Statistical Analysis on Order Structures

Mahito Sugiyama (ISIR, Osaka University)
（杉山麿人; 大阪大学産業科学研究所）
Summary

Probability distribution on posets (partially ordered sets)

Information geometry

\[ \log p(x) = \sum \theta(s) \]

Decomposition in the log-linear model

S. Amari, Information geometry on hierarchy of probability distributions, IEEE TIT 2001
Probability distribution on posets (partially ordered sets)

\[
\log p(x) = \sum \theta(s)
\]

Decomposition in the log-linear model

Numerical score (KL divergence) and the \( p \)-value for higher-order interactions

S. Amari, Information geometry on hierarchy of probability distributions, IEEE TIT 2001
Transaction database

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ID10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Itemset lattice

∅

{□} Frequency = 0.3

{□,□,□} Frequency = 0.3

{□,□} Frequency = 0.5

{□,□} Frequency = 0.5

{□,□} Frequency = 0.6

{□,□,□} Frequency = 0.9

{□,□,□} Frequency = 0.7

{□,□} Frequency = 0.7

{□,□} Frequency = 0.5

{□} Frequency = 1.0
| ID 1: | 1 | 1 | 1 | 0 |
| ID 2: | 1 | 1 | 1 | 1 |
| ID 3: | 1 | 1 | 0 | 0 |
| ID 4: | 1 | 1 | 1 | 1 |
| ID 5: | 1 | 1 | 0 | 1 |
| ID 6: | 1 | 0 | 1 | 1 |
| ID 7: | 1 | 0 | 1 | 1 |
| ID 8: | 1 | 1 | 1 | 1 |
| ID 9: | 1 | 0 | 0 | 0 |
| ID 10: | 0 | 1 | 1 | 0 |

### Itemset lattice

- Itemset frequency: 0.3

- Itemset lattice:
  - $\emptyset$ (frequency 1.0)
  - $\{\cdot\}$ (frequency 0.9)
  - $\{\cdot, \cdot\}$ (frequency 0.6)
  - $\{\cdot, \cdot, \cdot\}$ (frequency 0.5)
  - $\{\cdot, \cdot, \cdot\}$ (frequency 0.3)

- Example itemsets:
  - $\{\cdot, \cdot\}$ (frequency 0.6)
  - $\{\cdot, \cdot, \cdot\}$ (frequency 0.5)
  - $\{\cdot, \cdot\}$ (frequency 0.7)
  - $\{\cdot, \cdot, \cdot\}$ (frequency 0.3)
Itemset lattice

Positive class:
- \{\textcolor{cyan}{\textbullet}, \textcolor{red}{\textbullet}, \textcolor{green}{\textbullet}\}
- \{\textcolor{cyan}{\textbullet}, \textcolor{green}{\textbullet}\}
- \{\textcolor{red}{\textbullet}\}

Negative class:
- \{\textcolor{cyan}{\textbullet}\}
- \{\textcolor{red}{\textbullet}\}
- \{\textcolor{green}{\textbullet}\}

\(p\)-value = 1.0

Transaction database

<table>
<thead>
<tr>
<th>ID</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

LAMP (Terada et al. PNAS 2013)
Westfall-Young light (Llinares-López et al. KDD 2015)
Transaction database

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0
ID 4: 1 1 1
ID 5: 1 1 0
ID 6: 1 0 1
ID 7: 1 0 1
ID 8: 1 1 1
ID 9: 1 0 0
ID 10: 0 1 0

Itemset lattice

Frequency = 0.3
Probability = 0.3
Upward = Pattern mining

\[ \eta: \text{Frequency} \]
\[ p: \text{Probability} \]

\[ \eta(\{\bullet, \bullet\}) = p(\{\bullet, \bullet\}) + p(\{\bullet, \bullet, \bullet\}) \]
**Upward** = **Pattern mining**

**Downward** = **Log-linear analysis**

\[ \eta: \text{Frequency} \]
\[ p: \text{Probability} \]
\[ \theta: \text{Coefficient of log-linear model} \]

**Itemset lattice**

\[ \eta(\{\text{●, ●}\}) = p(\{\text{●, ●}\}) + p(\{\text{●, ●, ●}\}) \]

\[ \log p(\{\text{●, ●}\}) = \theta(\{\text{●, ●}\}) + \theta(\{\text{●}\}) + \theta(\{\text{●}\}) + \theta(\emptyset) \]
\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
$$\log p(x) = \sum_{s \leq x} \theta(s)$$

Natural parameter

Exponential family:

$$p(x) = \exp\left( \sum \theta(s) F_s(x) - \psi(\theta) \right)$$

E.g. Gaussian
\[ \eta(x) = \sum_{s \geq x} p(s) \]

\[ \eta(x) = \mathbb{E}[F_x(s)] \]

Sufficient statistics of exponential family

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]

Natural parameter

Exponential family:

\[ p(x) = \exp\left(\sum \theta(s) F_s(x) - \psi(\theta)\right) \]
Triple for each node

$$\eta(x) = \sum_{s \geq x} p(s)$$

$$\log p(x) = \sum_{s \leq x} \theta(s)$$
Triple for each node

\[ p, \eta, \theta \]

\[
\{\text{●, ●} \} \]

\[
\{\text{●} \} \]

\[
\{\text{●} \} \]

\[
\emptyset \]

\[
0.2 \]

\[
0.2 \]

\[
-1.79 \]

\[
0.3 \]

\[
0.5 \]

\[
1.10 \]

\[
0.1 \]

\[
1.0 \]

\[
-2.30 \]

\[
0.4 \]

\[
0.6 \]

\[
1.39 \]

\[
\log p(x) = \sum_{s \leq x} \theta(s) \]

\[ \eta(x) = \sum_{s \geq x} p(s) \]
Triple for each node

\[ p, \eta, \theta \]

\{ \bullet, \bigcirc \} \quad 0.2 \quad 0.2

\{ \bullet \} \quad 0.3 \quad 0.5

\{ \bigcirc \} \quad -1.79

\emptyset \quad 0.1 \quad 1.10

Probability distribution is a "point" in 3D space

\[ \eta(x) = \sum_{s \geq x} p(s) \]

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
Triple for each node

\[ \eta(x) = \sum_{s \geq x} p(s) \]

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]

Probability distribution is a “point” in 3D space
Triple for each node

\[ p \]
\[ \eta \]
\[ \theta \]

\[
\left\{ \, \bullet, \, \bullet \, \right\}
\]

\[
0.2
\]
\[
0.2
\]
\[
0.4
\]
\[
0.6
\]

\[
\{ \, \bullet \, \}
\]
\[
\{ \, \bullet \, \}
\]
\[
\{ \, \bullet, \, \bullet \, \}
\]

\[
\emptyset
\]
\[
0.3
\]
\[
0.1
\]
\[
1.10
\]
\[
1.0
\]
\[
-1.79
\]
\[
-2.30
\]

One-to-one

\[
\theta(\{ \, \bullet \, \})
\]

\[
\eta(\{ \, \bullet \, \})
\]

\[
\eta(x) = \sum_{s \geq x} p(s)
\]
\[
\log p(x) = \sum_{s \leq x} \theta(s)
\]

\[ \theta \text{ and } \eta \text{ are dually orthogonal} \]
<table>
<thead>
<tr>
<th>Dist. $P$</th>
<th>Dist. $Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$0.1$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>$0.1$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>$1.95$</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>1.10</td>
<td>0.0</td>
</tr>
<tr>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>-1.79</td>
<td>1.95</td>
</tr>
<tr>
<td>0.6</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>1.39</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Mixed distribution $R$
Mixed distribution $R$
Nonnegative decomposition of the KL divergence

\[ KL(P, Q) = KL(P, R) + KL(R, Q) \]
Nonnegative decomposition of the KL divergence

Dist. $P$

Choose $\eta$

$\rho$

$\eta$

$\theta$

Dist. $Q$

Choose $\theta$

Mixed distribution $R$

$0.4390 = 0.3946 + 0.0444$
Dist. $P$

Mixed distribution $R$

Uniform dist. $P_0$

Log-linear model

$$\log p(x) = \sum_{s \leq x} \theta(s)$$
Log-linear model

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]

Mixed distribution \( R \)

Choose \( \eta \)

Choose \( \theta \) (KNOCK DOWN)

Contribution of the node

\[ = KL(P, R) = 0.086 \]
The statistics $\lambda$:

$$\lambda = 2 \cdot [\text{sample size}] \cdot KL(P, R)$$

follows $\chi^2$-distribution with d.f. $[\#\text{nodes} - 1]$

$\Rightarrow$ $p$-value can be obtained!

Mixed distribution $R$

Log-linear model

$$\log p(x) = \sum_{s \leq x} \theta(s)$$
Make a Poset from Data

Dataset

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0
ID 4: 1 1 1
ID 5: 1 1 0
ID 6: 1 0 1
ID 7: 1 0 1
ID 8: 1 1 1
ID 9: 1 0 0
ID10: 0 1 0

Number of nodes = 2

#features \Rightarrow \text{combinatorial explosion!}
Make a Poset from Data

Dataset

ID 1:  1  1  0
ID 2:  1  1  1
ID 3:  1  1  0
ID 4:  1  1  1
ID 5:  1  1  0
ID 6:  1  0  1
ID 7:  1  0  1
ID 8:  1  1  1
ID 9:  1  0  0
ID10: 0  1  0

Frequency = 0.3
Probability = 0.3

Probability $\geq 0.2$
(user specified threshold)
Remove Nodes with Probability 0

Dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ID10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
p \eta \theta = 1 - \sum p(s) = 0.2 = 1 - 0.405 - 0.3 - 0.0 - 0.6 - 1.0 = -1.61
\]
Example on Real Data (kosarak)

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0 ...
ID 4: 1 1 1
ID 5: 1 1 0
...<br>

Sample size: 990,002

# features: 41,270

# nodes: 3,253
(Threshold: $10^{-5}$)

# significant interactions: 583

Single feature: 537
Pairwise interactions: 41
Triple interactions: 5

Total runtime: 4.95 seconds
Example on Real Data (accidents)

# features: 468

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0 ...
ID 4: 1 1 1
ID 5: 1 1 0

# nodes: 281
(Threshold: $5 \times 10^{-6}$)

Sample size: 340,183

# significant interactions: 280

# features in each interaction is between 26 to 41

Total runtime: 4.95 seconds
Conclusion

• We build information geometry for posets (partially ordered sets)
  – Natural connection between the information geometric dual coordinates and the partial order structure

• We can decompose a probability distribution and assess the significance of any-order interactions beyond pairwise interactions