อัลกอริทึ่มที่ใช้

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

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

ข้อมูลที่ใช้ (pirot)

มี 13 features
มี 2 คลาส คือ 1. คลาสของผู้ป่วยที่ควรจะจัดฟัน (Yes) 2. คลาสของผู้ป่วยที่ไม่จำเป็นต้องจัดฟัน (No)
คลาส Yes มีจำนวนข้อมูล 20 ข้อมูล
คลาส No มีจำนวนข้อมูล 15 ข้อมูล
รวมทั้งหมดมี 35 ข้อมูล

ขั้นตอนการเตรียมความพร้อมของข้อมูล

จัดการกับ missing value อย่างไร?

เลือกค่าที่เกิดขึ้นบ่อยที่สุดในคลาสนั้นๆ เอาไปใส่แทน โดยมีสมมุติฐานว่า ความน่าจะเป็นของค่าที่หายไป

แปลงข้อมูล continuous เป็น discrete

เนื่องจากข้อมูลเป็น continuous เพื่อให้อัลกอริทึ่มที่ใช้ไม่ซับซ้อนจนเกินไป จึงได้มีการแบ่งช่วงของข้อมูล และแปลงข้อมูลทุก feature ให้เป็น discrete โดยมีสูตรการแปลงดังนี้

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

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

การทดลอง

แบ่งออกเป็น 5 ครั้ง เพื่อหาค่าเฉลี่ยของตัววัดผลการทดลอง โดยแต่ละครั้งจะสุ่มเลือกข้อมูลมาสอนระบบ 80% และเอาไว้ทดสอบ 20% จากข้อมูล 35 ข้อมูล ดังนั้นจะได้ ข้อมูลที่จะนำไปสอนระบบจำนวน 28 ข้อมูล และที่จะนำไปทดสอบอีก 7 ข้อมูล

โดยจะเลือกข้อมูลจากคลาสที่ควรจะจัดฟันจำนวน 16 ข้อมูล และจากคลาสที่ไม่ต้องจัดฟันจำนวน 12 ข้อมูล นำไปสอนระบบ ข้อมูลที่เหลือจากคลาสที่ควรจะจัดฟันจำนวน 4 ข้อมูล และจากคลาสที่ไม่ต้องจัดฟันจำนวน 3 ข้อมูล นำไปทดสอบระบบ

การทดลองจะวัดผลจากค่า Precision และ Recall (http://en.wikipedia.org/wiki/Precision_and_recall) และค่าความแม่นยำ Accuracy ซึ่งจะได้ตามนี้

	Precision	**Recall**	**Accuracy**
Naive Bayes Classifier	0.96	0.9	0.9142858

สามารถคิดเป็น % ก็ได้ครับ จะได้ดังนี้

	Precision	**Recall**	**Accuracy**
Naive Bayes Classifier	96%	90%	91.43%