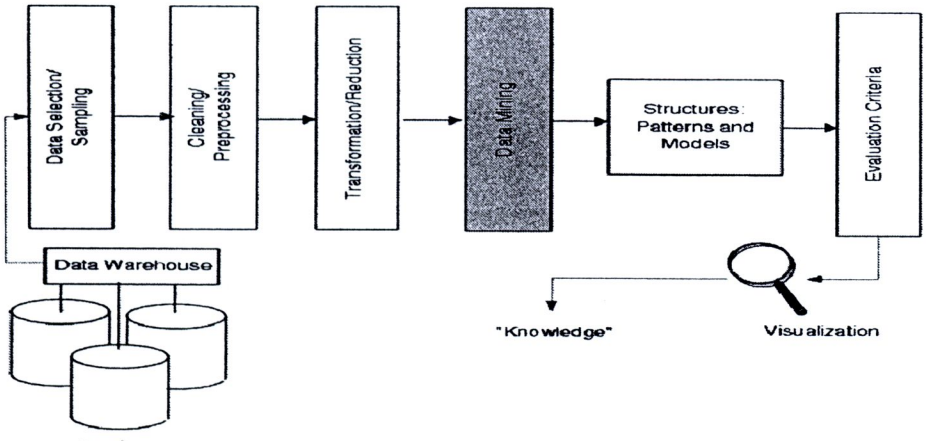
บทที่ 2
ทฤษฎี แนวคิดและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีและแนวคิดที่ใช้ในการศึกษา

2.1.1 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) คือกระบวนการสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases หรือ KDD) ซึ่งเป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วดึงความรู้ หรือสิ่งที่สำคัญออกมาเพื่อใช้ในการวิเคราะห์ หรือทำนายสิ่งต่างๆที่จะเกิดขึ้น ในการค้นหาความรู้นั้นความจริงที่แฝงอยู่ในข้อมูล (Knowledge Discovery) ซึ่งเป็นกระบวนการขุดค้นสิ่งที่น่าสนใจในกองข้อมูลที่เรามีอยู่เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลกระบวนการ KDD แบ่งออกเป็น 6 ขั้นตอน ดังนี้



ที่มา: [ระบบออนไลน์] std.kku.ac.th

รูป 2.1 ขั้นตอนต่างๆของกระบวนการ KDD

ขั้นตอนที่ 1: Data selection / sampling เป็นขั้นตอนการคัดเลือกข้อมูลเฉพาะกลุ่มที่สนใจออกมาจากฐานข้อมูลหรือคลังข้อมูลซึ่งเก็บข้อมูลหลายอย่างไว้เป็นจำนวนมหาศาลโดยทั่วไปเรามักจะต้องการความรู้ในบางเรื่องที่น่าสนใจเท่านั้นจึงไม่จำเป็นต้องใช้ข้อมูลทั้งหมดในฐานข้อมูลหรือคลังข้อมูล

ขั้นตอนที่ 2: Data cleaning / preprocessing เนื่องจากข้อมูลที่คัดเลือกมาอาจจะมีบางส่วนที่ผิดพลาด (เรียกว่า noise) เช่นข้อมูลที่ระบุเพศของบุคคลแทนที่จะปรากฏรหัส F หรือ M กลับปรากฏรหัส W เป็นต้นหรือค่าของข้อมูลอาจจะมีลักษณะที่ผิดปกติเช่นอายุ 130 ปีนอกจากนี้ข้อมูลบางส่วนอาจจะขาดหายไปและข้อมูลบางรายการมีค่าเปลี่ยนแปลงไปตามช่วงเวลาซึ่งข้อมูลที่มีลักษณะดังกล่าวข้างต้นนั้นจะต้องได้รับการแก้ไขเปลี่ยนแปลงก่อนที่จะถูกนำไปใช้เพื่อการค้นหาความรู้ในขั้นตอนต่อไป

ขั้นตอนที่ 3: Data transformation / reduction ถ้ารูปแบบของข้อมูลไม่เหมาะสมหรือไม่ตรงกับรูปแบบที่โปรแกรม Data mining (ซึ่งจะเกิดขึ้นในขั้นตอนลำดับถัดไป) ต้องการจะต้องมีการเปลี่ยนรูปแบบของข้อมูลให้ถูกต้องและพร้อมกันนั้นจะมีการพิจารณาลดขนาดของข้อมูลโดยการตัดส่วนที่จะไม่เป็นประโยชน์ในขั้นตอน Data mining ออกไปการลดขนาดของข้อมูลนี้ถ้าทำอย่างมีประสิทธิภาพจะช่วยให้การทำงานของขั้นตอน Data mining รวดเร็วขึ้น

ขั้นตอนที่ 4: Data mining เป็นขั้นตอนเลือกอัลกอริธึมที่จะใช้ค้นหารูปแบบหรือโมเดลจากข้อมูลที่ผ่านการคัดเลือกและกลั่นกรองมาแล้วในขั้นตอนที่ 1-3 ในขั้นตอนนี้ผู้ใช้มักจะต้องระบุรูปแบบการแสดงผลที่ต้องการเช่นแสดงผลลัพธ์ในลักษณะ classification rules, ลักษณะ decision tree, regression, clustering หรืออื่นๆ

ขั้นตอนที่ 5: Interpretation / evaluation คือ การตรวจสอบและแปลผลที่ได้จากขั้นตอน Data mining ถ้าผลที่ได้ยังมีความถูกต้องต่ำเกินไปอาจจะต้องย้อนกลับไปปรับพารามิเตอร์บางตัวของโปรแกรม Data mining หรือในบางครั้งอาจจะต้องย้อนกลับไปถึงขั้นตอนที่ 1 โดยไปสุ่มเลือกข้อมูลชุดใหม่มาใช้ถ้าหากข้อมูลชุดแรกที่ใช้ไม่แสดงรูปแบบใดออกมาให้เห็นได้ชัดเจน

ขั้นตอนที่ 6: Consolidating discovered knowledge เป็นขั้นตอนในการรวบรวมและสรุปความรู้ที่ได้จากการทำ KDD เพื่อนำเสนอต่อผู้บริหารหรือหน่วยงานที่เกี่ยวข้องในบางครั้งความรู้ที่ค้นพบใหม่นี้ขัดแย้งกับความรู้เดิมที่มีอยู่อาจจะต้องมีการตรวจสอบเพื่อหาข้อสรุปว่าความรู้ใดเป็นความรู้ที่ถูกต้อง

ในกระบวนการ KDD ทั้ง 6 ขั้นตอนนี้ขั้นตอนที่เป็นหัวใจสำคัญคือการทำ Data mining จุดมุ่งหมายหลักของการทำ Data mining มีสองประการคือ

การทำ Data mining เพื่อการทำนาย เป็นการนำความรู้ที่เรียนรู้มาจากข้อมูลที่มีอยู่เพื่อประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคตเช่นจากข้อมูลตัวแปรทางเศรษฐกิจต่างๆ การทำ Data mining สามารถเรียนรู้จากข้อมูลเหล่านี้และค้นหาโมเดลที่สามารถใช้อธิบายลักษณะของตัวแปรที่มีผลต่อการเพิ่มขึ้นหรือลดลงของราคาหุ้นได้และจากโมเดลที่ได้นี้สามารถนำไปใช้ทำนายได้ว่า ราคาหุ้นในอนาคตน่าจะมีทิศทางไปในแนวทางใด

การทำ Data mining เพื่อการอธิบาย เป็นการค้นหารูปแบบที่น่าสนใจจากกลุ่มข้อมูล รูปแบบนี้มักจะเป็นความสัมพันธ์หรือลักษณะที่เชื่อมโยงกันของข้อมูลการทำแบบนี้ต่างจากแบบแรกตรงที่ ผู้ใช้ไม่ได้กำหนดล่วงหน้าว่าจะให้โปรแกรม Data mining ค้นหารูปแบบหรือโมเดลของอะไร แต่ให้ค้นหาทุกรูปแบบที่น่าสนใจจากข้อมูลซึ่งจะทำให้สามารถค้นพบความรู้ใหม่ๆ เช่น ความสัมพันธ์ระหว่างตัวแปร การเกาะกลุ่มไปในทิศทางเดียวกัน เป็นต้น

ขั้นตอนการทำเหมืองข้อมูล

ในการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลนี้ มีกระบวนการมาตรฐานที่เรียกว่า "Cross-Industry Standard Process for Data Mining" หรือเรียกย่อๆ ว่า "CRISP-DM" ซึ่งเกิดจากความร่วมมือระหว่าง บริษัท DaimlerChrysler บริษัท SPSS และบริษัท NCR ซึ่ง กระบวนการ CRISP-DM ประกอบด้วย 6 ขั้นตอน (ที่มา: [ระบบออนไลน์] www.open-miner.com) ได้แก่

1. Business Understanding

เป็นขั้นตอนแรกสุดในกระบวนการ CRISP-DMขั้นตอนนี้เป็นการทำความเข้าใจระบุปัญหาหรือโอกาสเชิงธุรกิจ จากนั้นทำการแปลงโจทย์ที่ได้ให้อยู่ในรูปแบบที่เหมาะสมต่อการนำมาวิเคราะห์ข้อมูล

2. Data Understanding

ข้อมูลเป็นปัจจัยที่สำคัญที่สุดที่ขาดไม่ได้ในการทำเหมืองข้อมูลในขั้นตอนนี้เป็นการรวบรวมข้อมูลที่เกี่ยวข้องเพื่อใช้ในการวิเคราะห์ด้วยเทคนิคการทำเหมืองข้อมูลในการรวบรวมข้อมูลนั้นควรพิจารณาด้วยว่าเป็นข้อมูลที่ได้มาจากแหล่งข้อมูลที่ถูกต้องน่าเชื่อถือข้อมูลที่ได้มีปริมาณมากพอหรือยัง และเป็นข้อมูลที่เหมาะสมมีรายละเอียดเพียงพอต่อการนำไปใช้ในการวิเคราะห์

3. Data Preparation

ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุดเนื่องจากโมเดลที่ได้จากการทำเหมืองข้อมูลจะให้ผลลัพธ์ที่ถูกต้องหรือไม่นั้นขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ กล่าวคือถ้าข้อมูลที่ใช้นั้นไม่ถูกต้อง มีผิดพลาดย่อมสะท้อนถึงผลลัพธ์ที่ได้ ซึ่งอาจทำให้ตีความผลลัพธ์ได้คลาดเคลื่อนเช่นกัน โดยการเตรียมข้อมูลนั้น สามารถแบ่งออกได้เป็น 3 ขั้นตอนย่อยคือ

- ทำการคัดเลือกข้อมูล (Data Selection) เราควรกำหนดเป้าหมายก่อนว่าเราจะทำการวิเคราะห์อะไรแล้วจึงเลือกใช้เฉพาะข้อมูลที่เกี่ยวข้องกับสิ่งที่เราจะทำการวิเคราะห์

- การกลั่นกรองข้อมูล (Data Cleaning) ในบางกรณีอาจพบข้อมูลที่ไม่ถูกต้อง อันเนื่องมาจากปัญหาในระหว่างการจัดเก็บข้อมูลเช่นการกรอกข้อมูลไม่ครบบ้าง กรอกข้อมูลซ้ำซ้อนบ้างในขั้นตอนนี้เราจะทำการกรองข้อมูลที่ไม่ถูกต้องหรือซ้ำซ้อนออกหรืออาจทำการซ่อมข้อมูลที่ขาดหายไปด้วยวิธีการบางอย่างเช่นการพิจารณาจากค่าเฉลี่ยของข้อมูลส่วนใหญ่เป็นต้น

- การแปลงรูปข้อมูล (Data Transformation) เป็นขั้นตอนการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมนำไปใช้ในการวิเคราะห์ตามอัลกอริทึมของการทำเหมืองข้อมูลที่เลือกใช้

4. Modeling

เป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล ได้แก่การสร้างตัวทำนาย (Prediction model) ในบางครั้งพบว่ามีการนำเทคนิคการทำเหมืองข้อมูลหลายเทคนิคมาใช้ในการวิเคราะห์ข้อมูลเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ดังนั้นเมื่อทำขั้นตอนนี้แล้วอาจมีการย้อนกลับไปที่ขั้นตอน Data preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วยนอกจากนี้ยังมีการประเมินโมเดลวิเคราะห์ข้อมูลที่ได้ ในรูปแบบความถูกต้องของโมเดลเพื่อเป็นตัวบ่งชี้ความน่าเชื่อถือของโมเดลที่ได้

5. Evaluation

การประเมินหรือวัดประสิทธิภาพของโมเดลวิเคราะห์ข้อมูลในขั้นตอนก่อนหน้านั้นเป็นเพียงการวัดความน่าเชื่อถือของโมเดลเท่านั้นในขั้นตอนนี้เป็นการประเมินประสิทธิภาพของผลลัพธ์จากโมเดลวิเคราะห์ข้อมูลว่าครอบคลุมและสามารถตอบโจทย์ทางธุรกิจที่ตั้งไว้ในขั้นตอนแรกหรือไม่ในกรณีที่มีการสร้างโมเดลวิเคราะห์ข้อมูลหลายโมเดลในขั้นตอนนี้จะทำการประเมินแต่ละโมเดลด้วยว่ามีส่วนดีส่วนด้อยอย่างไรและควรเลือกใช้โมเดลใดการทำงานในส่วนนี้ต้องอาศัยทักษะในการวิเคราะห์ข้อมูลและธุรกิจเพื่อช่วยให้การวิเคราะห์ทำได้สะดวกและรวดเร็วขึ้นจึงมีการใช้เครื่องมือทางด้านกราฟิก เช่นการแสดงผลการวิเคราะห์ด้วยกราฟรายงานรูปแบบต่างๆ เป็นต้น

6. Deployment

ผลลัพธ์หรือองค์ความรู้ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลจะไม่มีประโยชน์เลยถ้าไม่ถูกนำไปใช้งานจริง ตัวอย่างเช่นการนำองค์ความรู้ที่ได้ไปใช้ในการจัดโปรโมชั่นส่งเสริมการขายสินค้าใช้ในการทำนายแนวโน้มการทุจริตในระบบการเงินของธนาคารหรือตรวจจับความผิดปกติในการซื้อขายหุ้นในตลาดหลักทรัพย์เป็นต้น

เทคนิคต่าง ๆ ของการทำเหมืองข้อมูล (Data Mining) (คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1)

1. Classification Trees

แบบจำลอง Classification Trees เป็นแบบจำลองที่มีพื้นฐานมาจากแผนภูมิต้นไม้ ซึ่งเป็นวิธีที่ง่ายและมีประสิทธิภาพในการทำนายผลของการจำแนกประเภทของตัวแปรตอบรับที่ขึ้นอยู่กับกลุ่มของตัวแปรที่ใช้ในการทำนาย โดยข้อมูลจะถูกแบ่งออกเป็น 2 กลุ่ม ขึ้นอยู่กับตัวแปรที่ใช้ทำนาย (ตัวแปรอิสระ) ซึ่งจะถูกทดสอบซ้ำไปซ้ำมาจนกระทั่งได้ค่าตัวแปรตอบรับ (ตัวแปรตาม) ที่เป็น Homogenous และลำดับของการแบ่งกลุ่มของตัวแปรที่ใช้ในการทำนายนั้น จะถูกแสดงออกมาในรูปของ Binary Tree ตามชื่อของแบบจำลอง

แบบจำลอง Classification Trees นั้น สามารถอธิบายได้ด้วยชุดของกฎการทำนาย สำหรับค่าตัวแปรตาม y และ เซ็ตของตัวแปรที่ใช้ในการทำนาย $x_1, x_2, ..., x_p$ ดังนั้น กฎของแบบจำลอง Classification Trees จึงสามารถอธิบายได้ด้วยรูปแบบที่ถูกกำหนด เช่น

$$\text{ถ้า } x_1 < 23 \text{ และ } x_2 \in \{A, B\} \text{แล้ว y จะจัดอยู่ในประเภทที่ 2} \qquad (1)$$

ด้วยความเรียบง่ายของการแสดงผลของแบบจำลองและกฎของการทำนาย ทำให้แบบจำลอง Classification Trees เป็นเครื่องมือหนึ่งในกระบวนการทำเหมืองข้อมูลที่น่าสนใจ ประโยชน์ในด้านอื่นๆของแบบจำลองต้นไม้ ได้แก่

- ไม่มีการผันแปรในการแสดงตัวแปรที่ใช้ในการทำนายซ้ำในทิศทางเดียวกัน
- สามารถนำไปจับพฤติกรรมที่ไม่เป็นเส้นตรงของตัวแปรที่ใช้ในการทำนายเช่นเดียวกับค่าผลกระทบระหว่างตัวแปรที่ใช้ในการทำนาย
- วิธีนี้แตกต่างจากแบบจำลองการวิเคราะห์การถดถอยโลจิสติกตรงที่ สามารถทำนายผลของการจำแนกประเภทของตัวแปรตอบรับ (ตัวแปรตาม) ที่มีค่ามากกว่า 2 ระดับได้

2. Logistic regression

กระบวนการการวิเคราะห์การถดถอยโลจิสติก (Logistic regression) ผู้วิเคราะห์จะเป็นผู้กำหนดคุณสมบัติของแบบจำลองเหตุการณ์ 2 เหตุการณ์ ที่เกิดขึ้นในรูปของฟังก์ชันเส้นตรง (Linear function) ของเซ็ตของตัวแปรอิสระ (Independent variables)

แบบจำลองการถดถอยโลจิสติก (Logistic regression model) นั้น เป็นแบบจำลองเชิงเส้น (linear model) ชนิดพิเศษ ซึ่งตัวแปรตามจะอยู่ในรูปของระดับชั้นและมีเพียง 2 ระดับเท่านั้น คือ การยอมรับและปฏิเสธเงื่อนไข ตัวอย่างที่พบโดยทั่วไปได้แก่ การยอมรับหรือปฏิเสธข้อเสนอ

สำนักงานคณะกรรมการวิจัยแห่งชาติ ห้องสมุดงานวิจัย

ทางการตลาด การทำนายการเกิดหรือไม่เกิดขึ้นของโรคต่างๆ และการหาความเป็นไปได้ที่จะประสบความสำเร็จหรือไม่ของแผนวงจรอิเล็กทรอนิกส์เป็นต้น

จากแบบจำลองเชิงเส้น (linear model) จะให้ค่าประมาณของตัวแปรตาม Y ซึ่งจะเป็นไปตามเงื่อนไขของฟังก์ชันเส้นตรง (linear function) ของตัวแปรอิสระ $X_1, X_2, X_3, ..., X_p$ ซึ่งสามารถเขียนในรูปสมการทางคณิตศาสตร์ได้ดังนี้

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon \qquad (2)$$

จากสมการ β_i คือ ค่า coefficient ของแบบจำลองเชิงเส้น

β_0 คือ ค่า intercept ของแบบจำลอง

ε คือ ค่า residual

ค่า estimate ของ coefficient ($\hat{\beta}_i$) จะคำนวณได้จากข้อมูลฝึกสอน (training data) ซึ่งได้มาจากการประมาณค่าของตัวแปรตาม ส่วนค่า $\hat{Y}$ คำนวณได้จากการแทนค่า estimate ของ coefficient ลงในสมการ(1) และค่า estimate ของ residual($\hat{\varepsilon}$) จะได้มาจากผลต่างระหว่างค่าสังเกต (observe) ของตัวแปรตาม และค่าประมาณ (estimate) ของตัวมันเอง

ในแบบจำลอง logistic regression ค่าตัวแปรตามจะถูกแบ่งเป็น 2 ค่าโดยจะถูกกำหนดให้เป็นค่า 0 กับ 1 ค่า $\hat{Y}$ จะเป็นค่าประมาณของความน่าจะเป็นของระดับซึ่งถูกแทนด้วย 1 ซึ่งแบบจำลอง logistic regression นั้น จะใช้ฟังก์ชันโลจิสติก เพื่อแสดง $\hat{Y}$ ในรูปของฟังก์ชันเส้นตรง (linear function) ของเซ็ตของตัวแปรอิสระ เขียนในรูปสมการทางคณิตศาสตร์ได้ดังนี้

$$g\left(\hat{Y}\right) = \log\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i X_i \qquad (3)$$

การประมาณค่าพารามิเตอร์ในที่นี้จะใช้วิธีการหาค่า Maximum Likelihood ด้วยวิธี Interactive Re-weighted Least Square (IRLS) ซึ่งเป็นวิธีมาตรฐานและเป็นเทคนิคที่เข้าใจง่ายในการ fitting แบบจำลองการถดถอยโลจิสติกค่า log-likelihood $l(\beta, Y)$ จะถูก maximize ด้วยการหาผลลัพธ์จากสมการการให้คะแนน (score equation)

$$\partial l(\beta, Y)/\partial\beta = 0 \qquad (4)$$

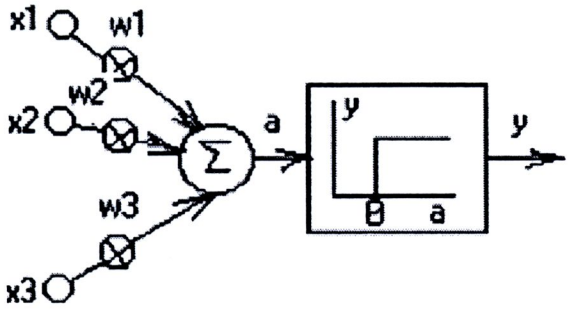
เช่นเดียวกันกับสมการที่ (2) ค่า β ในที่นี้คือ ค่า coefficient และค่า Y คือ ค่าตัวแปรตาม สำหรับแบบจำลองการถดถอยโลจิสติก ค่าสมการการให้คะแนน (score equation) จะเป็น nonlinear ใน β ด้วยเหตุนี้จึงต้องถูกนำมาหาผลลัพธ์โดยการใช้ IRSL ซึ่งสามารถอ่านรายละเอียดเพิ่มเติมได้จาก Chambers and Hasties (1992) หรือ McCullagh and Nelder (1989)

สำนักงานคณะกรรมการวิจัยแห่งชาติ
ห้องสมุดงานวิจัย
วันที่ 17 พ.ย. 2555
เลขทะเบียน 248663
เลขเรียกหนังสือ

3. Classification Neural Networks

กระบวนการ Classification Neural Networks เป็นรูปแบบของการจำแนกกลุ่มด้วยแบบจำลองกล่องดำ (Black Box Classification) เพื่อใช้สำหรับการคำนวณความน่าจะเป็นและการทำนายประเภทของข้อมูล คำว่า Black Box ในที่นี้จะหมายถึง ข้อเท็จจริงซึ่งมีลักษณะเป็นกลุ่มข้อมูลเล็กๆซึ่งสามารถแปลความหมายได้ ในด้านของความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ ผ่านโครงสร้างและการใช้ตัวแปรสมมติ (estimate parameter) ของแบบจำลอง จุดมุ่งหมายของเทคนิคนี้คือ การประมาณค่าความน่าจะเป็นที่มีความสอดคล้องกับแต่ละระดับซึ่งขึ้นอยู่กับข้อมูลในเซ็ตของข้อมูลในเทคนิคนี้ ตัวแปรตามที่ใช้จะเป็นหมวดหมู่ย่อยซึ่งประกอบด้วย 2 ระดับขึ้นไป

Warren Mcculloch และ Walter Pitt ได้เป็นผู้เสนอนิวรอนเทียมตัวแรกในปี ค.ศ.1943 โดยสร้างเป็นวงจรอีเล็คโทรนิกส์ ดังรูป



ที่มา: [ระบบออนไลน์] std.kku.ac.th

รูป 2.2 แบบจำลองนิวรอนเทียมตัวแรก โดย Warren Mcculloch และ Walter Pitt

โดยนิวรอนจะนำสัญญาณอินพุต $x_1, x_2, ..., x_n$ ไปคูณกับค่าน้ำหนัก $w_1, w_2, ..., w_n$ แล้วนำผลคูณทั้งหมดมารวมกันให้ได้ผลรวมเป็น a เพื่อจะนำไปตรวจสอบกับค่า Threshold θ ถ้า $a \geq \theta$ ก็จะให้เอาท์พุต y = 1 ออกมา แต่ถ้า $a < \theta$ อยู่ก็จะให้ y = 0 ดังนั้นการทำงานจึงเป็นไปตามสมการดังนี้

$$a = w_1x_1 + w_2x_2 +, \ldots, w_nx_n \tag{5}$$

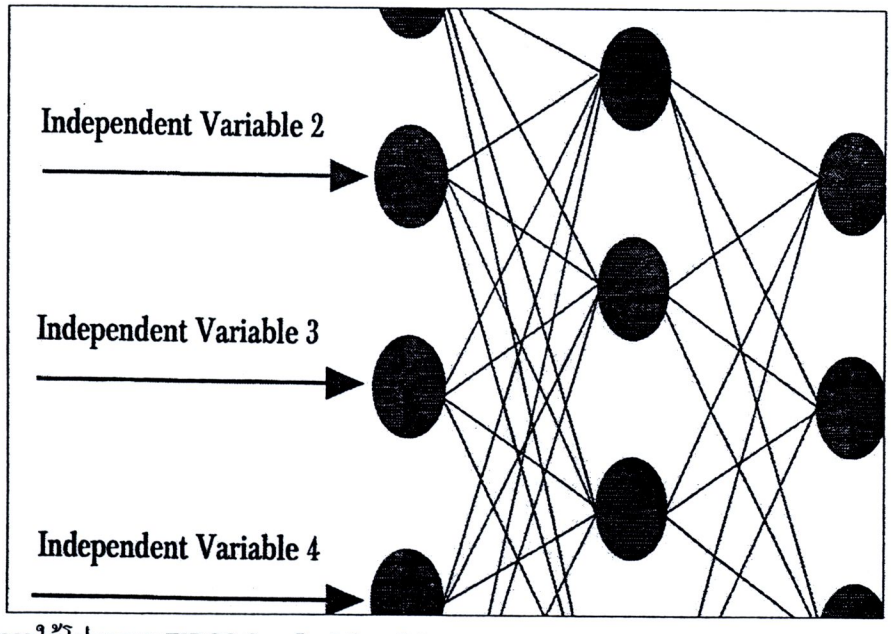
$$y = f(a) = \begin{cases} 0, a<\theta \\ 1, a \geq \theta \end{cases} \tag{6}$$

x_1 เป็นตัวแทนของสัญญาณกระตุ้นที่รับมาจากนิวรอนข้างเคียง

w_1 คือ ค่าน้ำหนัก (weight)

f (a) คือ ตัวแทนของฟังก์ชันทำนบ (Threshold Function) ที่จะส่งสัญญาณออกไปให้แก่เซลล์ข้างเคียงโดยมีลักษณะของฟังก์ชั่นเป็นแบบที่เรียกว่า Binary Function หรือ Hard Limit Function

กระบวนการ Classification Neural Network ในโปรแกรมสำเร็จรูปที่ใช้ในงานวิจัยอิสระนี้จะเป็นแบบจำลอง classification ชนิด 2-stage ซึ่งแนวคิดหลักของเทคนิคนี้ คือ การคำนวณหาการเกาะกลุ่มเชิงเส้น (linear combination) ของตัวแปรอิสระ และสร้างแบบจำลองในแต่ละระดับของตัวแปรตามในรูปของฟังก์ชันที่ไม่เป็นเส้นตรง (non-linear function) ของการรวมกลุ่ม แสดงได้โดยรูป 4ซึ่ง output ของ network จะเป็นค่าความน่าจะเป็นซึ่งแต่ละรูปแบบของ input จะเป็นชั้นของรูปแบบตัวแปรตามที่จำเพาะเจาะจง โดยค่าของตัวแปรอิสระจะถูกใส่ลงใน input node ผ่านชั้น hidden layer แล้วออกสู่ output node ซึ่งตัวเชื่อมต่อแต่ละตัวในไดอะแกรมนี้ จะแสดงการเกาะกลุ่มเชิงเส้นและ output node จะให้ค่าความน่าจะเป็นที่ในแต่ระดับของตัวแปรตาม



ที่มา: คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1

รูป 2.3 ไดอะแกรมของกระบวนการ Classification Neural Network

ในเซ็ตที่อยู่ตรงกลางของ node จะแสดงการเกาะกลุ่มเชิงเส้นของตัวแปรอิสระ node เหล่านี้จะถูกเรียกว่า hidden layer เนื่องจากชั้นนี้จะมีการรวมค่าตัวแปรที่ไม่ได้ observe โดยตรงเข้าไว้ด้วย ซึ่งโปรแกรมที่ใช้ในงานวิจัยอิสระครั้งนี้ สามารถที่จะมี hidden layer ได้ถึง 3 ชั้น แต่ละชั้นจะรวมเซ็ตต่างๆของการเกาะกลุ่มเชิงเส้นของ output จากชั้นที่ผ่านมา ซึ่งก็หมายความว่าถ้าหากใช้ 0 ชั้น network ก็จะล้มเหลวและกลับเข้าสู่ standard linear model

ค่า unknown parameter ใน Classification Neural Network จะเรียกว่า weight ซึ่งเป็นค่า coefficient อย่างง่าย ที่สอดคล้องกับ linear combination ค่า unknown parameter เหล่านี้ คือค่าถ่วงน้ำหนักของแต่ละการเชื่อมโยงในไดอะแกรมด้านบน Neural Network จะคำนวณค่าประมาณของตัวถ่วงน้ำหนักด้วยการทดสอบของโปรแกรมผ่านเข้าสู่ข้อมูลแล้วคูณด้วยระยะเวลา จึงทำให้ได้มาซึ่งค่าการเกาะกลุ่มเชิงเส้น และการปรับปรุงค่าถ่วงน้ำหนักนั้นๆ

ในแต่ละการทดสอบของโปรแกรมผ่านเข้าสู่ข้อมูลจะเรียกว่า epoch ซึ่งในที่นี้ Neural Network จะศึกษาผ่านข้อมูลนั้นๆ แต่สิ่งที่ไม่สามารถแสดงได้ในรูป 2.3 คือ ค่า bias node ซึ่งจะถูกถ่วงน้ำหนักในแต่ละโหนดในชั้น hidden layer และ output node แต่ค่าถ่วงน้ำหนักนี้จะแสดงค่า intercept ในแบบจำลอง

4. Naive Bayes

Naive Bayes คือ กระบวนการทาง classification อย่างง่ายซึ่งเป็นเทคนิคที่สามารถให้ผลที่ดีกว่าเทคนิคที่ซับซ้อนอื่นๆ โดยขึ้นอยู่กับกฎของเบย์

จากกฎของเบย์

$$P(A|B) = P(B|A)\left(\frac{P(A)}{P(B)}\right) \tag{7}$$

ซึ่ง $P(A|B)$ คือ โอกาสของเหตุการณ์ A ที่จะทำให้เหตุการณ์ B สำหรับค่าความน่าจะเป็นแบบมีเงื่อนไขและการคาดการณ์ความเป็นอิสระของตัวแปร จะให้ค่าจำเพาะของหมวดหมู่ย่อยของตัวแปรตามซึ่งกระบวนการนี้อธิบายได้จากตัวอย่างดังต่อไปนี้

สมมติว่า สมาคมศิษย์เก่าของมหาวิทยาลัยแห่งหนึ่งมีข้อมูลในการในการรับบริจาคในครั้งที่ผ่านมาดังในตาราง (แต่ในที่นี้จะขอยกตัวอย่างในจำนวนที่ค่อนข้างน้อยเพื่อความสะดวก)

ตารางที่ 2.1 ภาพรวมการบริจาคทั้งหมดจากสมาชิกของชมรมศิษย์เก่า มีสมาชิกที่บริจาคให้มหาวิทยาลัย 7 คน ส่วนอีก 16 คนที่เหลือไม่บริจาค

ภาครวมการบริจาคทั้งหมด	
บริจาค	ไม่บริจาค
7	16

ที่มา: คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1

ตารางที่ 2.2 การบริจาคจำแนกโดยระดับปริญญาที่สมาชิกชมรมศิษย์เก่าได้รับ

การบริจาคจำแนกโดยระดับปริญญาของสมาชิก		
ระดับปริญญา	บริจาค	ไม่บริจาค
ปริญญาตรี	4	12
ปริญญาโท	1	3
ปริญญาเอก	2	1

ที่มา: คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1

ตารางที่ 2.3 การบริจาคจำแนกโดยเพศของสมาชิกของชมรมศิษย์เก่า

การบริจาคจำแนกโดยเพศของสมาชิก		
เพศ	บริจาค	ไม่บริจาค
หญิง	3	7
ชาย	4	9

ที่มา: คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1

ตารางที่ 2.4 การบริจาคจำแนกโดยที่อยู่ของสมาชิกของชมรมศิษย์เก่า

การบริจาคจำแนกโดยที่อยู่ของสมาชิก		
ที่อยู่	บริจาค	ไม่บริจาค
ในประเทศ	5	10
ต่างประเทศ	2	6

ที่มา: คู่มือการใช้โปรแกรม TIBCO Spotfire Miner 8.1

สมมติว่าสมาชิกใหม่ของชมรมศิษย์เก่ามีคุณสมบัติเป็นผู้หญิง ได้รับปริญญาโท และอาศัยอยู่ภายในประเทศแล้ว แนวโน้มที่เธอจะสนับสนุนการบริจาคคือ

จากกฎของเบย์

$$P(A|B) = P(B|A)\left(\frac{P(A)}{P(B)}\right) \tag{8}$$

$P(A|B)$ คือ โอกาสของเหตุการณ์ A ที่จะทำให้เหตุการณ์ B เกิดขึ้น ถ้าเราคาดการณ์เบื้องต้นว่า สถานการณ์บริจาคขึ้นอยู่กับตัวแปรที่เกี่ยวข้อง คือ ระดับปริญญา, เพศ และที่อยู่ปัจจุบัน ซึ่งเป็นตัวแปรอิสระ จากนั้นเราสามารถที่จะได้ค่าความน่าจะเป็นรวมออกมา ด้วยการคูณด้วยค่าความน่าจะเป็นของแต่ละค่า ดังนี้

$$P(Yes|M.S., Female, OutofStage)$$

$$= P(M.S., Female, OutofStage|Yes)\left(\frac{P(Yes)}{P(M.S., Female, OutofStage)}\right) \tag{9}$$

$$= P(M.S.|Yes)P(Female|Yes)P(OutofStage|Yes)\left(\frac{P(Yes)}{P(M.S., Female, OutofStage)}\right) \tag{10}$$

$$= \frac{\frac{1}{7}\times\frac{3}{7}\times\frac{2}{7}\times\frac{7}{23}}{P(M.S., Female, OutofStage)} \tag{11}$$

ในขั้นตอนนี้เราจะยังไม่สนใจตัวหาร สำหรับค่าความน่าจะเป็นของสมาชิกใหม่ซึ่งจะไม่บริจาคให้กับชมรมศิษย์เก่าก็เช่นเดียวกันคือ

$$P(No|M.S., Female, OutofStage)$$

$$= P(M.S.|No)P(Female|No)P(OutofStage|No)\left(\frac{P(No)}{P(M.S., Female, OutofStage)}\right) \tag{12}$$

$$= \frac{\frac{3}{16}\times\frac{7}{16}\times\frac{6}{16}\times\frac{16}{23}}{P(M.S., Female, OutofStage)} \tag{13}$$

ค่าความน่าจะเป็นที่ได้จากทั้งสมการที่(9) และ (11) เมื่อรวมกันแล้วต้องมีค่าเท่ากับ 1 ดังนั้นเราจึงหลีกเลี่ยงการคำนวณในส่วนของตัวหารด้วยการ normalization ดังนี้

$$P(Yes|M.S., Female, OutofStage) = \frac{\frac{1}{7}\times\frac{3}{7}\times\frac{2}{7}\times\frac{7}{23}}{\left(\frac{1}{7}\times\frac{3}{7}\times\frac{2}{7}\times\frac{7}{23}\right)+\left(\frac{3}{16}\times\frac{7}{16}\times\frac{6}{16}\times\frac{16}{23}\right)} \tag{14}$$

ค่าที่ได้จะมีค่าเท่ากับ 0.1992 แสดงว่า ค่าความน่าจะเป็นที่สมาชิกใหม่จะบริจาคให้กับชมรมศิษย์เก่ามีค่าประมาณ 20%

$$P(No|M.S., Female, OutofStage) = \frac{\left(\frac{3}{16}\times\frac{7}{16}\times\frac{6}{16}\times\frac{16}{23}\right)}{\left(\frac{1}{7}\times\frac{3}{7}\times\frac{2}{7}\times\frac{7}{23}\right)+\left(\frac{3}{16}\times\frac{7}{16}\times\frac{6}{16}\times\frac{16}{23}\right)} \quad (15)$$

ค่าที่ได้จะมีค่าเท่ากับ 0.8008 แสดงว่า ค่าความน่าจะเป็นที่สมาชิกใหม่จะไม่บริจาคให้กับชมรมศิษย์เก่ามีค่าประมาณ 80%

สมมติฐานซึ่งมีคุณสมบัติเป็นอิสระจะให้ผลที่มีความรวบรัด แต่สามารถนำไปใช้งานได้ดีในการแก้ปัญหาในทาง Classification แต่ถ้าคุณสมบัติบางตัวมีความยืดยาวจนเกินไปก็จะไม่มีความเป็นอิสระ และเทคนิคนั้นก็จะไม่สามารถให้ผลที่ดีได้ดังนั้นกระบวนการคัดเลือกตัวแปรอย่างชาญฉลาดจึงมีความจำเป็นเพื่อที่จะนำมาใช้กรองเอาตัวแปรที่ไม่ดีออกไปจากเซ็ตของตัวแปรให้มากที่สุด

ปัญหาที่มักเกิดกับกระบวนการ Naive Bayes จะเกิดขึ้นเมื่อหนึ่งในค่าคุณสมบัติของเหตุการณ์ที่ทำให้เกิดกลุ่มของข้อมูลนั้นไม่สามารถจัดให้เข้าได้กับระดับใดระดับหนึ่งในชั้นของตัวแปรตาม จากตัวอย่างด้านบน ถ้าหากไม่มีสมาชิกคนใดซึ่งได้รับปริญญาเอกเคยให้การบริจาคกับสมาคมศิษย์เก่า ดังนั้นค่าความน่าจะเป็นของผู้ที่ได้รับปริญญาเอก และให้การบริจาคกับสมาคมจะเท่ากับ 0 ซึ่งในสถานการณ์นี้จะให้ค่าความน่าจะเป็นที่ศิษย์เก่ารุ่นใหม่ซึ่งมีการศึกษาในระดับปริญญาเอกและให้การบริจาคกับสมาคมศิษย์เก่าก็จะเท่ากับ 0 ไปตลอดด้วย เนื่องจากค่าความน่าจะเป็นอื่นๆทั้งหมดเมื่อนำมาคูณกับ 0 แล้วก็จะให้ค่าเท่ากับ 0 ดังนั้นโปรแกรมจึงหลีกเลี่ยงปัญหาดังกล่าวด้วยการเริ่มต้นที่ 1 แทนค่า 0

2.1.2 ทฤษฎีการวิเคราะห์การถดถอยโลจิสติกและการวิเคราะห์การถดถอยโพรบิท

การวิเคราะห์การถดถอยโลจิสติก (Logistic regression) เป็นการนำตัวแปรอิสระหลายตัวมาวิเคราะห์ความสัมพันธ์พร้อมๆกันกับตัวแปรตัวแปรตามที่อยู่ในระดับ นามบัญญัติ การวิเคราะห์ประเภทนี้สามารถบอกได้ว่าปัจจัยใดที่ทำให้เกิดเหตุการณ์ที่คาดหวัง หรือเป็นการวิเคราะห์ถึงโอกาสในการเกิดเหตุการณ์ที่สนใจการวิเคราะห์โลจิสติก แบ่งเป็น

1) Binary Logistic Regression (Binary Regression) เมื่อตัวแปรตาม อยู่ในลักษณะ dichotomous ที่มีค่า 1 กับ 0

2) Multinomial Logistic Regression (Multinomial Regression) เมื่อตัวแปรตาม อยู่ในลักษณะ เป็นตัวแปรเชิงกลุ่ม หรือนามบัญญัติ ที่มีค่าตั้งแต่ 2 ค่าขึ้นไป

ในการวิเคราะห์โลจิสติก จะต้องกำหนดโมเดลโลจิสติก เรียกว่า Logit Model ซึ่งเป็นแบบจำลองที่นำมาใช้วิเคราะห์ข้อมูลว่าตัวแปรอิสระ (X) ส่งผลต่อโอกาสการเกิดเหตุการณ์ที่สนใจ (Y) หรือไม่ ซึ่งความน่าจะเป็นของการเกิดเหตุการณ์จะมีค่าในช่วง 0 ถึง 1

รูปแบบของโมเดลจะอยู่ในรูปลอกกาลิทึมของสัดส่วนความน่าจะเป็น การวิเคราะห์ในรูปแบบนี้ใช้เมื่อฟังก์ชั่นการกระจายสะสม (Cumulative distribution function) ของตัวแปรสุ่มมีการกระจายแบบ Logistic และมีการแจกแจงของค่าความคลาดเคลื่อนเป็นแบบ Logistic (การกระจายแบบโลจิสติก หมายถึง รูปแบบการกระจายของข้อมูลที่ไม่เป็นโค้งปกติ แต่ตัวแปรจะมีการกระจายเป็นรูปตัว S (Sigmoid Curve)

การวิเคราะห์การถดถอยโพรบิท (Probit regression) เป็นการนำตัวแปรอิสระหลายตัวมาวิเคราะห์ความสัมพันธ์พร้อมๆกันกับตัวแปรตัวแปรตามที่อยู่ในระดับนามบัญญัติ เช่นเดียวกับการวิเคราะห์ถดถอยโลจิสติก แต่แตกต่างกันที่รูปแบบของโมเดลจะอยู่ในรูปแบบโลจิทมาตรฐาน การวิเคราะห์ในรูปแบบนี้ใช้เมื่อฟังก์ชั่นการกระจายสะสม (Cumulative distribution function) ของตัวแปรสุ่มมีการกระจายแบบปกติ Normal Distribution

การกำหนดแบบจำลองโลจิทและโพรบิต

รูปแบบทั่วไปของแบบจำลองโลจิทและโพรบิตคือ

$$P(Y=1|X) = G(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k) = G(\beta_0 + X\beta) \qquad (16)$$

ความน่าจะเป็นที่จะเกิด Y=1 เป็นฟังก์ชันของตัวแปรอิสระ X โดยที่ G คือฟังก์ชันที่มีค่าจำกัดอยู่ระหว่าง 0 ถึง 1 ($0 < G(z) < 1$ สำหรับทุกค่าของ z) โดย G จะเป็นฟังก์ชันที่ไม่เป็นเส้นตรง (Nonlinear function) รูปแบบที่ที่มักนำมาใช้คือ

(1) ฟังก์ชันลอจิสติกส์สำหรับแบบจำลองโลจิท

$$G(z) = \exp(z)/[1+\exp(z)] = \Lambda(z) \qquad (17)$$

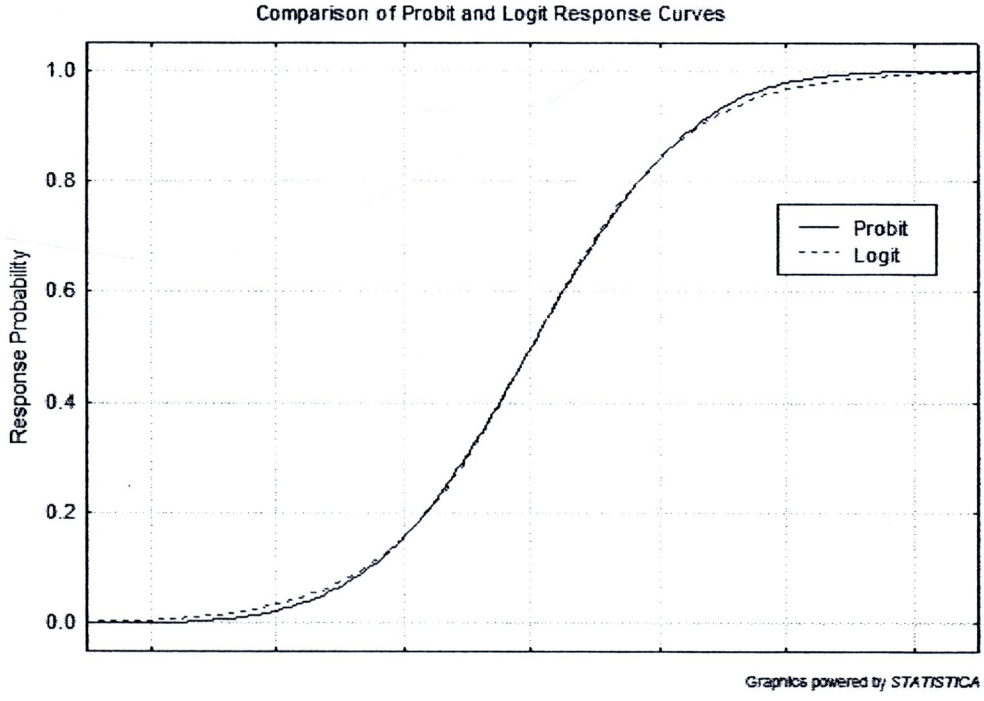
ค่าของฟังก์ชันจะอยู่ระหว่าง 0 และ 1 สำหรับทุกค่าของ z โดยฟังก์ชันดังกล่าวเป็นฟังก์ชันการกระจายสะสม (Cumulative distribution function) ของตัวแปรสุ่มที่มีการกระจายแบบลอจิสติกส์มาตรฐาน (Standard Logistic distribution)

(2) ฟังก์ชันโพรบิตสำหรับแบบจำลองโพรบิต

$$G(z) = \Phi(z) \equiv \int_{-\infty}$$ (18)

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2) \qquad (19)$$

ค่าของฟังก์ชันจะอยู่ระหว่าง 0 และ 1 สำหรับทุกค่าของ z โดยฟังก์ชันดังกล่าวเป็นฟังก์ชันการกระจายสะสม (Cumulative distribution function) ของตัวแปรสุ่มที่มีการกระจายปกติมาตรฐาน (Standard normal distribution) ลักษณะการกระจายของตัวแปรที่มีการกระจายแบบลอจิสติกส์ และแบบปกติมาตรฐานสามารถแสดงได้ดังรูป



ที่มา: [ระบบออนไลน์] std.kku.ac.th

รูป 2.4 การกระจายแบบลอจิสติกส์และการกระจายแบบปกติมาตรฐาน

เงื่อนไขการวิเคราะห์ด้วยแบบจำลองโลจิทและโพรบิทคือ

(1) ตัวแปรอิสระอาจจะเป็น Dummy Interval หรือ Ratio

(2) ค่าความคลาดเคลื่อนไม่สัมพันธ์กัน

(3) ค่าความคลาดเคลื่อนเป็นอิสระจากตัวแปรอิสระ (ค่าความคลาดเคลื่อนไม่สัมพันธ์กับตัวแปรอิสระ)

(4) ตัวแปรอิสระแต่ละตัวต้องไม่มีความสัมพันธ์กันสูง

(5) จำนวนกลุ่มตัวอย่างต้องมีจำนวนมากเพียงพอ (อย่างน้อย 30*p)

วิธีการวิเคราะห์ทั้งสองแบบ ใช้การประมาณค่าแบบจำลอง โดยวิธี Maximum Likelihood Estimation (MLE) แทนวิธีการประมาณแบบจำลองเชิงเส้น ที่ใช้วิธีกำลังสองน้อยที่สุดแบบ (Ordinary Least Squares: OLS)

การวิเคราะห์ผลของการประมาณค่า

(1) ค่าสัมประสิทธิ์การพยากรณ์ (B) ใช้บอกทิศทางและปริมาณของผลกระทบของตัวแปรอิสระตัวนั้นๆ (X) ที่มีต่อโอกาสการเกิดเหตุการณ์ (Y) ค่านี้นำไปคำนวณค่า Odd ratio ซึ่งจะบอกถึงการเพิ่มขึ้น หรือลดลง ของโอกาสที่จะเกิดเหตุการณ์ และค่านี้นำไปใช้เขียนสมการโลจิสติกหรือโพรบิทในการทำนายการการเกิดเหตุการณ์ที่สนใจ

(2) Standard error ของสัมประสิทธิ์แต่ละตัว เป็นตัวกำหนดค่านัยสำคัญทางสถิติของตัวแปรอิสระตัวนั้นๆ

(3) Log likelihood เป็นค่าที่บอกถึงความเหมาะสมของโมเดล ถ้ามีค่าเข้าใกล้ 0 แสดงว่าโมเดลมีความเหมาะสม

(4) ค่าไคสแควร์ เป็นค่าที่บอกถึงความสอดคล้องระหว่างโมเดลและข้อมูลเชิงประจักษ์ ถ้าหากมีความสอดคล้องค่าไคสแควร์ไม่ควรแตกต่างอย่างมีนัยสำคัญทางสถิติ (ไม่มีนัยสำคัญทางสถิติ)

(5) Odd ratio หรือ ค่าสัดส่วนของเหตุการณ์ที่สนใจต่อเหตุการณ์ที่ไม่สนใจศึกษา ถ้า Odd ratio มีค่าเท่ากับ 1 แสดงว่าปัจจัยนั้นไม่มีความสัมพันธ์กับเหตุการณ์ ถ้ามากกว่า 1 แสดงว่าปัจจัยนั้นสัมพันธ์กับเหตุการณ์ และถ้าน้อยกว่า 1 แสดงว่าปัจจัยนั้นทำให้เกิดเหตุการณ์ได้น้อยกว่าปกติ

ในการวิเคราะห์ตามแบบจำลองโลจิทและโพรบิท มีผลการศึกษาที่ไม่แตกต่างกันมากนัก นอกจากขนาดกลุ่มตัวอย่างจะต่างกันมากๆจึงจะทำให้เห็นความแตกต่างช่วงปลายการแจกแจง แต่การวิเคราะห์รูปแบบโพรบิท มีความซับซ้อนมากกว่า การวิเคราะห์แบบโลจิท ดังนั้นในการศึกษาส่วนใหญ่จึงนิยมใช้แบบจำลองโลจิทแทนแบบจำลองโพรบิท รวมทั้งในการวิจัยอิสระครั้งนี้ด้วย

2.1.3 การวิเคราะห์ปัจจัยพื้นฐาน (Fundamental Analysis)

การศึกษาปัจจัยพื้นฐานที่มีอิทธิพลต่อการเพิ่มขึ้นของราคาหุ้นในงานวิจัยอิสระครั้งนี้ประกอบไปด้วย

- อัตราส่วนราคาปิดต่อกำไรต่อหุ้น (Pricé-Earning Ratio: P/E)

$$P/E = \frac{\text{ราคาตลาดของหุ้นสามัญ (P)}}{\text{กำไรสุทธิต่อหุ้นประจำงวด 12 เดือนของหุ้นสามัญ (E)}} \qquad (20)$$

- อัตราส่วนราคาปิดต่อมูลค่าหุ้นทางบัญชี (Price-Book Value: P/BV)

$$P/BV = \frac{\text{ราคาของหุ้นสามัญ}}{\text{มูลค่าตามบัญชีของหุ้นสามัญต่อหุ้น}} \quad (21)$$

- มูลค่าหุ้นทางบัญชี (Book Value Per Share : BVPS)

$$BVPS = \frac{\text{ส่วนของผู้ถือหุ้นสามัญของบริษัท}}{\text{จำนวนหุ้นสามัญ}} \quad (22)$$

- อัตราส่วนราคาปิดต่อมูลค่าทรัพย์สินสุทธิ (Price-Net Asset Value: P/NAV)

$$P/NAV = \frac{\text{ราคาปิดของหุ้นสามัญ}}{\text{มูลค่าทรัพย์สินสุทธิต่อหุ้น}} \quad (23)$$

- มูลค่าทรัพย์สินสุทธิ (Net Asset Value: NAV)

$$NAV = \frac{\text{สินทรัพย์สุทธิ}}{\text{จำนวนหน่วยลงทุนที่จำหน่ายแล้ว}} \quad (24)$$

- อัตราส่วนเงินปันผลตอบแทน (Dividend Yield: DIY)

$$DIY = \frac{\text{มูลค่าปันผลต่อหุ้น} \times 100}{\text{ราคาตลาดของหุ้น}} \quad (25)$$

- อัตราหมุนเวียนการซื้อขายหลักทรัพย์ (Turnover Ratio)

$$\text{Turnover Ratio} = \frac{\text{ผลรวมปริมาณการซื้อขายหลักทรัพย์ในช่วงเวลานั้น x 100}}{\text{ค่าเฉลี่ยปริมาณหุ้นจดทะเบียนกับตลาดหลักทรัพย์ในช่วงเวลานั้น}} \quad (26)$$

- มูลค่าหลักทรัพย์ตามราคาตลาด (Market Capitalization)

$$\text{Market Capital} = \text{ราคาปิดของหุ้น X ปริมาณหุ้นจดทะเบียนกับตลาดหลักทรัพย์} \quad (27)$$

- อัตราผลตอบแทนส่วนของผู้ถือหุ้น(Return on Equity: ROE)

$$ROE = \frac{\text{กำไร (ขาดทุน) สุทธิ} \times 100}{\text{รวมส่วนของผู้ถือหุ้นของบริษัทใหญ่ (เฉลี่ย)}} \quad (28)$$

- **อัตราส่วนราคาปิดต่อมูลค่าหุ้นทางบัญชี (Return on Asset: ROA)**

$$ROA = \frac{\text{กำไร (ขาดทุน) ก่อนภาษีเงินได้} \times 100}{\text{รวมสินทรัพย์ (เฉลี่ย)}} \quad (29)$$

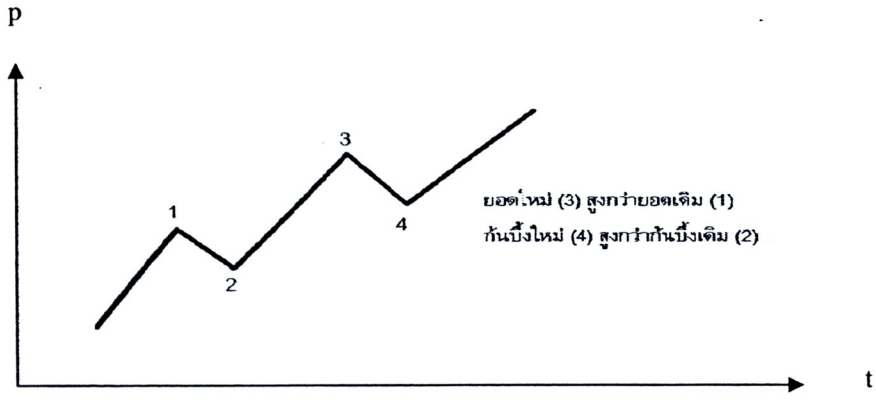
- **ดัชนีราคาหลักทรัพย์ (Stock Price Index)**

$$\text{ดัชนีราคาหุ้น} = \frac{\text{มูลค่าตลาดรวม ณ วันปัจจุบัน (Current Market Value) x 100}}{\text{มูลค่ารวม ณ วันฐาน (Base Market Value)}} \quad (30)$$

- **แนวโน้มราคาหุ้น (Trend)**

แนวโน้มที่ราคาหุ้นจะเคลื่อนตัวไปมี2รูปแบบคือ uptrendและdowntrend โดยมีรายละเอียดดังนี้

- Uptrend คือแนวโน้มขาขึ้นเป็นการชี้ถึงราคาหุ้นมีแนวโน้มสูงขึ้นในทางเทคนิค uptrend จะมีลักษณะดังรูป

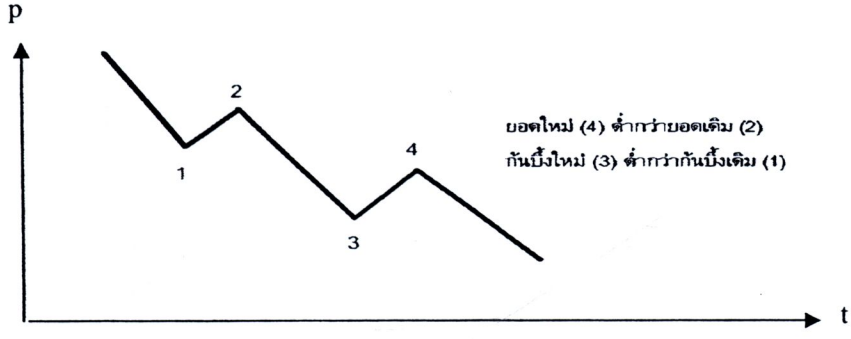


ที่มา: [ระบบออนไลน์] std.kku.ac.th

รูป 2.5 ภาพ uptrend

- Downtrend คือแนวโน้มขาลงเป็นการชี้ถึงราคาหุ้นมีแนวโน้มต่ำลงในทาง

เทคนิค downtrend จะมีลักษณะดังรูป



ที่มา: [ระบบออนไลน์] std.kku.ac.th

รูป 2.6 ภาพ downtrend

2.2 งานวิจัยที่เกี่ยวข้อง

ปิยะธิดา ดวงแก้ว (2552) ศึกษาเรื่องการวิเคราะห์ความสามารถในการชำระหนี้ของลูกค้าเช่าซื้อรถยนต์โดยใช้แบบจำลองโลจิท (Logistic regression) มีวัตถุประสงค์เพื่อศึกษาข้อมูลทั่วไปของลูกหนี้รวมถึงปัจจัยที่มีผลต่อความสามารถในการชำระหนี้และวิเคราะห์ความสามารถในการชำระหนี้ของลูกหนี้ในธุรกิจเช่าซื้อรถยนต์ของธนาคารธนชาติ จำกัด (มหาชน) สำนักงานภาคเหนือ 1 (เชียงใหม่) เพื่อใช้ในการวางแผนและกลยุทธ์ในการบริหารความเสี่ยงของธุรกิจเช่าซื้อรถยนต์ของธนาคารธนชาติ จำกัด (มหาชน) สำนักงานภาคเหนือ 1 (เชียงใหม่)โดยใช้ข้อมูลจากฐานข้อมูลลูกค้าสินเชื่อเช่าซื้อรถยนต์ ของธนาคาร โดยใช้ข้อมูลระบบงานสินเชื่อเช่าซื้อรถยนต์ในช่วงปี 2549 – 2552 ซึ่งเป็นข้อมูลรายตัวของลูกค้า ได้แก่ อายุของผู้เช่าซื้อรถยนต์ อาชีพ รายได้เฉลี่ยต่อเดือน ยี่ห้อรถยนต์ ราคารถยนต์ จำนวนเงินดาวน์ อัตราดอกเบี้ย ประเภทการให้สินเชื่อเช่าซื้อรถยนต์ ประเภทการจดทะเบียนรถยนต์ วงเงินให้สินเชื่อ ค่างวด จำนวนงวดสัญญา สัดส่วนรายได้ต่อค่างวดและสัดส่วนเงินดาวน์ต่อราคารถยนต์ผลการศึกษาพบว่าที่ระดับนัยสำคัญ 0.05 ตัวแปรอิสระที่มีผลต่อความสามารถในการชำระหนี้ของลูกค้าเช่าซื้อรถยนต์ประกอบด้วย 14 ปัจจัย คือ อายุของลูกค้าผู้ขอกู้ อาชีพของลูกค้าผู้ขอกู้ รายได้เฉลี่ยต่อเดือนราคารถยนต์จำนวนเงินดาวน์ อัตราดอกเบี้ย ประเภทการให้สินเชื่อรถยนต์ ประเภทการจดทะเบียนของรถยนต์ ค่างวด จำนวนงวดสัญญา สัดส่วนรายได้ต่อค่างวด และสัดส่วนเงินดาวน์ต่อราคารถยนต์ ส่วนยี่ห้อรถยนต์และวงเงินให้สินเชื่อนั้น ไม่มีผลต่อความสามารถในการผ่อนชำระของลูกค้าเช่าซื้อรถยนต์สำหรับการวิเคราะห์ความสามรถในการชำระหนี้ของลูกค้าเช่าซื้อรถยนต์โดยพิจารณาจากค่าความน่าจะเป็น

พบว่า กลุ่มลูกค้าที่เป็นลูกค้าปกติมีค่าความน่าจะเป็นอยู่ระหว่าง 0.61-1.00 สำหรับลูกหนี้ที่ไม่ก่อให้เกิดรายได้มีค่าความน่าจะเป็นอยู่ระหว่าง 0.00-0.30โดยมีการจำแนกถูกต้องร้อยละ 98.75

พิชญ์ฌาภรณ์ เมืองงาม ปัทมา ผาโคตร และ กวีพจน์ บันลือวงศ์ (2552) ศึกษาการทำนายผลสำเร็จการศึกษา โดยอาศัยเทคนิคทางการทำเหมืองข้อมูลคือเทคนิคข่ายงานเบย์มาใช้เพื่อวิเคราะห์ถึงตัวแปรที่มีผลต่อการทำนายผลสำเร็จการศึกษาของนักศึกษาระดับปริญญาตรีและในการทดสอบแบบจำลองที่ได้จะทำการทดสอบผลบนพื้นฐานวิธี k - fold Cross Validation ผลการทดลองแสดงให้เห็นว่าเทคนิคของข่ายงานเบย์สามารถค้นพบตัวแปรสำคัญสำหรับการทำนายผลการสำเร็จการศึกษาได้และให้ความแม่นยำในการทำนายสูงจากแบบจำลองที่ได้ทำการพัฒนาขึ้นทำให้ทราบตัวแปรสำคัญที่มีผลต่อการสำเร็จการศึกษาของนักศึกษาระดับปริญญาตรีคือรายได้รวมของครอบครัวอาชีพของมารดาและเกรดเฉลี่ยที่เข้ามาศึกษาในชั้นปีแรกซึ่งผลที่ได้มีความสอดคล้องกับผลการวิเคราะห์ความถดถอยเชิงพหุคูณ

รพีพรรณ ดวงคำสวัสดิ์ (2550) ศึกษาการวิเคราะห์เพื่อการพยากรณ์หนี้ที่ไม่ก่อให้เกิดรายได้ของธนาคารกรุงไทยจำกัด (มหาชน) ในเขตอำเภอเมืองเชียงใหม่ โดยใช้แบบจำลองโลจิต การศึกษาครั้งนี้ใช้ข้อมูลจากอัตราส่วนทางการเงินของนิติบุคคลที่มีวงเงินสินเชื่อมากกว่า 5 ล้านบาทในช่วงปี 2546 – 2549 จำนวน 43 บริษัท 133ข้อมูล ผลการศึกษาพบว่าอัตราผลตอบแทนของสินทรัพย์ อัตราส่วนความสามารถในการจ่ายดอกเบี้ย อัตราทุนหมุนเวียนและอัตราส่วนกระแสเงินสดจากการดำเนินงานต่อดอกเบี้ยจ่าย มีผลในทิศทางตรงกันข้ามกับการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ ส่วนปัจจัยอื่นๆคือ อัตรากำไรสุทธิ มีผลในทิศทางเดียวกันกับการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ และถ้าสัดส่วนทางการเงินมีค่าสูง โอกาสที่จะเกิดหนี้ที่ไม่ก่อให้เกิดรายได้จะมีน้อยในทางตรงกันข้ามเมื่ออัตราส่วนทางการเงินมีค่าต่ำโอการที่จะเกิดหนี้ที่ไม่ก่อให้เกิดรายได้จะมีสูง

นฤมล ตันตระวิวัฒน์ (2547) ศึกษาแนวโน้มหุ้นโดยใช้เทคนิคการจำแนกข้อมูล (Data Classification) โดยศึกษาเฉพาะหุ้นในกลุ่มอสังหาริมทรัพย์สำหรับข้อมูลที่นำมาใช้ในการวิเคราะห์นั้นประกอบด้วย 2 ส่วนโดยส่วนแรกเป็นข้อมูลการซื้อขายหุ้นจากตลาดหลักทรัพย์แห่งประเทศไทยและส่วนที่สองคือดัชนีชี้วัดแนวโน้มทางเทคนิคซึ่งได้จากการนำข้อมูลการซื้อขายหุ้นมาคำนวณโดยผลที่ได้จากการวิเคราะห์ถูกนำเสนอในรูปแบบของโปรแกรมประยุกต์ที่ผู้ใช้งานสามารถเลือกหุ้นที่สนใจได้รวมทั้งสามารถแสดงผลราคาหุ้นในรูปกราฟเพื่อให้ผู้ใช้งานสามารถมองแนวโน้มโดยรวมได้อีกด้วยผลลัพธ์จากการทำเหมืองข้อมูลคือรูปแบบ (Model) ที่ใช้สำหรับการวิเคราะห์แนวโน้มราคาหุ้นในช่วงเวลาดังนี้ 15 วัน , 1 เดือนและ 3 เดือนและจากการทดสอบโปรแกรมประยุกต์พบว่ามีค่าความถูกต้องมากกว่า 70 เปอร์เซ็นต์

กรรณิการ์ จรัญธัญกุล (2543) ศึกษาปัจจัยที่ก่อให้เกิดหนี้มีปัญหาของธุรกิจเช่าซื้อรถยนต์ของบริษัทลิสซิ่งแห่งหนึ่งในจังหวัดลำปาง ผลของการศึกษาจะใช้เป็นแนวทางในการปรับปรุงมาตรการการให้เงินกู้ของบริษัทต่อไป ข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่รวบรวมมาจากเอกสารข้อมูลลูกหนี้ของบริษัทสยามพาณิชย์ลิสซิ่ง จำกัด (มหาชน) สาขาลำปาง จำนวน 200 ราย และทำการวิเคราะห์แบบ Probit analysis ผลการศึกษาพบว่ามีตัวแปรที่มีนัยสำคัญ 7 ตัวแปร ได้แก่ รายได้ ประสบการณ์การทำงาน ค่างวด วงเงินให้สินเชื่อยอดหนี้คงเหลือ อัตราร้อยละของเงินดาวน์ และอาชีพของผู้เช่าซื้อ ผลการศึกษาพบว่า ลูกหนี้ที่มีรายได้สูง มีประสบการณ์การทำงานมานาน ค่างวดสูง ร้อยละของเงินดาวน์สูงและมีอาชีพรับราชการหรือรัฐวิสาหกิจจะทำให้โอกาสการเกิดหนี้มีปัญหามีต่ำ แต่ถ้าหากยอดหนี้คงเหลือมากจะมีโอกาสเกิดปัญหาหนี้สูง

พัฒนา กันยานนท์ (2543) ศึกษาปัจจัยที่ทำให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ของธนาคารพาณิชย์แห่งหนึ่งในจังหวัดเชียงราย ผลการศึกษาพบว่าลูกหนี้ส่วนใหญ่มีอายุ 41-50 ปี มีประสบการณ์การทำงานต่ำกว่า 11 ปี ประกอบอาชีพพาณิชยกรรมทั่วไปและค้าขาย มีรายได้ต่ำกว่า 20,000 บาทต่อเดือน ระยะเวลาการทำสัญญากู้อยู่ในช่วง 6-10 ปี ภาระหนี้คงเหลืออยู่ระหว่าง 1-5 ล้านบาทและมีวัตถุประสงค์กู้เพื่อทำธุรกิจอสังหาริมทรัพย์ สำหรับการทดสอบปัจจัยที่ทำให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้ โดยใช้สมการถดถอย Logistic Regression เพื่อประมาณค่าความน่าจะเป็นของตัวแปรอิสระว่าเป็นเป็นตัวแปรที่ทำให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้หรือไม่ ผลการศึกษาพบว่า ตัวแปรที่สามารถนำมาอธิบายปัจจัยที่ทำให้เกิดหนี้ที่ไม่ก่อให้เกิดรายได้อย่างมีนัยสำคัญ ได้แก่ อาชีพ ประสบการณ์การทำงาน ระดับรายได้ วงเงินให้สินเชื่อ ภาระหนี้คงเหลือ ภาระหนี้ในสถาบันการเงินอื่น จำนวนกิจการของลูกหนี้และวัตถุประสงค์การกู้ยืม ส่วนระยะเวลาการกู้และอายุของผู้ขอกู้นั้นไม่มีผลการทบต่อการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้

กรวรรณ วัฒนชัย (2539) ศึกษาปัจจัยที่มีผลกระทบต่อความต้องการสินเชื่อเพื่อการเช่าซื้อรถยนต์นั่งในจังหวัดเชียงใหม่ โดยมีวัตถุประสงค์หลักเพื่อวิเคราะห์ปัจจัยที่มีผลต่อความต้องการสินเชื่อเพื่อการเช่าซื้อรถยนต์นั่ง โดยใช้เทคนิคการวิเคราะห์ด้วยสมการถอดถอย (Regression analysis) ผลการศึกษาพบว่าปัจจัยที่มีผลกระทบต่อความต้องการสินเชื่อเพื่อการเช่าซื้อรถยนต์นั่ง ได้แก่ ราคาประเมินรถยนต์นั่ง รายได้เฉลี่ยต่อเดือนและงวดการชำระหนี้ ส่วนปัจจัยที่มีความสัมพันธ์ในทิศทางตรงกันข้ามกับปริมาณความต้องการสินเชื่อเช่าซื้อรถยนต์นั่ง คืออัตราดอกเบี้ยเรียกเก็บ สำหรับการวิเคราะห์ความยืดหยุ่นของอุปสงค์สินเชื่อเช่าซื้อรถยนต์นั่ง พบว่าปัจจัยทางด้านราคามีอิทธิพลต่อความต้องการสินเชื่อเช่าซื้อมากกว่าปัจจัยด้านรายได้