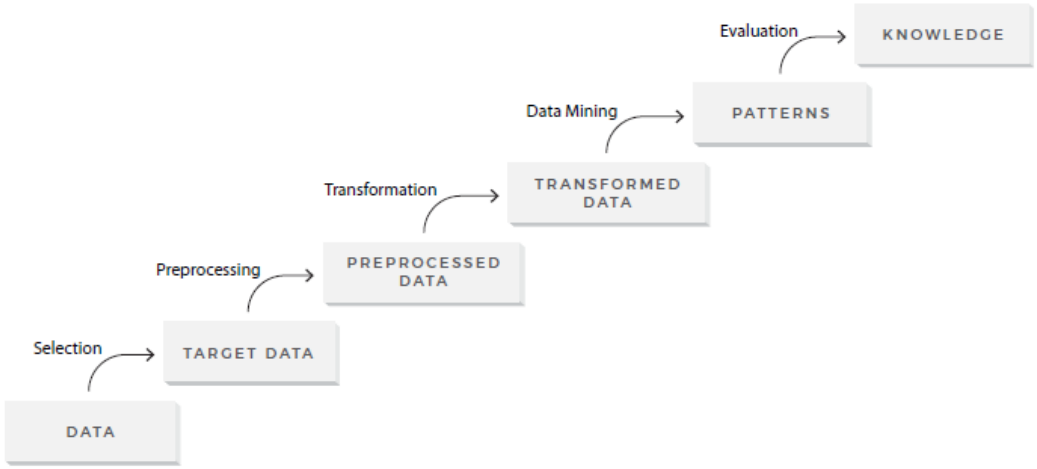
ข้อสอบมิดเทอมวิชา Data Mining

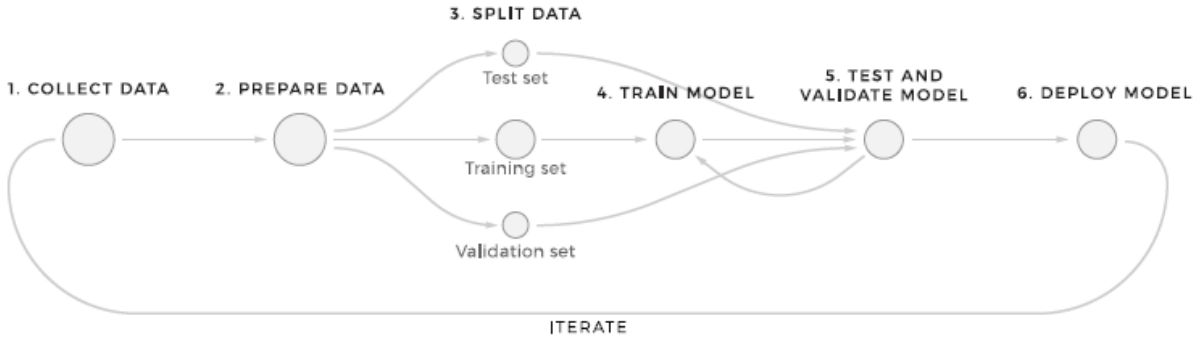
1. บอกความหมายของคำว่า Data Mining และอธิบาย Knowledge Discovery in Databases (KDD)
   - Data Mining เป็นการค้นหาส่วนที่เป็นข้อมูลสำคัญหรือมีประโยชน์จากชุดข้อมูลที่เราให้ความสนใจ ซึ่งเป็นส่วนหนึ่งในเทคนิค KDD
   - KDD เป็นเทคนิคเพื่อค้นหารูปแบบที่พอจะนำไปประมวลผลต่อได้ (digestible pattern) ของข้อมูลจากข้อมูลจำนวนมหาศาลโดยใช้ขั้นตอนทาง Statistics, Machine Learning และ Pattern Recognition



Knowledge Discovery in Databases

2. บอกความหมายของคำว่า Machine Learning
   - เป็นการสร้างอัลกอริทึมเพื่อดึงเอาส่วนที่เป็นข้อมูลสำคัญหรือข้อมูลที่เราสนใจจากข้อมูลดิบที่มีขนาดใหญ่ และมีการเปลี่ยนแปลงของข้อมูลอยู่เสมอ โดยการเรียนรู้นั้นสามารถปรับเปลี่ยนตามข้อมูลที่เข้ามาใหม่ได้โดยอัตโนมัติ

3. อธิบายขั้นการทำงานของ Machine Learning ทีละขั้นตอน



Machine Learning Workflow

- Collect data หาแหล่งข้อมูลเพื่อใช้ในการรวบรวมข้อมูลดิบที่คิดว่ามีประโยชน์กับเป้าหมายของเราให้มากที่สุดเท่าที่ทำได้ แล้วจัดเก็บเป็นชุดข้อมูลหนึ่ง
- Prepare data ชุดข้อมูลอาจมีบางส่วนที่ขาดหายไป หรือมีความไม่ถูกต้องของข้อมูลอยู่ จึงต้องจัดเตรียมข้อมูลให้สมบูรณ์ถูกต้อง
- Split data แบ่งชุดข้อมูลเป็น 3 ส่วน โดยจำนวนข้อมูลต้องมีจำนวนที่มากพอในการแบ่งด้วย
  1) Training set
  2) Test set
  3) Validation set
- Train a model ใช้ training set ใส่ลงไปในอัลกอริทึมหนึ่งของโมเดลเพื่อให้เกิดการเรียนรู้รูปแบบของข้อมูล
- Test and validate a model ประเมินโมเดลว่ามีความแม่นยำขนาดไหนในการทำนายด้วย test set และ validation set ถ้ายังได้ความแม่นยำที่ยังไม่เป็นที่พอใจตามเป้าหมายก็สามารถย้อนกลับไปปรับโมเดลในขั้นตอน train a model ได้
- Deploy a model นำโมเดลที่พอใจไปใช้ในการช่วยตัดสินใจของงานนั้นๆ ตามความต้องการ
- Iterate วนกลับไปรวบรวมข้อมูลใหม่ และทำขั้นตอนในการ train a model อีกครั้ง เพื่อปรับปรุงโมเดลให้ดีขึ้น

4. ในขั้นตอนการ Train a model เราก็สามารถรู้ accuracy ของ model ได้แล้วทำไมถึงต้องทำการ Test and validate a model เพื่อวัดผลอีกที
   - การ Train a model นั้นใช้เพียง training set ให้โมเดลเกิดการเรียนรู้ซึ่งโมเดลนั้นจะเรียนรู้เพียงรูปแบบของข้อมูลชุดที่เทรนเท่านั้น ทำให้ต้องใช้ Test and validate a model ที่ใช้ data set อีกชุดหนึ่งที่โมเดลไม่เคยเห็นมาก่อนมาทดสอบเพื่อทำให้มั่นใจว่าโมเดลนั้นเกิดความแม่นยำจากการเรียนรู้จริง ไม่ใช่การจำ และยังทำเพื่อปรับปรุงให้โมเดลนั้นมีความแม่นยำเพิ่มมากขึ้นได้อีกด้วย
5. เมื่อไหร่เราถึงจำเป็นให้ Model มีการเรียนรู้แบบ Unsupervised learning
   - ใช้กับชุดข้อมูลที่ยังไม่ถูกจัดกลุ่มหรือยังไม่ทราบแน่ชัดว่าควรจัดกลุ่มอย่างไร
6. อธิบายการเรียนรู้แบบ Reinforcement learning
   - เป็นการเรียนรู้ด้วยอัลกอริทึมหนึ่ง โดยเมื่อมีข้อมูลเข้ามาให้ตัดสินใจหรือทำนายข้อมูลนั้นๆ ถ้าตัดสินใจหรือทำนายได้ถูกต้องจะได้รับรางวัล (awarded) แต่ถ้าตัดสินใจหรือทำนายพลาดจะถูกลงโทษ (punished) ซึ่งการเรียนรู้ด้วยอัลกอริทึมแบบนี้ต้องกำหนดกฎ เป้าหมาย และสิ่งแวดล้อมที่ตายตัว ทำให้ส่วนใหญ่เหมาะกับการถูกใช้งานในด้านเกม
7. Classification ใช้กับปัญหาแบบไหน และยกตัวอย่างปัญหามา 3 ข้อ
   - ใช้กับปัญหาที่รู้แล้วว่าต้องการจัดกลุ่มชุดข้อมูลหนึ่งไปเป็นกลุ่มอะไรได้บ้าง (binary classification problems หรือ multiclass problems) หรืออาจใช้กับปัญหาที่ต้องการบอกว่าใช่กลุ่มนี้หรือไม่ (one-class classification)

   <u>ตัวอย่าง</u> Is this email spam or not?

   Is this picture a cat, a dog, or a bird?

   Can we spot unusual behaviors among our bank clients?

8. Cluster analysis ใช้กับปัญหาแบบไหน และยกตัวอย่างปัญหามา 3 ข้อ
   - ใช้กับปัญหาที่ชุดข้อมูลยังไม่ถูกจัดกลุ่มหรือยังไม่ทราบแน่ชัดว่าควรจัดกลุ่มอย่างไร ส่วนใหญ่จะเป็นปัญหาที่มีชุดข้อมูล domain ใหม่ๆ

   <u>ตัวอย่าง</u> How can we classify the keywords that people use to reach our website?

   Is there any relationship between default risks of some bank clients and their behaviors?

What are the main segments of customers we have considering their demographics and behaviors?

9. Regression ใช้กับปัญหาแบบไหน และยกตัวอย่างปัญหามา 3 ข้อ
   - ใช้กับปัญหาที่ชุดข้อมูลเป็นตัวเลขและต้องการทำนายออกมาเป็นตัวเลขเช่นกัน ส่วนใหญ่จะถูกใช้ในการทำนายความต้องการสินค้า ตัวเลขยอดขาย และค่าตอบแทนกำไรจากตลาด

   <u>ตัวอย่าง</u> How many items of this product will we be able to sell next month?

   What's will the airfare be for this destination?

   What's going to be the rental price for this house?

10. Prepare data จำเป็นต้องทำหรือไม่ ส่งผลต่อ Model อย่างไร จงอธิบาย
    - จำเป็น เนื่องจากชุดข้อมูลอาจมีบางส่วนที่ขาดหายไป หรือมีความไม่ถูกต้องของข้อมูลอยู่ จึงต้องจัดเตรียมข้อมูลให้สมบูรณ์ถูกต้องก่อนนำไปเทรนโมเดลเพื่อให้โมเดลนั้นเกิดการเรียนรู้ที่ถูกต้องตรงตามความต้องการ
11. อธิบายคำว่า siloed data และ anonymized data
    - Siloed data เป็นข้อมูลที่ไม่มีความสัมพันธ์กันและถูกเก็บแยกส่วนไว้คนละที่ เช่น ข้อมูลของแต่ละแผนก
    - Anonymized data เป็นข้อมูลที่ถูกทำให้เป็น unknown เพื่อปกป้องไม่ให้มีการใช้หรือเปิดเผยข้อมูลของเจ้าของข้อมูลโดยตรง ตามกฎหมาย