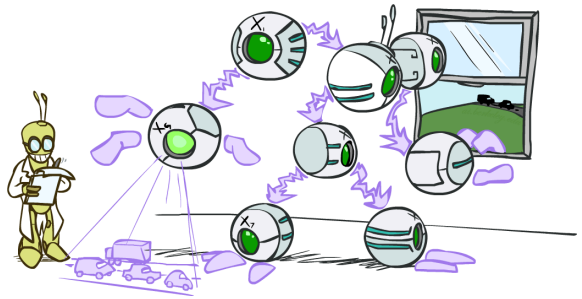
ความรู้เบื้องต้นเกี่ยวกับปัญญาประดิษฐ์

บรรยายครั้งที่ 5: การให้เหตุผลเชิงความน่าจะเป็น (Probabilistic Reasoning)

ผศ. ดร. อิทธิพล ฟองแก้ว
[ittipon@g.sut.ac.th]

หัวข้อในวันนี้

- โครงข่ายเบย์ (Bayesian networks)
  - ความหมาย (Semantics)
  - การสร้าง (Construction)
  - ความสัมพันธ์เชิงอิสระ (Independence relations)
- การอนุมาน (Inference)
- การเรียนรู้พารามิเตอร์ (Parameter learning)



—

Credits: CS188, UC Berkeley.

การแทนความรู้ที่ไม่แน่นอน

การแทนค่าการแจกแจงความน่าจะเป็นร่วม (joint probability distribution) อย่างชัดเจน จะมีขนาดเพิ่มขึ้นแบบทวีคูณ (exponentially) ตามจำนวนตัวแปร

สมมติฐานเรื่อง ความเป็นอิสระ (Independence) และ ความเป็นอิสระแบบมีเงื่อนไข (conditional independence) ช่วยลดจำนวนความน่าจะเป็นที่ต้องกำหนด เราสามารถแทนความสัมพันธ์เหล่านี้ได้อย่างชัดเจนในรูปแบบของ โครงข่ายเบย์ (Bayesian network)

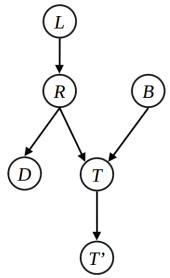
โครงข่ายเบย์ (Bayesian networks)

โครงข่ายเบย์ คือ **กราฟมีทิศทางที่ไม่มีวงจร (directed acyclic graph)** โดยที่

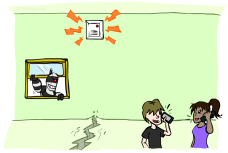
- แต่ละ โหนด (node) แทนตัวแปรสุ่ม (random variable)
  - ซึ่งอาจจะสังเกตได้ (observed) หรือสังเกตไม่ได้ (unobserved)
  - และอาจเป็นได้ทั้งแบบไม่ต่อเนื่อง (discrete) หรือต่อเนื่อง (continuous)
- แต่ละ เส้นเชื่อม (edge) จะมีทิศทางและบ่งชี้ถึงความสัมพันธ์เชิงความน่าจะเป็นโดยตรงระหว่างสองตัวแปร
- แต่ละโหนด $\mathbf{X_i}$ จะถูกกำกับด้วย การแจกแจงความน่าจะเป็นแบบมีเงื่อนไข (conditional probability distribution)

$$\mathbf{P}(\mathbf{X_i}|\text{parents}(\mathbf{X_i}))$$

ซึ่งกำหนดการแจกแจงของ $\mathbf{X_i}$ เมื่อกำหนดค่าของโหนดพ่อแม่ (parents) ของมันในโครงข่าย



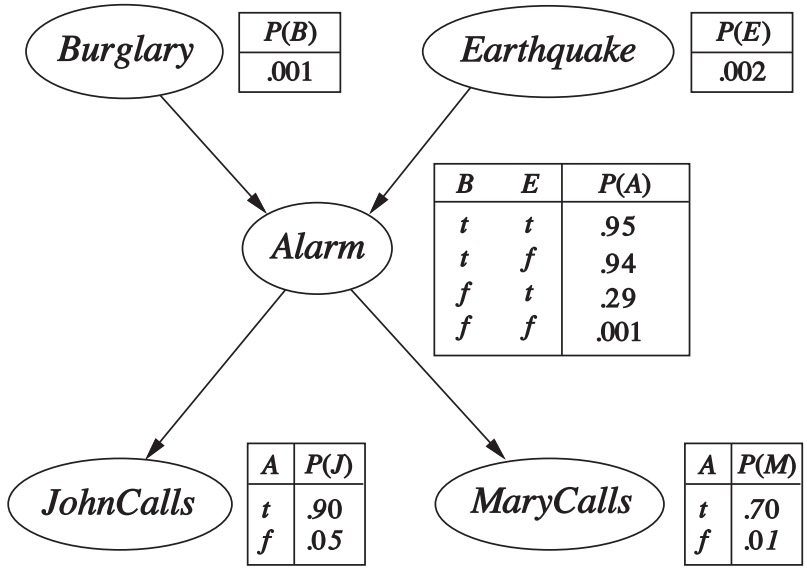
ในนกรณีที่ง่ายที่สุด การแจกแจงแบบมีเงื่อนไขจะถูกแสดงในรูปแบบตารางความน่าจะเป็นแบบมีเงื่อนไข (Conditional Probability Tables - CPTs)



ตัวอย่างที่ 1

- ตัวแปร: Burglar (ขโมย), Earthquake (แผ่นดินไหว), Alarm (สัญญาณเตือน), JohnCalls (จอห์นโทร), MaryCalls (แมรี่โทร)
- โครงสร้างของเครือข่ายสามารถกำหนดได้จากความรู้ในสถานการณ์นั้นๆ:
  - ขโมยสามารถทำให้สัญญาณเตือนดังได้
  - แผ่นดินไหวสามารถทำให้สัญญาณเตือนดังได้
  - สัญญาณเตือนอาจทำให้แมรี่โทรมา
  - สัญญาณเตือนอาจทำให้จอห์นโทรมา สถานการณ์: ฉันอยู่ที่ทำงาน เพื่อนบ้านชื่อจอห์นโทรมาบอกว่าสัญญาณเตือนที่บ้านฉันดัง แต่เพื่อนบ้านอีกคนชื่อแมรี่ไม่ได้โทรมา บางครั้งสัญญาณเตือนก็ดังจากแผ่นดินไหวเล็กน้อย คำถามคือ มีขโมยเข้าบ้านหรือไม่?

—
Credits: CS188, UC Berkeley.


Burglary
P(B)
.001
Earthquake
P(E)
.002
Alarm
B E P(A)
t t .95
t f .94
f t .29
f f .001
JohnCalls
A P(J)
t .90
f .05
MaryCalls
A P(M)
t .70
f .01

ความหมาย (Semantics)

โครงข่ายเบย์เข้ารหัสการแจกแจงความน่าจะเป็นร่วมทั้งหมด (full joint distribution) โดยปริยาย ในรูปแบบผลคูณของการแจกแจงเฉพาะที่ (local distributions) นั่นคือ

$$P(x1, ..., x_n) = \prod{i=1}^n P(x_i | \text{parents}(X_i))$$

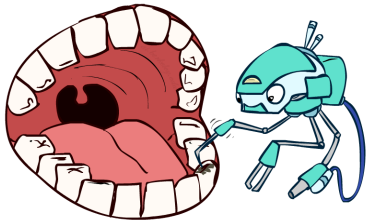
การพิสูจน์:

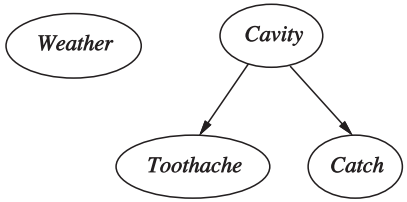
- จากกฎลูกโซ่ (chain rule), $P(x1, ..., x_n) = \prod{i=1}^n P(xi | x_1, ..., x{i-1})$
- หากเราตั้งสมมติฐานว่า $Xi
  เป็นอิสระแบบมีเงื่อนไข(conditionallyindependent)จากโหนดรุ่นก่อนหน้าในลำดับที่กำหนดเมื่อให้ค่าของโหน
  \text{parents}(X_i) \subseteq { X_1, ..., X{i-1}}**เราจะได้ว่า**P(xi | x_1, ..., x{i-1}) = P(x_i | \text{parents}(X_i))$
- ดังนั้น, $P(x1, ..., x_n) = \prod{i=1}^n P(x_i | \text{parents}(X_i))$

ตัวอย่างที่ 1 (ต่อ)

$$P(j, m, a, \neg b, \neg e) = P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \quad = 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \approx 0.00063$$

ตัวอย่างที่ 2





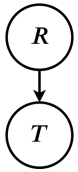
สถานการณ์ของทันตแพทย์สามารถจำลองเป็นโครงข่ายเบย์ที่มี 4 ตัวแปร ดังที่แสดงทางด้านขวา

โดยโครงสร้างของเครือข่าย จะเข้ารหัสสมมติฐานความเป็นอิสระแบบมีเงื่อนไขไว้ในตัว แต่ละตัวแปรจะเป็นอิสระจากโหนดที่ไม่ใช่ลูกหลาน (non-descendants) ของมัน เมื่อกำหนดค่าของโหนดพ่อแม่:

- Weather (สภาพอากาศ) เป็นอิสระจากตัวแปรอื่นๆ
- Toothache (ปวดฟัน) และ Catch (ตรวจพบ) เป็นอิสระแบบมีเงื่อนไขจากกัน เมื่อกำหนดค่า Cavity (ฟันผุ) ทันตแพทย์กำลังตรวจฟันของคนไข้ คนไข้มีฟันผุ แต่ทันตแพทย์ยังไม่ทราบ อย่างไรก็ตาม คนไข้มีอาการปวดฟัน ซึ่งทันตแพทย์สังเกตได้

ตัวอย่างที่ 3

เส้นเชื่อม (Edges) อาจสอดคล้องกับความสัมพันธ์เชิงสาเหตุ (causal relations)



$\mathbf{P}(\mathrm{R})$

R	P
r	0.25
¬r	0.75

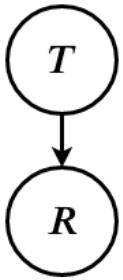
$\mathbf{P}(\mathrm{T}\,|\mathrm{R})$

R	T	P
r	t	0.75
r	¬t	0.25
¬r	t	0.5
¬r	¬t	0.5

แบบจำลองเชิงสาเหตุ (Causal model): ฝนตก (R) ทำให้รถติด (T)

ตัวอย่างที่ 3 (อีกแบบ)

... แต่เส้นเชื่อมไม่จำเป็นต้องเป็นความสัมพันธ์เชิงสาเหตุเสมอไป!



$\mathbf{P}(\mathrm{T})$

T	P
t	9/16
¬t	7/16

$\mathbf{P}(\mathrm{R}|\mathrm{T})$

T	R	P
t	r	1/3
t	¬r	2/3
¬t	r	1/7
¬t	¬r	6/7

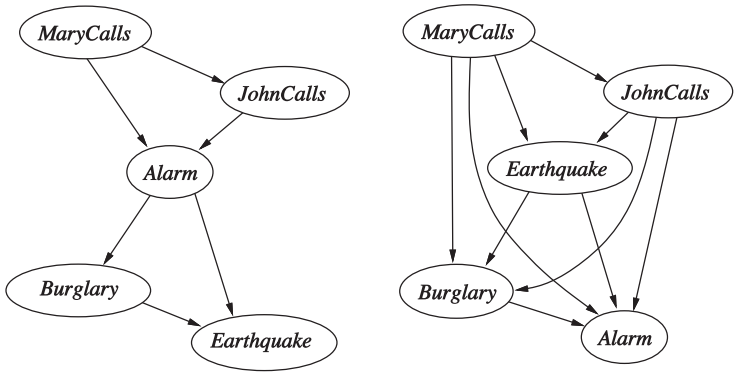
Credits: CS188, UC Berkeley.

การสร้าง (Construction)

เราสามารถสร้างโครงข่ายเบย์โดยใช้ลำดับของตัวแปรแบบใดก็ได้ ตราบใดที่ยังเคารพสมมติฐานความเป็นอิสระแบบมีเงื่อนไข

อัลกอริทึม

1. เลือก ลำดับ (ordering) ของตัวแปร $X_1, ..., X_n$
2. สำหรับ $i = 1$ ถึง n:
    1. เพิ่ม X_i เข้าไปในเครือข่าย
    2. เลือกเซตของโหนดพ่อแม่ (parents) ที่เล็กที่สุดจาก $X1, ..., X{i-1}**ที่ทำให้**P(xi | x_1, ..., x{i-1}) = P(x_i | \text{parents}(X_i))$
    3. สำหรับพ่อแม่แต่ละโหนด ให้ลากเส้นเชื่อมจากพ่อแม่ไปยัง X_i
    4. สร้างตาราง CPT (Conditional Probability Table)



โครงข่ายทั้งสองนี้แทนการแจกแจงเดียวกันหรือไม่? และมีความกระชับ (compact) เท่ากันหรือไม่?

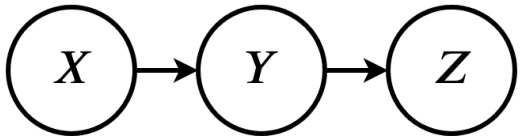
สำหรับโครงข่ายด้านซ้าย:

- P(J|M) = P(J)? ไม่
- P(A|J, M) = P(A|J)? P(A|J, M) = P(A)? ไม่
- P(B|A, J, M) = P(B|A)? ใช่
- P(B|A, J, M) = P(B)? ไม่
- P(E|B, A, J, M) = P(E|A)? ไม่
- P(E|B, A, J, M) = P(E|A, B)? ใช่

โครงข่ายด้านขวา (โครงข่ายเดิม) มีความกระชับกว่ามาก เพราะต้องการพารามิเตอร์ใน CPT น้อยกว่า

ความสัมพันธ์เชิงอิสระ (Independence relations)

เนื่องจากโครงสร้าง (topology) ของโครงข่ายเบย์เข้ารหัสสมมติฐานความเป็นอิสระแบบมีเงื่อนไข เราจึงสามารถใช้มันเพื่อตอบคำถามเกี่ยวกับความเป็นอิสระของตัวแปรต่างๆ เมื่อมีข้อมูลบางอย่าง (evidence)



ตัวอย่าง: X และ Z จำเป็นต้องเป็นอิสระจากกันหรือไม่?

แนวคิดเบื้องหลัง d-separation

d-separation (ย่อมาจาก direction-dependent separation) เป็นวิธีการตรวจสอบว่ากลุ่มของตัวแปร $\mathbf{X}$ เป็นอิสระแบบมีเงื่อนไข (conditionally independent) จากกลุ่มของตัวแปร $\mathbf{Y}$ หรือไม่ เมื่อกำหนดกลุ่มของตัวแปรที่เป็นหลักฐาน (evidence) $\mathbf{Z}$ ให้

แนวคิดหลักคือการพิจารณา เส้นทาง (path) ทั้งหมดระหว่างโหนดใน $\mathbf{X}$ และโหนดใน $\mathbf{Y}$

- หากทุกเส้นทางถูก 'ปิดกั้น' (blocked) โดยหลักฐาน $\mathbf{Z}$ เราจะสรุปได้ว่า $\mathbf{X}$ และ $\mathbf{Y}$ เป็นอิสระจากกันเมื่อมี $\mathbf{Z}$
- หากมีอย่างน้อยหนึ่งเส้นทางที่ 'เปิด' (active) แสดงว่าอิทธิพลอาจส่งผ่านได้ และเราไม่สามารถรับประกันความเป็นอิสระได้

เส้นทางจะถูกพิจารณาว่าเป็น 'active' (ข้อมูลไหลผ่านได้) หรือ 'inactive' (ถูกปิดกั้น) ขึ้นอยู่กับโครงสร้างของเส้นทางและโหนดหลักฐานที่อยู่บนเส้นทางนั้น

Cascades (แบบอนุกรม)

X เป็นอิสระจาก Z หรือไม่? ไม่

ตัวอย่างแย้ง:

- ความกดอากาศต่ำ (X) ทำให้เกิดฝน (Y), ฝนทำให้รถติด (Z)
- ในทางตัวเลข:
  - $P(y|x) = 1, P(z|y) = 1$
  - $P(\neg y|\neg x) = 1, P(\neg z|\neg y) = 1$
- ข้อมูลเกี่ยวกับ X ส่งผลต่อความเชื่อของเราเกี่ยวกับ Z

X: ความกดอากาศต่ำ Y : ฝน Z: รถติด

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

—

Credits: CS188, UC Berkeley.

X เป็นอิสระจาก Z เมื่อกำหนด Y หรือไม่? ใช่

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

X: ความกดอากาศต่ำ Y : ฝน Z: รถติด

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

เรากล่าวว่าหลักฐาน (evidence) Y ตามเส้นทางแบบอนุกรมนี้ ปิดกั้น (blocks) อิทธิพล

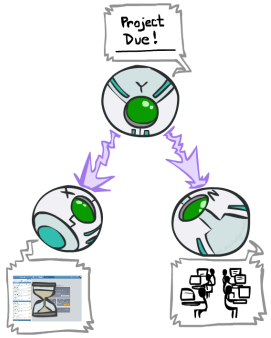
—

Common parent (พ่อแม่ร่วม)

X เป็นอิสระจาก Z หรือไม่? ไม่

ตัวอย่างแย้ง:

- มีโปรเจกต์ที่ต้องส่ง (Y) ทำให้ฟอรัมคนเยอะ (X) และห้องแล็บเต็ม (Z)
- ในทางตัวเลข:
  - $P(x|y) = 1, P(\neg x|\neg y) = 1$
  - $P(z|y) = 1, P(\neg z|\neg y) = 1$
- การสังเกตว่าฟอรัมคนเยอะ (X) ทำให้เราเชื่อมากขึ้นว่ามีโปรเจกต์ต้องส่ง (Y) ซึ่งก็ทำให้เราเชื่อว่าห้องแล็บจะเต็ม (Z)



X: ฟอรัมคนเยอะ Y : โปรเจกต์ต้องส่ง Z: ห้องแล็บเต็ม

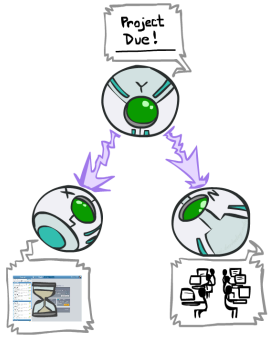
$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

—

X เป็นอิสระจาก Z เมื่อกำหนด Y หรือไม่? ใช่

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

การสังเกตโหนดพ่อแม่ (Y) จะปิดกั้นอิทธิพลระหว่างโหนดลูก (X, Z)



X: ฟอรัมคนเยอะ Y : โปรเจกต์ต้องส่ง Z: ห้องแล็บเต็ม

$$P(x,y,z) = P(y)P(x|y)P(z|y)$$

—

Credits: CS188, UC Berkeley.

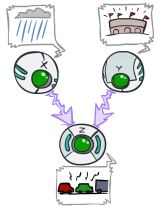
v-structures (โครงสร้างรูปตัว V)

X และ Y เป็นอิสระจากกันหรือไม่? ใช่

- ฝนตก (X) และการแข่งขันเบสบอล (Y) ทำให้รถติด (Z) แต่การเกิดฝนกับการมีแข่งเบสบอลเป็นอิสระจากกัน
- (พิสูจน์!)

X และ Y เป็นอิสระจากกันเมื่อกำหนด Z หรือไม่? ไม่!

- การที่เรารู้ว่ารถติด (Z) ทำให้ฝนตก (X) และการแข่งเบสบอล (Y) กลายเป็นคำอธิบายที่ "แข่งขัน" กันเอง (explaining away) เช่น ถ้ารู้ว่ารถติดแต่ไม่มีแข่งเบสบอล ก็จะทำให้เราเชื่อมากขึ้นว่าฝนตก
- นี่คือสิ่งที่ ตรงกันข้าม กับกรณีก่อนหน้า การสังเกตโหนดลูกจะ เปิด (activates) อิทธิพลระหว่างโหนดพ่อแม่]]



X: ฝน Y : แข่งเบสบอล Z: รถติด

$$P(x, y, z) = P(x)P(y)P(z|x, y)$$

พิสูจน์: $P(x,y,z) = P(x)P(y)P(z|x,y)$ และ $P(x,y,z) = P(x,y)P(z|x,y)$ ดังนั้น $P(x,y) = P(x)P(y)$

d-separation

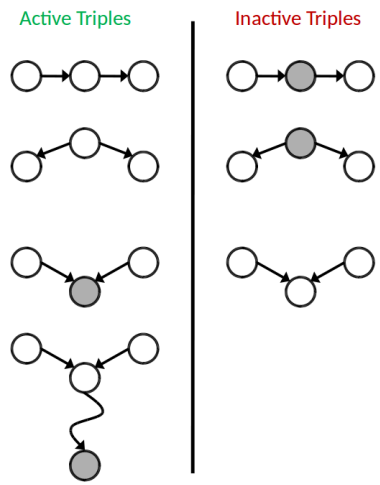
สมมติว่าเรามีโครงข่ายเบย์ที่สมบูรณ์ X_i และ X_j เป็นอิสระแบบมีเงื่อนไขหรือไม่ เมื่อกำหนดหลักฐาน $Z_1 = z_1, ..., Z_m = z_m$?

พิจารณาเส้นทาง (ที่ไม่กำหนดทิศทาง) ทั้งหมดจาก X_i ไปยัง X_j:

- ถ้ามีเส้นทางที่ active (เปิด) อย่างน้อยหนึ่งเส้นทาง จะไม่สามารถรับประกันความเป็นอิสระได้
- หากทุกเส้นทางเป็น inactive (ปิดกั้น) จะสามารถรับประกันความเป็นอิสระได้

เส้นทางจะ active (เปิด) ถ้าแต่ละสามโหนด (triple) ตลอดเส้นทางนั้น active:

- Cascade $A \rightarrow B \rightarrow C$ โดยที่ B ไม่ถูกสังเกต (unobserved) (ไม่ว่าจะทิศทางใด)
- Common parent $A \leftarrow B \rightarrow C$ โดยที่ B ไม่ถูกสังเกต
- v-structure $A \rightarrow B \leftarrow C$ โดยที่ B หรือลูกหลาน (descendent) ของมัน ถูกสังเกต (observed)

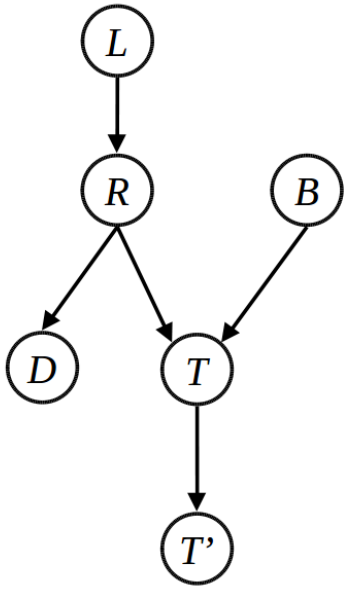


—

Credits: CS188, UC Berkeley.

ตัวอย่าง

- $\mathrm{L} \perp \mathrm{T}' | \mathrm{T}$?
- $\mathrm{L} \perp \mathrm{B}$?
- $\mathrm{L} \perp \mathrm{B} | \mathrm{T}$?
- $\mathrm{L} \perp \mathrm{B} | \mathrm{T}'$?
- $\mathrm{L} \perp \mathrm{B} | \mathrm{T}, \mathrm{R}$?



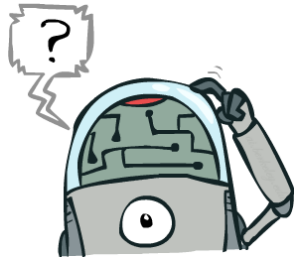
- Yes (เส้นทาง $\mathrm{L} \to \mathrm{T} \to \mathrm{T}'$ ถูกปิดกั้นโดย T)
- Yes (เส้นทาง $\mathrm{L} \to \mathrm{T} \leftarrow \mathrm{R} \to \mathrm{B}$ เป็น v-structure ที่ T ไม่ถูกสังเกต จึงปิดกั้น)
- No (การสังเกต T ทำให้เส้นทาง $\mathrm{L} \to \mathrm{T} \leftarrow \mathrm{R}$ เปิดออก)
- No (การสังเกต T' ซึ่งเป็นลูกของ T ทำให้เส้นทาง $\mathrm{L} \to \mathrm{T} \leftarrow \mathrm{R}$ เปิดออก)
- Yes (เส้นทาง $\mathrm{L} \to \mathrm{T} \leftarrow \mathrm{R} \to \mathrm{B}$ ถูกปิดกั้นโดย R ซึ่งเป็น common parent ที่ถูกสังเกต)

การอนุมาน (Inference)

การอนุมาน (Inference) คือกระบวนการในการ **คำนวณการแจกแจงความน่าจะเป็นแบบ marginal และ/หรือ conditional** จากการแจกแจงความน่าจะเป็นร่วม (joint probability distribution):

คำถามอย่างง่าย:	$\mathbf{P}(X_i \mid e)$ (ความน่าจะเป็นของ X_i เมื่อมีหลักฐาน e)
คำถามแบบเชื่อม:	$\mathbf{P}(X_i, X_j \mid e) = \mathbf{P}(X_i \mid e)\mathbf{P}(X_j \mid X_i, e)$
คำอธิบายที่เป็นไปได้ที่สุด (Most Likely Explanation):	$\arg\max_q P(q \mid e)$ (หาชุดค่า q ที่มีความน่าจะเป็นสูงสุดเมื่อมี e)

การตัดสินใจที่เหมาะสมที่สุด $\arg\max_a \mathbb{E}_{p(s' \mid s,a)}[V(s')]$



—

Credits: CS188, UC Berkeley.

อธิบายว่า $\arg\max$ หมายถึงอะไร (การหาอาร์กิวเมนต์ที่ทำให้ฟังก์ชันมีค่าสูงสุด)

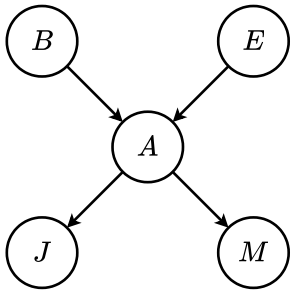
เน้นย้ำความสำคัญของการอนุมาน: การอนุมาน <=> การให้เหตุผล

การอนุมานโดยการแจกแจง (Inference by enumeration)

เริ่มต้นจากการแจกแจงร่วม $\mathbf{P}(\mathrm{Q}, \mathrm{E}_1, \ldots, \mathrm{E}_\mathrm{k}, \mathrm{H}_1, \ldots, \mathrm{H}_\mathrm{r})$ (Q: Query, E: Evidence, H: Hidden)

1. เลือกเฉพาะรายการที่สอดคล้องกับหลักฐาน (evidence) $\mathrm{E}_1, \ldots, \mathrm{E}_\mathrm{k} = \mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{k}$
2. ทำการ Marginalize out (บวกรวมข้าม) ตัวแปรซ่อนเร้น (hidden variables) เพื่อให้ได้การแจกแจงร่วมของตัวแปรที่สนใจ (query) และตัวแปรหลักฐาน: $${\bf P}(Q,e1,...,e_k) = \sum{h_1, ..., h_r} \mathbf{P}(Q, h_1, ..., h_r, e_1, ..., e_k)$$
3. ทำให้เป็นปกติ (Normalize):

$$\mathrm{Z} = \sum_{\mathrm{q}} \mathrm{P}\left(\mathrm{q}, \mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{k}\right) \quad \mathbf{P}(\mathrm{Q} | \mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{k}) \quad = \frac{1}{\mathrm{Z}} \mathbf{P}(\mathrm{Q}, \mathrm{e}_1, \ldots, \mathrm{e}_\mathrm{k})$$

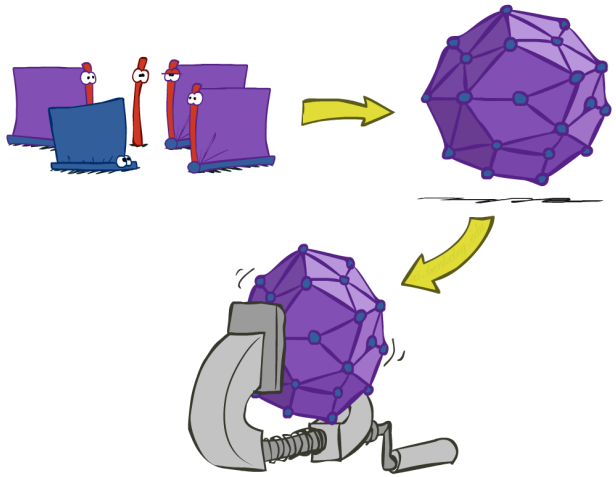


พิจารณาโครงข่ายสัญญาณเตือนและคำถาม $\mathbf{P}(\mathrm{B}|\mathrm{j},\mathrm{m})$ (ความน่าจะเป็นที่จะมีขโมย เมื่อจอห์นและแมรี่โทรมา)

$$\mathbf{P}(\mathrm{B}|\mathrm{j},\mathrm{m}) = \frac{1}{\mathrm{Z}}\sum_{\mathrm{e}}\sum_{\mathrm{a}}\mathbf{P}(\mathrm{B},\mathrm{j},\mathrm{m},\mathrm{e},\mathrm{a}) \quad \propto \sum_{\mathrm{e}}\sum_{\mathrm{a}}\mathbf{P}(\mathrm{B},\mathrm{j},\mathrm{m},\mathrm{e},\mathrm{a})$$

เมื่อใช้โครงข่ายเบย์ เราสามารถเขียนการแจกแจงร่วมในรูปผลคูณของค่าจาก CPT ได้:

$$\mathbf{P}(\mathrm{B}|\mathrm{j},\mathrm{m}) \propto \sum_{\mathrm{e}}\sum_{\mathrm{a}}\mathbf{P}(\mathrm{B})\mathrm{P}\left(\mathrm{e}\right)\mathbf{P}(\mathrm{a}|\mathrm{B},\mathrm{e})\mathrm{P}\left(\mathrm{j}|\mathrm{a}\right)\mathrm{P}\left(\mathrm{m}|\mathrm{a}\right)$$



การอนุมานโดยการแจกแจงนั้นช้า เพราะต้องสร้างการแจกจ่ายร่วมทั้งหมดขึ้นมาก่อนที่จะบวกรวมตัวแปรซ่อนเร้นออกไป

—

Credits: CS188, UC Berkeley.

เราสามารถดึงตัวคูณ (factors) ที่ไม่ขึ้นกับตัวแปรในผลรวมออกมาได้ ซึ่งหมายความว่าการทำ marginalization ไม่จำเป็นต้องทำตอนท้ายสุดเสมอไป และช่วยลดการคำนวณลงได้

สำหรับโครงข่ายสัญญาณเตือน เราจะได้:

$$\mathbf{P}(\mathrm{B}|\mathrm{j},\mathrm{m}) \propto \sum_{\mathrm{e}}\sum_{\mathrm{a}} \mathbf{P}(\mathrm{B})\mathrm{P}(\mathrm{e})\mathbf{P}(\mathrm{a}|\mathrm{B},\mathrm{e})\mathrm{P}(\mathrm{j}|\mathrm{a})\mathrm{P}(\mathrm{m}|\mathrm{a}) \quad = \mathbf{P}(\mathrm{B})\sum_{\mathrm{e}}\mathrm{P}(\mathrm{e})\sum_{\mathrm{a}}\mathbf{P}(\mathrm{a}|\mathrm{B},\mathrm{e})\mathrm{P}(\mathrm{j}|\mathrm{a})\mathrm{P}(\mathrm{m}|\mathrm{a})$$

function ENUMERATION-ASK(X, **e**, bn) **returns** a distribution over X
  inputs: X, the query variable
    e, observed values for variables **E**
    bn, a Bayes net with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ /* **Y** = *hidden variables* */

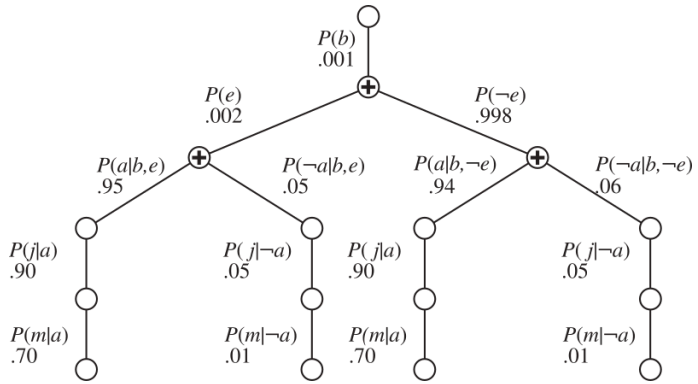
  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(bn.VARS, $\mathbf{e}_{x_i}$)
      where $\mathbf{e}_{x_i}$ is **e** extended with $X = x_i$
  return NORMALIZE($\mathbf{Q}(X)$)

function ENUMERATE-ALL($vars$, **e**) **returns** a real number
  if EMPTY?($vars$) **then return** 1.0
  $Y \leftarrow$ FIRST($vars$)
  if Y has value y in **e**
    then return $P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), **e**)
    else return $\sum_y P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is **e** extended with $Y = y$

มีความซับซ้อนเช่นเดียวกับ DFS: $O(n)$ ในด้านพื้นที่ และ $O(d^n)$ ในด้านเวลา

- n คือจำนวนตัวแปร
- d คือขนาดของโดเมนของตัวแปร

แผนภูมิต้นไม้การประเมินสำหรับ $P(b|j,m)$



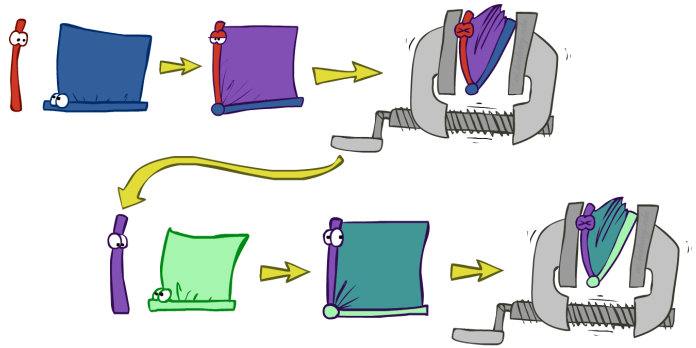
แม้จะมีการแยกตัวคูณออกมา การอนุมานโดยการแจกแจงก็ยังคง ไม่มีประสิทธิภาพ เพราะมีการคำนวณซ้ำซ้อน!

- เช่น $P(j|a)P(m|a)$ ถูกคำนวณสองครั้ง ครั้งหนึ่งสำหรับ e และอีกครั้งสำหรับ $\neg e$
- เราสามารถหลีกเลี่ยงปัญหานี้ได้โดยการเก็บ ผลลัพธ์ระหว่างกลาง ไม่มีประสิทธิภาพเพราะผลคูณถูกประเมินจากซ้ายไปขวาใน ลักษณะของ DFS

การอนุมานโดยการกำจัดตัวแปร (Variable Elimination)

อัลกอริทึม **Variable Elimination** จะทำการบวกรวมจากขวาไปซ้าย และเก็บตัวคูณระหว่างกลาง (intermediate factors) เพื่อหลีกเลี่ยงการคำนวณซ้ำ อัลกอริทึมจะทำงานสลับกันระหว่าง:

- การรวมตารางย่อย (Joining sub-tables)
- การกำจัดตัวแปรซ่อนเร้น (Eliminating hidden variables)



—

Credits: CS188, UC Berkeley.

Variable Elimination

คำถาม: $\mathbf{P}(Q|e_1, ..., e_k)$

1. เริ่มต้นด้วยตัวคูณเริ่มต้น (initial factors) (คือ CPTs เฉพาะที่ ซึ่งถูกแทนค่าด้วยหลักฐานแล้ว)
2. ตราบใดที่ยังมีตัวแปรซ่อนเร้น:
    1. เลือกตัวแปรซ่อนเร้น H
    2. รวม (Join) ตัวคูณทั้งหมดที่เกี่ยวข้องกับ H
    3. กำจัด (Eliminate) H
3. รวมตัวคูณที่เหลือทั้งหมด
4. ทำให้เป็นปกติ (Normalize)

ตัวคูณ (Factors)

- ตัวคูณ $\mathbf{f}_i$ แต่ละตัวคืออาร์เรย์หลายมิติที่ถูกทำดัชนีด้วยค่าของตัวแปรที่เป็นอาร์กิวเมนต์ของมัน เช่น:

$$\mathbf{f}_4 = \mathbf{f}_4(\mathrm{A}) = \left(\mathrm{P}\left(\mathrm{j}|\mathrm{a}\right)\ \mathrm{P}\left(\mathrm{j}|\neg\mathrm{a}\right)\right) = \left(0.90\ 0.05\right)\ \ \mathbf{f}_4(\mathrm{a}) \quad = 0.90\ \mathbf{f}_4(\neg\mathrm{a}) = 0.05$$

- ตัวคูณจะถูกกำหนดค่าเริ่มต้นด้วย CPTs ที่กำกับโหนดของโครงข่ายเบย์ โดยปรับตามหลักฐาน (evidence) ที่มี

การรวม (Join)

ผลคูณแบบจุดต่อจุด (pointwise product) $\times$, หรือ *join, ของสองตัวคูณ $\mathbf{f}_1$ และ $\mathbf{f}_2$ จะได้ตัวคูณใหม่ $\mathbf{f}_3$

- เหมือนกับการทำ database join!
- ตัวแปรของ $\mathbf{f}_3$ คือ ยูเนียน ของตัวแปรใน $\mathbf{f}_1$ และ $\mathbf{f}_2$
- สมาชิกของ $\mathbf{f}_3$ ได้มาจากการคูณสมาชิกที่สอดคล้องกันใน $\mathbf{f}_1$ และ $\mathbf{f}_2$

A	B	$\mathbf{f}_1(A,B)$	B	C	$\mathbf{f}_2(B,C)$	A	B	C	$\mathbf{f}_3(A,B,C)$
T	T	.3	T	T	.2	T	T	T	$.3 \times .2 = .06$
T	F	.7	T	F	.8	T	T	F	$.3 \times .8 = .24$
F	T	.9	F	T	.6	T	F	T	$.7 \times .6 = .42$
F	F	.1	F	F	.4	T	F	F	$.7 \times .4 = .28$
						F	T	T	$.9 \times .2 = .18$
						F	T	F	$.9 \times .8 = .72$
						F	F	T	$.1 \times .6 = .06$
						F	F	F	$.1 \times .4 = .04$

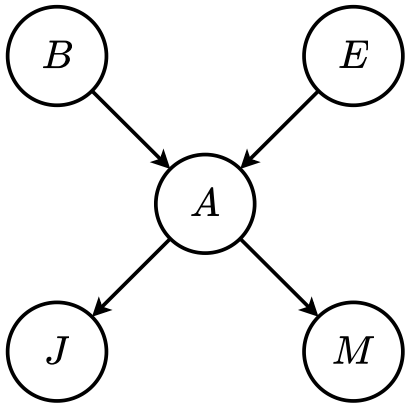
Figure 14.10 Illustrating pointwise multiplication: $\mathbf{f}_1(A,B) \times \mathbf{f}_2(B,C) = \mathbf{f}_3(A,B,C)$.

การกำจัด (Elimination)

การบวกรวมข้าม (Summing out), หรือ การกำจัด (eliminating), ตัวแปรออกจากตัวคูณ ทำได้โดยการบวกรวมอาร์เรย์ย่อยที่เกิดจากการกำหนดค่าของตัวแปรนั้นให้เป็นแต่ละค่าที่เป็นไปได้

ตัวอย่างเช่น เพื่อกำจัด A ออกจาก $\mathbf{f}_3(\mathrm{A}, \mathrm{B}, \mathrm{C})$, เราจะเขียนว่า:

$$\mathbf{f}(\mathrm{B}, \mathrm{C}) = \sum_{\mathrm{a}} \mathbf{f}_3(\mathrm{a}, \mathrm{B}, \mathrm{C}) = \mathbf{f}_3(\mathrm{a}, \mathrm{B}, \mathrm{C}) + \mathbf{f}_3(\neg\mathrm{a}, \mathrm{B}, \mathrm{C}) \quad = (0.06 \quad 0.24\ 0.42 \quad 0.28) + (0.18 \quad 0.72\ 0.06 \quad 0.04) = (0.24$$



ลองใช้อัลกอริธึม variable elimination สำหรับคำถาม $\mathbf{P}(\mathrm{B}|\mathrm{j},\mathrm{m})$

ความเกี่ยวข้อง (Relevance)

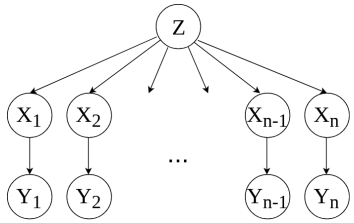
พิจารณาคำถาม $\mathbf{P}(\mathrm{J}|\mathrm{b})$:

$$\mathbf{P}(\mathrm{J}|\mathrm{b}) \propto \mathrm{P}(\mathrm{b}) \sum_{\mathrm{e}} \mathrm{P}(\mathrm{e}) \sum_{\mathrm{a}} \mathrm{P}(\mathrm{a}|\mathrm{b},\mathrm{e}) \mathbf{P}(\mathrm{J}|\mathrm{a}) \sum_{\mathrm{m}} \mathrm{P}(\mathrm{m}|\mathrm{a})$$

- $\sum_{\mathrm{m}} \mathrm{P}(\mathrm{m}|\mathrm{a}) = 1$ (เพราะผลรวมความน่าจะเป็นของทุกค่าที่เป็นไปได้ของ M ต้องเท่ากับ 1) ดังนั้น M จึง ไม่เกี่ยวข้อง (irrelevant) กับคำถามนี้
- กล่าวอีกนัยหนึ่ง $\mathbf{P}(\mathrm{J}|\mathrm{b})$ จะไม่เปลี่ยนแปลงถ้าเรานำ M ออกจากเครือข่าย

ทฤษฎีบท. ตัวแปร H จะไม่เกี่ยวข้องกับ $\mathbf{P}(\mathrm{Q}|\mathrm{e})$ เว้นแต่ว่า H จะอยู่ในเซตของบรรพบุรุษ (ancestors) ของ $\mathrm{Q} \cup \mathrm{E}$

ความซับซ้อน (Complexity)



พิจารณาคำถาม $\mathbf{P}(\mathrm{X_n}|\mathrm{y_1}, ..., \mathrm{y_n})$

ลองพิจารณาลำดับการกำจัดสองแบบ:

- $Z, X1, ..., X{n-1}$
- $X1, ..., X{n-1}, Z$

ขนาดของตัวคูณที่ใหญ่ที่สุดที่สร้างขึ้นสำหรับแต่ละลำดับคือเท่าใด?

- คำตอบ: 2^{n+1} เทียบกับ 2^2 (สมมติว่าเป็นตัวแปรแบบบูลีน)

ความซับซ้อนทางเวลาและพื้นที่ของ variable elimination ถูกกำหนดโดยขนาดของตัวคูณที่ใหญ่ที่สุด

- ลำดับ การกำจัดตัวแปรมีผลอย่างมากต่อขนาดของตัวคูณที่ใหญ่ที่สุด
- การหาลำดับที่ดีที่สุดเป็นปัญหา NP-hard ซึ่งหมายความว่ายังไม่มีอัลกอริทึมที่หาคำตอบได้ในเวลาพหุนาม (polynomial-time)

การอนุมานเชิงประมาณ (Approximate inference)

การอนุมานแบบแม่นตรง (Exact inference) นั้น ไม่สามารถทำได้ในทางปฏิบัติ (intractable) สำหรับแบบจำลองความน่าจะเป็นส่วนใหญ่ที่น่าสนใจในโลกแห่งความเป็นจริง (เช่น แบบจำลองที่มีตัวแปรจำนวนมาก, มีทั้งตัวแปรต่อเนื่องและไม่ต่อเนื่อง, มีวงจรแบบไม่มีทิศทาง ฯลฯ)

เราจึงต้องหันไปใช้อัลกอริทึมการอนุมาน เชิงประมาณ (approximate):

- วิธีสุ่มตัวอย่าง (Sampling methods): สร้างคำตอบโดยการสุ่มตัวอย่างซ้ำๆ จากการแจกแจงที่สนใจ
- วิธีเชิงการแปรผัน (Variational methods): กำหนดปัญหาการอนุมานให้เป็นปัญหาการหาค่าที่เหมาะสมที่สุด (optimization problem)
- วิธี Belief propagation: กำหนดการอนุมานเป็นอัลกอริทึมการส่งผ่านข้อความ (message-passing)
- วิธี Machine learning: เรียนรู้การประมาณค่าของการแจกแจงเป้าหมายจากข้อมูลตัวอย่าง

การเรียนรู้พารามิเตอร์ (Parameter learning)

เมื่อสร้างแบบจำลองสำหรับโดเมนใดๆ เราสามารถเลือกแบบจำลองความน่าจะเป็นที่ระบุเป็นโครงข่ายเบย์ได้ อย่างไรก็ตาม การกำหนดค่าความน่าจะเป็นแต่ละค่าด้วยตนเองมักเป็นเรื่องยาก

ทางออกคือการใช้กลุ่มของแบบจำลองที่ มีพารามิเตอร์ (parameterized) $\mathbf{P}(\mathbf{X}|\theta)$ (บางครั้งเขียนเป็น $\mathbf{P}_\theta(\mathbf{X})$) และ ประมาณค่า (estimate) พารามิเตอร์ θ จากข้อมูล

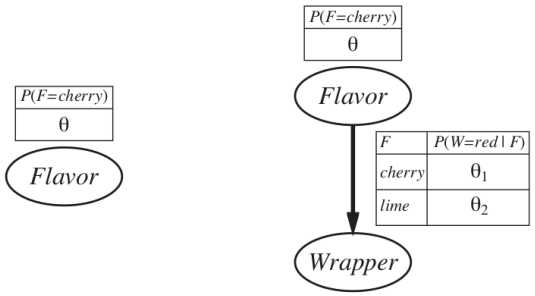


Figure 20.2 (a) Bayesian network model for the case of candies with an unknown proportion of cherries and limes. (b) Model for the case where the wrapper color depends (probabilistically) on the candy flavor.

การประมาณค่าด้วยความควรจะเป็นสูงสุด (Maximum Likelihood Estimation)

สมมติว่าเรามีชุดข้อมูลสังเกต N ชุดที่เป็นอิสระและมีการแจกแจงเหมือนกัน (i.i.d.) $\mathbf{d} = \mathrm{x}_1, ..., \mathrm{x}_N$

ความควรจะเป็น (likelihood) ของพารามิเตอร์ θ คือความน่าจะเป็นของข้อมูลเมื่อกำหนดพารามิเตอร์

$$\mathrm{P}(\mathbf{d}|\theta) = \prod_{j=1}^{N} \mathrm{P}(\mathrm{x}_j|\theta)$$

ค่าประมาณความควรจะเป็นสูงสุด (Maximum Likelihood Estimate - MLE)

θ ของพารามิเตอร์ คือค่าของ θ ที่ทำให้ค่า likelihood สูงที่สุด

$$\theta^* = \arg\max_{\theta} \mathrm{P}(\mathbf{d}|\theta)$$

ในทางปฏิบัติ:

1. เขียนฟังก์ชัน log-likelihood $\mathbf{L}(\theta) = \log \mathbf{P}\left(\mathbf{d}|\theta\right)$ ของพารามิเตอร์ θ
2. หาอนุพันธ์ $\frac{\partial \mathbf{L}}{\partial \theta}$ ของ log-likelihood เทียบกับพารามิเตอร์ θ
3. หาค่าพารามิเตอร์ $\theta*$ ที่ทำให้อนุพันธ์เป็นศูนย์ (และตรวจสอบว่า Hessian เป็น negative definite เพื่อให้แน่ใจว่าเป็นค่าสูงสุด)

กรณี (a)

สัดส่วน θ ของลูกอมรสเชอร์รี่เป็นเท่าใด?

สมมติว่าเราแกะลูกอม N เม็ด ได้รสเชอร์รี่ c เม็ด และรสมะนาว $l = N - c$ เม็ด ข้อมูลเหล่านี้เป็น i.i.d. ดังนั้น

$$P(\mathbf{d}|\theta) = \prod_{j=1}^{N} P(x_j|\theta) = \theta^c(1-\theta)^l$$

หาค่าสูงสุดของฟังก์ชันนี้เทียบกับ θ ซึ่งจะง่ายกว่าถ้าทำกับ log-likelihood:

$$L(\mathbf{d}|\theta) = \log P(\mathbf{d}|\theta) = c\log\theta + l\log(1-\theta) \quad \frac{\partial L(\mathbf{d}|\theta)}{\partial\theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0$$

ดังนั้น $\theta = \frac{c}{N}$

กรณี (b)

เปลือกลูกอมสีแดงและสีเขียวขึ้นอยู่กับรสชาติอย่างน่าจะเป็น เช่น likelihood ของลูกอมรสเชอร์รี่ในเปลือกสีเขียวคือ

$$\mathrm{P}\left(\mathrm{cherry}, \mathrm{green} | \theta, \theta_1, \theta_2\right) = \mathrm{P}\left(\mathrm{cherry} | \theta, \theta_1, \theta_2\right) \mathrm{P}\left(\mathrm{green} | \mathrm{cherry}, \theta, \theta_1, \theta_2\right) = \theta(1-\theta_1)$$

likelihood ของพารามิเตอร์ เมื่อมีลูกอม N เม็ด, เชอร์รี่เปลือกแดง $\mathrm{r_c}$ เม็ด, เชอร์รี่เปลือกเขียว $\mathrm{g_c}$ เม็ด, ฯลฯ คือ

$$\mathrm{P}\left(\mathbf{d} | \theta, \theta_1, \theta_2\right) = \theta^{\mathrm{c}}(1-\theta)^{\mathrm{l}} \theta_1^{\mathrm{r_c}}(1-\theta_1)^{\mathrm{g_c}} \theta_2^{\mathrm{r_l}}(1-\theta_2)^{\mathrm{g_l}} \quad \mathrm{L} = \quad \mathrm{c} \log \theta + \mathrm{l} \log(1-\theta) + \quad \mathrm{r_c} \log \theta_1 + \mathrm{g_c} \log(1-\theta_1) + \quad \mathrm{r_l} \log \theta_2 +$$

อนุพันธ์ของ L จะให้ผลลัพธ์:

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \Rightarrow \theta = \frac{c}{c+l} \quad \frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \Rightarrow \theta_1 = \frac{r_c}{r_c + g_c} \quad \frac{\partial L}{\partial \theta_2} = \frac{r_l}{\theta_2} - \frac{g_l}{1-\theta_2} = 0 \Rightarrow \theta_2$$

ในกรณี (a) ถ้าเราแกะลูกอม 1 เม็ดและได้รสเชอร์รี่ ค่า MLE คืออะไร? เรามั่นใจในค่าประมาณนี้แค่ไหน?

- เมื่อมีข้อมูลน้อย การประมาณค่าด้วย MLE อาจนำไปสู่การ overfitting (การปรับตัวเข้ากับข้อมูลมากเกินไป)
- MLE ไม่ได้ให้ข้อมูลเกี่ยวกับความไม่แน่นอน (uncertainty) ของพารามิเตอร์

การเรียนรู้พารามิเตอร์แบบเบย์ (Bayesian parameter learning)

เราสามารถมองปัญหาการเรียนรู้พารามิเตอร์เป็นปัญหา การอนุมานแบบเบย์ (Bayesian inference) ได้:

- ทำให้พารามิเตอร์ θ เป็นตัวแปรสุ่ม (random variables) และมองว่าเป็นตัวแปรซ่อนเร้น
- กำหนดการแจกแจง ก่อนหน้า (prior) $\mathbf{P}(\theta)$ ให้กับพารามิเตอร์
- จากนั้น เมื่อมีข้อมูลเข้ามา เราจะอัปเดตความเชื่อของเราเกี่ยวกับพารามิเตอร์เพื่อหาการแจกแจง ภายหลัง (posterior) ${f P}(\theta|\mathbf{d})$

ในมุมมองนี้ พารามิเตอร์ไม่ใช่ค่าคงที่ที่เราพยายามจะหาค่าที่ดีที่สุดค่าเดียวอีกต่อไป แต่เป็นตัวแปรสุ่มที่มีการแจกแจงความน่าจะเป็นของตัวเอง ซึ่งจะเปลี่ยนแปลงไปเมื่อเราได้รับข้อมูลมากขึ้น

กรณี (a)

สัดส่วน θ ของลูกอมรสเชอร์รี่เป็นเท่าใด?

เราสมมติ prior แบบ Beta:

$$\mathrm{P}(\theta) = \mathrm{Beta}(\theta|\mathrm{a},\mathrm{b}) = \frac{1}{\mathrm{Z}}\theta^{\mathrm{a}-1}(1-\theta)^{\mathrm{b}-1}$$

โดยที่ Z คือค่าคงที่เพื่อให้เป็นปกติ (normalization constant)

จากนั้น เมื่อสังเกตเห็นลูกอมรสเชอร์รี่ จะได้ posterior:

$$\mathrm{P}(\theta|\mathrm{cherry}) \propto \mathrm{P}(\mathrm{cherry}|\theta)\mathrm{P}(\theta) \quad = \theta \cdot \mathrm{Beta}(\theta|\mathrm{a},\mathrm{b}) \propto \theta \cdot \theta^{\mathrm{a}-1}(1-\theta)^{\mathrm{b}-1} \quad = \theta^{\mathrm{a}}(1-\theta)^{\mathrm{b}-1} = \mathrm{Beta}(\theta|\mathrm{a}+1,\mathrm{b})$$

การสังเกตข้อมูลใหม่ จะเป็นการอัปเดตพารามิเตอร์ของ Beta distribution

กรณี (b)

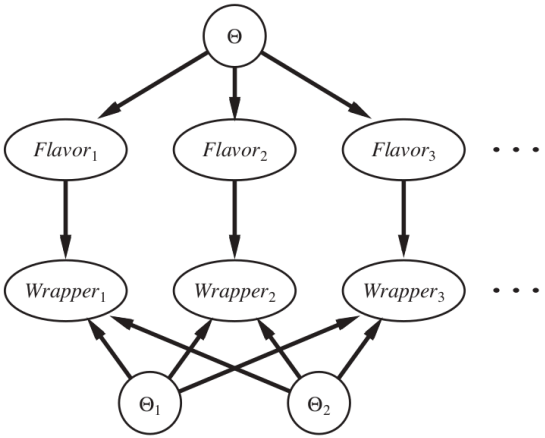


Figure 20.6 A Bayesian network that corresponds to a Bayesian learning process. Posterior distributions for the parameter variables Θ, Θ_1, and Θ_2 can be inferred from their prior distributions and the evidence in the $Flavor_i$ and $Wrapper_i$ variables.

การประมาณค่า Posterior สูงสุด (Maximum a posteriori - MAP)

เมื่อไม่สามารถคำนวณ posterior แบบ (analytically) ได้ เราสามารถใช้การประมาณค่า Maximum a posteriori (MAP) \ ซึ่งประกอบด้วยการประมาณค่า posterior ด้วยจุดประมาณ θ^* ที่ทำให้การแจกแจง posterior มีค่าสูงสุด นั่นคือ

$$ \begin{aligned} \theta^* = \arg \max\theta P(\theta|\mathbf{d}) = \arg \max\theta P(\mathbf{d}|\theta) P(\theta) \end{aligned}$$

MAP คล้ายกับ MLE แต่เพิ่มพจน์ของ prior $\mathrm{P}\left(\theta\right)$ เข้ามา ซึ่งทำหน้าที่เหมือน "regularizer" ป้องกันไม่ให้ค่าพารามิเตอร์สุดโต่งเกินไปเมื่อมีข้อมูลน้อย

สรุป

- โครงข่ายเบย์ (Bayesian Network) ระบุการแจกแจงร่วมทั้งหมด และมักจะมีขนาดเล็กกว่าการแจกแจงร่วมที่แจกแจงอย่างชัดเจนแบบทวีคูณ
- โครงสร้าง (Topology) ของโครงข่ายเบย์เข้ารหัสสมมติฐานความเป็นอิสระแบบมีเงื่อนไขระหว่างตัวแปรสุ่ม
- การอนุมาน (Inference) คือปัญหาของการคำนวณการแจกแจงความน่าจะเป็นแบบ marginal และ/หรือ conditional
  - การอนุมานแบบแม่นตรง (Exact inference) สามารถทำได้สำหรับโครงข่ายเบย์อย่างง่าย แต่ไม่สามารถทำได้ในทางปฏิบัติสำหรับแบบจำลองส่วนใหญ่
  - อัลกอริทึมการอนุมานเชิงประมาณ (Approximate inference) ถูกนำมาใช้ในทางปฏิบัติ
- พารามิเตอร์ ของโครงข่ายเบย์สามารถเรียนรู้ได้จากข้อมูลโดยใช้ Maximum Likelihood Estimation (MLE) หรือ Bayesian inference

จบ