Statistical Relational Models and Learning

150SRL: Spring 2009, Tufts University
Instructor: Roni Khardon

What the course is about:

- World has objects and relations among them
- Rules about behavior in world are probabilistic
- Want to model both relational aspect and probabilities
- Useful to understand domain and for task (e.g., prediction)

SRL: Some Applications

- University: professors, students, courses, projects, registration, grades (predict success?)
- Citation database: articles, authors, topics, venues, citations (predict topics; predict citations; object identity)
- Movie database: movies, actors, producers, directors, showing location, user ratings, profits: (predict user rating; predict profit)
- Financial database: brokers, customers, info disclosures (predict success; predict fraud)

SRL: Course Info

- See Web Page
- Same model as 150AML Spring 2007 (Learning planning and acting . . .)

Probabilities and Bayesian Inference

This unit gives basic notions of probability theory, and our framework of working in probability spaces defined by a joint distribution over a number of random variables.

Slides use material from [RN] text and slides.

Probabilities: Example and Definitions

- We throw 2 dice (each with 6 sides and uniform construction).
- The result of the experiment is a pair of numbers \((a, b)\).
  Each such outcome is called an elementary event.
- The sample space is the set of elementary events.
  In our case it is \(\{(1, 1), (1, 2), \ldots, (6, 6)\}\).
- An event is a subset of the sample space. For example:
  Event \(F1\): outcomes where the first die has value 1.
  Event \(SE\): outcomes where the sum of the two numbers is even.
  Event \(S11\): outcomes where the sum of the two numbers is at least 11.
• Events $A$ and $B$ are Mutually Exclusive if $A \cap B = \emptyset$ (the empty set).
• All pairs of elementary events are mutually exclusive.
• Events $F_1$ and $S_{11}$ are mutually exclusive.

Probability Distributions

A Probability Distribution on the sample space $S$ is a mapping from events to real numbers that satisfies the axioms of probability.

1. $\Pr\{A\} \geq 0$ for any event $A$
2. $\Pr\{S\} = 1$
3. If $A$ and $B$ are mutually exclusive then $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\}$

Properties of Probabilities

- $\Pr\{\emptyset\} = 0$
- $A \subseteq B$ implies $\Pr\{A\} \leq \Pr\{B\}$
- For $\overline{A} = S \setminus A$, the complement of $A$, $\Pr\{\overline{A}\} = 1 - \Pr\{A\}$.

Event $\overline{F_1}$: outcomes where the first die has value not equal to 1.
$\Pr\{\overline{F_1}\} = 1 - (1/6) = 5/6$

Conditional Probability

- If $\Pr\{B\} \neq 0$ then the probability of $A$ given $B$ is $\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$
- $\Pr\{F_1 \cap S_{11}\} = \Pr\{(1,1)\} + \Pr\{(1,3)\} + \Pr\{(1,5)\} = 1/12$
- $\Pr\{S_{11}\} = \Pr\{(6,6)\} = 1/36$
- $\Pr\{F_1 \cap S_{11}\} = \Pr\{F_1\} = \Pr\{\overline{F_1}\} = 0$
- $\Pr\{F_1 \cap S_{11}\} = \Pr\{F_1\} = \Pr\{\overline{F_1}\} = 0$
- $\Pr\{SE\} = \Pr\{SE\} = \frac{1/12}{1/12} = 1/6$
- $\Pr\{SE\} = \Pr\{SE\} = \frac{1/12}{1/12} = 1/3$
- $\Pr\{E\} = \Pr\{E\} = \frac{1/12}{1/12} = 0$

Bayes' Theorem

- $\Pr\{A \cap B\} = \Pr\{B\} \Pr\{A|B\}$
- Reorganizing we get:
  $\Pr\{A|B\} = \frac{\Pr\{A\} \Pr\{B|A\}}{\Pr\{B\}}$
- $\Pr\{S_{11}\} = \Pr\{S_{11}\} = \frac{1/12(1/6)}{1/6} = 1/2$
- This will be the basis of our Bayesian inference and learning procedures!
Statistical Independence

- Events $A$ and $B$ are statistically independent iff
  \[ \Pr\{A \cap B\} = \Pr\{A\}\Pr\{B\} \]
- This is equivalent to the condition $\Pr\{A|B\} = \Pr\{A\}$
- Events $F1$ and $SE$ are statistically independent
  Events $SE$ and $S11$ are not
- Independence can be used to simplify computations!

Product Probability Spaces

- The setting we will look at will normally have a set of random variables $X_1, \ldots, X_n$.
- Each variable $X_i$ ranges over a finite set of values $v_{1,i}, \ldots, v_{k,i}$.
- An elementary event is an assignment of values to all variables.
- In our example, we have two variables, where $X_1$ is the value of the first die and $X_2$ of the second. The values in both cases are $1, \ldots, 6$.
- We can write a big table with $n$ columns and $\prod_{i} v_{i,k}$ rows describing the probability of every elementary event.
- In our example we have 2 columns and 36 rows.

The Joint Distribution

<table>
<thead>
<tr>
<th>Cold</th>
<th>Cat</th>
<th>Allergy</th>
<th>Sneez</th>
<th>(\Pr{})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.84645</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.004455</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.010095</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.0018</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.006465</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.008672</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0002475</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.000094</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.00072</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.001725</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.000002</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.00018</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.000004</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.00076</td>
</tr>
</tbody>
</table>

Marginal Distribution

- Given the joint distribution we can get an induced probability distribution for any one variable (or any subset of them).
- This is the set of probabilities we get if we ignore the other variables.
- $\Pr(X_1)$ is the distribution of values for the first die ignoring the second. It can be described in a table with 1 column and 6 rows (one for each value).
- We describe this generically as:
  \[ \Pr(X_1) = \sum \Pr\{X_1 = v_{1,i}\} \]
  This gives a probability distribution over $X_1$.
For particular values:
\[ \Pr \{ X_1 = v_{1,j} \} = \sum_{v_2} \Pr \{ X_1 = v_{1,j} \text{ and } X_2 = v_{2,n} \} \]
\[ \Pr \{ X_1 = 1 \} = \sum_{v_1} \Pr \{ X_1 = 1 \text{ and } X_2 = v_1 \} = 1/6 \]
\[ \Pr \{ \text{Cold} = 0 \} = \sum_{v_1} \Pr \{ \text{Cold} = 0 \text{ and } \text{Sneeze} = v_1 \} \]
\[ \Pr \{ \text{Cold} = 1 \} = \sum_{v_1} \Pr \{ \text{Cold} = 1 \text{ and } \text{Sneeze} = v_1 \} = 1 \]
\[ \Pr \{ \text{Sneeze} = 1 \} = \sum_{v_1} \Pr \{ \text{Sneeze} = 1 \text{ and } \text{Allergy} = v_1 \} \]

### Inference using the Joint
- **Compute probabilities of events**
  \[ \Pr \{ \text{Cold} = 1 \text{ and } \text{Sneeze} = 1 \} = \sum \ldots = 0.0902875 \]
- **Causal Inference**
  \[ \Pr \{ \text{Sneeze} = 1 | \text{Cold} = 1 \} = 0.0902875 = 0.902875 \]
- **Diagnostic Inference**
  \[ \Pr \{ \text{Cold} = 1 | \text{Sneeze} = 1 \} = 0.902875 = 0.66 \]
- **Inter-Causal Inference**
  \[ \Pr \{ \text{Cold} = 1 | \text{Sneeze} = 1 \text{ and } \text{Allergy} = 1 \} \]

### Normalization
- \[ \Pr \{ \text{Cold} = 1 | \text{Sneeze} = 1 \} = \frac{\Pr \{ \text{Sneeze} = 1 | \text{Cold} = 1 \} \Pr \{ \text{Cold} = 1 \}}{\Pr \{ \text{Sneeze} = 1 \}} = \frac{A}{2} \]
- \[ \Pr \{ \text{Cold} = 0 | \text{Sneeze} = 1 \} = \frac{\Pr \{ \text{Sneeze} = 1 | \text{Cold} = 0 \} \Pr \{ \text{Cold} = 0 \}}{\Pr \{ \text{Sneeze} = 1 \}} = \frac{B}{2} \]
- \[ \frac{A}{2} + \frac{B}{2} = 1 \text{ so } \alpha = A + B \text{ and} \]
- \[ \Pr \{ \text{Cold} = 1 | \text{Sneeze} = 1 \} = \frac{A}{A+B} \]
- Will often use normalization to simplify the computation.

### Conditional Independence
- Events \( A \) and \( B \) are statistically independent given event \( C \) iff
  \[ \Pr \{ A \cap B | C \} = \Pr \{ A | C \} \Pr \{ B | C \} \]
- This is equivalent to the condition \( \Pr \{ A \cap B \} = \Pr \{ A \} \)
- As in standard independence this can simplify the computations.
- Events \( \{ S \} \) outcomes where at least one die has value 6.
- \[ \Pr \{ \text{SE} | \text{S11} \} = 1/3 \]
- \[ \Pr \{ \text{SE} | \text{S11 and A6} \} = 1/3 \]

### Independence in Product Distributions
- In product distributions we can express a more general form of independence.
- \( X_1 \) and \( X_2 \) are independent iff \( \Pr \{ X_1 \cap X_2 \} = \Pr \{ X_1 \} \Pr \{ X_2 \} \)
- This means that for all \( v_1 \) and \( v_2 \)
  \[ \Pr \{ X_1 = v_1 \text{ and } X_2 = v_2 \} = \Pr \{ X_1 = v_1 \} \Pr \{ X_2 = v_2 \} \]
- And similarly for conditional independence
  \[ \Pr \{ X_1 \text{ and } X_2 \} = \Pr \{ X_1 | X_2 \} \Pr \{ X_2 \} \]
- \[ \Pr \{ X_1 = v_1 \text{ and } X_2 = v_2, \text{ and } X_3 = v_3 \} = \Pr \{ X_1 = v_1 | X_2 = v_2, X_3 = v_3 \} \Pr \{ X_2 = v_2, X_3 = v_3 \} \]

### Bayesian Networks
This unit introduces Bayesian networks and basic (ad hoc) inference using them.
Slides use material from [RN] text and slides.
Product Probability Spaces

- We have \( n \) variables, \( X_1, \ldots, X_n \).
- Each variable \( X_i \) ranges over a finite set of values \( v_{i,1}, \ldots, v_{i,k} \).
- We can write a big table with \( n \) columns and \( \prod_i k_i \) rows describing the probability of every elementary event.
- Table grows exponentially with \( n \).
  Not feasible unless \( n \) is very small.

Bayesian Networks

- Allow us to represent distributions more compactly.
- Take advantage of the structure available in a domain.
- Basic idea: represent dependence and independence explicitly.
- If \( \Pr(X_1 \cap X_2) = \Pr(X_1)\Pr(X_2) \) then we can use two 1-dimensional tables instead of a 2-dimensional table.
  - If each has 6 values, this means 12 entries instead of 36!

Example

- Edges represent "direct influence".
- Assume that \( \text{Cat} \) and \( \text{Cold} \) do not depend on other variables.
- \( \text{Allergy} \) depends only on \( \text{Cat} \).
- \( \text{Sneeze} \) depends on \( \text{Cold} \) and \( \text{Allergy} \).
- \( \text{Sneeze} \) depends on \( \text{Cat} \) BUT only through \( \text{Allergy} \).
  - For each node we associate a conditional probability table.

The network structure expresses independence of variables.

- \( \Pr(\text{Cat} | \text{Cold}) = \Pr(\text{Cat}) \)
- \( \Pr(\text{Allergy} | \text{Cat, Cold}) = \Pr(\text{Allergy} | \text{Cat}) \)
- \( \Pr(\text{Sneeze} | \text{Allergy, Cat, Cold}) = \Pr(\text{Sneeze} | \text{Allergy, Cold}) \)

More generally, the joint distribution can be expressed as the product of the distributions in the network:

\[
\Pr(\text{Cat, Cold, Allergy, Sneeze}) = \Pr(\text{Cat})\Pr(\text{Cold})\Pr(\text{Allergy} | \text{Cat})\Pr(\text{Sneeze} | \text{Allergy, Cold})
\]
Pr\{Cold=1, Cat=0, Allergy=0, Sneeze=1\} = Pr\{Cat=0\}Pr\{Cold=1\}Pr\{Allergy=0\} = 0.99 \cdot 0.1 \cdot 0.9 = 0.0891

Pr\{Sneeze=1\} = Pr\{Cold=1\}Pr\{Allergy=0\} = 0.99 \cdot 0.1 = 0.099

Pr\{Sneeze=1\} = Pr\{Cat=0\}Pr\{Cold=1\}Pr\{Allergy=0\} = 0.99 \cdot 0.1 \cdot 0.9 = 0.0891

So to represent a distribution we need to represent the network and CPTs.

If for all nodes the number of parents is small then all CPTs are small and we have a compact representation.

---

**d-Separation**

A set of nodes \( E \) d-separates \( X \) and \( Y \) iff every path is blocked. A path is blocked iff some \( z \) on path has one of 3 configurations:

1. \( z \in E \) and path goes through \( z \)
2. \( z \notin E \) and both arrows out of \( z \)
3. \( z \notin E \), children(\( z \)) \( \notin E \), and both arrows into \( z \)

**Theorem:** if \( E \) d-separates \( X, Y \) then \( X \) is independent of \( Y \) given \( E \).

---

**Conditional Independence**

Each node is conditionally independent of all others given its Markov blanket: parents + children + children’s parents.

---

**How to Construct a Network?**

- Can represent any distribution using a network.
- By repeated application of \( \Pr\{A, B\} = \Pr\{A|B\}\Pr\{B\} \)
- Choose ordering of variables \( X_1, \ldots, X_n \).
- For \( i = 1 \) to \( n \)
  - Add \( X_i \) to network with \( \Pr(X_i|X_1, \ldots, X_{i-1}) \).
- The joint distribution is:
  \( \Pr(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i|X_1, \ldots, X_{i-1}) \).
- But this is no improvement as \( X_n \) is connected to all predecessors (so we need a huge table for it).

---

**Example**

You are at work, neighbour John calls to say your home alarm is ringing, but neighbour Mary doesn’t call. Sometimes alarm set off by minor earthquakes. Is there a burglar?

Variables and ordering:

- Burglar, Earthquake, Alarm, JohnCalls, MaryCalls

Network topology reflects “causal” knowledge:

---

**Instead, at each stage choose a subset such that parents(\( X_i \)) \( \subseteq \{X_1, \ldots, X_{i-1}\} \).

- Choice of parents must satisfy \( \Pr(X_i|X_1, \ldots, X_{i-1}) = \Pr(X_i|\text{parents}(X_i)) \)
  - so that \( X_i \) is independent of other predecessors given its parents.
  - If parents(\( X_i \)) is small then representation is compact.
  - For example if for all \( X_i \), \( |\text{parents}(X_i)| \leq 3 \) then instead of \( 2^n \) entries we have \( n2^3 = 8n \) entries !

- How should we order the variables ?
  - Causal links tend to produce small representations.
  - Choose “root causes” first. Then continue with causal structure as much as possible.
What if we choose another ordering?

Variable ordering (1):
Mary Calls, John Calls, Alarm, Burglary, Earthquake

Variable ordering (2):
Mary Calls, John Calls, Earthquake, Burglary, Alarm,

How to Construct a Network

• May work with domain experts to decide on structure and then find values of probabilities.
• Choosing a "bad order" can have adverse effect: Large probability tables.
• "Unnatural" dependencies, that are in turn hard to estimate.
• Luckily, experts are often good at identifying causal structure, and prefer giving probability judgments for causal rules.

Compact conditional distributions

• Deterministic nodes can be represented explicitly.
• Other structure can be used: Noisy Or is common.

Parents \( U_1 \ldots U_k \) include all causes (can add leak node)
Independent failure probability \( q_i \) for each cause alone
\[
P(X|U_1, \ldots, U_j, \bar{U}_{j+1}, \ldots, U_k) = 1 - \prod_{i=1}^k q_i
\]
Using \( k \) parameters instead of \( 2^k \) parameters.

Real Valued Nodes

• Use parametric distributions for the nodes: Normal, Uniform
• With real parents: linear Gaussian is common:
\[
Y|X \sim \mathcal{N}(Ax + b, \Sigma)
\]
• If all nodes are Gaussian or linear Gaussian then the joint is Gaussian and inference is relatively easy.
• If parents are discrete can use Conditional Gaussian, that is Gaussian for each assignment of parent values.

Real Valued Nodes

• Discrete child of continuous parent: often use sigmoid (logit):
\[
P(\text{Buys} = \text{true}|\text{Cost} = c) = \frac{1}{1 + \exp(-c + \mu \sigma)}
\]
Computing with Bayes Nets

- We have seen how to reconstruct the joint distribution from the network.
- Can we compute other probabilities efficiently?
  \[
  \Pr\{\text{Cold} = 1 \text{ and } \text{Sneeze} = 1\} = ? \\
  \Pr\{\text{Sneeze} = 1 | \text{Cold} = 1\} = ? \\
  \Pr\{\text{Cold} = 1 | \text{Sneeze} = 1\} = ? \\
  \Pr\{\text{Cold} = 1 | \text{Sneeze} = 1 \text{ and } \text{Allergy} = 1\} = ?
  \]

Computing a Marginal Distribution

- To compute \(\Pr(\text{Cold}, \text{Sneeze})\)
- First compute \(\Pr(\text{Allergy})\) by summing \(\text{Cat}\) out of \(\Pr(\text{Allergy, Cat}) = \Pr(\text{Cat}) \Pr(\text{Allergy} | \text{Cat})\)
- Then compute \(\Pr(\text{Sneeze} | \text{Cold})\) by summing \(\text{Allergy}\) out of \(\Pr(\text{Sneeze, Allergy} | \text{Cold}) = \Pr(\text{Sneeze} | \text{Allergy, Cold}) \Pr(\text{Allergy})\)
- Finally compute
  \[
  \Pr(\text{Cold, Sneeze}) = \Pr(\text{Sneeze} | \text{Cold}) \Pr(\text{Cold})
  \]
- Can similarly compute \(\Pr\{\text{Cold} = 1 \text{ and } \text{Sneeze} = 1\}\)

In general, “sum out” variables that do not appear in the question. Try to maintain small tables along the way.

Causal Inference

- To compute \(\Pr(\text{Sneeze} = 1 | \text{Cold} = 1)\)
- First compute \(\Pr(\text{Allergy})\) as in previous example.
- Then compute
  \[
  \Pr(\text{Sneeze} = 1 | \text{Cold} = 1) = \sum_v \Pr(\text{Sneeze} = 1, \text{Allergy} = v, \text{Cold} = 1) \Pr(\text{Allergy} = v)
  \]

Diagnostic Inference

- Use Bayes’ Rule to compute \(\Pr(\text{Cold} = 1 | \text{Sneeze} = 1)\)
  \[
  A_1 = \Pr(\text{Cold} = 1 | \text{Sneeze} = 1) = \frac{\Pr(\text{Sneeze} = 1 | \text{Cold} = 1) \Pr(\text{Cold} = 1)}{\Pr(\text{Sneeze} = 1)} = \frac{N_1}{N_0 + N_1}
  \]
  \[
  A_0 = \Pr(\text{Cold} = 0 | \text{Sneeze} = 1) = \frac{\Pr(\text{Sneeze} = 1 | \text{Cold} = 0) \Pr(\text{Cold} = 0)}{\Pr(\text{Sneeze} = 1)} = \frac{N_0}{N_0 + N_1}
  \]
- But \(A_0 + A_1 = 1\)
- \(\Pr(\text{Cold} = 1 | \text{Sneeze} = 1) = \frac{N_1}{N_0 + N_1} = \frac{N_0 + N_1}{N_0 + N_1} \cdot \frac{N_1}{N_0 + N_1} = \frac{N_1}{N_0 + N_1}
  \]
- \(\text{Normalize}(N_0, N_1) = \frac{N_0 + N_1}{N_0 + N_1}
  \]

- From the CPT we have:
  \[
  \begin{align*}
  \Pr(\text{Cold} = 0) &= 0.90 \quad \text{and} \quad \Pr(\text{Cold} = 1) = 0.10 \\
  \Pr(\text{Sneeze} = 1 | \text{Cold} = 0) &= 0.05175 \\
  \Pr(\text{Sneeze} = 1 | \text{Cold} = 1) &= 0.902875
  \end{align*}
  \]
- computed as in previous example using \(\Pr(\text{Allergy})\)
  \[
  \begin{align*}
  N_0 &= 0.05175 \cdot 0.90 = 0.046575 \\
  N_1 &= 0.902875 \cdot 0.10 = 0.0902875 \\
  \Pr(\text{Cold} = 1 | \text{Sneeze} = 1) &= \frac{N_1}{N_0 + N_1} = 0.66
  \end{align*}
  \]
**Inference in Burglary Example**

- Use $B, E, A, J, M$ to denote variables
- Compute $\Pr\{A = 1 | E = 1, J = 1\}$
- By computing and normalizing

\[
\Pr\{J = 1 | A = 1, E = 1, J = 1\} \cdot \Pr\{A = 1 | E = 1\} = 0.001 \cdot 0.95 + 0.999 \cdot 0.29 = 0.29066
\]

**Poly-Tree Networks**

A singly connected network. There are no cycles even ignoring the direction of edges.

**Inference — Overview**

- In general the situation is more complicated than in our simple network. The problem is NP-Hard.
- Can always compute via the joint but this may not be efficient.
- Efficient Algorithms are known for graphs with poly-tree structure.
- Otherwise, try to turn graph into a tree, or use simulation methods to approximate the probability.
- Simulation is not guaranteed to give good answers but works well in some cases.
- Details next time: using [RN] slides and extra slides.

**Familiar Problems as Bayes Nets**

- Naive Bayes Algorithm
- Mixture of Gaussians
- Linear Regression
- Bayesian Linear Regression
- ...