

\[ f_b(y_i, y_j) = \begin{cases} 2 & y_i = y_j \\ -2 & y_i \neq y_j \end{cases} \]
Noisy and clean images are related

- **a**: the values of $x_i$ and $y_i$

The colors in the clean image is smooth.

- **b**: the values of the neighboring $y_i$

Realize $F(x, y)$ easily from the factor graph

$$F(x, y) = \sum_{i=1}^{4} f_a(x_i, y_i) + f_b(y_1, y_2) + f_b(y_1, y_3) + f_b(y_2, y_4) + f_b(y_3, y_4)$$
**Factor Graph - Example**

**Factor:**
- **c:** the values of $x_i$ and the values of the neighboring $y_i$
- **d:** the values of the neighboring $x_i$ and the values of $y_i$

$$f_c(x_i, y_i, y_{i-1})$$

$$f_d(x_i, x_{i-1}, y_i)$$

$$f_e(x_i, x_{i-1}, y_i, y_{i-1})$$
Markov Random Field (MRF)

**Clique**: a set of components connecting to each other

**Maximum Clique**: a clique that is not included by other cliques
Each maximum clique on the graph corresponds to a factor.
**Evaluation Function**

\[ f_a(A, B) + f_b(A, D) + f_c(B, C) + f_d(C, D, E) + f_e(E, F, G) \]
Training

\[ F(x, y) = f_a(x_1, x_2, y_1) + f_b(y_1, y_2) \]
\[ = w_a \cdot \phi_a(x_1, x_2, y_1) + w_b \cdot \phi_b(y_1, y_2) \]
\[ = \begin{bmatrix} w_a \\ w_b \end{bmatrix} \begin{bmatrix} \phi_a(x_1, x_2, y_1) \\ \phi_b(y_1, y_2) \end{bmatrix} \]
\[ = w \cdot \phi(x, y) \]

Simply training by

**structured perceptron**
or **structured SVM**

Max-Margin Markov Networks (M3N)
Training

\[ F(x, y) = f_a(x_1, x_2, y_1) + f_b(y_1, y_2) \]
\[ = w_a \cdot \phi_a(x_1, x_2, y_1) + w_b \cdot \phi_b(y_1, y_2) \]

\( y_1, y_2 \epsilon \{+1, -1\} \)

<table>
<thead>
<tr>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( f_b(y_1, y_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>+1</td>
<td>( w_1 )</td>
</tr>
<tr>
<td>+1</td>
<td>-1</td>
<td>( w_2 )</td>
</tr>
<tr>
<td>-1</td>
<td>+1</td>
<td>( w_3 )</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>( w_4 )</td>
</tr>
</tbody>
</table>

\[
\phi_b(+1, +1) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \\
\phi_b(+1, -1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \\
\phi_b(-1, +1) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \\
\phi_b(-1, -1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}
\]

\[
w_b = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}
\]
Now can you interpret this?

\[ \phi(x, y) \]
Probability Point of View

• $F(x, y)$ can be any real number
• If you like probability
  
  Between 0 and 1
  
  $P(x, y) = \frac{e^{F(x, y)}}{\sum_{x', y'} e^{F(x', y')}}$ To be positive

  normalization

$P(y|x) = \frac{P(x, y)}{P(x)}$

$P(x, y) = \frac{P(x, y)}{\sum_{y''} P(x, y'')} = \frac{e^{F(x, y)}}{\sum_{x', y'} e^{F(x', y')}}$

$= \frac{e^{F(x, y)}}{\sum_{x', y'} e^{F(x', y')}}$
Gibbs Sampling

Inference for the dumb
Given input noisy image $x$

$x_1, x_2, x_3, x_4 = -1, -1, -1, 1$

Design an efficient algorithm to do that is not always easy.
Sampling?

**Probability point of view:**

\[
P(x, y) = \frac{e^{F(x,y)}}{\sum_{x',y'} e^{F(x',y')}}
\]

\[
P(y|x) = \frac{e^{F(x,y)}}{\sum_{y''} e^{F(x,y'')}}
\]

Independent of \( y \)

\[
\hat{y} = \text{arg max}_y F(x,y) = \hat{y} = \text{arg max}_y P(y|x) \propto F(x, y)
\]
Sampling?

- \( P(y|x) \) is a distribution

Given \( x_1, x_2, x_3, x_4 = -1, -1, -1, 1 \)

\[
\hat{y} = \arg \max_y P(y|x)
\]

Sample from the distribution ......

\(-1, -1, -1, -1\)  
\(1,1,1,1\)  
\(-1, -1, -1, 1\)  
\(-1, -1, -1, -1\)  
\(1,1,1,1\)  
\(-1, -1, -1, -1\)  
......  
\(-1, -1, -1, -1\)

Max probability  
Inference result
$P(y|x) = \frac{e^{F(x,y)}}{\sum_{y''} e^{F(x,y'')}}$

- $P(y|x)$ is a distribution

Given $x_1, x_2, x_3, x_4 = -1, -1, -1, 1$

- If we know the distribution, why bother with the sampling?
- It is hard to know the distribution.

Sample from the distribution ......

- $y_1, y_2, y_3, y_4 = -1, -1, -1, -1$
- $y_1, y_2, y_3, y_4 = 1, 1, 1, 1$
- $y_1, y_2, y_3, y_4 = -1, -1, -1, 1$
- $y_1, y_2, y_3, y_4 = -1, -1, -1, -1$
- $y_1, y_2, y_3, y_4 = 1, 1, 1, 1$
- $y_1, y_2, y_3, y_4 = -1, -1, -1, -1$
- $y_1, y_2, y_3, y_4 = -1, -1, -1, -1$

Max probability Inference result
Gibbs Sampling

• There is a probability distribution \( P(y|x) \)
  • \( y = \{y_1, y_2, ..., y_N\} \)
• We want to sample from \( P(y|x) \), but it is too complex to do that
• However, \( P(y_i|y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_N, x) \) can be computed
• We can sample from \( P(y|x) \) by Gibbs sampling
**Gibbs Sampling**

\[ y^0 = \{y_1^0, y_2^0, \ldots, y_N^0\} \quad \text{Initialization} \]

For \( t = 1 \) to \( T \):

\[
egin{align*}
    y_1^t & \sim P(y_1 | y_2 = y_2^{t-1}, y_3 = y_3^{t-1}, \ldots, y_N = y_N^{t-1}, x) \\
    y_2^t & \sim P(y_2 | y_1 = y_1^t, y_3 = y_3^{t-1}, \ldots, y_N = y_N^{t-1}, x) \\
    y_3^t & \sim P(y_3 | y_1 = y_1^t, y_2 = y_2^t, \ldots, y_N = y_N^{t-1}, x) \\
    & \vdots \\
    y_N^t & \sim P(y_N | y_1 = y_1^t, y_2 = y_2^t, \ldots, y_{N-1} = y_{N-1}^t, x)
\end{align*}
\]

Get a sample: \( y^t = \{y_1^t, y_2^t, \ldots, y_N^t\} \)

\[ y^1, y^2, y^3, \ldots, y^T \quad \text{As sampling from } P(y|x) \]
Gibbs Sampling

- Is \( P(y_i | y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_N, x) \) easy to be computed?

\[
P(y|x) = \frac{e^{F(x,y)}}{\sum_{y''} e^{F(x,y'')}}\]

\( y_i \epsilon \{+1, -1\} \rightarrow 2^N \) possible \( y \)

Enumerate all possible \( y \) may not be tractable

\[
P(y_i | y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_N, x)
\]

\[
= \frac{e^{F(x,y_{-i},y_i)}}{\sum_{y'_i} e^{F(x,y_{-i},y'_i)}}
\]

\( y_i \epsilon \{+1, -1\} \rightarrow 2 \) possible \( y_i \)

Enumerate all possible \( y_i \) may be tractable
Initialization \[ y_1, y_2, y_3, y_4 = -1, -1, -1, 1 \]

Sample from \( P(y|x) \) by Gibbs sampling
Sample $y_1$ given all the other variables

$$y_1 \sim P(y_1 | y_{-1}, x) \quad y_{-1} = \{y_2, y_3, y_4\}$$

Compute $P(y_1 = 1 | y_{-1}, x)$ and $P(y_1 = -1 | y_{-1}, x)$

$$P(y_1 = 1 | y_{-1}, x) = \frac{P(x, y_1 = 1, y_{-1})}{P(x, y_1 = 1, y_{-1}) + P(x, y_1 = -1, y_{-1})}$$

$$= \frac{e^{F(x,y_1=1,y_{-1})}}{e^{F(x,y_1=1,y_{-1})} + e^{F(x,y_1=-1,y_{-1})}}$$

$$= -1.8$$
Sample $y_1$ given all the other variables

$$y_1 \sim P(y_1 | y_{-1}, x) \quad y_{-1} = \{y_2, y_3, y_4\}$$

Compute $P(y_1 = 1 | y_{-1}, x)$ and $P(y_1 = -1 | y_{-1}, x)$

$$P(y_1 = 1 | y_{-1}, x) = \frac{P(x, y_1 = 1, y_{-1})}{P(x, y_1 = 1, y_{-1}) + P(x, y_1 = -1, y_{-1})}$$

$$= \frac{e^{F(x,y_1=1,y_{-1})}}{e^{F(x,y_1=1,y_{-1})} + e^{F(x,y_1=-1,y_{-1})}} = -1.8$$

$$= 0.10 \quad \text{Random sample} \quad y_1 = -1$$
Sample $y_2$ given all the other variables

$$y_2 \sim P(y_2 | y_{-2}, x)$$

$$P(y_2 = 1 | y_{-2}, x) = \frac{P(x, y_2 = 1, y_{-2})}{P(x, y_2 = 1, y_{-2}) + P(x, y_2 = -1, y_{-2})}$$

$$= \frac{e^{F(x, y_2 = 1, y_{-2})}}{e^{F(x, y_2 = 1, y_{-2})} + e^{F(x, y_2 = -1, y_{-2})}} = 0.2$$
Sample $y_2$ given all the other variables

$$y_2 \sim P(y_2 | y_{-2}, x)$$

$$P(y_2 = 1 | y_{-2}, x) = \frac{P(x, y_2 = 1, y_{-2})}{P(x, y_2 = 1, y_{-2}) + P(x, y_2 = -1, y_{-2})}$$

$$= \frac{e^{F(x, y_2 = 1, y_{-2})}}{e^{F(x, y_2 = 1, y_{-2})} + e^{F(x, y_2 = -1, y_{-2})}}$$

$$= 0.2$$

$$= 0.4$$
Sample \( y_2 \) given all the other variables

\[
y_2 \sim P(y_2 | y_{-2}, x)
\]

\[
P(y_2 = 1 | y_{-2}, x) = \frac{P(x, y_2 = 1, y_{-2})}{P(x, y_2 = 1, y_{-2}) + P(x, y_2 = -1, y_{-2})}
\]

\[
e^{F(x, y_2 = 1, y_{-2})} = 0.2
\]

\[
e^{F(x, y_2 = 1, y_{-2})} + e^{F(x, y_2 = -1, y_{-2})} = 0.2 + 0.4 = 0.6
\]

Random sample

\[
y_2 = 1
\]
Sample $y_3$ given all the other variables

$$y_3 \sim P(y_3 | y_{-3}, x)$$

$$P(y_3 = 1 | y_{-3}, x) = ? \quad 0.45$$

$y_3 = -1$

Random sample
Sample $y_4$ given all the other variables

$$y_4 \sim P(y_4 | y_{-4}, x)$$

Get 1-st sample $y_1 = -1, y_2 = 1, y_3 = -1, y_4 = -1$
Get **1-st** sample \( y_1=-1, y_2=1, y_3=-1, y_4=-1 \)

Get **2-nd** sample \( y_1=-1, y_2=-1, y_3=-1, y_4=-1 \)
Get **1-st** sample $y_1=-1$, $y_2=1$, $y_3=-1$, $y_4=-1$

Get **2-nd** sample $y_1=-1$, $y_2=-1$, $y_3=-1$, $y_4=-1$

Get **3-rd** sample $y_1=1$, $y_2=1$, $y_3=-1$, $y_4=1$

Get **4-th** sample $y_1=-1$, $y_2=1$, $y_3=-1$, $y_4=1$

Get **5-th** sample $y_1=1$, $y_2=1$, $y_3=1$, $y_4=1$

\[\vdots\]

**Until you want to stop**
P(y^A|x) \approx 0.33 \quad P(y^B|x) \approx 0.22 \quad P(y^C|x) \approx 0.004

<table>
<thead>
<tr>
<th>No. of samples</th>
<th>y^A</th>
<th>y^B</th>
<th>y^C</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>23</td>
<td>37</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>315</td>
<td>230</td>
<td>8</td>
</tr>
<tr>
<td>10000</td>
<td>3307</td>
<td>2225</td>
<td>40</td>
</tr>
<tr>
<td>100000</td>
<td>32637</td>
<td>22129</td>
<td>422</td>
</tr>
</tbody>
</table>

From sampling: y^A would be the results of inference.
How about starting from different initialization?

Not really change the final results.

<table>
<thead>
<tr>
<th>No. of samples</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>1000</td>
<td>331</td>
<td>237</td>
<td>2</td>
</tr>
<tr>
<td>10000</td>
<td>3251</td>
<td>2176</td>
<td>31</td>
</tr>
<tr>
<td>100000</td>
<td>32911</td>
<td>21845</td>
<td>385</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. of samples</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>28</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>318</td>
<td>226</td>
<td>2</td>
</tr>
<tr>
<td>10000</td>
<td>3277</td>
<td>2169</td>
<td>46</td>
</tr>
<tr>
<td>100000</td>
<td>32319</td>
<td>21751</td>
<td>393</td>
</tr>
</tbody>
</table>
All rivers run into the sea.

Practical Suggestion

• “burn-in”
  • “burn-in” period: The first few of samples would be influenced by the initialization
  • Discard the samples in the “burn-in” period

• Modify the sampling distribution

\[
P(y_i | y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_N, x) = \frac{e^{F(x, y_i - y_i'y_i')} X c}{\sum_{y_i'} e^{F(x, y_i - y_i'y_i')} X c}
\]

\( c > 1 \)

Increase \( c \) after each interaction
Gibbs Sampling

A little bit of theory
Gibbs Sampling

Gibbs sampling from a distribution $P(z)$ ( $z = \{z_1, \ldots, z_N\}$ )

$z^0 = \{z^0_1, z^0_2, \ldots, z^0_N\}$

For $t = 1$ to $T$:

$z^t_1 \sim P(z_1 | z_2 = z^t_{2-1}, z_3 = z^t_{3-1}, z_4 = z^t_{4-1}, \ldots, z_N = z^t_{N-1})$

$z^t_2 \sim P(z_2 | z_1 = z^t_1, z_3 = z^t_{3-1}, z_4 = z^t_{4-1}, \ldots, z_N = z^t_{N-1})$

$z^t_3 \sim P(z_3 | z_1 = z^t_1, z_2 = z^t_2, z_4 = z^t_{4-1}, \ldots, z_N = z^t_{N-1})$

$\vdots$

$z^t_N \sim P(z_N | z_1 = z^t_1, z_2 = z^t_2, z_3 = z^t_3, \ldots, z_{N-1} = z^t_{N-1})$

Output: $z^t = \{z^t_1, z^t_2, \ldots, z^t_N\}$

As sampling from $P(z)$

Why?
Three cities A, B and C

The traveler recorded the cities he visited each day.

A B C A A A ......

This is a Markov chain state
Markov Chain

With sufficient samples ......
A : B : C = 0.6 : 0.2 : 0.2
(independent of the starting city)
Markov Chain

P(A) = 0.6
P(B) = 0.2
P(C) = 0.2

Stationary Distribution

The distribution will not change.

\[
\begin{align*}
PT(A|A)P(A) + PT(A|B)P(B) + PT(A|C)P(C) &= P(A) \\
PT(B|A)P(A) + PT(B|B)P(B) + PT(B|C)P(C) &= P(B) \\
PT(C|A)P(A) + PT(C|B)P(B) + PT(C|C)P(C) &= P(C)
\end{align*}
\]

The distribution will not change.
Markov Chain

A Markov Chain can have multiple stationary distributions.

- Reaching which stationary distribution depends on starting state.

The Markov Chain fulfill some conditions will have unique stationary distribution.

\[ P_T(s' | s) \text{ for any states } s \text{ and } s' \text{ is not zero} \]

(sufficient but not necessary condition)
Markov Chain from Gibbs Sampling

Gibbs sampling from a distribution $P(z)$ ($z = \{z_1, ..., z_N\}$)

$z^0 = \{z_1^0, z_2^0, ..., z_N^0\}$

For $t = 1$ to $T$:

$z_1^t \sim P(z_1 | z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, ..., z_N = z_N^{t-1})$

$z_2^t \sim P(z_2 | z_1 = z_1^t, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, ..., z_N = z_N^{t-1})$

$z_3^t \sim P(z_3 | z_1 = z_1^t, z_2 = z_2^t, z_4 = z_4^{t-1}, ..., z_N = z_N^{t-1})$

\vdots

$z_N^t \sim P(z_N | z_1 = z_1^t, z_2 = z_2^t, z_3 = z_3^t, ..., z_{N-1} = z_{N-1}^t)$

Output: $z^t = \{z_1^t, z_2^t, ..., z_N^t\}$

This is a **Markov Chain**

- $z^t$ only depend on $z^{t-1}$
Markov Chain from Gibbs Sampling

Gibbs sampling from a distribution $P(z)$ ($z = \{z_1, ..., z_N\}$)

$z^0 = \{z_1^0, z_2^0, \ldots, z_N^0\}$

For $t = 1$ to $T$:

$z_1^t \sim P(z_1 | z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \ldots, z_N = z_N^{t-1})$

$z_2^t \sim P(z_2 | z_1 = z_1^t, z_3 = z_3^{t-1}, z_4 = z_4^{t-1}, \ldots, z_N = z_N^{t-1})$

$z_3^t \sim P(z_3 | z_1 = z_1^t, z_2 = z_2^t, z_4 = z_4^{t-1}, \ldots, z_N = z_N^{t-1})$

$\vdots$

$z_N^t \sim P(z_N | z_1 = z_1^t, z_2 = z_2^t, z_3 = z_3^t, \ldots, z_{N-1} = z_{N-1}^t)$

Output: $z^t = \{z_1^t, z_2^t, \ldots, z_N^t\}$

Proof that the Markov chain has unique stationary distribution which is $P(z)$. 
Markov Chain from Gibbs Sampling

- Markov chain from Gibbs sampling has unique stationary distribution? Yes
  - $P_T(z'|z) > 0$, for any $z$ and $z'$

\[
\begin{align*}
  z_1^t &\sim P(z_1 | z_2 = z_2^{t-1}, z_3 = z_3^{t-1}, \ldots, z_N = z_N^{t-1}) \\
  z_2^t &\sim P(z_2 | z_1 = z_1^t, z_3 = z_3^{t-1}, \ldots, z_N = z_N^{t-1}) \\
  z_3^t &\sim P(z_3 | z_1 = z_1^t, z_2 = z_2^t, \ldots, z_N = z_N^{t-1}) \\
  &\quad \vdots \\
  z_N^t &\sim P(z_N | z_1 = z_1^t, z_2 = z_2^t, \ldots, z_{N-1} = z_{N-1}^t)
\end{align*}
\]

None of the conditional probability is zero

can be any $z^t$
Markov Chain from Gibbs Sampling

• Show that $P(z)$ is a stationary distribution

$$\sum_z P_T(z' | z) \ P(z) = P(z')$$

$$P_T(z' | z) = P(z_1' | z_2, z_3, z_4, \ldots, z_N) \times P(z_2' | z_1', z_3, z_4, \ldots, z_N) \times P(z_3' | z_1', z_2', z_4, \ldots, z_N) \cdots \times P(z_N' | z_1', z_2', z_3', \ldots, z_{N-1}')$$

Please do the math yourself

There is only one stationary distribution for Gibbs sampling, so we are done.
Thank you for your attention!
Acknowledgement

• 感謝 陳乃群 同學於上課時糾正投影影片上的錯誤