คำอธิบายจักรกลเรียนรู้แบบข้อต้านความจริง

(Counterfactual explanations)

ศิระกร ลำใย
กลุ่มวิจัยเชิงทฤษฎี ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

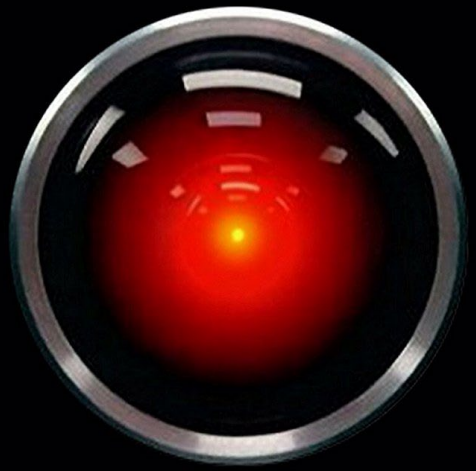
ทำไมเราต้องการคำอธิบายให้ปัญญาประดิษฐ์

กรณีศึกษา: Apple Card

- Apple Card ให้วงเงินบัตรเครดิตกับผู้ชายมากกว่าผู้หญิง
  - กรณีนี้เกิดขึ้นแม้กับคู่ชาย-หญิงที่ใช้เอกสารทางการเงินร่วมกันทั้งหมดด้วยซ้ำ
- บริษัท Goldman Sachs ผู้ออกบัตรบอกว่า “In fact, we do not know your gender or marital status during the Apple Card application process.”

กรณีศึกษา: ทำไมทิ้งกัน

- มาตรการเยียวยา 5,000 บาท ของเราไม่ทิ้งกัน เคลมว่าใช้ AI คัดคน 27 ล้านคน
  - เป็น AI ปริศนา?
  - Dataset มาจากไหน?
  - การคัดแบบนี้ควรจะเป็น data driven จริงๆ หรือ?



สิทธิ์ต่อคำอธิบาย

(Rights to explanation)

สรุปต้นตอแห่งความ*กระเสือกกระสน*

ความเป็นธรรม

ความเป็นส่วนตัว

ความน่าเชื่อถือ (reliability)

ความเป็นเหตุเป็นผล

ความเชื่อใจ (trust)

ว่าด้วยการอธิบายจักรกลเรียนรู้

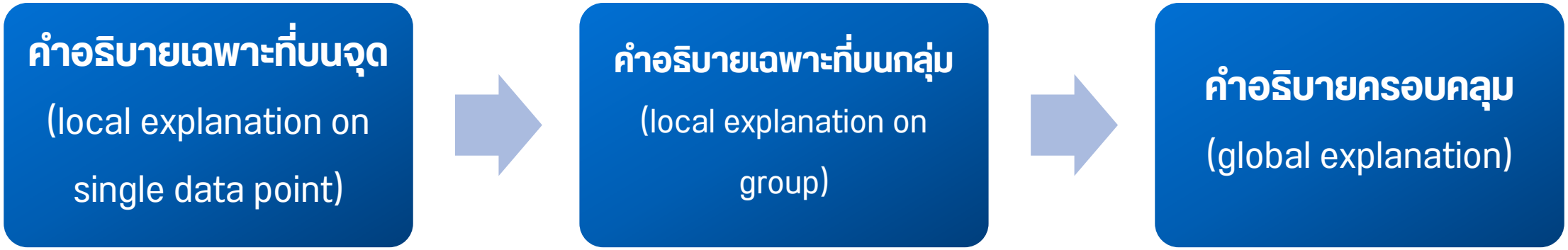
เพราะเรา**ตีความ**จักรกลเรียนรู้ได้

- จักรกลเรียนรู้**ไม่ใช่กล่องดำเสมอไป**–บางครั้งคราว เราสามารถ “แงะ” กล่องดำออกมาส่องดูข้างในได้เช่นกัน
- ศาสตร์แห่งจักรกลเรียนรู้ที่ตีความได้มุ่งเน้นศึกษาวิธี “งัด” กล่องดำข้างในออกมาดู

การตีความจักรกลเรียนรู้

- ว่าด้วยวิธีการตีความ
  - **ภายใน (intrinsic):** สร้างข้อจำกัดให้แบบจำลองไม่ซับซ้อนเกินไป จะได้อธิบายได้
  - **ภายหลัง (post-hoc):** สร้างๆ แบบจำลองมาก่อน เดี๋ยวค่อยอธิบาย
- ว่าด้วยการนำขั้นตอนวิธีไปใช้
  - **ไม่ขึ้นกับแบบจำลอง (model agnostic):** วิธีนี้ปรับใช้ได้กับแบบจำลองทุกประเภท
  - **ขึ้นกับแบบจำลอง (model dependent):** วิธีนี้ใช้ได้กับแบบจำลองบางประเภทเท่านั้น

ระดับของขั้นตอนวิธีการอธิบาย



- ไม่ใช่ว่าทุกคำอธิบายจะมีค่าเท่ากันหมด
  - คำอธิบายบางรูปแบบ ใช้อธิบายแบบจำลองได้แค่นิดๆ หน่อย
  - คำอธิบายบางรูปแบบ อธิบายได้เฉพาะพฤติกรรมของ instance หนึ่งของข้อมูลเท่านั้น
- ปัญหาการอธิบายให้ได้ครอบคลุม (เราอยากได้ local explanation ยังคงเป็นปัญหาที่นักวิจัยพยายามตีให้แตกอยู่เหมือนกัน)

รูปแบบผลการตีความจักรกลเรียนรู้

- ชุดพารามิเตอร์
- ชุดฟีเจอร์
- จุดข้อมูล

คำอธิบาย*ที่ดี*

- บ่งบอกความแตกต่าง
- ถูกเลือกมาเป็นคำอธิบาย
- อิงกับบริบททางสังคม
- อิงกับกรอบวิถีแห่งความปกติ

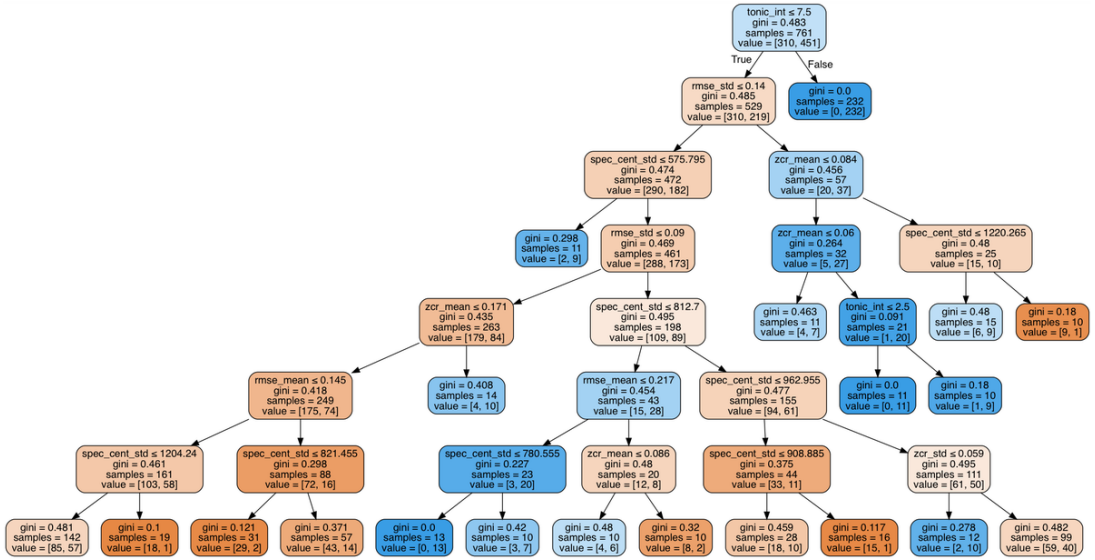
มาอธิบายแบบจำลองกันเถอะ!

การถดถอยเชิงเส้น (linear regression)

- อธิบายฟีเจอร์ด้วยพารามิเตอร์แบบจำลอง
  - กรณีฟีเจอร์เป็นค่าต่อเนื่อง
    - w_i บอกผลกระทบที่ y จะเปลี่ยนเมื่อ x_i เพิ่มขึ้นหนึ่งหน่วย
  - กรณีฟีเจอร์เป็นข้อมูลหมวดหมู่
    - w_i บอกผลกระทบที่หากมี c_i แล้วทำให้เกิดการเปลี่ยนแปลงของ y_i
  - ไบแอส (b, w_0)
- อธิบายความสำคัญของฟีเจอร์ ผ่านการทดสอบ t-statistic

ต้นไม้ตัดสินใจ (decision tree)

- Rule-based AI คือ AI ที่อธิบายได้ในระดับหนึ่ง
- ถ้าเรากำหนดชั้นของ tree น้อยๆ เราก็ย่อมพออธิบายแบบจำลองของเราผ่านการแตก tree ได้
  - อย่าลืมว่าจำนวนใบที่เป็นไปได้มากที่สุด โตเป็น exponential ของจำนวนชั้นของต้นไม้

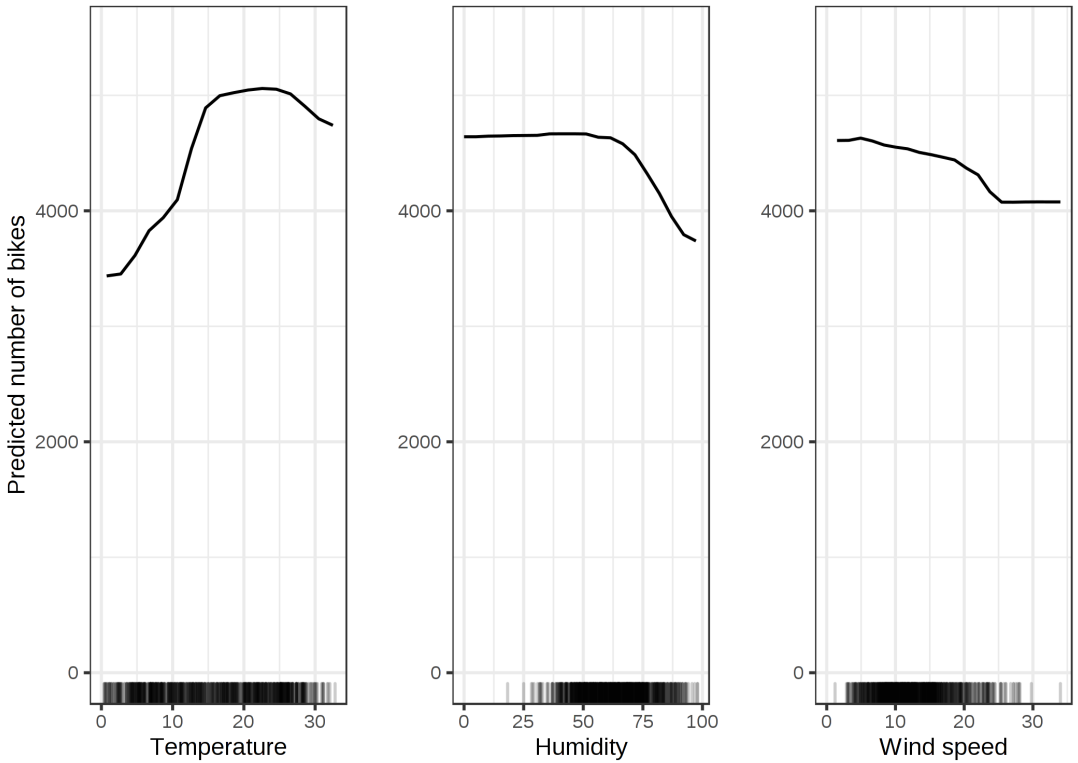


ขั้นตอนวิธีการอธิบายแบบจำลอง

Partial Dependency Plot

- หากมีแบบจำลอง พยายามหาความเกี่ยวโยงว่าการเปลี่ยนแปลงค่าของฟีเจอร์หนึ่งส่งผลต่อแบบจำลองอย่างไร
- ใช้วิธีการมอนที-คาร์โล ในการประมาณฟีเจอร์ตัวที่เราไม่ได้สนใจ

$$\hat{f}_{x_s}(x_s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(x_s, x_c^{(i)}\right)$$

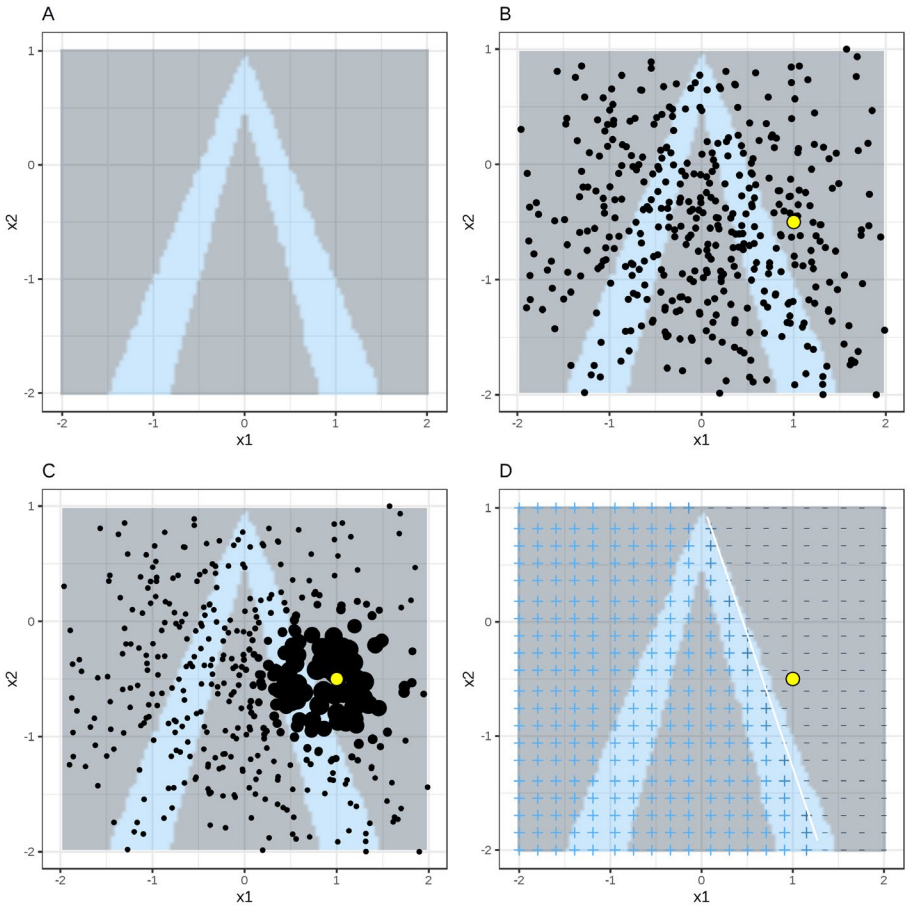


การปฏิสัมพันธ์ของฟีเจอร์ (feature interaction)

- Friedman เสนอ H-statistic มาวัดการปฏิสัมพันธ์ของฟีเจอร์สองตัว

LIME (Local Interpretable Model-Agnostic Explanation)

- คำอธิบายเกิดจากการฝึกสอนแบบจำลองที่ซับซ้อนน้อยกว่าและเข้าใจได้ง่ายกว่า
  - พยายาม minimise ค่าสองค่า: loss ของแบบจำลองนั้นรอบจุดที่จะอธิบาย และฟังก์ชันสำหรับพิจารณาความซับซ้อนของแบบจำลอง
- **Local explanation: วิษณุ เครืองาม**
  - ทำไมเรายังอยู่เฟสสอง อ๋อเพราะเราใช้เกณฑ์ของเราเอง...
  - ทำไมจุด x ตอบแบบนี้ อ๋อเพราะรอบๆ จุด x มันเป็นแบบนี้...
  - ทำไมจุด y ตอบแบบนี้ อ๋อเพราะจริงๆ ถ้าเราดูใกล้ๆ จุด y...



คำอธิบายข้อต้านความจริง
(Counterfactual examples)

คำอธิบายข้อต้านความจริง

“โดนปฏิเสธสินเชื่อเพราะมีเงินเดือน 50,000 บาท ถ้ามีเงินเดือนสัก 70,000 บาทก็ไม่โดนปฏิเสธแล้ว”

- ให้นึกถึง if clause แบบที่สามในไวยากรณ์ภาษาอังกฤษ
  - “ถ้า” กับเรื่องที่ไม่ได้เป็นจริง และ “แล้ว” ไม่ได้เกิดขึ้นจริง
- เสนอความขึ้นกัน (dependency) กับข้อมูลภายนอก (ในที่นี้คือเงินเดือน) ที่ส่งผลต่อการตัดสินใจของแบบจำลอง

ผมสร้าง**โลก**ขึ้นมาสองใบ

“โดนปฏิเสธสินเชื่อเพราะมีเงินเดือน 50,000 บาท ถ้ามีเงินเดือนสัก 70,000 บาทก็ไม่โดนปฏิเสธแล้ว”

- โลกที่เกิดขึ้นจริง (มีเงินเดือน 50,000 บาทจริง) และโลกสมมติ (ที่เรามีเงินเดือน 70,000 บาท)
- โลกสมมติอยู่บนความเปลี่ยนแปลงที่ทำให้เกิดผลลัพธ์ที่แตกต่างจากที่เป็นอยู่
- คำอธิบายเสนอได้ทั้ง “โลกที่ใกล้ที่สุด” และ “โลกที่เป็นไปได้”
  - ผมสร้างโลกขึ้นมาหลายใบ...

บนปรัชญาแห่งความรู้

ผมรู้ว่าเค้าน่ารัก

(แต่ผมไม่รู้ว่าเค้าไหน)

ผู้รู้ (subject)

ความรู้ (knowledge)

บนปรัชญาแห่งความรู้

Propositional knowledge

- ความรู้คือ “ความเชื่อที่เป็นจริงและมีเหตุอันสมควร” (justified true beliefs)
  - Beliefs: ต้องเชื่อ
  - True: สิ่งที่เชื่อต้องเป็นจริง
  - Justified: มีเหตุผลที่ดีในการเชื่อ

Sensitivity

- “ถ้า p เป็นเท็จ S จะไม่เชื่อ p”
  - “ถ้า p เป็นเท็จ” คือคำอธิบายข้อต้านความจริง!
    - เสนอ “โลก” ที่ p เป็นเท็จ และพยายามอธิบายโลกนั้นให้ได้

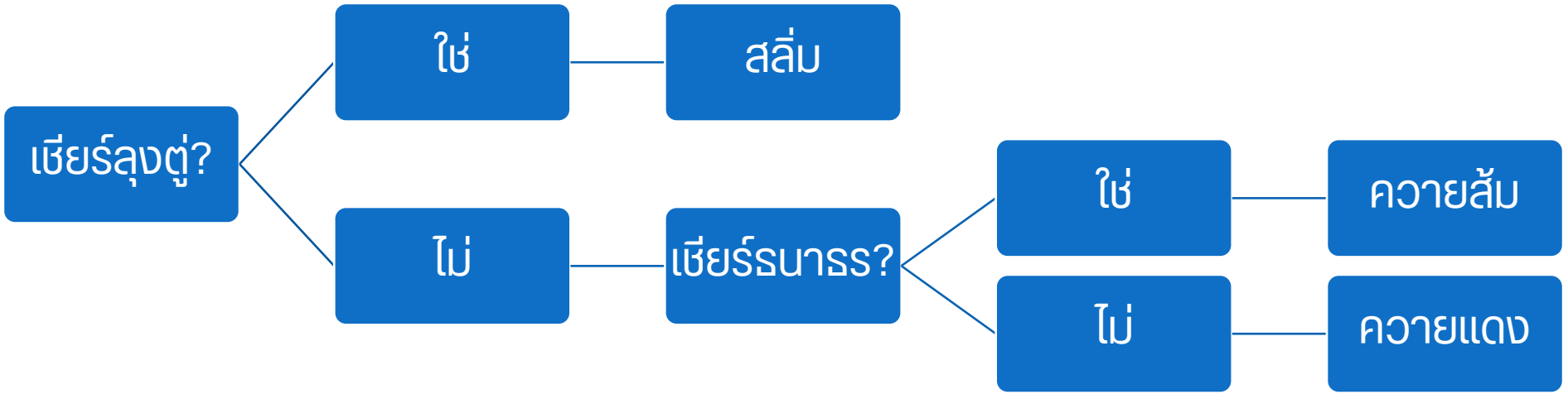
นิยามของคำอธิบายข้อต้านความจริง

นิยามของคำอธิบายต่อต้านความจริง

- Result p was returned because of the values V.
- If V instead had values V' and other variables had remained constant, p' would have been returned.

- จะสังเกตเห็นว่ามีค่า V' มากมายที่เป็นไปได้
  - ถ้ามีเงินเดือน 2.2 ล้านล้านบาทก็กู้สินเชื่อได้ (แต่ใครมันจะมี?)
- เราสนใจหา "โลกที่ใกล้ที่สุด" หรือไม่ก็ "โลกหลายๆ ใบ"

คำอธิบายข้อต้านความจริงบนแบบจำลองการเรียนรู้แบบเขียนกฎ (rule-based)



- จากต้นไม้ตัดสินใจนี้ เอามาเขียนเป็นกฎได้ว่า “ถ้าเชียร์ลุงตู่ = N และเชียร์ธนาธร = Y แล้ว เป็นควายส้ม”
- คำอธิบายจะอยู่ในรูปแบบคล้ายๆ กับ “ถ้าเชียร์ลุงตู่ = N หรือเชียร์ธนาธร = N ก็ไม่เป็นควายส้มหรอก”
  - เสนอ “โลกอีกใบ”

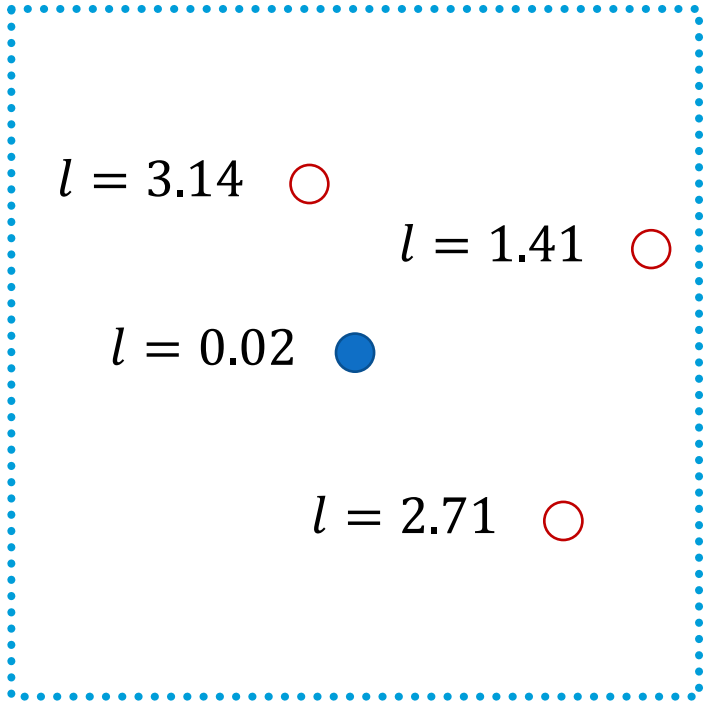
การเรียนรู้ประสงค์ร้าย
(Adversarial learning)

การเรียนรู้ประสงค์ร้าย 101

https://www.facebook.com/photo?fbid=10218655842060241&set=gm.3018514041545705

- หมาโดนเงาตกใส่ โดนมองผิดว่าเป็นเสือ
- เงาคือ**สัญญาณโจมตี (perturbations)** ที่ใส่ในหมาแล้วทำให้แบบจำลองคืนค่าคำตอบที่ไม่ควรเป็น

การคำนวนสัญญาณโจมตี



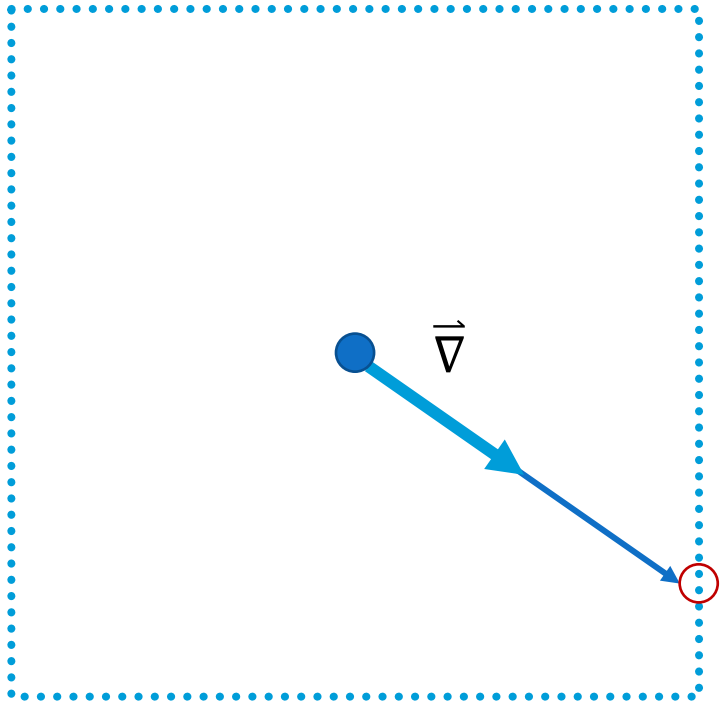
กรอบของนอร์ม

Given an input to be attacked that lies in an input space...

- Define the “invisibility” measurement
  - **Norm or other constraints**
- Find the perturbation which maximise such loss function within the constrained norm
  - **Optimisation problem**
- There exists many perturbations, but their *power* may not be equal

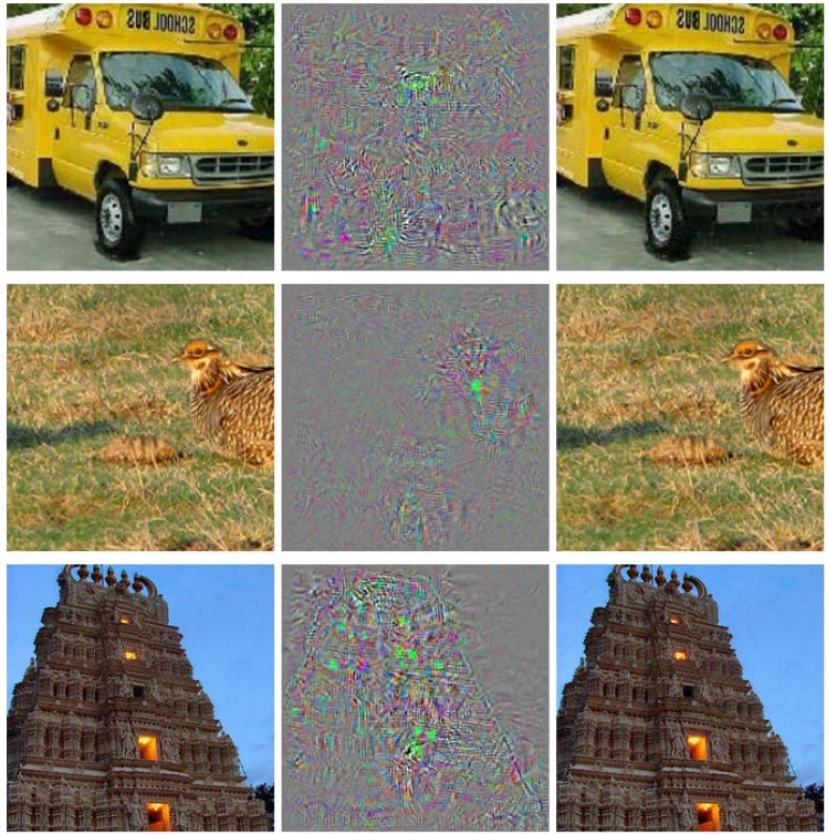
Fast Gradient Sign Method (FGSM)

[Goodfellow+ 2014, arXiv: 1412.6572]



- คำนวนเกรเดียนต์ของข้อมูลรับเข้าเทียบกับฟังก์ชันสูญเสีย
- ฉายไปสุดกรอบของนอร์ม
- เป็นการประมาณการเพิ่มค่า loss ให้สูงที่สุดที่เป็นไปได้
- เวลารันคงที่ ไม่มีกระบวนการวนซ้ำ (iterative)

Adversarial VS Counterfactual



Szegedy et al. *Intriguing Properties of Neural Networks*. ICLR '14 (https://arxiv.org/abs/1312.6199v4)

- การโจมตีประสงค์ร้ายไม่สามารถใช้อธิบายอะไรได้เท่าไหร่มากนัก
  - ขั้นตอนวิธีไม่ได้ penalise จำนวนตัวแปรที่ถูกแก้ไข ทำให้เกิดการแก้ไขตัวแปรจำนวนเยอะมากๆ
    - ในที่นี้คือแก้พิกเซลของรูปเท่าไหร่ก็ได้
  - การแก้ไขลักษณะนี้ทำให้เสียคุณลักษณะของการอธิบายได้ไป

คำอธิบายข้อต้านความจริงเพื่อตรวจสอบความยุติธรรม

- คำอธิบายข้อต้านความจริงอาจใช้ในการตรวจสอบ**ความเป็นธรรม (fairness)** ของขั้นตอนวิธีได้
  - อย่างน้อยก็ใช้ตรวจสอบการเลือกปฏิบัติ (discrimination) บนข้อมูลบางชุด (เช่นเชื้อชาติ) ได้

การสร้างคำอธิบาย ข้อต้านความจริง

เทรนแบบจำลอง

$$\underset{\theta}{\operatorname{argmin}}\, l(f_\theta(x_i), y_i\,) + p(\theta)$$

หาพารามิเตอร์แบบจำลอง θ ที่ลดผลรวมค่าสูญเสีย (loss) และค่าการปรับปกติ (regularise)

หาโลกอีกใบ

$$\underset{x'}{\operatorname{argmin}} \max_{\lambda} \lambda(f_\theta(x') - y')^2 + d(x_i, x')$$

- อยากได้ x' ที่ $f_\theta(x') = y'$
- พยายามหา x' ที่ใกล้กับ x_i มากที่สุดด้วยการใช้ฟังก์ชันระยะทาง d
- เทอม λ ทำอะไร?
  - ยิ่ง λ มาก เทอม $\lambda(f_\theta(x') - Y)^2$ จะมีค่ามากตาม ดังนั้นระยะทาง d จะต้องน้อย และทำให้ x_i กับ x' ใกล้กันขึ้น
  - ในทางปฏิบัติ อาจทำได้โดยการค่อยๆ เพิ่ม λ แล้ว solve ค่า x' จนถึงจุดที่เพิ่ม λ ไม่ได้แล้ว
    - หรือเพิ่ม λ จนเราได้ x_i, x' ที่ใกล้กันมากพอ
    - เดี๋ยวเรามาพูดถึงกันอีกที

ฟังก์ชันระยะทาง

$$\underset{x'}{\operatorname{argmin}} \max_{\lambda} \lambda(f_\theta(x') - y')^2 + d(x_i, x')$$

- ฟังก์ชัน d ควรจะมีความหมายในแง่ใดบ้าง?
  - คิดอะไรไม่ออกนอกจากคิดอยู่ใช้นอร์ม L_1
- สมมติพิจารณาฟีเจอร์ k บนข้อมูลของเรา...
  - ถ้าหากว่าฟีเจอร์ k มีความหลากหลาย (vary) บนข้อมูลของเรา จุด x_i และ x' ก็อาจจะอยู่ใกล้กันได้แม้ว่าค่า k จะต่างกันมากๆ
- เราจะผสานสองตรงนี้กันได้หรือเปล่านะ?

MAD (Median Average Distant)

$$\mathrm{MAD}_k = \underset{j \in P}{\mathrm{median}} \left(\left| X_{j,k} - \underset{l \in P}{\mathrm{median}} \left(X_{l,k} \right) \right| \right)$$

- MAD ของฟีเจอร์ k บนเซตของจุด P คำนวนได้จากฟังก์ชันดังแสดง

Distant Function

$$d(x, x') = \sum_{k \in P} \frac{|x_{i,k} - x'_k|}{\mathrm{MAD}_k}$$

- อย่าลืมว่าเราปรับ distant function ให้เหมาะสมเองได้นะ

การหาคำอธิบาย

$$\underset{x'}{\operatorname{argmin}} \max_{\lambda} \lambda(f_\theta(x') - y')^2 + d(x_i, x')$$

- เลือก x_i มาอธิบาย ถามว่าจะทำอย่างไรให้ได้ผลลัพธ์ y'
- ให้ $\lambda = 0$
- วนซ้ำ...
  - เพิ่มค่า λ
  - Optimise ค่า x'
  - Terminal statement: ไม่สามารถเพิ่มค่า λ ได้อีกแล้ว
- Return x' ตัวล่าสุด (หรือ x' ที่เคยหาได้ทั้งหมด)

Technical implementation

$$\underset{x'}{\operatorname{argmin}} \max_{\lambda} \lambda(f_\theta(x') - y')^2 + d(x_i, x')$$

- ถ้าเราสามารถหาเกรเดียนต์ของฟังก์ชันนี้เทียบกับ x' ได้ เราสามารถใช้ตัว optimiser ที่ขึ้นกับเกรเดียนต์ได้ (ในเปเปอร์แนะนำ Adam–อาห์ ความมหัศจรรย์ของโลก)
- ถ้าเราไม่สามารถหาเกรเดียนต์ดังกล่าวได้ ก็ใช้วิธีอื่น (อย่าง Nelder-Mead)

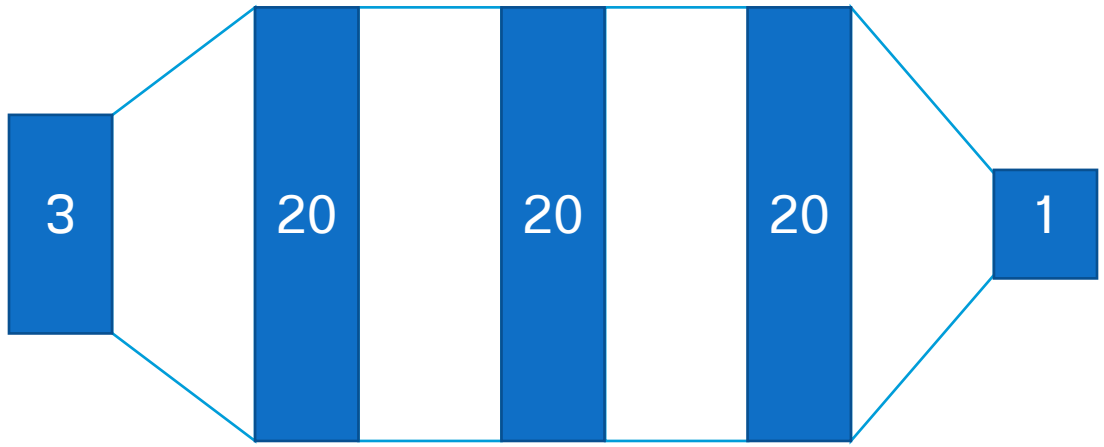
ตัวอย่างการตรวจสอบอุคติในแบบจำลองการเรียนรู้

LSAT Dataset (แบบตัดทอน)

- พยายามทำนายเกรดนิสิตปีหนึ่งจากแฟกเตอร์สามตัว
  - GPA ก่อนเป็นนิสิต
  - คะแนนสอบเข้า
  - เชื้อชาติ
- ต้องการจะ “อธิบาย” ว่าทำอย่างไรจึงจะได้คะแนน (normalised) เป็น 0?
  - ก็คือทำอย่างไรถึงจะได้คะแนนเป็นค่าเฉลี่ย

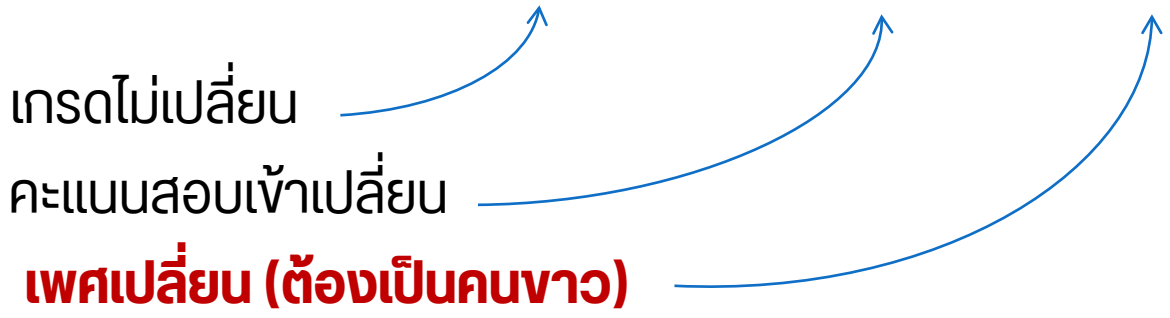
แบบจำลอง

- FCNN แบบ hidden layer สามชั้น
  - แต่ละชั้นมีนิวรอน 20 ตัว
- จะมีน้ำหนักทั้งหมด 880 ค่า และไบแอสอีก 64 ค่า–รวมเป็น 941 ค่า
  - แค่แบบจำลองที่ไม่ได้ซับซ้อนมากแบบนี้ ก็ยากที่จะแงะดูข้างในแล้ว



ผลลัพธ์

Score	GPA	LSAT	Race	GPA x’	LSAT x’	Race x’
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0



ลองเล่นกับแพคเกจ Alibi ในไพทอน

ทำความรู้จักกับ Alibi

- ไลบรารีไพทอนสำหรับการตีความ (interpret) และสำรวจ (inspect) แบบจำลอง
- เน้นการพิจารณาแบบจำลองที่เป็น black box
- เน้นการอธิบายแบบ instance based

ALIBI

เสา

อ้างอิง

- Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.
- Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *ArXiv:1711.00399 [Cs]*. http://arxiv.org/abs/1711.00399.