อัลกอริที่มที่ใช้

ชื่อ Naive Bayes Classifie อ้างอิงเนื้อหาจากบล็อก http://thaiml.org/?p=161

ทำไมถึงใช้ Naive Bayes?

เนื่องจากค่า feature แต่ละค่าไม่ขึ้นต่อกัน (independent) ซึ่งสอดคล้องกับสมมุติฐานของ Naive Bayes ที่ว่า เมื่อเรารู้ว่าข้อมูลนั้นๆ อยู่ในคลาสอะไรแล้ว ค่า feature แต่ละค่าของข้อมูลนั้นๆ จะไม่ขึ้นต่อกัน

ค่า feature ในที่นี้หมายถึงค่า attribute แต่ละค่า เช่น ค่า missing ค่า overjet ค่า overbite และอื่นๆ
คลาสในที่นี้แบ่งออกเป็น 2 คลาส คือ 1. คลาสของผู้ป่วยที่ควรจะจัดฟัน 2. คลาสของผู้ป่วยที่ไม่จำเป็นต้องจัดฟัน

ลักษณะข้อมูลที่ใช้

มี 15 features
มี 2 คลาส คือ 1. คลาสของผู้ป่วยที่ควรจะจัดฟัน (Yes) 2. คลาสของผู้ป่วยที่ไม่จำเป็นต้องจัดฟัน (No)
คลาส Yes มีจำนวนข้อมูล 226 ข้อมูล
คลาส No มีจำนวนข้อมูล 176 ข้อมูล
รวมทั้งหมดมี 401 ข้อมูล

ขั้นตอนการเตรียมความพร้อมของข้อมูล

จัดการกับ missing value อย่างไร?

จริงๆ แล้วมีหลายวิธี แต่ในการทดลองนี้จะเลือกค่าที่เกิดขึ้นบ่อยที่สุดในคลาสนั้นๆ นำไปใส่แทนค่าที่หายไป

แปลงข้อมูล continuous เป็น discrete

เนื่องจากข้อมูลเป็น continuous เพื่อให้อัลกอริที่มที่ใช้ไม่ซับซ้อนจนเกินไป จึงได้มีการแบ่งช่วงของข้อมูล และแปลงข้อมูลทุก feature ให้เป็น discrete โดยมีสูตรการแปลงดังนี้

$$ค่า\ discrete\ ที่ได้ = \frac{((ค่า\ feature\ ในขณะนั้น - ค่าต่ำสุดของ\ feature\ นั้น) x\ จำนวนช่วงของข้อมูลที่ต้องการ)}{(ค่าสูงสุดของ\ feature\ นั้น - ค่าต่ำสุดของ\ feature\ นั้น)}$$

ค่าข้อมูลที่เป็น Y หรือ N จะแปลงเป็น 1.0 และ 0.0 ตามลำดับ การเลือกจำนวนช่วงของข้อมูล จะเลือกตามความเหมาะสมของข้อมูลนั้นๆ เป็นค่าที่กำหนดเองตาม domain knowledge

การทดลอง

แบ่งออกเป็น 5 ครั้ง เพื่อหาค่าเฉลี่ยของตัววัดผลการทดลอง โดยแต่ละครั้งจะสุ่มเลือกข้อมูลมาสอนระบบ 80% และ

เอาไว้ทดสอบ 20% จากข้อมูล 401 ข้อมูล ดังนั้นจะได้ ข้อมูลที่จะนำไปสอนระบบจำนวน 321 ข้อมูล และที่จะนำไปทดสอบอีก 80 ข้อมูล

การทดลองจะวัดผลจากค่า Precision และ Recall (http://en.wikipedia.org/wiki/Precision_and_recall) และค่าความแม่นยำ Accuracy แบบเฉลี่ย 5 ครั้ง จะได้ตามนี้

	Precision	**Recall**	**Accuracy**
ข้อมูลสุ่มชุดที่ **1**	1	0.939394	0.95
ข้อมูลสุ่มชุดที่ **2**	1	0.947368	0.9625
ข้อมูลสุ่มชุดที่ **3**	0.9375	0.957447	0.9375
ข้อมูลสุ่มชุดที่ **4**	1	0.942308	0.9625
ข้อมูลสุ่มชุดที่ **5**	0.9375	0.9375	0.925
ค่าผลการทดลองเฉลี่ย	**0.975**	**0.9448034**	**0.9475**

สามารถคิดเป็น % ก็ได้ครับ จะได้ดังนี้

	Precision	**Recall**	**Accuracy**
ค่าผลการทดลองเฉลี่ย	97.5%	94.48%	94.75%