Uncertainty in Data Mining
MTAT.03.183 Data Mining

Meelis Kull
meelis.kull@ut.ee
meelis.kull@bristol.ac.uk

May 19, 2016
21st Century Healthcare Challenges

- UK: 1.4 million people aged over 85, by 2035 there will be 3.6 million.
- Japan: will have the oldest population in human history by 2050 (52 yrs).
- China: a retired population larger than Europe.
- Ageing populations living with long term health conditions:
  - obesity
  - diabetes
  - depression
  - heart disease
  - dementia ...
- Technology will have to fill the gap between expectations and reality of healthcare.
The SPHERE Project

- Temperature, light level, humidity, air quality
- Water & electricity consumption
- Video: emotion, gate, activity, interaction
- Wearables: activity, sleep, etc.
- Contextual information
- Feedback
SPHERE WP5: Data fusion and data mining

• Goals:
  • Combine information from all sensors to recognise activities in multi-resident homes, provide summaries and make decisions

• Some key challenges:
  • Differences across households and residents
  • Overlapping and hierarchical activities
  • Missing data
  • Uncertainty
  • Summarisation
  • Decision making
SPHERE WP5: Data fusion and data mining

• Goals:
  • Combine information from all sensors to recognise activities in multi-resident homes, provide summaries and make decisions

• Some key challenges:
  • Differences across households and residents
  • Overlapping and hierarchical activities
  • Missing data
  • Uncertainty
  • Summarisation
  • Decision making

May 19, 2016
Meelis Kull, University of Tartu, MTAT.03.183 Data Mining
SPHERE WP5: Data fusion and data mining

• **Goals:**
  • Combine information from all sensors to recognise activities in multi-resident homes, provide summaries and make decisions

• **Some key challenges:**
  • Differences across households and residents
  • Overlapping and hierarchical activities
  • Missing data
  • Uncertainty
  • Summarisation
  • Decision making
Lecture on Uncertainty in Data Mining

- Introduction to uncertain data
- Representing uncertainty in a relational database
- Possible worlds
- U-relation
- Data mining algorithms with uncertainty
- Gaussian processes
Lecture on Uncertainty in Data Mining

• Introduction to uncertain data
• Representing uncertainty in a relational database
• Possible worlds
• U-relation
• Data mining algorithms with uncertainty
• Gaussian processes
Uncertainty Is (Almost) Everywhere

• Uncertainty is often caused by our limited perception and understanding of reality
  – Limited observation equipment
  – Limited resource to collect, store, transform, analyze, and understand data

• Uncertainty can be inherent in nature
  – How much do you like/dislike McCain and Obama?
Data Collection Using Sensors

• Sensors are often used to collect data
  – Thermal, electromagnetic, mechanical, chemical, optical radiation, acoustic,
  – Applications: environment surveillance, security, manufacture systems,

• Ideal sensors
  – Ideal sensors are designed to be linear: the output signal of a sensor is linearly proportional to the value of the measured property
  – Sensitivity: the ratio between output signal and measured property

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yifei Tao, Xuemin Lin
Measurement Errors – Certain

- Sensitivity error: the sensitivity differs from the value specified
- Offset (bias): the output of a sensor at zero input
- Nonlinearity: the sensitivity is not constant over the range of the sensor

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertain (Dynamic) Errors

- **Dynamic error**: deviation caused by a rapid change of the measured property over time
- **Drift**: the output signal changes slowly independent of the measured property
  - Long term drift: a slow degradation of sensor properties over a long period
- **Noise**: random deviation of the signal varying in time
- **A sensor may to some extent be sensitive to properties (e.g., temperature) other than the one being measured**
- **Dynamic error due to sampling frequency of digital sensors**

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertainty in Survey Data

- Social security number: 185 or 785
  - Exclusiveness: SSN should be unique

- Is Smith married?
  - Single or married, but not both

Antova et al. ICDE’07

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertainty due to Data Granularity

- Which state is p9 in?
- What is the total repair cost for F150’s in the East?

<table>
<thead>
<tr>
<th></th>
<th>Auto</th>
<th>Loc</th>
<th>Repair</th>
<th>Text</th>
<th>Brake</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>F-150</td>
<td>NY</td>
<td>$200</td>
<td>...</td>
<td>(0.8, 0.2)</td>
</tr>
<tr>
<td>p2</td>
<td>F-150</td>
<td>MA</td>
<td>$250</td>
<td>...</td>
<td>(0.9, 0.1)</td>
</tr>
<tr>
<td>p3</td>
<td>F-150</td>
<td>CA</td>
<td>$150</td>
<td>...</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p4</td>
<td>Sierra</td>
<td>TX</td>
<td>$300</td>
<td>...</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p5</td>
<td>Camry</td>
<td>TX</td>
<td>$325</td>
<td>...</td>
<td>(0.7, 0.3)</td>
</tr>
<tr>
<td>p6</td>
<td>Camry</td>
<td>TX</td>
<td>$175</td>
<td>...</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p7</td>
<td>Civic</td>
<td>TX</td>
<td>$225</td>
<td>...</td>
<td>(0.3, 0.7)</td>
</tr>
<tr>
<td>p8</td>
<td>Civic</td>
<td>TX</td>
<td>$120</td>
<td>...</td>
<td>(0.2, 0.8)</td>
</tr>
<tr>
<td>p9</td>
<td>F150</td>
<td>East</td>
<td>$140</td>
<td>...</td>
<td>(0.5, 0.5)</td>
</tr>
<tr>
<td>p10</td>
<td>Truck</td>
<td>TX</td>
<td>$500</td>
<td>...</td>
<td>(0.9, 0.1)</td>
</tr>
</tbody>
</table>

Burdick et al. VLDB’05

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
## Uncertainty in Data Integration

- **Schema 1**: `(pname, email-addr, permanent-addr, current-addr)`
- **Schema 2**: `(name, email, mailing-addr, home-addr, office-addr)`
- **How to map the two schemas?**

<table>
<thead>
<tr>
<th>Possible Mapping</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_1 = (pname, name), (email-addr, email), (current-addr, mailing-addr), (permanent-addr, home-address) )</td>
<td>0.5</td>
</tr>
<tr>
<td>( m_2 = (pname, name), (email-addr, email), (permanent-addr, mailing-addr), (current-addr, home-address) )</td>
<td>0.4</td>
</tr>
<tr>
<td>( m_3 = (pname, name), (email-addr, mailing-addr), (current-addr, home-addr) )</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Dong et al. VLDB’07

---

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin

Ambiguous Entities

- Entity identification is a challenging task.
Disguised Missing Data

Information about "State" is missing "Alabama" is used as the disguise

Hua and Pei KDD’07

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Disguised Missing Data

- Disguised missing data is the missing data entries that are not explicitly represented as such, but instead appear as potentially valid data values.
- Disguised missing data also introduces uncertainty.
Why Uncertain Data Is Still Useful?

• For a temperature sensor, suppose the difference between the real temperature and the sensed temperature follows normal distribution

• The real temperature can be modeled by a probability density function

• What is the real temperature? Uncertain

• What is the probability that the real temperature is over 50°C? Certain!

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertainty and Confidence

• Uncertain data can provide probabilistic answers to aggregate questions
  – How can we estimate the percentage of married voters supporting Obama from survey data?
  – What is the total repair cost for F150’s in the East?

• An answer derived from uncertain data may often be a function on probability or confidence
Reducing Uncertainty

• Removing uncertain entries
  – Removing uncertain attribute values
  – Removing uncertain records
  – Cons: reducing available data

• Generalization
  – Remove attribute city if some entries on the attribute is uncertain
  – Can accurately answer questions at level city or above
  – Still cannot answer questions at level city or below
Being Certain or Uncertain?

• Answering questions on uncertain data in general can be more complicated
  – Probability is a new (and often difficult) dimension

• Simplifying uncertain data to certain data may not use the full potential of data
  – Many details may be lost

• Probabilistic answers on uncertain data are often interesting and useful

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertain Data Analysis Framework

- Mining uncertain/probabilistic data
- Query answering on uncertain/probabilistic data
- Uncertainty assessment and estimation

Data sources:
- Sensor network
- Survey form

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertain Data Acquisition

- Statistics-based, model-driven approaches are often used
- Misrepresentations of data in sensor networks
  - Impossible to collect all relevant data – potentially infinite
  - Samples are non-uniform in time and space due to non-uniform placement of sensors in space, faulty sensors, high packet loss rates,

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Levels of Uncertainty

• Uncertainty can exist in object/tuple level and attribute level

• Object/tuple level uncertainty
  – An object/tuple takes a probability to appear (existing probability)

• Attribute level uncertainty
  – An attribute of an object/tuple takes a few possible values
Lecture on Uncertainty in Data Mining

- Introduction to uncertain data
- Representing uncertainty in a relational database
- Possible worlds
- U-relation
- Data mining algorithms with uncertainty
- Gaussian processes
Survey Data Example

Object 1: Smith
- Smith.SSN=785, 60%
- Smith.SSN=185, 40%

Object 2: Brown
- Brown.SSN=185, 50%
- Brown.SSN=186, 50%

Constraints:
“Smith.SSN=185” ⊕ “Brown.SSN=185”

Antova et al. ICDE’07

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin

May 19, 2016
Meelis Kull, University of Tartu, MTAT.03.183 Data Mining
Survey Data Example

Social Security Number: 185
Name: Smith
Marital Status: (1) single ☑ (2) married ☐ (3) divorced ☑ (4) widowed ☐

Social Security Number: 185
Name: Brown
Marital Status: (1) single ☐ (2) married ☐ (3) divorced ☐ (4) widowed ☐

<table>
<thead>
<tr>
<th>TID</th>
<th>Name</th>
<th>SSN</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>Smith</td>
<td>185</td>
<td>40%</td>
</tr>
<tr>
<td>t2</td>
<td>Smith</td>
<td>785</td>
<td>60%</td>
</tr>
<tr>
<td>t3</td>
<td>Brown</td>
<td>185</td>
<td>50%</td>
</tr>
<tr>
<td>t4</td>
<td>Brown</td>
<td>186</td>
<td>50%</td>
</tr>
</tbody>
</table>

Generation rules:
t1 ⊕ t2, (constraints)
t3 ⊕ t4,
t1 ⊕ t3

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Uncertainty of Mobile Objects

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Prob Table vs. Uncertain Objects

• A probabilistic table can be represented as a set of uncertain objects
  – All tuples in a generation rule are modeled as an uncertain object
  – Use NULL instances to make the sum of membership probabilities in one object to 1
• Uncertain objects with discrete instances can be represented using a probabilistic table
  – One record per instance
  – All instances of an object are constrained by one generation rule
  – Uncertain objects with continuous probability density functions cannot be represented using a finite probabilistic table
• More complicated constraints may not be captured in the transformation
Probrabilistic Database Model

Speed of cars detected by radar

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Location</th>
<th>Car make</th>
<th>Plate No.</th>
<th>Speed</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
</tr>
</tbody>
</table>

Generation rules: (t2⊕t3), (t4⊕t5)

- The values of each tuple are certain
- Each tuple carries an existing/membership probability
- Generation rules: constraints specifying exclusive tuples

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin

May 19, 2016
Meelis Kull, University of Tartu, MTAT.03.183 Data Mining
Prob Table vs. Uncertain Objects

A probabilistic table

A set of uncertain objects

A tuple

An instance

A generation rule

An uncertain object

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
</tr>
</tbody>
</table>

Rules: (t2 \oplus t3), (t4 \oplus t5)

This slide has been adapted from the KDD'08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin

May 19, 2016
Meelis Kull, University of Tartu, MTAT.03.183 Data Mining
Lecture on Uncertainty in Data Mining

- Introduction to uncertain data
- Representing uncertainty in a relational database
  - Possible worlds
  - U-relation
- Data mining algorithms with uncertainty
- Gaussian processes
Possible Worlds

• A possible world
  – a possible snapshot that may be observed

• Probabilistic database model
  – A possible world = a set of tuples
  – At most one tuple per generation rule in a possible world

• Uncertain object model
  – A possible world = a set of instances of uncertain objects
  – At most one instance per object in a possible world

• A possible world carries an existence probability
An Example of Possible Worlds

0.4 = 0.112 + 0.168 + 0.048 + 0.072

### A probabilistic table

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>11:45</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
</tr>
<tr>
<td>t2</td>
<td>11:50</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
</tr>
<tr>
<td>t3</td>
<td>11:35</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
</tr>
<tr>
<td>t4</td>
<td>12:10</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
</tr>
<tr>
<td>t5</td>
<td>12:25</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
</tr>
<tr>
<td>t6</td>
<td>12:15</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
</tr>
</tbody>
</table>

### Possible worlds

<table>
<thead>
<tr>
<th>World</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW₁={t1,t2,t6,t4}</td>
<td>0.112</td>
</tr>
<tr>
<td>PW₂={t1,t2,t5,t6}</td>
<td>0.168</td>
</tr>
<tr>
<td>PW₃={t1,t6,t4,t3}</td>
<td>0.048</td>
</tr>
<tr>
<td>PW₄={t1,t5,t6,t3}</td>
<td>0.072</td>
</tr>
<tr>
<td>PW₅={t2,t6,t4}</td>
<td>0.168</td>
</tr>
<tr>
<td>PW₆={t2,t5,t6}</td>
<td>0.252</td>
</tr>
<tr>
<td>PW₇={t6,t4,t3}</td>
<td>0.072</td>
</tr>
<tr>
<td>PW₈={t5,t6,t3}</td>
<td>0.108</td>
</tr>
</tbody>
</table>

### Rules:

(t₂ ⊕ t₃), (t₄ ⊕ t₅)

T2 and T3 never appear in the same possible world!

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin


May 19, 2016

Meelis Kull, University of Tartu, MTAT.03.183 Data Mining
Possible Worlds and Rules

- Possible worlds are governed by rules

<table>
<thead>
<tr>
<th>Time</th>
<th>Radar Loc</th>
<th>Car Make</th>
<th>Plate No</th>
<th>Speed</th>
<th>Conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>L1</td>
<td>Honda</td>
<td>X-123</td>
<td>130</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>L2</td>
<td>Toyota</td>
<td>Y-245</td>
<td>120</td>
<td>0.7</td>
</tr>
<tr>
<td>t3</td>
<td>L3</td>
<td>Toyota</td>
<td>Y-245</td>
<td>80</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>L4</td>
<td>Mazda</td>
<td>W-541</td>
<td>90</td>
<td>0.4</td>
</tr>
<tr>
<td>t5</td>
<td>L5</td>
<td>Mazda</td>
<td>W-541</td>
<td>110</td>
<td>0.6</td>
</tr>
<tr>
<td>t6</td>
<td>L6</td>
<td>Nissan</td>
<td>L-105</td>
<td>105</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Worlds**

- PW¹ = {t1, t2, t6, t4} 0.16
- PW² = {t1, t2, t5, t6} 0.24
- PW³ = {t2, t6, t4} 0.12
- PW⁴ = {t2, t5, t6} 0.18
- PW⁵ = {t6, t4, t3} 0.12
- PW⁶ = {t5, t6, t3} 0.18

*Rules*: \((t2 \oplus t3), \ (t4 \oplus t5) \ (t1 \rightarrow t2)\)

A new rule

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Correlation and Dependencies

• An example of correlated tuples

<table>
<thead>
<tr>
<th>TID</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>0.4</td>
</tr>
<tr>
<td>t2</td>
<td>0.42</td>
</tr>
<tr>
<td>t3</td>
<td>0.468</td>
</tr>
</tbody>
</table>

A probabilistic table

• Factored representations

\[
\Pr(t1 = x1, t2 = x2, t3 = x3) = f_1(t1 = x1) f_{12}(t1 = x1, t2 = x2) f_{23}(t2 = x2, t3 = x3)
\]

Dependencies among tuples

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Correlation and Dependencies

• An example of correlated tuples

This is a Bayesian network
More general term: Graphical model

Dependencies among tuples

This slide has been adapted from the KDD'08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
### Possible Worlds

- Compute the joint probability of possible world assignments (Details in [Sen and Deshpande, ICDE’07])

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
<th>( \text{Pr}(t_1,t_2,t_3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.378</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.162</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.018</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.042</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.028</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.012</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.108</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.252</td>
</tr>
</tbody>
</table>

Joint probability of \((t_1,t_2,t_3)\)

<table>
<thead>
<tr>
<th>World</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW1=(\emptyset)</td>
<td>0.378</td>
</tr>
<tr>
<td>PW2={t_3}</td>
<td>0.162</td>
</tr>
<tr>
<td>PW3={t_2}</td>
<td>0.018</td>
</tr>
<tr>
<td>PW4={t_2,t_3}</td>
<td>0.042</td>
</tr>
<tr>
<td>PW5={t_1}</td>
<td>0.028</td>
</tr>
<tr>
<td>PW6={t_1,t_3}</td>
<td>0.012</td>
</tr>
<tr>
<td>PW7={t_1,t_2}</td>
<td>0.108</td>
</tr>
<tr>
<td>PW8={t_1,t_2,t_3}</td>
<td>0.252</td>
</tr>
</tbody>
</table>

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Infer.net

- A language to express probabilistic facts and perform probabilistic inference
- In other words, it can calculate probabilities
- Discrete & continuous

http://research.microsoft.com/infernet

```csharp
// Model
open MicrosoftResearch.Infer.Fun.FSharp.Syntax

// Model
let coins () =
    let c1 = random (Bernoulli(0.5))
    let c2 = random (Bernoulli(0.5))
    let bothHeads = c1 && c2
    observe (bothHeads = false)
    c1, c2, bothHeads

// Sampling

// Sampling does not take observations into account.
printf "Sample: %O\n" (coins ()

// Inference

open MicrosoftResearch.Infer.Fun.FSharp.Inference

let (c1D,c2D,bothD) = inferFun3 <> coins ()
printf "coinsD: \n%O\n%O\n%O\n" c1D c2D bothD
```
Conceptual Query Answering

Probabilistic database → Query Q → Result probabilistic database

Expand → Possible worlds

Summarize → Possible worlds

Adapted from Singh et al. ICDE’08

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Lecture on Uncertainty in Data Mining

• Introduction to uncertain data
• Representing uncertainty in a relational database
• Possible worlds
• U-relation
• Data mining algorithms with uncertainty
• Gaussian processes
Attribute Level Uncertainty

- An aerial photograph of a battlefield

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Attribute Level Uncertainty

• A relation R(ID, Type, Faction) with uncertain attributes
  – ID = { 1, 2, 3, 4 }
  – Type = { Tank, Transport }
  – Faction = { Friend, Enemy }

• Uncertainty in data
  – Vehicle 1 is a friendly tank a
  – Vehicle 2 and 3 are either
    • a friendly transport b, or
    • an enemy tank c
  – Vehicle 4 is unknown vehicle d

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Representing Uncertainty

- ID of vehicle b and c
  - “b’s ID is 2 and c’s ID is 3”, or “b’s ID is 3 and c’s ID is 2”?  
  - Random variable $x=\{1, 2\}$

- Type of Vehicle d
  - “Tank” or “Transport”? 
  - Random variable $y=\{1, 2\}$

- Faction of Vehicle d
  - “Friend” or “Enemy”? 
  - Random variable $z=\{1, 2\}$

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin  
U-Relation

- Vertical Representation
  - Use a U-relation to represent each attribute of relation R

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>x=1</td>
<td>a</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>3</td>
</tr>
<tr>
<td>x=2</td>
<td>b</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>d</td>
<td>4</td>
</tr>
</tbody>
</table>

U-relation for “ID”

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>Tank</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Transport</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Tank</td>
</tr>
<tr>
<td>y=1</td>
<td>d</td>
<td>Tank</td>
</tr>
<tr>
<td>y=2</td>
<td>d</td>
<td>Transport</td>
</tr>
</tbody>
</table>

U-relation for “Type”

<table>
<thead>
<tr>
<th>D</th>
<th>Vehicle</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>Friend</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>Friend</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>Enemy</td>
</tr>
<tr>
<td>z=1</td>
<td>d</td>
<td>Friend</td>
</tr>
<tr>
<td>z=2</td>
<td>d</td>
<td>Enemy</td>
</tr>
</tbody>
</table>

U-relation for “Faction”

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
### Possible Worlds of U-Relations

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>World</th>
<th>d.Type (y)</th>
<th>d.Faction(z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>b.ID=2, c.ID=3</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>b.ID=2, c.ID=3</td>
<td>Tank</td>
<td>Enemy</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>b.ID=2, c.ID=3</td>
<td>Transport</td>
<td>Friend</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>b.ID=2, c.ID=3</td>
<td>Transport</td>
<td>Enemy</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>b.ID=3, c.ID=2</td>
<td>Tank</td>
<td>Friend</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>b.ID=3, c.ID=2</td>
<td>Tank</td>
<td>Enemy</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>b.ID=3, c.ID=2</td>
<td>Transport</td>
<td>Friend</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>b.ID=3, c.ID=2</td>
<td>Transport</td>
<td>Enemy</td>
<td></td>
</tr>
</tbody>
</table>

**Possible worlds**

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Transformation of U-Relation

- U-Relations can be transformed to a probabilistic table

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
</tr>
<tr>
<td>b</td>
<td>?</td>
<td>Transport</td>
<td>Friend</td>
</tr>
<tr>
<td>c</td>
<td>?</td>
<td>Tank</td>
<td>Enemy</td>
</tr>
<tr>
<td>d</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

b.ID=2, c.ID=3 (30%)
b.ID=3, c.ID=2 (70%)
d.Type=Tank(50%), Transport(50%)
d.Faction=Friend (50%), Enemy(50%)

<table>
<thead>
<tr>
<th>TID</th>
<th>Vehicle</th>
<th>ID</th>
<th>Type</th>
<th>Faction</th>
<th>Conf.</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>a</td>
<td>1</td>
<td>Tank</td>
<td>Friend</td>
<td>1</td>
</tr>
<tr>
<td>t2</td>
<td>b</td>
<td>2</td>
<td>Transport</td>
<td>Friend</td>
<td>0.3</td>
</tr>
<tr>
<td>t3</td>
<td>c</td>
<td>3</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.3</td>
</tr>
<tr>
<td>t4</td>
<td>b</td>
<td>3</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.7</td>
</tr>
<tr>
<td>t5</td>
<td>c</td>
<td>2</td>
<td>Transport</td>
<td>Friend</td>
<td>0.7</td>
</tr>
<tr>
<td>t6</td>
<td>d</td>
<td>4</td>
<td>Tank</td>
<td>Friend</td>
<td>0.25</td>
</tr>
<tr>
<td>t7</td>
<td>d</td>
<td>4</td>
<td>Tank</td>
<td>Enemy</td>
<td>0.25</td>
</tr>
<tr>
<td>t8</td>
<td>d</td>
<td>4</td>
<td>Transport</td>
<td>Friend</td>
<td>0.25</td>
</tr>
<tr>
<td>t9</td>
<td>d</td>
<td>4</td>
<td>Transport</td>
<td>Enemy</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Generation rules: t2→t3, t4→t5, t2⊕t4, t3⊕t5, t6⊕t7⊕t8⊕t9

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
Lecture on Uncertainty in Data Mining

- Introduction to uncertain data
- Representing uncertainty in a relational database
- Possible worlds
- U-relation
- Data mining algorithms with uncertainty
- Gaussian processes
Data mining algorithms where uncertainty can be introduced

- Apriori (frequent itemset mining)
- KNN (classification)
- K-means (clustering)
- ... many others
Probabilistic Transactions

- A transaction t contains a number items where each item x is associated with a positive probability \( P_t(x) \)
  - Assuming items in a transaction are independent
  - Itemset xyz has probability \( P_t(x)P_t(y)P_t(z) \) to happen in t

- In a probabilistic transaction database D of d transactions, an itemset X is frequent if its expected support is at least \( \rho d \), where \( \rho \) is a user-specified support threshold
  - [Chui et al., PAKDD’07]
Possible Worlds of Transactions

- Enumerating all possible worlds to compute the expected supports is computationally infeasible for large transaction databases
Independent Transactions

• If transactions are independent, expected support can be calculated efficiently transaction by transaction

\[ S_e(X) = \sum_{j=1}^{d} \prod_{x \in X} P_{t_j}(x) \]

• Anti-monotonicity still holds: if \( X \) is infrequent, then every super set of \( X \) cannot be frequent

• U-Apriori: extending Apriori straightforwardly

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
1NN Classification on Certain Data

- Point x will be classified using point y since $\text{dist}(x, y) < \text{dist}(x, z)$

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yufei Tao, Xuemin Lin
1NN on Uncertain Data

- Object $x$ may have a good chance to be classified using $z$
  - Instances of $x$ have a high probability to lie in the error boundary of $z$
- When classification on uncertain data, it is important to use the relative errors of different data points over the different dimensions in order to improve the accuracy

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yafei Tao, Xuemin Lin
Fuzzy Clustering

- Each data point is certain
- Clusters are fuzzy (uncertain to some extent)
  - No sharp boundary between clusters, often perform better in some applications
  - Each point is assigned to a cluster with a probability (membership degree)
Clustering Uncertain Objects

• Objects are fuzzy/uncertain, clusters can be certain or fuzzy
  – A fuzzy object can be represented by a probability density function or a set of instances
  – All instances of an object are in the same space, different objects may have a different number of instances

• In clustering, the distribution of the distance between two objects and the probability that an object is a cluster center should be considered

\[
\Pr[a \leq \text{dist}(o, o') \leq b] = \int_{a}^{b} \Pr[\text{dist}(o, o') = x] \, dx
\]

Kriegel and Pfeifle, KDD’05, ICDM’05

This slide has been adapted from the KDD’08 tutorial by Jian Pei, Ming Hua, Yifei Tao, Xuemin Lin
K-means on Uncertain Data

- Run k-means, use expectation of distance to assign objects/probabilistic points to clusters
- Computation can be sped up by using bounding rectangles or other polygon to bound PDF regions and approximate distance calculation
Lecture on Uncertainty in Data Mining

• Introduction to uncertain data
• Representing uncertainty in a relational database
• Possible worlds
• U-relation
• Data mining algorithms with uncertainty
• Gaussian processes
What about continuous variables?

• Continuous attributes?
  – Instead of a small number of different possible values we have infinitely many
  – Can be represented as a probability distribution function (PDF)
Continuous Uncertain Model

- An attribute may take a continuous PDF as the value
- A table \( T = (\Sigma_T, \Delta_T) \)
  - \( \Sigma_T \): a relational schema
  - \( \Delta_T \): dependency information including pdfs or joint pdfs
  - For each dependent group of uncertain attributes, store history \( \Lambda \). When a new tuple is added, check whether the dependency remains

<table>
<thead>
<tr>
<th>Car-id</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Gaussian(mean 18, variance 6)</td>
</tr>
<tr>
<td>C2</td>
<td>Uniform(center (32, 26), radius 7)</td>
</tr>
</tbody>
</table>
What about continuous variables?

• Continuous attributes?
  – Instead of a small number of different possible values we have infinitely many
  – Can be represented as a probability distribution function (PDF)

• Regression?
  – Gaussian processes (GPs)
    • the rest of the lecture
Why GPs?

• Here are some data points! What function did they come from?

• I have no idea.

• Oh. Okay. Uh, you think this point is likely in the function, too?

• I still have no idea.

This slide has been adapted from the presentation by Barnabas Poczos
Why GPs?

Under certain assumptions GPs can answer the following questions

• Here are some data points, and here’s how I rank the likelihood of functions.

• Here’s where the function will most likely be. (expected function)

• Here are some examples of what it might look like. (sampling from the posterior distribution)

• Here is a prediction of what you’ll see if you evaluate your function at $x'$, with confidence

This slide has been adapted from the presentation by Barnabas Poczos
1D Gaussian Distribution

- Parameters
  - Mean, $\mu$
  - Variance, $\sigma^2$

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

This slide has been adapted from the presentation by Barnabas Poczos
Multivariate Gaussian

\[ P(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{ \frac{-1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \]

This slide has been adapted from the presentation by Barnabas Poczos
The Multivariate Gaussian

- A 2-dimensional Gaussian is defined by
  - a mean vector \( \mu = [\mu_1, \mu_2] \)
  - a covariance matrix: \( \Sigma = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{2,1}^2 \\ \sigma_{1,2}^2 & \sigma_{2,2}^2 \end{bmatrix} \)

where \( \sigma_{ij}^2 = \mathbb{E}[ (x_i - \mu_i)(x_j - \mu_j) ] \) is (co)variance

- Note: \( \Sigma \) is symmetric,
  "positive semi-definite": \( \forall x: x^T \Sigma x \geq 0 \)
Multivariate Gaussian examples

\[ \mu = (0,0) \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \]
Visualizing Draws from GPs

• Visualizing draws from 2-D Gaussian:

• Three draws from a 6-D Gaussian:

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Visualizing Draws from GPs

- Three draws from 25-D Gaussian

To generate these, the mean was set to zero: zeros(25,1)

The covariance was set using a covariance function: $\sum_{nm} = k(x_n, x_m)$.

The x’s are the positions that are planted the tics on the axis.

We can visualize draws from a GP iterative sampling $f(x_n) \mid f(x_1),...,f(x_{n-1})$ on a sequence of input points $x_1, x_2, \ldots x_n$.

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Samples from GPs

Squared-exponential kernel

\[ k(x_n, x_m) = \exp \left( -\frac{\theta}{2} (x_n - x_m)^2 \right) \]

Exponential kernel

\[ k(x_n, x_m) = \exp \left( -\theta |x_n - x_m| \right) \]

- Ornstein-Uhlenbeck process that describes Brownian motion.

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Covariance Function

- One widely used covariance (kernel) function for GP regression is given by the squared-exponential plus constant and linear terms:

\[ k(x_n, x_m) = \theta_0 \exp \left( -\frac{\theta_1}{2} \|x_n - x_m\|^2 \right) + \theta_2 + \theta_3 x_n^T x_m \]

- Note that the last term corresponds to a parametric model that is a linear function of the input variables.

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Covariance Function

- One widely used covariance (kernel) function for GP regression is given by the squared-exponential plus constant and linear terms:

\[ k(x_n, x_m) = \theta_0 \exp\left( -\frac{\theta_1}{2} ||x_n - x_m||^2 \right) + \theta_2 + \theta_3 x_n^T x_m \]

- Note that the last term corresponds to a parametric model that is a linear function of the input variables.

\[ \theta_2 = 0, \theta_3 = 0 \quad \text{(1.00, 4.00, 0.00, 0.00)} \]
\[ \theta_2 = 10, \theta_3 = 0 \quad \text{(1.00, 4.00, 10.00, 0.00)} \]
\[ \theta_2 = 0, \theta_3 = 5 \quad \text{(1.00, 4.00, 0.00, 5.00)} \]

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Why stop there?

- We indexed before with $\mathbb{R}$, why not with $\mathbb{R}^D$?
- Need functions $\mu(x), k(x, z), \forall x, z \in \mathbb{R}^D$
Gaussian Process

Definition:

• Probability distribution *indexed by* an arbitrary set

• Each element gets a Gaussian distribution over the reals with mean $\mu(x)$

• These distributions are dependent/correlated as defined by $k(x,z)$

• Any finite subset of indices defines a multivariate Gaussian distribution
  • Crazy mathematical statistics and measure theory ensures this

This slide has been adapted from the presentation by Barnabas Poczos
Gaussian Process

• Distribution over functions

• Domain of the functions (index set) can be pretty much whatever
  • Reals
  • Real vectors
  • Graphs
  • Strings
  • Sets
  • …

• Most interesting structure is in $k(x, z)$, the ‘kernel.’

This slide has been adapted from the presentation by Barnabas Poczos
Bayesian Updates for GPs

• How do Bayesians use a Gaussian Process?
  • Start with GP prior
  • Get some data
  • Compute a posterior

• Ask interesting questions about the posterior

This slide has been adapted from the presentation by Barnabas Poczos
Samples from the prior distribution

This slide has been adapted from the presentation by Barnabas Poczos
Samples from the posterior distribution

Picture is taken from Rasmussen and Williams

This slide has been adapted from the presentation by Barnabas Poczos
• Illustration of GP regression applied to the sinusoidal data set.

• The green curve shows the true function.

• The blue data points are samples from the true function plus some additive Gaussian noise.

• The red curve shows the mean of the GP predictive distribution, with shaded region corresponding to +/- 2 standard deviations.

This slide has been adapted from the presentation by Ruslan Salakhutdinov
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^T K^{-1} y}{2} \]
Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^T K^{-1} y}{2} \]
Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[
E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2}
\]

This slide has been adapted from the presentation by Trevor Cohn and Daniel Beck
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^T K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]

This slide has been adapted from the presentation by Trevor Cohn and Daniel Beck
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^TK^{-1}y}{2} \]

This slide has been adapted from the presentation by Trevor Cohn and Daniel Beck
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Summary of Gaussian processes

• GPs are infinite-dimensional Gaussian distributions
• GPs can be used to solve regression tasks
• Fitting a GP requires specification of mean and covariance functions (kernel)
  — This choice matters, otherwise can over- or under-fit
• GPs are universal function approximators
  — It means they can approximate any smooth function
Summary of the Lecture

• Taking uncertainty into account
  – Requires analysis to identify uncertainties
  – Requires techniques to be stored in the data
  – Allows richer queries
  – Provides richer descriptions
  – Provides better predictions
  – Is computationally more expensive

• Uncertainty is (almost) everywhere!!!
  – We have only touched very few aspects
Lecture on Uncertainty in Data Mining

• Introduction to uncertain data
• Representing uncertainty in a relational database
• Possible worlds
• U-relation
• Data mining algorithms with uncertainty
• Gaussian processes

THANK YOU!
QUESTIONS?

Bach, Master, PhD thesis topics available, ask me!
meelis.kull@ut.ee