Lecture 2.

Causal Reasoning

as an example of
Bayes Nets (DAG's)
Parameter Inference.
Model Selection.

Ocam's Razor

Introducing:
Noisy-or, Noisy-And,
Noisy-Logical probability model
Intervention.

Lectures:
Griffiths → Graphical Models & Human Causal Learning.
HongJing Lu → Model Estimation & Parameter Estimation in Causal Learning
Causal Models

Wu (Friday)

Causality: what is the causal effect of one variable on another?

Suppose we are given the joint statistics $P(A,B)$ of two variables. Can we tell which variable "caused" the other?

No, "correlation is not causation."

$$\frac{P(A,B)}{P(B|A)P(A)} = \frac{P(A|B)P(B)}{P(A)P(B)}$$

These are equivalent ways of expressing probability distributions.

In causal models, the arrow is used to determine the direction of causation.

E.g. $A$ - bomb put in city at 10 am
$B$ - explosion in city at 11 am

Clearly (laws of physics) $B \rightarrow A$ $B$ causes $A$
Causal Models

Goals: (i) clarify causal and causal networks
(ii) describe the role of intervention (experiment)
(iii) describe the "causal effect" and when we can infer it with, or without, intervention.

Work in this area has been dominated by Rubin (statistics) & Pearl (Computer Science).

Rubin → counterfactuals & missing data.
A is a smoker and he got lung cancer.
Counterfactual: what would happen to A if he didn't smoke?
We don't know the answer → God, Oracle knows.

This question can be addressed by considering a population of smokers and non-smokers — with model assumptions.

Another counterfactual — would World War II have happened if Hitler had died in a car crash in 1932? is harder to answer.
Causal Models: How to address the counterfactual about A not smoking?

Rubin says we should think in terms of missing data.

<table>
<thead>
<tr>
<th>People</th>
<th>Effect of smoking</th>
<th>Effect of not-smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (smoker)</td>
<td>1</td>
<td>0*</td>
</tr>
<tr>
<td>B (non-smoker)</td>
<td>1*</td>
<td>1</td>
</tr>
<tr>
<td>C (smoker)</td>
<td>0</td>
<td>1*</td>
</tr>
<tr>
<td>D (non-smoker)</td>
<td>1*</td>
<td>0</td>
</tr>
</tbody>
</table>

1 - cancer
0 - not-cancer
x - known to oracle (but not to us).

To address the counterfactual we must assume a simple causal model:

- \( S \) & \( C \) binary variables
  - \( S = 1 \) smoker, \( S = 0 \) non-smoker
  - \( C = 1 \) cancer, \( C = 0 \) non-cancer.

Observe data \( \{(S_m, C_m) : m = 1:n\} \)

Parameterize the model:

\[
\begin{align*}
P(C=1|S=1) &= \delta_1, & P(C=0|S=1) &= 1 - \delta_1, \\
P(C=1|S=0) &= \delta_0, & P(C=0|S=0) &= 1 - \delta_0 \\
P(S=1) &= \gamma, & P(S=0) &= 1 - \gamma \\
P(C|S)P(S) &= P(C,S) \\
\end{align*}
\]

Independent assumption: \( P(\{S_m, C_m\}) = \prod_m P(S_m, C_m) \)
Causal Models. Then estimate \( \lambda, d_0, \delta \) by maximum likelihood (ML):

\[
P(\{S^i, C^j\} | x, d_0, d_1, \delta) = \left( \frac{N}{1-N} \right)^{N_1} \left( 1 - \frac{N}{1-N} \right)^{N_0} \left( 1 - d_0 \right)^{N_1} \left( d_0 \right)^{N_0} \ 
\]

with \( N_1 + N_0 + N_{10} + N_{00} = N \), \( N_1 + N_0 = N_1 \), \( N_{ij} \rightarrow \) no of cases with \( C=i \& S=j \).

ML gives:

\[
\hat{\lambda}_1 = \frac{N_{11}}{N_1 + N_{10}} \quad \hat{\lambda}_0 = \frac{N_{00}}{N_0 + N_{01}} \quad \hat{\delta} = \frac{N_{10}}{N_1 + N_{10}}
\]


Estimate of \( P(\{C=11, S=0\} | N_{10}/N_0 + N_{00} \) of not-smoking.

Rubin says you must carefully distinguish between the observational and the counterfactual, what happen to A if he doesn’t smoke — and the method for estimating this.

Here — the estimation assumed the simple model as. But what if there are other risk factors for cancer?
Causal Models:

\[ \text{S - some (unknown) genetic factor} \]
\[ \text{T - type of cigarette.} \]
\[ \text{non-filtered, filtered} \]

or

\[ \text{G - predisposes people to smoke} \]
\[ \text{it also predisposes people to get cancer.} \]

In these cases, what is the causal effect of smoking on getting cancer?

In
\[ \text{G - there is a direct effect of smoking to cause cancer. But also an indirect (backwards) effect.} \]
Causal Models

What is causal effect of $S$ on $C$?

**Intervention** — countfactual intervention

Cut all causal links into $S$.

This is a "conceptual cut". It may be impossible due to physical or ethical reasons.

But estimating causal effect requires us to estimate $p(C|S)$ on the graph with the cuts.
Causal Model

Question: If we can observe sample from the original graph $G$, can we estimate the conditional distribution $P(C|S)$ for the cut graph $\overline{G}$?

Answer: No, not for this graph. Why, because we will mix up the direct effect of $S$ on $C$ with the indirect (backward) effect of $S$ on $G$.

For $s, x, y, G$: $P(S, Y, C, G) = \frac{\sum P(S, Y, C, G)}{\sum P(S, Y, G, G')}$

For $s, x, y, G$: $P(S, Y, C, G) = \sum P(S, Y, C, G') P(G | C, S) = \sum P(S, Y, C, G') P(C | S, G) P(G)$
Causal Model.

But, it is easy to check that we can estimate the causal effect of $S$ on $C$ from the graph:

\[ S \rightarrow G \quad \text{or} \quad G \rightarrow S \]

Pearl's identifying two criteria:

\begin{itemize}
  \item Backdoor criterion
  \item Frontdoor criterion
\end{itemize}

Backdoor is the most standard (Frontdoor occurs less often).

Alternative notation:

\[ S \rightarrow X \rightarrow \epsilon \]

\[ C = f(S, G, \epsilon_1) \]
\[ S = f(G, \epsilon_2) \]
\[ \epsilon_1, \epsilon_2 \text{ are noise variables} \]
\[ \text{structural equation} \]

\[ \text{e.g. } X = \mu + \epsilon \quad \text{and } \epsilon_1 \sim \mathcal{N}(\mu, \sigma) \]

Intervention means that the value of $S$ is fixed from outside the system. It cannot influence $C$.

\[ S = s, \quad C = f(s, G, \epsilon_1) \]

The intervention cannot affect upstream variables $G$, hence requiring $G$ to cut.
Causal Models

Conclusion:

For some causal models, we can determine the causal effects of variables without intervening (i.e. by observing the original system).

But for others, e.g. if there are causes upstream that have independent ways to influence the results, then we have to intervene.

Intervening is like doing an experiment. Experimental studies (e.g. Gopnik) suggest that infants help learn the structure of the world by performing experiments. The "theory theory", infants as scientists.

For more, read Causality by Pearl.
Griffiths & Huangyong Lu

Causal Reasoning

Basic Experiments (Cheng et al.)

Before Medicine

States: $E_1$ $E_0$  

$E_1$: Smiley
$E_0$: Sad

Background Cause $B = 1$ always.

Cause $C = 0$: no medicine
$C = 1$: medicine.

Causal model: $\downarrow B \downarrow C$

Typical experiment:

show $P(E = 1 | B = 1, C = 0)$

Before Medicine

$P(E = 1 | B = 1, C = 1)$

After Medicine

Ask: what is $P(E = 1 | C = 1)$ how effective is medicine?
Strictly speaking, there is no correct answer to this question. You have not given enough information to determine \( P(E=1 \mid B=0, C=1) \) or \( P(E=1 \mid C=1) \).

But people give consistent results.

Power-PC theory (Cheng) as formulated by Griffiths & Tenenbaum.

Hypothesis: people assume a special type of probability distribution:
- noisy-or
- or noisy-and
- noisy-or
- whether B or C can independently cause the effect E.
- (i) but each only causes the effect probabilistically.
Griffiths & Cu. \( P(E=1 \mid B, C) = 1 - \omega_B(1 - \omega_C) \)
\( \omega_B \in [0, 1], \omega_C \in [0, 1] \).

\( P(E=1 \mid B=1, C=0) = \omega_B \)
\( P(E=1 \mid B=0, C=1) = \omega_C \) \( \rightarrow\) causal power.
\( P(E=1 \mid B=1, C=1) = \omega_B + \omega_C - \omega_B \omega_C \)

Noisy-or model is called
"generative" because both causes \((B \& C)\)
can generate the effect.
(Misleading terminology)

Noisy-or has two parameters \( \omega_B, \omega_C \)

Hence data \( \text{P(E} \mid B=1, C=0) \& \text{P(E} \mid B=1, C=1) \)
is sufficient to learn the distribution

\( \sum_{E^m, B^m, C^m} \prod_{\mu} P(E^m \mid B^m, C^m, \omega_B, \omega_C) \)

\( \hat{\omega_B}, \hat{\omega_C} = \text{Max} \prod_{\mu} \frac{P(E^m \mid B^m, C^m, \omega_B, \omega_C)}{P(E^m \mid B^m, C^m, \omega_B, \omega_C)} \)

Predicts \( \hat{\omega}_C \) caused effect of medication
Agrees with experiments.
Griffiths & Yu. Noisy-or means that presence of the cause (i.e., \( \omega = 1 \)) makes it more likely that the effect occurs.

Attention \( \rightarrow \) preventative cause

Noisy-And-Not model.

\( B \) can cause the effect.
\( C \) can prevent the effect from occurring.

\[
P(E = 1 \mid B, C, \omega_B, \omega_C) = \omega_B B (1 - \omega_C C)
\]

The cause \( C \) makes it less likely that the effect occurs.

Can estimate \( \omega_B, \omega_C \) from data by maximum likelihood.

Use noisy-or if data shows that the effect happens more frequently when cause is present, noisy-and otherwise.

Again, this prediction seems to fit data.
Claim: people assume (unconsciously) noisy or or noisy-and-not models of the data. This enables them to estimate the causal effect: $P(C=1 \mid C=1,0)$.

This answers the question "what is the strength of the medicine?" or "how effective is the medicine?"

But there is an alternative question that the experimenters can ask.

→ "does the medicine have any effect?"

This is a model selection task:

\[
\begin{align*}
\beta \cdot 0 \cdot C & \quad \text{or} \quad \beta \cdot 0 \cdot C \\
\text{effective} & \quad \text{or} \quad \text{ineffective}
\end{align*}
\]
Recall model selection
Compute probability that the model generates
the data
\[ P(D|M_i) = \int d\omega_1 \int d\omega_c P(D|\omega_1, \omega_c) P(\omega_1, \omega_c|M_i) \]

model \( M_i \) can assume that the
prior parameter have uniform prior
\[ p(\omega_1, \omega_c|M_i) = 1 \]

\[ P(D|\omega_1, \omega_c) = \prod_{(E^m,b^m,c^m)} P(E^m|b^m,c^m,\omega_1, \omega_c) P(b^m|c^m) \]

\( \{(E^m,b^m,c^m)\}, \) observed samples

Compute log \( P(D|M) \) to the confidence
that human subjects say the medicine
has an effect.
Lu et al. make the clear distinction between a

- strength question: how effective is the medicine?

and a

- selection question: does the medicine have any effect?

Lu et al. also suggest that humans make a "sparse & sufficient" assumption. This is encoded in the prior

\[ p(\omega_b|\omega_c) \]

If \( \omega_b = 1 \), then \( \omega_c = 0 \)

\( \omega_c = 1 \), then \( \omega_b = 0 \).

"Sparse" as few cancer as possible.

Lu et al. fit a large amount of experimental data to this model.
Bottom line so far.

→ People seem to use noisy-or and noisy-and distributions.

A bias towards logic.

Why? Computationally simple? Easily interpretable?

Evidence that infants prefer simple logical explanations to statistical ones.

Wild speculation: May be the structure of the world is well-described by probabilistic logical distributions?

How to extend noisy-or, noisy-and-not to a richer class of distributions?

What happens if experiments specify more condition? E.g.

\[ P(\neg B \mid 1, \neg 0), P(E \mid B=1, \neg 1), P(E \mid B=0, \neg 1) \]
Generalize to complete distribution for two causes (naively, based on degree)

\[ \psi \ \begin{array}{cccc}
C_1 & C_2 & C_1\wedge C_2 \\
\psi_0 & \psi_1 & \psi_2 & \psi_{12} \\
\end{array} \]

\[ H_0, H_1, H_2, H_{12} \]

Binary-valued “hidden states”

\[ H_0 \quad H_1 \quad H_2 \quad H_{12} \]

Final output \( E \) is a logical function of the hidden states.

\[ P(H_0=1 | \psi) = \psi_0 \]
\[ P(H_1=1 | C_1) = \psi_1 C_1 \]
\[ P(H_2=1 | C_2) = \psi_2 C_2 \]
\[ P(H_{12}=1 | C_1 C_2) = \psi_{12} C_1 C_2 \]

To get noisy-or use “causal feature” \( C_1, C_2 \)
set \( E = H_1 \lor H_2 \)

To get noisy-and-not use “causal feature” \( C_1, C_2 \)
set \( E = H_1 \land \neg H_2 \)
Theorem: Any conditional distribution \( P(E|C_1, C_2) \) can be represented in noisy-logical form by suitable choice of \( \omega_0, \omega_1, \omega_2, \ldots \) and the logical function relating \( E \) to the hidden states \( H_0, H_1, H_2, \ldots \).

The noisy-logical distribution can be extended to arbitrarily many causes \( C_1, \ldots, C_n \), by introducing conjunctive factors \( C_1 \land C_2, \ C_1 \land C_2 \land C_3, \ldots \) and so on.

Again, the theorem shows that any distribution \( P(E|C_1, \ldots, C_n) \) can be expressed in this form.

This gives a framework to explore causal reasoning with more than 2 causes, hypothesizing that people do not use conjunctive causes unless there is sufficient data.