Top 10 algorithms in data mining

1. C4.5

- เป็นอัลกอริทึมประเภท classification ในรูปแบบ decision tree

-มีความต่างกับ CART ที่ CART มี output ได้แค่ 2 รูปแบบแต่ C4.5 มีได้มากกว่านั้น, CART ใช้ gini index ส่วน C4.5 ใช้ information gain, มีการคำนวณความซับซ้อนที่แตกต่างกัน

- มีข้อเสียคือ ใช้ memory สูง

- ปัญหา error rate ของการนำไปใช้จริงจะเพิ่มขึ้น, การลดความซับซ้อนโดยที่ยังคงความแม่นยำ

2. k-Means

- เป็นอัลกอริทึมประเภท clustering

- มีหลักการ คือ กำหนดจุดศูนย์กลาง->จัดกลุ่ม->หาจุดศูนย์กลาง->จัดกลุ่มวนไปเรื่อยๆ

- ข้อจำกัดสามารถจัดกลุ่มได้แค่ในรูปแบบกลุ่มวงกลมเท่านั้น

3. SVM

- เป็นอัลกอริทึมประเภท classification

- มีหลักการโดยหาเส้นแบ่งกลุ่ม

4. Apriori

- เป็นอัลกอริทึมสำหรับหาความสัมพันธ์ของข้อมูล

- หลักการจัดคู่ข้อมูลแล้วหาโอกาสเกิดนำคู่ที่มีโอกาสเกิดมากกว่าที่กำหนดมาจับคู่อีก ผลลัพธ์จะเป็นรูปแบบของข้อมูลที่มีความเกี่ยวข้องกัน

5. EM

- เป็นอัลกอริทึมประเภท clustering

- จัดกลุ่มโดยใช้หลักการความน่าจะเป็น

- สามารถจัดการกับข้อมูลสูญหายได้

6. PageRank

- เป็นอัลกอที่ใช้ใน search engine

- มีหลักการ คือ หาหน้าที่ถูกอ้างอิงโยหน้าอื่นมากที่สุด

7. AdaBoost

- เป็น classification ที่ใช้โมเดล classification อื่น ๆ มาประมวลผล

- หลักการคือ นำโมเดลทั้งหมดมาประมวลผลคำตอบโดยแต่ละโมเดลจะมีค่าถ่วงน้ำหนักที่ขึ้นอยู่กับความแม่นยำของโมเดลนั้น ผลลัพธ์ที่ได้จะเป็นคำตอบที่มีค่าถ่วงน้ำหนักมากที่สุด

8. kNN

- เป็นอัลกอริทึมประเภท classification

- มีหลักการ คือ หาข้อมูลที่ใกล้เคียงกับข้อมูลที่ต้องการหา n ตัว ใน n ตัวนั้นมีผลลัพธ์อยู่ใน class ไหนมากที่สุดก็จัดให้อยู่ class นั้น ๆ

- ข้อเสีย อ่อนแอต่อ noise

9. Naive Bayes

- เป็นอัลกอริทึมประเภท classification

- หลักการ หาความน่าจะเป็นแต่ละคลาสในแต่ละแอตทริบิวต์แล้วคำนวณความน่าจะเป็นในการเป็น class นั้น(โดยแต่ละแอตทริบิวต์เป็นอิสระต่อกัน)ผลลัพธ์เป็น class ที่มีความน่าจะเป็นมากที่สุด

10. CART

- เป็นอัลกอริทึมที่มีรูปแบบเป็น decision tree

- ในแต่ละ root มี leaf node เพียงสอง node