การใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ

Based on E-Commerce Data to Model Personal Loan from Considering Credit Scoring

กนกพล หลักบุญ

จุฑาทิพย์ สุวรรณสิงห์

ณัฏฐนิชา วิวรรธนาภิรักษ์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์)

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2561

Based on E-Commerce Data to Model Personal Loan from Considering Credit Scoring

KANOKPOL LAKBOON
JUTHATHIP SUWANNASING
NATTHANICHA WIWATTANAPIRAK

A SPECIAL PROBLEM SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR
THE DEGREE OF BACHELOR OF SCIENCE (APPLIED MATHEMATICS)
DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
ACADEMIC YEAR 2018

หัวข้อปัญหาพิเศษ	การใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ Based on E- Commerce Data to Model Personal Loan from Considering Credit Scoring
ชื่อนักศึกษา	นายกนกพล หลักบุญ รหัสนักศึกษา 58050002 นางสาวจุฑาทิพย์ สุวรรณสิงห์ รหัสนักศึกษา 58050028 นางสาวณัฏฐนิชา วิวรรธนาภิรักษ์ รหัสนักศึกษา 58050052
ปริญญา	วิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์)
ภาควิชา	คณิตศาสตร์
ปีการศึกษา	2561
อาจารย์ที่ปรึกษา	ดร.บุษยมาส พิมพ์พรรณชาติ
อาจารย์ที่ปรึกษาร่วม	ดร.กัมปนาท นามงาม

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์) ประจำปีการศึกษา 2561

คณะกรรมการสอบ	ลายมือชื่อ
ผศ.ดร.ธวัชชัย คำประภัสสร ประธานกรรมการ	
ดร.จิรภัทร์ หยกรัตนศักดิ์ กรรมการ	
ดร.บุษยมาส พิมพ์พรรณชาติ กรรมการและที่ปรึกษา	
ดร.กัมปนาท นามงาม กรรมการและที่ปรึกษาร่วม	

หัวข้อปัญหาพิเศษ	การใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ
ชื่อนักศึกษา	นายกนกพล หลักบุญ รหัสนักศึกษา 58050002
	นางสาวจุฑาทิพย์ สุวรรณสิงห์ รหัสนักศึกษา 58050028
	นางสาวณัฏฐนิชา วิวรรธนาภิรักษ์ รหัสนักศึกษา 58050052
ปริญญา	วิทยาศาสตรบัณฑิต (คณิตศาสตร์ประยุกต์)
ภาควิชา	คณิตศาสตร์
คณะ	วิทยาศาสตร์
มหาวิทยาลัย	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.)
ปีการศึกษา	2561
อาจารย์ที่ปรึกษา	ดร.บุษยมาส พิมพ์พรรณชาติ
อาจารย์ที่ปรึกษาร่วม	ดร.กัมปนาท นามงาม

บทคัดย่อ

ปัญหาพิเศษนี้ศึกษาการสร้างแบบจำลองการให้คะแนนสินเชื่อส่วนบุคคลโดยใช้ข้อมูลจากร้านค้าออนไลน์ มีวัตถุประสงค์เพื่อให้กลุ่มคนประเภทพ่อค้าแม่ค้าออนไลน์สามารถเข้าถึงแหล่งเงินทุนง่ายขึ้นจากการได้รับการอนุมัติการขอสินเชื่อจากธนาคารโดยใช้ข้อมูลพฤติกรรมการใช้อินเทอร์เน็ตบนหน้าเว็บไซต์ของร้านมาช่วยยืนยันว่าผู้กู้มีความสามารถในการชำระหนี้ ซึ่งข้อมูลที่นำมาวิเคราะห์ได้จากเทคนิคการดึงข้อมูลบนหน้าเว็บไซต์ด้วยภาษาอาร์ ผ่านขั้นตอนการทำความสะอาดข้อมูลโดยใช้เทคนิคทาง Data Mining และสร้างแบบจำลองฮิดเดนมาร์คอฟสำหรับร้านค้าร้านนั้นเพื่อพิจารณาการให้คะแนนสินเชื่อส่วนบุคคล

คำสำคัญ : การดึงข้อมูล คะแนนสินเชื่อส่วนบุคคล แบบจำลองฮิดเดนมาร์คอฟ

Title	Based on E-Commerce Data to Model Personal Loan from Considering Credit Scoring
Students	Mr. Kanokpol Lakboon Student ID 58050002
	Miss Juthathip Suwannasing Student ID 58050028
	Miss Natthanicha Wiwattanapirak Student ID 58050052
Degree	Bachelor of Science (Applied Mathematics)
Department	Mathematics
Faculty	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2018
Advisor	Dr.Busayamas Pimpunchart
Co-advisor	Dr.Kampanat Namngam

Abstract

For this century, E-commerce business is growing unprecedentedly. Because the trend of smartphones have been increasing, it makes mobile applications run faster. A large of group people can own their business and work at home via mobile applications. Some of them are struggling with banking. This special problem studies the personal loan credit scoring model using information from online stores. The objective is to allow the group of online merchants easily access funding from the bank. Customer feedback analysis is the key factor to approve banking. The data can be analyzed from the technique of data scraping on the website with the R language programming through the data cleaning process by using data mining technique and apply a Hidden Markov model to consider personal loan credit scoring for each online store.

Keywords : Web Scraping , Credit Scoring , Hidden Markov Model

กิตติกรรมประกาศ

ปัญหาพิเศษเล่มนี้สำเร็จลุล่วงได้ด้วยความกรุณาและความช่วยเหลือจากบุคคลหลายท่าน ขอขอบพระคุณ ดร.บุษยมาส พิมพ์พรรณชาติ และ ดร.กัมปนาท นามงาม อาจารย์ที่ปรึกษาปัญหาพิเศษ ที่ท่านได้เสียสละเวลาในการให้คำแนะนำ คำติชม ข้อคิดเห็นต่าง ๆ ตลอดจนช่วยหาวิธีการแก้ไขปัญหาข้อบกพร่องต่าง ๆ ที่เกิดขึ้น ไม่ว่าจะเป็นปัญหาทางด้านการศึกษาหรือปัญหาทางด้านการทำงาน รวมทั้งให้กำลังใจผู้จัดทำตลอดการทำปัญหาพิเศษนี้

ทางคณะผู้จัดทำขอขอบพระคุณ ผศ.ดร.ธวัชชัย คำประภัสสร ประธานกรรมการ และ ดร.จิรภัทร์ หยกรัตนศักดิ์ กรรมการ ที่ได้ให้คำแนะนำการทำปัญหาพิเศษครั้งนี้ รวมทั้งให้แนวคิดใหม่ ๆ มาปรับปรุงและพัฒนาตลอดจนแก้ไขตรวจสอบให้สมบูรณ์มากยิ่งขึ้น นอกจากนี้ทางคณะผู้จัดทำขอขอบพระคุณ ดร.กุลสวัสดิ์ จิตขจรวานิช ที่ได้กรุณาให้ข้อคิดเห็น และคำแนะนำที่เป็นประโยชน์ต่อปัญหาพิเศษฉบับนี้ รวมทั้งคอยให้กำลังใจและคำปรึกษาในระหว่างการทำงาน

ขอกราบขอบพระคุณ ครอบครัว เพื่อนนักศึกษา ตลอดจนผู้ที่เกี่ยวข้องทุกท่านที่ไม่ได้กล่าวไว้ ณ ที่นี้ ที่ได้ให้การสนับสนุนและให้กำลังใจตลอดการดำเนินงานของปัญหาพิเศษนี้ให้สำเร็จลุล่วงไปด้วยดี

กนกพล หลักบุญ
จุฑาทิพย์ สุวรรณสิงห์
ณัฏฐนิชา วิวรรธนาภิรักษ์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

สารบัญรูปภาพ (ต่อ)

หน้า

สารบัญรูปภาพ (ต่อ)

หน้า

คำย่อ/สัญลักษณ์

คำย่อ/สัญลักษณ์	คำอธิบาย
t	เวลา
T	จำนวนเวลาทั้งหมด
q_t	สถานะ ณ เวลาที่ t
N	จำนวนสถานะของแบบจำลองที่เป็นไปได้ทั้งหมด
$A = [a_{ij}]$	เมทริกซ์การเปลี่ยนสถานะที่ i ไปอีกสถานะ j
π_i	ความน่าจะเป็นของสถานะเริ่มต้นที่สถานะ i
V_M	สัญลักษณ์ของผลการสังเกตที่ M
M	จำนวนสัญลักษณ์ของผลการสังเกตที่แตกต่างกัน
$B = [b_i(k)]$	เมทริกซ์ความน่าจะเป็นของสัญลักษณ์การสังเกต k ในสถานะ i
O_t	ผลการสังเกต ณ เวลาที่ t
λ	องค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟซึ่งประกอบด้วย A, B, π
$\alpha_t(j)$	ความน่าจะเป็นของสถานะที่ j ณ เวลาที่ t
$\beta_t(i)$	ความน่าจะเป็นของการมองเห็นค่าสังเกตจากเวลาที่ $t+1$ จนถึงเวลาสุดท้าย โดยกำหนดให้สถานะปัจจุบันอยู่ในสถานะที่ i ณ เวลาที่ t

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เมื่อการทำงานในปัจจุบันมีการปรับเปลี่ยนไปตามเทคโนโลยีและการดำเนินชีวิตของผู้คน อินเทอร์เน็ตจึงเป็นสื่อกลางในการติดต่อสื่อสาร การค้าขาย และการดำเนินธุรกิจ ทำให้เกิดการสร้างรายได้จากหลายช่องทาง อาจเป็นการทำงานประจำควบคู่กับงาน part-time หรืองานอิสระ หรือการทำธุรกิจออนไลน์เป็นอาชีพหลัก เพราะรายได้ที่มากกว่างานประจำ สามารถเริ่มทำได้ง่าย และได้เป็นเจ้านายตนเอง ผู้คนส่วนใหญ่จึงเลือกที่จะเปิดธุรกิจเองบนโลกดิจิทัล ซึ่งไร้ขีดจำกัดด้านต่าง ๆ ทั้งอายุ เพศ และประสบการณ์ คนกลุ่มนี้ถูกเรียกว่าฟรีแลนซ์ และเมื่อมีช่องทางการทำงานที่หลากหลายและเปิดกว้างมากยิ่งขึ้นบนอินเทอร์เน็ต ธุรกิจใหม่อย่าง Startup ที่เปิดบริษัทขึ้นเพื่อมารองรับธุรกิจด้านไอทีและทำให้ธุรกิจเติบโตอย่างก้าวกระโดด เป็นการสร้างรายได้จำนวนมาก ยกตัวอย่าง การค้าขายออนไลน์ , Facebook ที่เริ่มต้นธุรกิจมาจาก Startup แน่นอนว่า Startup ที่กำลังกล่าวถึงอยู่นั้นเป็นระยะเริ่มต้นของธุรกิจและยังไม่มีแผนธุรกิจที่ชัดเจน เป็นผลให้รายได้มีความไม่คงที่และชัดเจนเหมือนกับข้าราชการหรือพนักงานประจำ

กลุ่มคนที่ไม่มีรายได้ประจำ หรือเป็นนายจ้างตัวเอง จะเรียกคนกลุ่มนี้ว่า Self-Employed อย่างเช่น พ่อค้าแม่ค้าออนไลน์ , Startup เนื่องจากการทำงานที่ไม่ใหญ่มาก การทำงาน การบริหารงานด้วยตัวเองหรืออาจมีลูกทีมขนาดย่อม จึงไม่ถึงขั้นเรียกว่าเป็นธุรกิจได้เต็มตัว

การขอสินเชื่อกับทางธนาคาร โดยทั่วไปจะให้เซ็นยอมรับให้ธนาคารทำการตรวจสอบข้อมูลสินเชื่อ เพื่อพิจารณาว่าควรปล่อยสินเชื่อหรือไม่ ซึ่งหากคนกลุ่มประเภท Self-Employed ต้องการขอสินเชื่อ การใช้หลักฐานยืนยันจำพวก ใบรับรองเงินเดือน จึงเป็นเรื่องที่รับประกันไม่ได้ แม้คนกลุ่มนี้ไม่อาจก่อให้เกิดหนี้สูญแต่เพราะมีหลักฐานทางการเงินที่ไม่แน่นอน เป็นสาเหตุที่ทำให้เกิดปัญหาในการเข้าถึงเงินทุน หรืออีกนัยหนึ่ง เป็นการยากที่คนกลุ่มนี้จะเข้าถึงเงินทุน หรือไม่สามารถเข้าถึงเงินทุนได้ เพื่อเป็นการสนับสนุนธุรกิจของคนกลุ่มนี้และการใช้จ่ายของคนในประเทศ เพื่อเป็นการช่วยส่งเสริมผลิตภัณฑ์มวลรวมของประเทศ (GDP) จึงเป็นเป้าหมายสำคัญในการบรรลุผลดังกล่าว

สำหรับแนวทางการสนับสนุนการขอสินเชื่อของกลุ่มคนประเภท Self-Employed ผู้จัดทำเล็งเห็นจากการใช้โซเชียลมีเดียของผู้คนในปัจจุบัน ข้อมูลมหาศาลที่เกิดขึ้นในแต่ละวันทำให้ธนาคารสามารถเข้าถึงพฤติกรรมการเงินของกลุ่มคนเหล่านั้น และสามารถใช้เป็นหลักฐานทางการเงินเพื่อพิจารณาการขอสินเชื่อสำหรับคนกลุ่มนี้ได้ โดยใช้ข้อมูลการใช้งานทางโซเชียลมีเดีย และข้อมูลทางธนาคารให้เกิดประโยชน์สูงสุด

ในเวลานี้ต้องยอมรับว่านอกจากความสำคัญทางด้านอินเทอร์เน็ตที่เป็นตัวแปรในการดำรงชีวิตของคนในปัจจุบันแล้ว ไม่สามารถปฏิเสธได้เลยว่า Online Banking เป็นอีกทางเลือกหนึ่งสำหรับการจัดการเงินแบบเดิม ๆ การใช้จ่ายที่ง่าย สะดวกสบาย และรวดเร็ว ในอดีตมีหลายกรณีที่ผู้ขอสินเชื่อไม่ผ่านการอนุมัติและผลการอนุมัตินั้นสร้างคำถามให้แก่ผู้ขอสินเชื่อมาโดยตลอดเนื่องจากรายได้ของผู้ขอสินเชื่อมีมากพอที่จะขอสินเชื่อได้แต่ผลการอนุมัติกลับไม่ผ่านเพราะหลักฐานทางการเงินที่ไม่เพียงพอ หากสามารถหาวิธีการที่ดีที่สุดในการให้คะแนนสินเชื่อได้โดยใช้ข้อมูลที่มีอยู่ให้เกิดประโยชน์สูงสุดและผ่านกระบวนการต่าง ๆ เพื่อช่วยบุคคลเหล่านั้น นอกจากจะเป็นการสร้างฐานลูกค้าใหม่แล้ว ยังสร้างผลประโยชน์ในการเป็นฐานสำคัญในการพัฒนาเศรษฐกิจของประเทศต่อไป

การขอสินเชื่อโดยใช้การให้คะแนนสินเชื่อ (Credit Scoring) ไม่ใช่เทคโนโลยีใหม่ที่นำมาใช้แต่เป็นอัลกอริทึมที่แตกต่างจากการทำงานทั่วไป ในปัจจุบันอัลกอริทึมมีอยู่จำนวนมากแล้วอัลกอริทึมใดจึงจะเหมาะสมกับงานที่สุด ขึ้นอยู่กับความพึงพอใจของผู้พิจารณาว่าจะสามารถยอมรับเปอร์เซ็นต์ของผลลัพธ์ในอัลกอริทึมได้มากน้อยเพียงใด อีกทั้งเป็นประโยชน์ต่อธนาคารที่ไม่ต้องเสี่ยงว่าจะต้องเสียผลประโยชน์จากผู้ขอสินเชื่อ เนื่องจากผู้ขอสินเชื่อมีความสามารถในการชำระหนี้คืนได้ในอนาคต

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อสามารถทำให้กลุ่มคนประเภทพ่อค้าแม่ค้าออนไลน์ได้รับการอนุมัติการขอสินเชื่อจากธนาคารได้ง่ายยิ่งขึ้น

2. เพื่อให้ธนาคารสามารถเลือกสรรวิธีการเลือกลูกค้าที่มาขอสินเชื่อ โดยใช้ข้อมูลบนโซเชียลมีเดียและดึงพฤติกรรมการใช้ข้อมูลบนอินเทอร์เน็ตมาช่วยยืนยันว่าบุคคลนั้น ๆ มีความสามารถในการชำระหนี้ เพื่อนำมาพิจารณาความเสี่ยงทางการเงิน ลดโอกาสการสร้างหนี้เสีย และเป็นการสร้างเกณฑ์การตัดสินใจที่มีมาตรฐานสำหรับเจ้าหน้าที่ในการพิจารณาอนุมัติสินเชื่อได้อย่างมีประสิทธิภาพ โดยการคำนวณคะแนนสินเชื่อ (Credit Scoring) ด้วยอัลกอริทึมแมชชีนเลิร์นนิง (Machine Learning)

1.3 สมมติฐานของการศึกษา

1. ข้อมูลทางร้านค้าออนไลน์มีอิทธิพลต่อการคำนวณคะแนนสินเชื่อส่วนบุคคล

2. กลุ่มผู้ขอสินเชื่อประเภทพ่อค้าแม่ค้าออนไลน์ที่ได้รับการคำนวณคะแนนสินเชื่อจากข้อมูลร้านค้าออนไลน์มีโอกาสได้รับอนุมัติสินเชื่อ

1.4 ขอบเขตการวิจัย

การวิจัยครั้งนี้มีเป้าหมายในการสร้างเกณฑ์การตัดสินใจให้กับเจ้าหน้าที่ปล่อยสินเชื่อให้กับเหล่าพ่อค้าแม่ค้าออนไลน์ได้มีโอกาสในการเข้าถึงแหล่งเงินทุนจากธนาคาร โดยการคำนวณคะแนนสินเชื่อจากคุณสมบัติของตัวบุคคลที่มาขอสินเชื่อด้วยอัลกอริทึมแมชชีนเลิร์นนิง (Machine Learning) โดยจะใช้ข้อมูลจากร้านค้าออนไลน์ของผู้กู้

ในการศึกษานี้จะศึกษาข้อมูลของพ่อค้าแม่ค้าออนไลน์จากร้านค้าออนไลน์เพียงอย่างเดียวเท่านั้น โดยราคาของสินค้าต่อชิ้นนั้นเป็นราคาที่ติดบนป้ายของสินค้านั้น ๆ เท่านั้น ไม่รวมราคาสินค้าในช่วงที่มีการลดราคาตามเทศกาล หรือโปรโมชั่นต่าง ๆ ผลลัพธ์ที่ได้จากการคำนวณนั้นจะสามารถนำมาประกอบการตัดสินใจได้ว่าผู้ขอสินเชื่อรายนี้เหมาะสมกับการได้รับสินเชื่อหรือไม่ และมุ่งเน้นการพัฒนาประสิทธิภาพของการคำนวณคะแนนสินเชื่อโดยใช้ฐานข้อมูลจากโซเชียลมีเดีย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. กลุ่มผู้ขอสินเชื่อประเภทพ่อค้าแม่ค้าออนไลน์มีโอกาสเข้าถึงแหล่งเงินทุนง่ายขึ้น
2. เป็นอีกหนึ่งทางเลือกในการให้คะแนนสินเชื่อของลูกค้าประเภทพ่อค้าแม่ค้าออนไลน์โดยใช้ฐานข้อมูลจากร้านค้าออนไลน์
3. เพื่อให้สามารถปล่อยสินเชื่อกับลูกค้าประเภทพ่อค้าแม่ค้าออนไลน์โดยธนาคารได้รับเงินชำระหนี้จากลูกค้า

1.6 ขั้นตอนของการศึกษา

1. การศึกษาอัลกอริทึมและแบบจำลองที่เป็นไปได้สำหรับการดำเนินงาน
2. ศึกษาข้อมูล (Explore Data) และตัวแปรที่ส่งผลกับแบบจำลอง
3. สร้างโปรแกรมสำหรับการดึงข้อมูลสำหรับเว็บไซต์ Shopee
4. ออกแบบโครงสร้างของแบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model)
5. ดึงข้อมูลสำหรับการคำนวณคะแนนสินเชื่อ (Credit Score) ของร้านค้าออนไลน์
6. เตรียมข้อมูลและทำความสะอาดข้อมูลสำหรับการคำนวณในแบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model)
7. คำนวณคะแนนสินเชื่อ (Credit Score)
8. บันทึกผลการดำเนินงาน
9. ทดสอบความแม่นยำ

1.7 ระยะเวลาในการดำเนินการ

แผนการทำงานแต่ละสัปดาห์	2561					2562			
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1.ศึกษาอัลกอริทึมและแบบจำลองที่เป็นไปได้สำหรับการดำเนินงาน	✓	✓	✓						
2.ศึกษาข้อมูล (Explore Data) และตัวแปรที่ส่งผลกับแบบจำลอง				✓	✓				
3.สร้างโปรแกรมสำหรับการดึงข้อมูลสำหรับเว็บไซต์ Shopee					✓	✓			
4.ออกแบบโครงสร้างของ Hidden Markov Model						✓			
5.ดึงข้อมูลสำหรับการคำนวณ Credit Score ของร้านค้าออนไลน์							✓		
6.เตรียมข้อมูลและทำความสะอาดข้อมูลสำหรับการคำนวณใน Hidden Markov Model							✓	✓	
7.คำนวณ Credit Score							✓	✓	
8.บันทึกผลการดำเนินงาน								✓	
9.ทดสอบความแม่นยำ								✓	✓

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การดำเนินงานวิจัยเรื่อง การใช้ข้อมูลทางโซเชียลมีเดียในการสร้างแบบจำลอง เพื่อพิจารณาการให้คะแนนสินเชื่อส่วนบุคคล เป็นการเพิ่มโอกาสสำหรับกลุ่มคนประเภทพ่อค้าแม่ค้าออนไลน์ในการเข้าถึงแหล่งเงินทุน โดยใช้ข้อมูลจากโซเชียลมีเดียซึ่งเป็นอีกหนึ่งทางเลือกในการการเข้าถึงข้อมูลหลักฐานทางการเงินสำหรับการพิจารณาความเสี่ยงทางการเงินของผู้กู้ เนื่องจากในปัจจุบันนั้นไม่สามารถปฏิเสธได้ว่าโซเชียลมีเดียมีอิทธิพลต่อการดำเนินชีวิตและเป็นแรงขับเคลื่อนเศรษฐกิจเป็นอย่างมาก ผู้วิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องตามหัวข้อ

กล่าวถึงทฤษฎีที่เกี่ยวข้องทั้งหมดของงานวิจัย ซึ่งการดำเนินการใช้งานตั้งแต่ต้นจนจบงานวิจัย โดยประกอบไปด้วยหัวข้อ E-Commerce Web Scraping การแจกแจงแบบปกติ การแปลงข้อมูลโดยใช้ลอการิทึม แบบจำลองฮิดเดนมาร์คอฟ และ Confusion Matrix ตามลำดับ

สำหรับส่วนนี้จะกล่าวถึงวิธีการคำนวณทั้ง 2 ปัญหาของแบบจำลองฮิดเดนมาร์คอฟ ซึ่งประกอบด้วยสามขั้นตอนโดยเรียงตามลำดับ คือ การคำนวณ Likelihood และการถอดรหัสวิเทอร์บิ ซึ่งทั้ง 2 ขั้นตอนนั้น ได้นำแบบอย่างมาจากงานวิจัยเรื่อง A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition ของ Lawrence R. Rabiner

2.1 E-Commerce

Electronic Commerce หรือ การพาณิชย์อิเล็กทรอนิกส์ หมายถึง การทำธุรกรรมทางเศรษฐกิจที่ผ่านสื่ออิเล็กทรอนิกส์ เช่น การซื้อขายสินค้าและบริหาร การโฆษณาสินค้า การโอนเงินทางอิเล็กทรอนิกส์ เป็นต้น

จุดเด่นของ E-Commerce คือ ประหยัดค่าใช้จ่าย และเพิ่มประสิทธิภาพในการดำเนินธุรกิจ โดยลดความสำคัญขององค์ประกอบของธุรกิจที่มองเห็นจับต้องได้ เช่น อาคารที่ทำการ ห้องจัดแสดงสินค้า (Show Room) คลังสินค้า พนักงานขายและพนักงานให้บริการต้อนรับลูกค้า เป็นต้น ดังนั้น ข้อจำกัดทางภูมิศาสตร์ คือ ระยะทางและเวลาทำการแตกต่างกัน จึงไม่เป็นอุปสรรคต่อการทำธุรกิจอีกต่อไป

2.1.1 อุปกรณ์และวิธีการทำ E-Commerce

อุปกรณ์เทคโนโลยีสารสนเทศประกอบด้วย ระบบสื่อสารโทรคมนาคม ระบบคอมพิวเตอร์และระบบฐานข้อมูล ระบบสื่อสารอาจเป็นระบบพื้นฐานทั่วไป เช่น ระบบโทรศัพท์ โทรสาร หรือวิทยุ โทรทัศน์ แต่ระบบอินเทอร์เน็ตซึ่งเชื่อมโยงถึงกันได้ทั่วโลก เป็น

ระบบเปิดกว้าง โดยเป็นระบบเครือข่ายของเครือข่าย ที่เรียกว่า World Wide Web มาจากความเป็นเอกลักษณ์คือสามารถสร้างให้มี Hyperlink จากหน้าหนึ่งไปอีกหน้าหนึ่งไป Webpage อื่น หรือไป Website อื่นได้อย่างมีประสิทธิภาพ นอกจากนี้ยังสามารถสื่อได้ทั้งภาพ เสียง และภาษาหนังสือที่หลากหลายซับซ้อน สามารถมีปฏิสัมพันธ์โต้ตอบกันได้ทันทีทันใด ข้อมูลอิเล็กทรอนิกส์สามารถบันทึกเก็บไว้หรือนำใช้ต่อเนื่องได้ การประยุกต์ใช้ และกระแสตอบรับธุรกิจบนอินเทอร์เน็ตจึงแพร่หลายภายในระยะเวลาอันสั้น

E-Commerce ใช้ติดต่อกับลูกค้าได้หลายระดับ ธุรกิจกับลูกค้า ธุรกิจกับธุรกิจ ธุรกิจกับภาครัฐ ฯ สาระของการติดต่อจะมี 4 ประการ คือ

- การขาย รวมการโฆษณา แสดงสินค้า เสนอราคา สั่งซื้อ คำนวณราคา

- การชำระเงิน การตกลงวิธีชำระเงิน สั่งโอนเงิน ให้ข้อมูลบัญชีธนาคารที่ใช้ตัดบัญชี ตลอดจนเงินดิจิทัลรูปแบบใหม่ ๆ

- การขนส่ง แจ้งวิธีการส่งมอบของ ค่าขนส่ง และสถานที่ติดต่อและระบบติดตามสินค้าที่ส่ง

- บริการหลังการขาย การติดต่อภายในบริษัท เช่นระบบบัญชี คลังสินค้า ระบบสั่งซื้อสินค้าและวัตถุดิบ สั่งผลิต ตลอดจนบริการลูกค้าหลังการขาย

2.1.2 ประเภทของ E-Commerce

พาณิชย์อิเล็กทรอนิกส์มีหลายประเภทและมีวิธีการที่แตกต่างกันในการจำแนกกลุ่มเหล่านี้ นักวิชาการกำหนดกรอบการทำงานจำนวนหนึ่งเพื่อจัดประเภทการพาณิชย์อิเล็กทรอนิกส์ ประเภทของการพาณิชย์อิเล็กทรอนิกส์ที่แตกต่างที่สำคัญคือธุรกิจกับธุรกิจ (B2B) ธุรกิจกับผู้บริโภค (B2C), ผู้บริโภคสู่ผู้บริโภค (C2C), ผู้บริโภคสู่ธุรกิจ (C2B) และ Mobile Commerce (M-Commerce)

1. Business-to-Business: B2B

เป็นธุรกรรมการค้าประเภทหนึ่งที่มีอยู่ระหว่างธุรกิจหรือธุรกรรมที่เกิดขึ้นระหว่างบริษัทและบริษัทอื่นเพื่อโอนบริการและผลิตภัณฑ์ คำอธิบายที่เป็นไปได้สำหรับสิ่งนี้อาจเป็นไปได้ว่าธุรกิจกับธุรกิจรวมถึงการค้าส่งออนไลน์ซึ่งธุรกิจขายวัสดุผลิตภัณฑ์และบริการให้กับธุรกิจอื่น ๆ บนเว็บไซต์

2. Business-to-Consumer: B2C

ธุรกิจกับผู้บริโภคหมายถึงธุรกรรมระหว่างธุรกิจและผู้บริโภคขั้นปลายดังนั้นจึงสร้างร้านค้าอิเล็กทรอนิกส์ที่เสนอข้อมูลสินค้าและบริการระหว่างธุรกิจและผู้บริโภคใน

ธุรกรรมค้าปลีกหรือเป็นอินเทอร์เน็ตและรูปแบบพาณิชย์อิเล็กทรอนิกส์ที่บ่งบอกถึงการเงิน การทำธุรกรรมหรือการขายออนไลน์ระหว่างธุรกิจและผู้บริโภค

3. Consumer-to-Business: C2B

เป็นการโอนบริการสินค้าหรือข้อมูลจากบุคคลสู่ธุรกิจหรือเป็นรูปแบบธุรกิจที่ผู้ใช้ปลายทางสร้างผลิตภัณฑ์และบริการที่ใช้โดยธุรกิจและสถาบัน

4. Consumer-to-Consumer: C2C

เป็นสื่ออิเล็กทรอนิกส์อำนวยความสะดวกอินเทอร์เน็ตซึ่งเกี่ยวข้องกับการทำธุรกรรมระหว่างผู้ใช้และเป็นรูปแบบธุรกิจที่ผู้บริโภคสองรายทำธุรกิจร่วมกันโดยตรง [13]

ตัวอย่าง

ร้านค้าออนไลน์ Shopee

เป็นตลาดซื้อขายแบบโซเชียลที่มุ่งเน้นการใช้งานผ่านโทรศัพท์มือถืออันดับแรก (Mobile First) เพื่อให้ทุกคนสามารถเลือกดูและซื้อขายได้อย่างสะดวก โดยเป็นแพลตฟอร์มที่ออกแบบขึ้นเพื่อชาวเอเชียตะวันออกเฉียงใต้ ด้วยการผสานคุณลักษณะของตลาดซื้อขายแบบผู้บริโภคสู่ผู้บริโภค (C2C) เข้ากับระบบการชำระเงินและการสนับสนุนด้านโลจิสติกส์ เพื่อให้การช้อปปิ้งออนไลน์กลายเป็นเรื่องที่สะดวก ปลอดภัย และไร้ความยุ่งยาก

รูปที่ 2.1 ตัวอย่างร้านค้าออนไลน์ Shopee

สำหรับ Shopee ในประเทศไทยถือว่าเป็น 1 ในประเภทแถบภูมิภาคอาเซียนที่โตเร็วมาก ๆ ซึ่ง Garena ได้เปิดตัวเมื่อเดือนธันวาคม 2558 ปัจจุบันมีผู้ใช้มากกว่า 4 ล้านคน ภายใน 2 ปี มีสินค้าให้เลือกซื้อมากกว่า 3 ล้านรายการ ซึ่งถือว่าโตเร็วกกว่า Lazada ที่เข้ามาทำการตลาดก่อนหน้าเป็นอย่างมาก

สิ่งที่ทำให้ Shopee แตกต่างจากคู่แข่งรายอื่นในตลาดอยู่ตรงที่ความรวดเร็ว ความปลอดภัย และการนำเสนอสินค้าในราคาที่ถูกกว่า

- **ความรวดเร็ว** ผู้ซื้อสามารถติดต่อกับผู้ขายได้โดยตรงผ่านข้อความแช็ต (LiveChat) สามารถสอบถามรายละเอียดของสินค้าได้

• **ความปลอดภัย** การซื้อขายผ่านแอปพลิเคชันที่เชื่อถือได้โดยมี Shopee เป็นสื่อกลาง เมื่อเกิดปัญหาก็สามารถสอบถามทาง Shopee ได้โดยตรง

• **สินค้าที่ถูกกว่า** ทาง Shopee มีการแจกโค้ดส่วนลด รวมถึงผู้ขายเองก็สามารถลดได้เองเช่นกัน แต่ตรงส่วนนี้ถือว่าแข่งขันกันสูงไม่ว่าจะเป็นของ Lazada หรือ 11 Streets ซึ่งทางผู้ซื้อต้องลองเช็กราคาจากหลาย ๆ แหล่งเสียก่อน[1]

2.2 Web Scraping

2.2.1 วัตถุประสงค์ของการทำ Web Scraping

การเติบโตอย่างรวดเร็วของเวิลด์ไวด์เว็บได้เปลี่ยนวิธีการแบ่งปันรวบรวมและเผยแพร่ข้อมูลอย่างมีนัยสำคัญ ข้อมูลจำนวนมหาศาลถูกจัดเก็บออนไลน์ทั้งในรูปแบบที่มีโครงสร้างและไม่มีโครงสร้าง

การใช้ประโยชน์เหล่านี้มักเกิดขึ้นได้เนื่องจากการมี Web Scraping โดยอัตโนมัติ หากไม่มีเทคนิคเหล่านี้จะเป็นไปไม่ได้ที่จะรวบรวมปริมาณข้อมูลซ้ำ ๆ และในเวลาที่เหมาะสม[10]

ตัวอย่าง

ตัวอย่างโค้ดเป็นการแสดงตัวอย่างการทำ Web Scraping โดยภาษา R ผ่านโปรแกรม R Studio ซึ่งจะเป็นการดึงข้อมูลจากเว็บไซต์ www.amazon.in เพื่อทำการจัดเก็บข้อมูลที่สนใจ

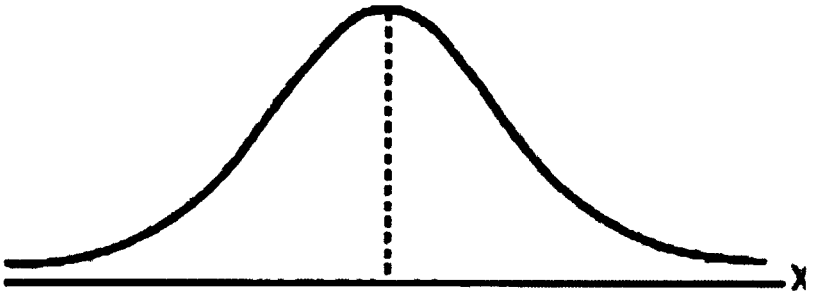
```
#loading the package:
library(xml2)
library(rvest)
library(stringr)
#Specifying the url for desired website to be scrapped
url <- "https://www.amazon.in/OnePlus-Mirror-Black-64GB-
Memory/dp/B0756Z43QS?tag=googinhydr18418-21&tag=googinkenshoo-
21&ascsubtag=aee9a916-6acd-4409-92ca-3bdbeb549f80"
#Reading the html content from Amazon
webpage <- read_html(url)
#scrape title of the product
title_html <- html_nodes(webpage, "h1#title")
title <- html_text(title_html)
head(title)
```

```
# remove all space and new lines
title <- str_replace_all(title, "[\r \n]" , "")
# scrape the price of the product
price_html <- html_nodes(webpage, "span#priceblock_ourprice")
price <- html_text(price_html)
#remove spaces and new line
str_replace_all(price,"[\r \n]" , "")
# print price value
head(price)
# scrape product description
desc_html <- html_nodes(webpage, "div#productDescription")
desc <- html_text(desc_html)
# replace new lines and spaces
desc <- str_replace_all(desc, "[\r \n \t]" , "")
desc <- str_trim(desc)
head(desc)
# scrape product rating
rate_html <- html_nodes(webpage, "span#acrPopover")
rate <- html_text(rate_html)
# remove spaces and newlines and tabs
rate <- str_replace_all(rate, "[\r \n]" , "")
rate <- str_trim(rate)
# print rating of the product
head(rate)
# Scrape size of the product
size_html <- html_nodes(webpage, "div#variation_size_name")
size_html <- html_nodes(size_html, "span.selection")
size <- html_text(size_html)
# remove tab from text
size <- str_trim(size)
```

```
# Print product size
head(size)
# Scrape product color
color_html <- html_nodes(webpage, "div#variation_color_name")
color_html <- html_nodes(color_html, "span.selection")
color <- html_text(color_html)
# remove tabs from text
color <- str_trim(color)
# print product color
head(color)
camera_html <- html_nodes(webpage , "div#feature-bullets")
camera_html <- html_nodes(camera_html , "span.a-list-item")
camera <- html_text(camera_html)
camera <- str_replace_all(camera, "[\r \n \t]" , "")
camera
#Combining all the lists to form a data frame
product_data <- list(Title = title, Price = price,Description = desc, Rating =
rate, Size = size , Color = color)
#Structure of the data frame
str(product_data)
# Include ?jsonlite? library to convert in JSON form.
library(jsonlite)
# convert dataframe into JSON format
json_data <- toJSON(product_data)
# print output
cat(json_data)
```

2.3 การแจกแจงแบบปกติ (Normal Distribution)

โดยทั่วไปการแจกแจงของตัวแปรที่ต่อเนื่องกันที่สำคัญที่สุด ได้แก่ การแจกแจงปกติ และทฤษฎีต่าง ๆ ในทางสถิติมักตั้งอยู่บนพื้นฐานของการแจกแจงแบบนี้ ลักษณะกราฟของการแจกแจงปกติ เรียกว่า เส้นโค้งปกติมีลักษณะเป็นรูประฆังที่สมมาตร มีความโค้งพอดี



รูปที่ 2.2 เส้นโค้งปกติ (Normal Curve)

<u>บทนิยาม</u> ตัวแปรสุ่ม (Random Variables)

เขียนแทนด้วย $X(\omega)$ หมายถึง ฟังก์ชันที่เกิดจากความสัมพันธ์จากปริภูมิตัวอย่าง S ไปยังจำนวนจริง นั่นคือ $X: S \to R$ หรือ $X(\omega) = x$ เมื่อ ω เป็นสมาชิกในปริภูมิตัวอย่าง S และ x เป็นจำนวนจริงใดๆ และมีโดเมน (Domain) ของ X มีค่าเป็น $\{\omega_1, \omega_2, .., \omega_{n(S)}\}$

ตัวแปรสุ่ม X ที่เป็นการแจกแจงปกตินี้ เรียกว่า ตัวแปรสุ่มปกติ (Normal Random Variable) มีรูปแบบคือ

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < \infty \qquad (2.1)$$

โดยที่ μ และ σ เป็นค่าคงที่

ถ้า X เป็นตัวแปรสุ่มที่มีการแจกแจงปกติ มีมัชฌิมเลขคณิตเท่ากับ μ ความแปรปรวนเท่ากับ σ^2 แทน ใช้สัญลักษณ์ $X \sim N(\mu, \sigma^2)$ แทน ซึ่งสามารถหาค่าของ $f(x)$ ได้ทุก ๆ ค่าของ x ที่เป็นจำนวนจริง[8]

2.3.1 สมบัติของเส้นโค้งปกติ

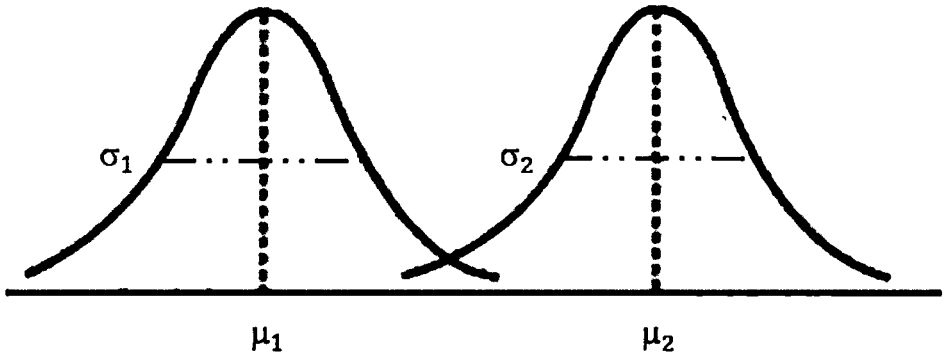
1. ค่ามัชฌิมเลขคณิต มัธยฐาน และฐานนิยม จะมีค่าเท่ากัน และจะอยู่ ณ จุดที่เส้นตรงลากผ่านจุดโด่งสุดเส้นโค้งนั้น ตั้งฉากกับแกน X

2. เส้นตรงที่ลากตั้งฉากกับแกน X ณ จุดที่เป็นค่ามัชฌิมเลขคณิตจะเป็นแกนสมมาตร และแกนสมมาตรจะแบ่งพื้นที่ใต้เส้นโค้งปกติออกเป็น 2 ส่วนเท่าๆ กัน

3. เมื่อลากปลายเส้นโค้งปกติทั้งสองข้างให้ห่างจากค่ามัชฌิมเลขคณิตออกไป เส้นโค้งจะเข้าใกล้แกน X แต่จะไม่ตัดแกน X

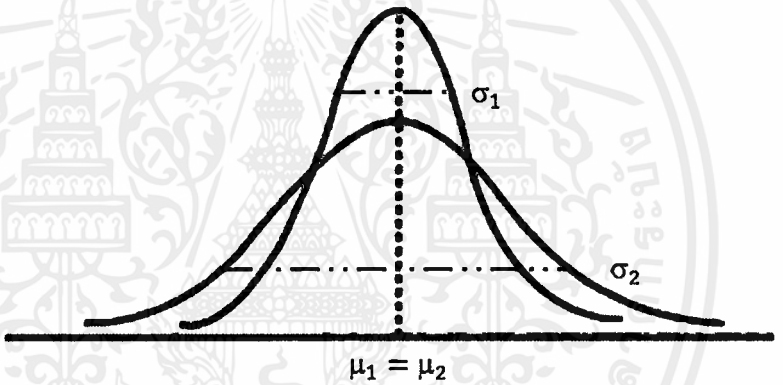
4. พื้นที่ใต้เส้นโค้งปกติแทนจำนวนความหนาแน่นของข้อมูล มีค่าเท่ากับ 1 เสมอ

5. ลักษณะเส้นโค้งปกติของข้อมูลสองชุดในลักษณะต่างๆ เช่น



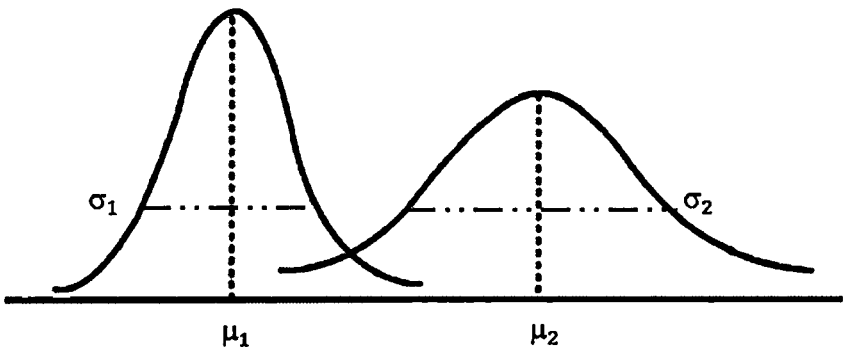
รูปที่ 2.3a ลักษณะเส้นโค้งปกติที่มี $\mu_1 \neq \mu_2$ และ $\sigma_1 = \sigma_2$

เส้นโค้งปกติที่มี $\mu_1 \neq \mu_2$ และ $\sigma_1 = \sigma_2$ จะมีลักษณะเหมือนกันทุกประการ ผิดตำแหน่งของจุดยอดของรูปทั้งสองรูปอยู่ต่างกัน



รูปที่ 2.3b ลักษณะเส้นโค้งปกติที่มี $\mu_1 = \mu_2$ และ $\sigma_1 < \sigma_2$

เส้นโค้งปกติที่มี $\mu_1 = \mu_2$ และ $\sigma_1 < \sigma_2$ จะมีลักษณะของเส้นทั้งสองจะแตกต่างกัน เส้นโค้งปกติที่มีส่วนเบี่ยงเบนมาตรฐานมากกว่าจะมีลักษณะที่แบน และลาดต่ำกว่าเส้นโค้งที่มีส่วนเบี่ยงเบนมาตรฐานน้อย แต่ตำแหน่งของจุดยอดอยู่ที่เดียวกัน



รูปที่ 2.3c ลักษณะเส้นโค้งปกติที่มี $\mu_1 \neq \mu_2$ และ $\sigma_1 \neq \sigma_2$

เส้นโค้งปกติที่มี $\mu_1 \neq \mu_2$ และ $\sigma_1 \neq \sigma_2$ ตำแหน่งของจุดยอดอยู่ต่างกัน และความโค้งของเส้นปกติของทั้งสองรูปก็ต่างกัน

จะเห็นได้ว่า เส้นโค้งปกติจะมีลักษณะโด่งมาก หรือโด่งน้อย ขึ้นอยู่กับส่วนเบี่ยงเบนมาตรฐาน ถ้าส่วนเบี่ยงเบนมาตรฐานใหญ่ เส้นโค้งจะมีลักษณะโด่งน้อยกว่า เส้นโค้งที่มีมาตรฐานเล็ก[6][8]

2.3.2 พื้นที่ใต้เส้นโค้งปกติ

เนื่องจากผลรวมของการทดลองสุ่มต้องเท่ากับหนึ่ง และพื้นที่ใต้เส้นโค้งปกติทั้งหมดเหนือแกน X มีค่าเท่ากับหนึ่ง ดังนั้นในการหาความน่าจะเป็นของตัวแปรสุ่มปกติอาจใช้พื้นที่ภายใต้เส้นโค้งปกติเป็นตัวแทนของตัวแปรสุ่มนั้น

ดังนั้น พื้นที่ใต้เส้นโค้งย่อมขึ้นอยู่กับค่ามัชฌิมเลขคณิต และส่วนเบี่ยงเบนมาตรฐานของประชากรหมู่นั้น ๆ สำหรับตัวแปรสุ่ม X ซึ่งเป็นแบบ $N\sim(\mu,\sigma^2)$ จะเห็นได้ว่า

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad (2.2)$$

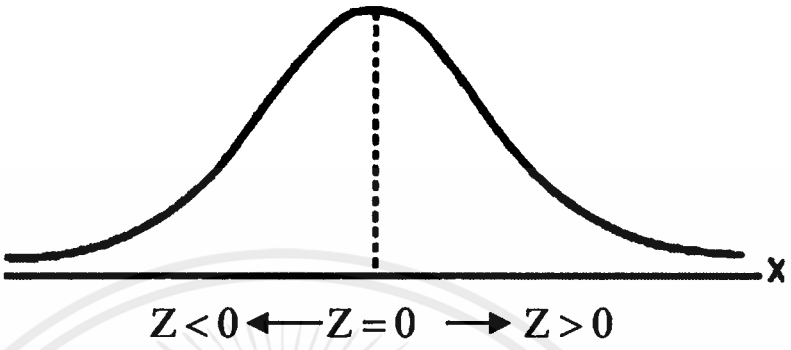
เพื่อให้ง่ายต่อการหาพื้นที่ใต้เส้นโค้ง สำหรับทุก ๆ ค่าของ μ และ σ ที่เปลี่ยนไปจึงมีการสร้างตารางมาตรฐานของพื้นที่ใต้เส้นโค้งปกติ (ดูตาราง Z) ซึ่งแสดงถึงพื้นที่ใต้เส้นโค้งระหว่าง $Z = 0$ ถึง $Z = 1$ โดยที่ Z เป็นคะแนนมาตรฐานที่มีการแจกแจงปกติ ซึ่งมีค่ามัชฌิมเลขคณิตเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 ตารางนี้จะใช้ได้กับข้อมูลทั่วไป ดังนั้นก่อนที่จะใช้ตารางนี้ จึงจำเป็นต้องเปลี่ยนข้อมูลดิบที่ได้มาให้เป็นคะแนนมาตรฐานเสียก่อน โดยใช้สูตร

$$Z = \frac{\text{ส่วนเบี่ยงเบนจากมัชฌิมเลขคณิต}}{\text{ส่วนเบี่ยงเบนมาตรฐาน}}$$

นั่นคือ

$$Z = \frac{X - \mu}{\sigma} \qquad (2.3)$$

<u>บทนิยาม</u> การแจกแจงปกติแบบมาตรฐาน (Standard Normal Distribution) คือ การแจกแจงของตัวแปรสุ่มปกติที่มีมัชฌิมเลขคณิตเป็นศูนย์ และส่วนเบี่ยงเบนมาตรฐานเป็นหนึ่ง ใช้สัญลักษณ์ $Z \sim N(0,1)$



รูปที่ 2.4 การแจกแจงปกติแบบมาตรฐาน (Standard Normal Distribution)

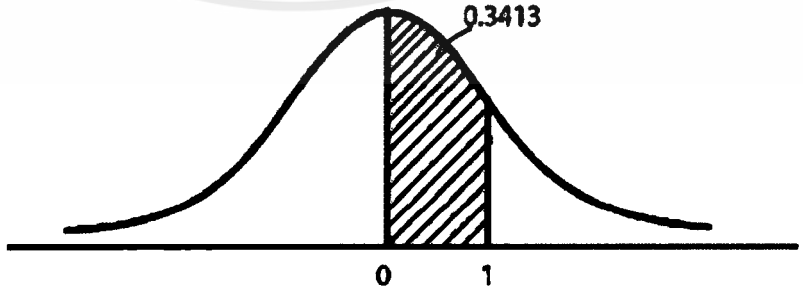
ถ้า $X = \mu$ แล้ว $Z = \frac{X-\mu}{\sigma} = 0$

ถ้า $X > \mu$ แล้ว $Z = \frac{X-\mu}{\sigma} > 0$

ถ้า $X < \mu$ แล้ว $Z = \frac{X-\mu}{\sigma} < 0$

โดยที่ส่วนโค้งสมมาตร (Symmetry) กันที่จุด $X = \mu$ ดังนั้น จึงไม่จำเป็นต้องสร้างตารางที่แสดงค่า Z เป็นลบไว้ เพราะข้างที่ Z เป็นบวกใช้ได้สำหรับข้างที่ Z เป็นลบด้วย

ตารางที่สร้างขึ้น จะบอกค่าพื้นที่ใต้เส้นโค้งนับจากจุดมัชฌิมเลขคณิตมายังที่มีค่า Z ตรงที่กำหนดไว้ สมมติว่าที่จุด Z มี $Z = 1$ จากตารางอ่านพื้นที่ได้ = 0.3413 พื้นที่นี้คือ พื้นที่ล้อมรอบด้วยเส้นโค้งแกนนอนและเส้นตรง $Z = 1$ ดังรูป หรือ $P(0 < Z < 1) = 0.3413$ นั่นเอง



รูปที่ 2.5 พื้นที่ใต้เส้นโค้งนับจากจุดมัชฌิมเลขคณิตมายังที่มีค่า Z กำหนดไว้

นั่นคือ พื้นที่ทั้งหมดทางขวามือ = พื้นที่ทั้งหมดทางซ้ายมือ = 0.5

ฉะนั้น $P(Z \leq 1) = 0.5 + P(0 < Z < 1)$

$= 0.5 + 0.3413$

$= 0.8413$

และ $P(Z \geq 1) = 0.5 - P(0 < Z < 1)$

$= 0.5 - 0.3413$

$= 0.1587$

2.3.3 เปอร์เซ็นไทล์ (Percentile)

การวัดตำแหน่งข้อมูล เป็นการแปลงข้อมูลแต่ละชุดให้อยู่ในลักษณะเดียวกัน เพื่อประโยชน์ในการเปรียบเทียบข้อมูลระหว่างข้อมูลคนละชุดกัน การแปลงข้อมูลมีลักษณะเป็นการแบ่งชุดข้อมูลออกเป็นส่วนย่อย ๆ มีทั้งแบ่งออกเป็น 4,10 และ 100 ส่วน

เปอร์เซ็นไทล์ (Percentile) คือ การวัดตำแหน่งที่แบ่งข้อมูลโดยเรียงจากน้อยไปมากออกเป็น 100 ส่วนเท่ากันๆ กัน

เนื่องจากข้อมูลแต่ละชุดมีลักษณะแตกต่างกัน ดังนั้นการจะนำคะแนนที่อยู่ต่างชุดกันมาเปรียบเทียบกันจึงจำเป็นต้องนำข้อมูลแต่ละชุดนั้นมาแปลงให้มีลักษณะเดียวกันเสียก่อน[12]

โดยการหาเปอร์เซ็นไทล์นั้น จะนำเรื่องพื้นที่ใต้กราฟมาช่วยในการหาสถิติที่เกี่ยวกับจำนวนข้อมูล และเงื่อนไขต่าง ๆ โดยที่รู้ค่ามัชฌิมเลขคณิต และส่วนเบี่ยงเบนมาตรฐาน สิ่งที่ต้องเข้าใจเกี่ยวกับพื้นที่ใต้กราฟ คือ

1. พื้นที่ใต้กราฟใช้อ้างอิงเปอร์เซ็นต์ของข้อมูลทั้งหมด
2. พื้นที่ใต้กราฟทั้งหมด คือ 1 ซึ่งหมายความว่าจำนวนข้อมูลทั้งหมด 100%
3. จุด $Z = 0$ จะแบ่งข้อมูลออกเป็นสองส่วนเท่าๆ กัน ซึ่งพื้นที่แต่ละส่วน คือ 0.5 นั่นคือแต่ละส่วนมีจำนวนข้อมูล 50% ของข้อมูลทั้งหมด
4. ตำแหน่งเปอร์เซ็นไทล์ที่ k ของข้อมูล คือ $Z = a$ ที่ทำให้พื้นที่ใต้กราฟของ $Z \leq a$ มีค่าเป็น $k\%$[4][8]

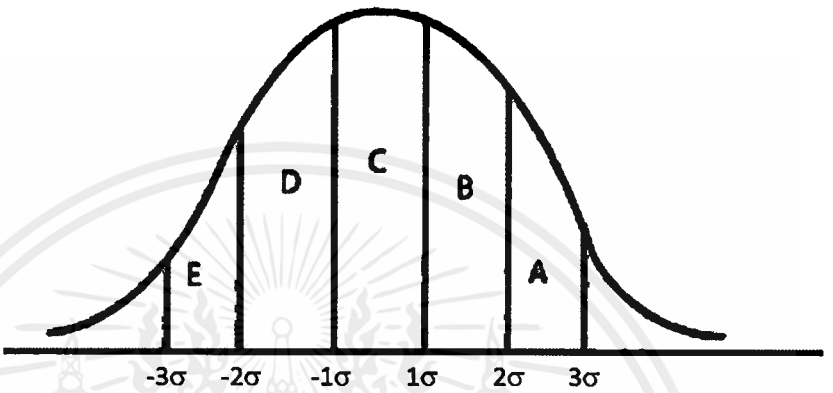
2.3.4 การกำหนดระดับ (Grading)

ในการพิจารณาเปรียบเทียบคะแนนผลการสอบของนิสิตแต่ละคนในแต่ละกลุ่ม โดยการจัดตำแหน่งของคะแนน กำหนดกลุ่มของคะแนนเป็นพวก ๆ และให้ระดับคะแนนในกลุ่มนั้น เป็นระดับคะแนน A,B,C,D หรือ E หรืออาจจะแบ่งย่อยเป็นระดับ A,B+,B,C+,C,D+,D หรือ E

สมมติให้การกระจายของคะแนนเป็นแบบโค้งปกติ และมีการให้ระดับคะแนนเป็น A,B,C,D หรือ E ขั้นแรกแบ่งการแจกแจงของโค้งปกตินี้เป็น 5 ส่วนเท่า ๆ กันตามแกนนอน

(X) โดยที่ X แทนคะแนน โดยให้คะแนนสูงสุดของระดับ A มากกว่าค่าเฉลี่ยอยู่ 3 เท่าของส่วนเบี่ยงเบนมาตรฐาน $(\mu + 3\sigma)$ และคะแนนต่ำสุดของระดับ E น้อยกว่าค่าเฉลี่ยอยู่ 3 เท่าของส่วนเบี่ยงเบนมาตรฐาน $(\mu - 3\sigma)$

นั่นคือ 99.74% ของคะแนนทั้งหมดจะอยู่ภายใต้ในช่อง $\mu - 3\sigma$ และ $\mu + 3\sigma$ (คะแนนสูงสุดของระดับ A ถึงคะแนนต่ำสุดของระดับ E จะกลุ่มคะแนนทั้งหมดไว้ถึง 99.74%) ดังรูป 2.6



รูปที่ 2.6 การกระจายของคะแนนเป็นแบบโค้งปกติ

พิสัยของแต่ละเกรด (คะแนนสูงสุด-คะแนนต่ำสุด) จะเท่ากับ 1.2 เท่าของส่วนเบี่ยงเบนมาตรฐาน (6 เท่าของส่วนเบี่ยงเบนมาตรฐาน ถูกแบ่งออกเป็น 5 ส่วนเท่า ๆ กันตามแนวแกน X) นั่นคือ

ระดับ A	คะแนนสูงสุด	$= \mu + 3\sigma$
	คะแนนต่ำสุด	$= \mu + 1.8\sigma$
ระดับ B	คะแนนสูงสุด	$= \mu + 1.8\sigma$
	คะแนนต่ำสุด	$= \mu + 0.6\sigma$
ระดับ C	คะแนนสูงสุด	$= \mu + 0.6\sigma$
	คะแนนต่ำสุด	$= \mu - 0.6\sigma$
ระดับ D	คะแนนสูงสุด	$= \mu - 0.6\sigma$
	คะแนนต่ำสุด	$= \mu - 1.8\sigma$
ระดับ E	คะแนนสูงสุด	$= \mu - 1.8\sigma$
	คะแนนต่ำสุด	$= \mu - 3\sigma$

2.4 การแปลงข้อมูลโดยใช้ลอการิทึม (Logarithm Transformation)

การแปลงข้อมูล หมายถึง การเปลี่ยนสภาพของข้อมูลที่ศึกษาให้มีการแจกแจงแบบปกติหรือทำให้ความแปรปรวนมีค่าเท่ากัน เนื่องจากข้อตกลงเบื้องต้นของการทดสอบสถิติบางตัวได้กำหนดไว้ เช่น การทดสอบค่าเฉลี่ย การทดสอบความแปรปรวน (Analysis of Variance) การวิเคราะห์การถดถอย (Regression Analysis) เป็นต้น การแปลงข้อมูลทำได้โดยใช้วิธีการทางสถิติ มีวัตถุประสงค์เพื่อแปลงข้อมูลเป็นมาตรใหม่แล้วข้อมูลมีการแจกแจงแบบปกติหรือใกล้เคียงกันแบบปกติ ค่าเฉลี่ยและความแปรปรวนของข้อมูลที่แปลงแล้วเป็นอิสระต่อกันจะทำให้ความแปรปรวนมีค่าเท่ากัน การแปลงข้อมูลกระทำได้โดยการยกกำลังข้อมูลเดิม โดยประมาณด้วยตัวยกกำลังตามเกณฑ์ที่กำหนดจากความสัมพันธ์ระหว่างความแปรปรวนที่เป็นสัดส่วนกับค่าเฉลี่ยของประชากรแต่ละทรีทเมนต์ ดังนี้

1. การแปลงโดยใช้รากที่สอง (Square Root Transformation)
2. การแปลงโดยใช้ลอการิทึม (Logarithmic Transformation)
3. การแปลงโดยใช้รากที่สองของการกลับเศษส่วน (Reciprocal Square Root Transformation)
4. การแปลงโดยใช้เศษส่วน (Reciprocal Transformation)[3]

การแปลงข้อมูลโดยใช้ลอการิทึม (Logarithm Transformation) เป็นการแปลงข้อมูลโดยใช้ค่าลอกาลิทึม อาจเรียกสั้นๆ ว่า ล็อก เป็นวิธีการแปลงข้อมูลที่มีประสิทธิภาพมากที่สุด ถ้าข้อมูลมีค่าส่วนเบี่ยงเบนมาตรฐาน (σ_i) และค่าเฉลี่ย (μ_i) เป็นสัดส่วนเท่ากัน นั่นคือ $\sigma_i = k\mu_i$ ผลจากการแปลงทำให้ข้อมูลมีความแปรปรวนเท่ากัน ข้อมูลที่มีค่าน้อยกว่า 0 เมื่อการแปลงข้อมูลนี้จะไม่สามารถทำได้ เพราะค่าล็อกไม่สามารถติดลบได้ กรณีที่ข้อมูลมีค่าเป็น 0 ให้ใช้ $\log(x+1)$ สำหรับการเลือกฐานของล็อก ฐาน 10 จะง่ายที่สุดสำหรับการแปลงข้อมูล

2.5 แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model)

ทฤษฎีพื้นฐานของแบบจำลองฮิดเดนมาร์คอฟถูกตีพิมพ์ครั้งแรกโดย Baum และคณะในช่วงปลายทศวรรษ 1960 ถึงต้นทศวรรษ 1970 หลังจากนั้น Baker กับ Jelinek และคณะได้นำแบบจำลองฮิดเดนมาร์คอฟมาประยุกต์ใช้กับการประมวลผลเสียงพูด อย่างไรก็ตามแบบจำลองฮิดเดนมาร์คอฟถูกใช้อย่างแพร่หลายในการประมวลผลเสียงพูดในราวทศวรรษที่ 1980 เนื่องจากทฤษฎีของแบบจำลองฮิดเดนมาร์คอฟถูกตีพิมพ์ในวารสารทางคณิตศาสตร์และในงานวิจัยแรก ๆ อธิบายถึงการนำแบบจำลองฮิดเดนมาร์คอฟไปประยุกต์กับการประมวลเสียงพูดไม่ละเอียดเพียงพอ ทำให้ผู้อ่านไม่สามารถเข้าใจและไม่สามารถนำแบบจำลองฮิดเดนมาร์คอฟไปประยุกต์ใช้ในงานวิจัยของตนได้ ด้วย

ปัญหาที่เกิดขึ้นทำให้มีงานวิจัยมากมายอธิบายถึงขั้นตอนและวิธีการประยุกต์ใช้แบบจำลองฮิดเดนมาร์คอฟกับการประมวลเสียงพูด

2.5.1 บทนิยามที่เกี่ยวข้อง

<u>บทนิยาม</u> กระบวนการเฟ้นสุ่ม (Stochastic Process)

หมายถึง เซตของสมาชิกทุกตัวในตัวแปรสุ่ม $\{X_t\}_{t\in T}$

<u>บทนิยาม</u> Discrete-time Markov Chain

ให้ $\{X_t\}_{t\in T}$ เป็นกระบวนการเฟ้นสุ่มที่มีสถานะปริภูมิพารามิเตอร์ T เมื่อ $T = \{0,1,2,\dots\}$ และมีคุณสมบัติมาร์คอฟ

$$P(X_{n+1} = j|X_0 = i_0, X_1 = i_1, X_2 = i_2, \dots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j|X_n = i)$$

เมื่อ $\forall n \in T$ และ $i_0, i_1, i_2, \dots, i_{n-1}, i, j \in S$

จะเห็นว่า ถ้า $\{X_t\}$ เป็นลูกโซ่มาร์คอฟ แล้วความน่าจะเป็นของการเกิดเหตุการณ์ที่เวลา $n+1$ ไม่ขึ้นกับเหตุการณ์ในอดีตที่เวลา $0,1,2,\dots,n-1$ แต่จะขึ้นกับเหตุการณ์ ณ ปัจจุบันที่เวลา n เท่านั้น

<u>บทนิยาม</u> เมทริกซ์เปลี่ยนสถานะ (Transition Probability Matrix)

ความน่าจะเป็นในการเปลี่ยนสถานะ 1 ขั้น ณ เวลา n (One-Step Transition Probability)

$$a_{ij}^{n,n+1} = P(X_{n+1} = j|X_n = i)$$

ถ้าตัวแปรเวลาเป็นอิสระต่อกัน (Stationary Transition Probability) จะได้ว่าความน่าจะเป็นในการเปลี่ยนแปลงสถานะ 1 ขั้น ณ เวลา คือ a_{ij} ถ้า $i,j = 0,1,2,\dots$ จะได้ว่า

1. $a_{ij} \geq 0, \forall i,j$
2. $\sum_{j=0}^{\infty} a_{ij} = 1, \forall i$

2.5.2 แบบจำลองมาร์คอฟไปสู่แบบจำลองฮิดเดนมาร์คอฟ

แบบจำลองมาร์คอฟเป็นแบบจำลองที่สามารถนำไปใช้ได้ประโยชน์ได้ แต่อย่างไรก็ตามแบบจำลองนี้มีข้อจำกัดมากเกินไป ทำให้ไม่สามารถนำไปประยุกต์ใช้กับการพิจารณาคะแนนสินเชื่อได้ ดังนั้น จึงจำเป็นที่จะต้องขยายแนวคิดแบบจำลองมาร์คอฟเพิ่มขึ้นเพื่อให้ครอบคลุม กรณีที่ผลการสังเกต (Observation) เป็นฟังก์ชันของความน่าจะเป็นสำหรับแต่ละสถานะ แบบจำลองลักษณะนี้คือแบบจำลองฮิดเดนมาร์คอฟซึ่งจะไม่ทราบถึงการเปลี่ยนแปลง

ของสถานะ มีเพียงแต่ลำดับผลการสังเกตเท่านั้นที่สามารถทราบได้ เพื่อให้เกิดเห็นภาพมากขึ้น พิจารณาปัญหาการโยนเหรียญต่อไปนี้

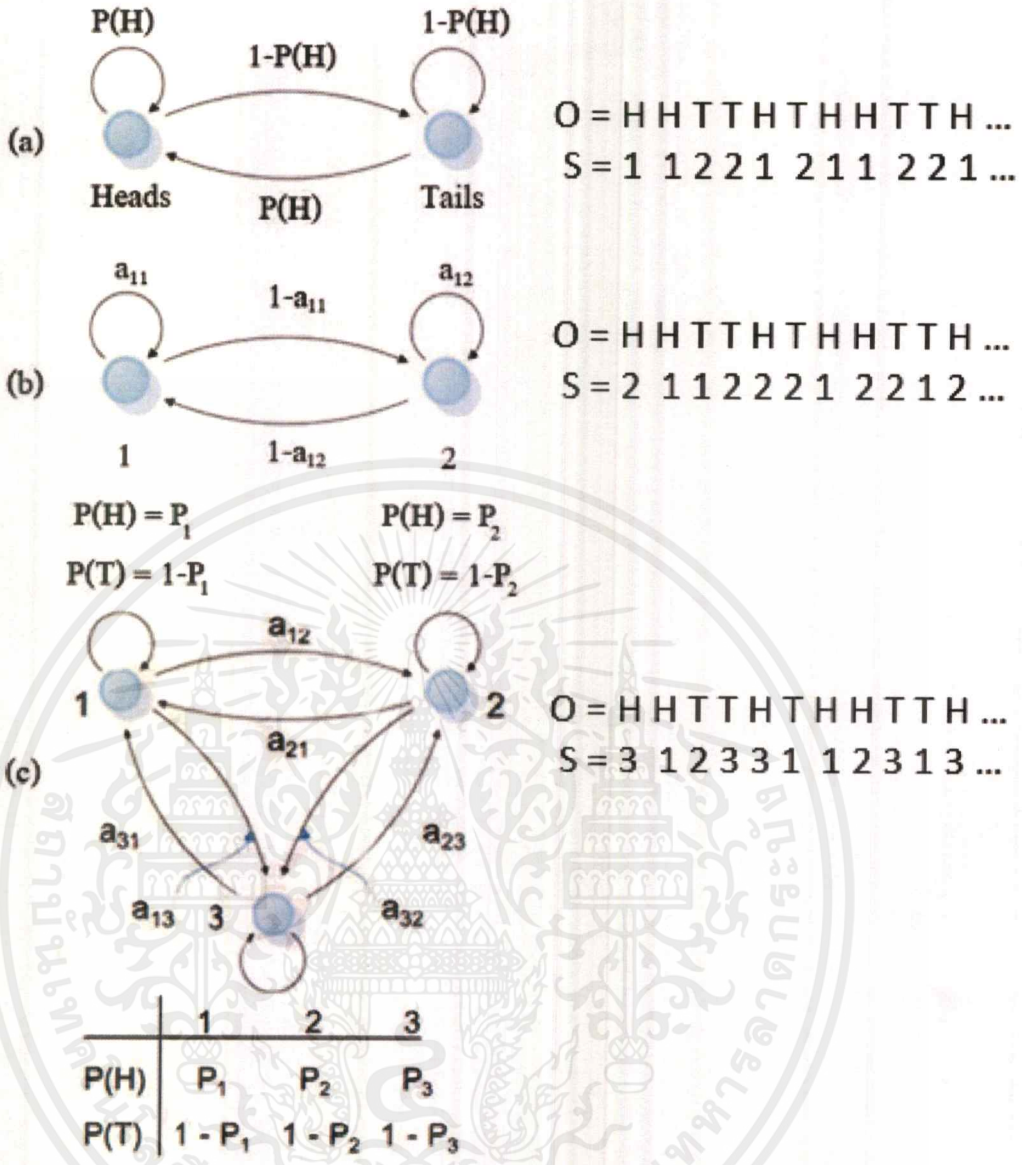
แบบจำลองการโยนเหรียญ (Coin Toss Model) สมมติว่า อยู่ในห้องซึ่งมีม่านกั้น ทำให้ไม่สามารถทราบได้ว่าเกิดอะไรขึ้น อีกฝั่งของม่านมีคนอีกคนหนึ่งกำลังทำการทดลองโยนเหรียญ (อาจใช้เหรียญมากกว่า 1 เหรียญ) ไม่ทราบวิธีการทดลองและคนที่ทำการทดลองจะบอกเพียงแค่ผลลัพธ์ของการโยนเหรียญในแต่ละครั้ง ซึ่งจะเป็น หัว (H) หรือก้อย (T) เท่านั้น ดังนั้นลำดับผลการสังเกตที่ทราบได้จึงเป็น

$$O = O_1, O_2, O_3, \dots, O_t \tag{2.4a}$$

$$= H, H, T, H, T, \dots, T \tag{2.4b}$$

(2.4b) เป็นรูปแบบหนึ่งของลำดับผลการสังเกตที่เป็นไปได้ การสร้างแบบจำลองฮิดเดนมาร์คอฟเพื่อที่จะอธิบายผลการสังเกตที่เห็น ต้องกำหนดว่าสถานะแต่ละสถานะของแบบจำลองฮิดเดนมาร์คอฟจะใช้แทนสิ่งใดและกำหนดว่าจะมีทั้งหมดกี่สถานะ ตัวเลือกหนึ่งที่เป็นไปได้สำหรับการโยนเหรียญ คือ ตั้งสมมติฐานว่า การโยนเหรียญนั้นใช้เหรียญเพียงแค่เหรียญเดียว ในกรณีนี้ จะสามารถจำลองเหตุการณ์ได้ด้วยแบบจำลอง 2 สถานะ แต่สถานะแทนหน้าของเหรียญ ดังแสดงในรูปที่ 2.7a ซึ่งแบบจำลองนี้ คือ แบบจำลองมาร์คอฟนั่นเอง

อีกวิธีหนึ่งที่เป็นไปได้ คือ ตั้งสมมติฐานว่าการโยนเหรียญใช้เหรียญมากกว่าหนึ่งเหรียญ เช่น พิจารณาว่ามีเหรียญทั้งหมด 2 เหรียญ ในสถานการณ์แบบนี้ ไม่สามารถประยุกต์ใช้แบบจำลองมาร์คอฟได้ เนื่องจากทราบว่า ผลการสังเกตที่เกิดขึ้นนั้นเกิดจากการโยนเหรียญ แต่แบบจำลองฮิดเดนมาร์คอฟสามารถจำลองเหตุการณ์นี้ได้โดยให้แต่ละสถานะคือ เหรียญแต่ละเหรียญที่จะถูกทอยและแต่ละสถานะจะมีการแจกแจงของความน่าจะเป็น (Probability Distribution) ของการเกิดหัวและก้อยเป็นของตัวเอง ใช้เมทริกซ์การเปลี่ยนสถานะเพื่อกำหนดการเปลี่ยนแปลงระหว่างสถานะ ซึ่งในที่นี้คือ การที่ผู้ทำการทดลองจะเปลี่ยนเหรียญที่ใช้โยนเหรียญโดยตั้งสมมติฐานว่ามีเหรียญ 2 เหรียญ และ 3 เหรียญสามารถแสดงได้ดังรูปที่ 2.7b และรูปที่ 2.7c ตามลำดับ



รูปที่ 2.7 สามแบบจำลองมาร์คอฟที่เป็นไปได้สำหรับการโยนเหรียญ

ในรูปที่ 2.7 สามารถอธิบายผลลัพธ์ที่ซ่อนอยู่ของการโยนเหรียญเพื่อทำการทดสอบ (a) แบบจำลอง 1 เหรียญ (b) แบบจำลอง 2 เหรียญ (c) แบบจำลอง 3 เหรียญ

ถึงแม้จะสามารถสร้างแบบจำลองได้แล้ว ปัญหาที่พบต่อมาก็คือ สามารถสร้างแบบจำลองได้หลายรูปแบบ แล้วแบบจำลองแบบใดเป็นแบบจำลองที่เหมาะสมที่สุดกับผลการสังเกตที่เกิดขึ้นจริง และหากพิจารณาถึงจำนวนพารามิเตอร์จะได้ว่า แบบจำลอง รูปที่ 2.7a มีพารามิเตอร์ที่ไม่ทราบค่า 1 พารามิเตอร์ แบบจำลอง รูปที่ 2.7b มี 4 พารามิเตอร์ที่ไม่ทราบค่า และ 9 พารามิเตอร์ที่ไม่ทราบค่าในแบบจำลองรูปที่ 2.7c ในทางปฏิบัติการมีจำนวนพารามิเตอร์มากขึ้นไม่ได้ส่งผลให้แบบจำลองดีขึ้น แบบจำลองที่ดีที่สุดคือ แบบจำลองที่มีความ

ใกล้เคียงกับเหตุการณ์ที่เกิดขึ้นจริงมากที่สุด เช่น ในตัวอย่างการโยนเหรียญ หากผู้ทดลองใช้เหรียญเพียงเหรียญเดียวเพื่อสร้างผลการสังเกตแต่ตั้งสมมติฐานว่ามี 3 เหรียญ ในกรณีนี้แบบจำลองที่สร้างขึ้นไม่สัมพันธ์เหตุการณ์ที่เกิดขึ้นจริง ทำให้การจำลองเหตุการณ์ไม่เหมาะสม เป็นต้น

2.5.3 องค์ประกอบของแบบจำลองฮิดเดนมาร์คอฟ

แบบจำลองฮิดเดนมาร์คอฟประกอบด้วยองค์ประกอบต่าง ๆ ดังต่อไปนี้

1. สถานะ (State) คือ สถานะของแบบจำลองที่เป็นไปได้ทั้งหมด โดยจะเป็นสถานะที่จะให้ค่าผลลัพธ์ออกมาได้ เช่น กำหนดให้แบบจำลองมี N สถานะที่เป็นไปได้ทั้งหมด นั่นคือสามารถให้ค่าผลลัพธ์ออกมาได้ N แบบ

หากพิจารณาปัญหาการโยนเหรียญที่กล่าวไป

- N คือ จำนวนเหรียญที่ตั้งสมมติฐานให้ผู้ทำการทดลองใช้โยน
- ใช้สัญลักษณ์ $S = \{S_1, S_2, \dots, S_N\}$ แทนสถานะที่ $1,2, \dots, N$
- ใช้สัญลักษณ์ q_t แทนสถานะ ณ เวลา t

2. สัญลักษณ์ของผลการสังเกต (Observation) คือ สิ่งที่เป็นข้อมูลที่จะนำเข้าแบบจำลอง นั่นคือข้อมูลเข้าของแบบจำลอง โดยจำนวนสัญลักษณ์ของผลการสังเกตที่แตกต่างกันเท่ากับ M

หากพิจารณาปัญหาการโยนเหรียญที่กล่าวไป

- สัญลักษณ์ของผลการสังเกตมี 2 แบบ คือ H และ T
- ใช้สัญลักษณ์ $V = \{V_1, V_2, \dots, V_M\}$ แทนสัญลักษณ์ของผลการสังเกต

3. การแจกแจงความน่าจะเป็นในการเปลี่ยนสถานะ (State Transition Probability Distribution) คือ การแจกแจงของค่าความน่าจะเป็นในการเปลี่ยนจากสถานะหนึ่งไปอีกสถานะหนึ่ง โดยกำหนดให้ $A = [a_{ij}]$ เป็นเมทริกซ์การเปลี่ยนสถานะที่ i ไปอีกสถานะ j โดยที่ $1 \leq i, j \leq N$

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \tag{2.5}$$

4. การแจกแจงของความน่าจะเป็นของสัญลักษณ์การสังเกต (Observation Symbol Probability Distribution) คือ การแจกแจงของค่าความน่าจะเป็นของสัญลักษณ์การสังเกต k สำหรับสถานะ S_j โดยกำหนดให้ $B = [b_j(k)]$ เป็นเมทริกซ์ความน่าจะเป็นของสัญลักษณ์การสังเกต k ในสถานะ j โดยที่ $1 \leq j \leq N, 1 \leq k \leq M$

$$b_j(k) = P(V_k \ at \ t | q_t = S_j) \quad (2.6)$$

5. การแจกแจงของสถานะเริ่มต้นของแต่ละสถานะ โดยกำหนดให้ $\pi = [\pi_i]$ เมทริกซ์ค่าความน่าจะเป็นเริ่มต้นที่สถานะ i โดยที่ $1 \leq i \leq N$

$$\pi_i = P(q_1 = S_i) \quad (2.7)$$

เมื่อมีค่าที่เหมาะสมของ N, M, A, B และ π แบบจำลองฮิดเดนมาร์คอฟจะสามารถใช้เป็นเครื่องผลิตลำดับผลการสังเกต

$$O = O_1, O_2, O_3, \dots, O_T \quad (2.8)$$

เมื่อผลการสังเกต O_t แต่ละตัวเป็นสมาชิกของเซต V และ t คือ เวลา ณ ขณะใดขณะหนึ่งได้ด้วยขั้นตอนต่อไปนี้

1. กำหนดสถานะเริ่มต้น $q_1 = S_i$ จาก π

2. ตั้งค่า $t = 1$

3. เลือก $O_t = V_k$ โดยใช้การแจกแจงของความน่าจะเป็นของสัญลักษณ์การสังเกตสำหรับสถานะ S_i ซึ่งก็คือ $b_i(k)$

4. เปลี่ยนสถานะไปสู่สถานะ $q_{t+1} = S_i$ ตามเมทริกซ์การเปลี่ยนสถานะของสถานะ S_i ซึ่งก็คือ a_{ij}

5. ค่า $t = t + 1$ และกลับไปสู่ขั้นตอนที่ 3 จนกระทั่ง $t = T$ จึงหยุด

จะเห็นว่า การระบุแบบจำลองฮิดเดนมาร์คอฟจำเป็นต้องมีองค์ประกอบของแบบจำลอง (N และ M) สัญลักษณ์ของผลการสังเกต และ A, B, π แต่ทั่วไปแล้วนิยมใช้สัญลักษณ์โดยย่อ

$$\lambda = (A, B, \pi) \quad (2.9)$$

เพื่อระบุองค์ประกอบที่สมบูรณ์ของแบบจำลองฮิดเดนมาร์คอฟและการที่จะนำไปประยุกต์ใช้ได้นั้น จำเป็นจะต้องแก้ปัญหา 2 ข้อ ดังต่อไปนี้ คือ

ปัญหาที่ 1: Likelihood

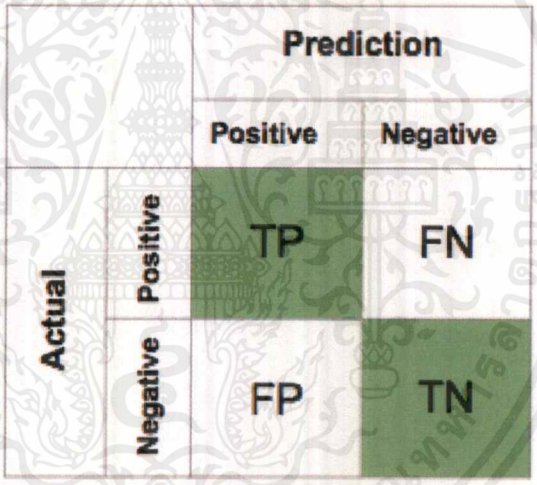
หากมีลำดับผลการสังเกต $O = O_1, O_2, \ldots, O_T$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะสามารถคำนวณหา $P(O|\lambda)$ ได้อย่างไร

ปัญหาที่ 2: Decoding

หากมีลำดับผลการสังเกต $O = O_1, O_2, \ldots, O_T$ และแบบจำลอง $\lambda = (A, B, \pi)$ จะสามารถเลือกลำดับของสถานะ $Q = q_1, q_2, \ldots, q_T$ ที่สัมพันธ์กับลำดับผลการสังเกตดังกล่าวได้อย่างไร

2.6 Confusion Matrix

Confusion Matrix คือ การประเมินผลลัพธ์การทำนาย (หรือผลลัพธ์จากโปรแกรม) เปรียบเทียบกับผลลัพธ์จริง ๆ



รูปที่ 2.8 ตาราง Confusion Matrix

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่าจริง และข้อมูลบอกว่าจริง

True Negative (TN) คือ สิ่งที่โปรแกรมทำนายว่าไม่จริง และข้อมูลบอกว่าไม่จริง

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่าจริง แต่ข้อมูลบอกว่าไม่จริง

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่าไม่จริง แต่ข้อมูลบอกว่าจริง

Accuracy คือ ค่าที่บอกว่าโปรแกรมสามารถทำนายได้แม่นยำขนาดไหน หาได้จาก

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

Recall (True Positive Rate) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าเป็นจริงได้ถูกต้องเป็นอัตราส่วนเท่าไรของจริงทั้งหมด หาได้จาก

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.11}$$

False Positive Rate (FP Rate) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าจริงแต่ไม่ถูกต้องเป็นอันตราส่วนเท่าไรของไม่จริงทั้งหมด หาได้จาก

$$FP\ Rate = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{2.12}$$

Precision คือ ค่าที่บอกว่าโปรแกรมทำนายแม่นยำถูกต้องเท่าไร หาได้จาก

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.13}$$

โดยที่ Recall + FNR = 1 และ TNR+TPR = 1

F Score คือ ค่าน้ำหนักเฉลี่ยของ Recall และ Precision สรุปแบบตามแนวทางทฤษฎี สู่การปฏิบัติคือ ค่าที่ดีที่สุดคือ 1 และค่าที่เลวที่สุดคือ 0

$$F\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{2.14a}$$

$$F\ Score = 2 \times \frac{Precision \times Recxall}{Precision + Recxall} \tag{2.14b}$$

$$F\ Score = 2 \times \frac{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}}{\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} + \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}} \tag{2.14c}$$

นัยยะสำคัญของ False Positive และ False Negative

FP (False Positive) คือ ผลทำนายไม่ถูกต้อง ตามค่าที่คาดหวัง ที่เป็น Negative

ยกตัวอย่างกรณีนี้คือ Actual เป็นมะเร็ง แล้วผลทำทายว่าไม่เป็นมะเร็ง (ทำนายไม่ถูกต้องและมีผลร้ายแรง) การแปลความหมายในลักษณะนี้ตีความว่า รับผลเสียเข้ามาอยู่ในกองของผลดี จากผลของการทำนายทำให้มีโอกาสเกิดความเสียหายได้สูงเชิงคุณภาพ

FN (False Negative) คือ ผลการทำนายไม่ถูกต้องตามค่าที่คาดหวังที่เป็น Positive

ยกตัวอย่างกรณีนี้คือ Actual ไม่เป็นมะเร็ง แล้วผลทำนายว่าเป็นมะเร็ง (ทำนายไม่ถูกต้อง) การแปลความหมายในลักษณะนี้ตีความว่า โยนผลดีเข้าไปกองในกองของเสียจากผลของการทำนายทำให้มีโอกาสเกิดความเสียหายได้สูงเชิงปริมาณ

นัยยะสำคัญของ Precision

Precision ในที่นี้ตีความเป็นความแม่นยำ จากสูตรจึงสรุปได้ว่า ถ้าต้องการค่าความแม่นยำเป็น 100% ก็ต้องให้ค่า FP ที่ทายผิดแบบเสียหายเชิงคุณภาพมีค่าเป็นศูนย์ คำถามสำคัญนั่นคือในงานจริง Precision ต้องเป็นเท่าไร บางงานที่เป็นต้องการผลที่ตรงมาก ๆ อาจจะต้องเป็น 100% แต่ก็ต้องแลกกับการทำแบบจำลองให้ดีมาก ๆ หรือใช้เวลาในการพัฒนานานมาก ๆ หรือใช้อัลกอริทึมที่เหมาะสมมาก ๆ เช่น ผลการทำนายเกี่ยวข้องกับชีวิต ขณะที่บางงานอาจจะมีการกำหนดค่าความคลาดเคลื่อน (%Error) ให้ยอมรับได้ แต่ทางปฏิบัติอาจจะต้องเอาผลการทำงานมาตรวจสอบและใช้ค่าบางค่าใน Feature หนึ่งมาเป็นตัวกลั่นกรองสำหรับการเลือกรายการนั้นมาตรวจสอบซ้ำด้วยคนทำงานอีกครั้ง

นัยยะสำคัญของ Recall

Recall ในที่นี้ตีความตามนัยยะ คือ ความอ่อนไหว โดยมีสูตรในการคำนวณดังนี้ จากสูตรจึงสรุปได้ว่า ถ้าความอ่อนไหว 100% ก็ต้องให้ค่า FN ที่ทายผิดแบบเสียหายเชิงปริมาณมีค่าเป็นศูนย์ แต่ถ้าหากค่า Recall = 95% นั้นหมายความว่าจะมีรายการที่ทำนายผิด แบบรายการจริงไม่เป็นแต่ผลทำนายว่าเป็น ตามตัวอย่างข้างต้น ถึง 5% ในเชิงปฏิบัติสรุปได้ว่า โยนของดีเข้าไปในกลุ่มของเสียถึง 5% ที่ต้องถูกโยนทิ้ง

นัยยะสำคัญของ Accuracy

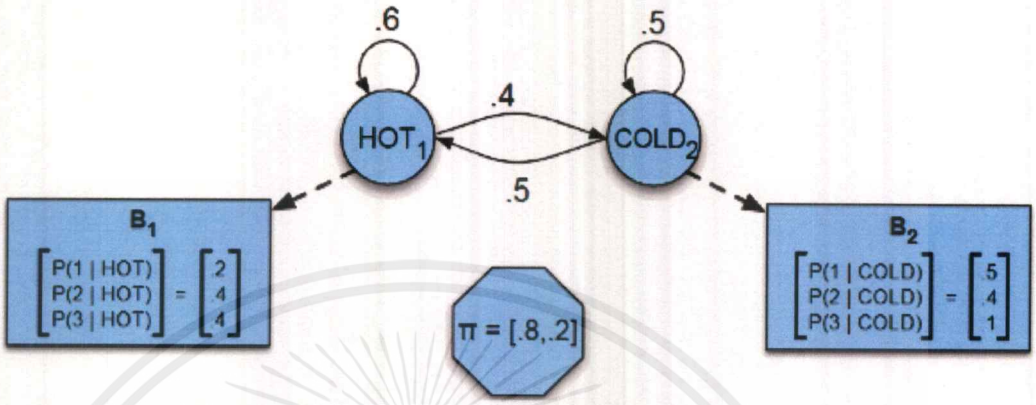
สำหรับ Accuracy ตีความตรงตัวซึ่งก็คือ ความถูกต้อง และความถูกต้องนี้คือทำนายถูกทั้งแบบ Positive และ Negative จากสูตรจึงสรุปได้ว่าถ้าทายถูกทั้งหมดแล้วค่า Accuracy = 100% เป็นค่าในอุดมคติ แต่ในทางปฏิบัติจริงมีโอกาสเกิดขึ้นได้ยากมาก ๆ [7][11][14][17]

2.7 การคำนวณ Likelihood โดยใช้ขั้นตอนวิธีแบบไปข้างหน้า (Forward algorithm)

ปัญหาที่พบในประการแรก คือการคำนวณ Likelihood ของลำดับผลการสังเกต

ตัวอย่าง

กำหนดให้ แบบจำลองฮิดเดนมาร์คอฟของการทานไอศกรีม ตามรูป 2.10 แล้วความน่าจะเป็นในการเกิดลำดับ [3 1 3] เป็นเท่าไร



รูปที่ 2.9 ตัวอย่างการแสดงความสัมพันธ์ของแบบจำลองฮิดเดนมาร์คอฟ

จากรูปที่ 2.9 เป็นจำนวนความสัมพันธ์ของไอศกรีมที่เจสันทาน (ผลการสังเกต) กับสภาพอากาศ (H หรือ C เป็นตัวแปรซ่อน) ของแบบจำลองฮิดเดนมาร์คอฟ

"การคำนวณ Likelihood : กำหนด แบบจำลองฮิดเดนมาร์คอฟ $\lambda = (A, B, \pi)$ *และมีลำดับผลการสังเกต* $O = O_1, O_2, \ldots, O_T$ *ทำให้สามารถคำนวณ Likelihood* $P(O|\lambda)$ *ได้อย่างไร"*

สำหรับสายโซ่มาร์คอฟ สิ่งที่ได้เห็นจากผลการสังเกตดูเหมือนมีเหตุการณ์บางอย่างจะถูกซ่อนเอาไว้ ซึ่งสามารถคำนวณความน่าจะเป็นของ [3 1 3] ได้ เพียงแค่ติดตามจากสถานะซ่อนที่ถูกติดป้ายไว้กับ [3 1 3] และคูณความน่าจะเป็นเหล่านั้นเข้าด้วยกัน สำหรับแบบจำลองฮิดเดนมาร์คอฟนี้ไม่ใช่เรื่องที่ง่าย ต้องกำหนดความน่าจะเป็นของการทานไอศกรีมจากลำดับของผลการสังเกต เช่น [3 1 3] และในทางกลับกันไม่มีทางรู้ว่าลำดับของสถานะซ่อนเป็นอะไร

เริ่มจากสถานการณ์ง่าย ๆ สมมติให้ตอนนี้ทราบว่าสภาพอากาศและต้องการที่จะทำนายว่าเจสันควรจะทานไอศกรีมจำนวนเท่าไร สิ่งนี้เป็นประโยชน์สำหรับปัญหาของแบบจำลองฮิดเดนมาร์คอฟในหลาย ๆ ปัญหา สำหรับการกำหนดลำดับของสถานะซ่อนให้แล้วสามารถคำนวณ Likelihood ของ [3 1 3] ได้อย่างง่ายดาย

วิธีทำอย่างแรกคือ จำไว้เสมอว่าสำหรับแบบจำลองฮิดเดนมาร์คอฟแต่ละสถานะซ่อนจะเกิดมาจากผลการสังเกตเพียงค่าเดียวเท่านั้น ดังนั้นลำดับของสถานะซ่อนและลำดับของผลการสังเกตต้องมีความยาวเท่ากัน

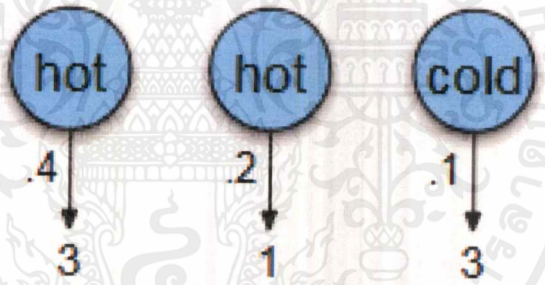
กำหนดให้นี้เป็นการจับคู่แบบหนึ่งต่อหนึ่ง สำหรับแต่ละลำดับสถานะซ่อน $Q = q_1, q_2, \ldots, q_T$ และลำดับของผลการสังเกต $O = O_1, O_2, \ldots, O_T$ และสมมติฐานของมาร์คอฟ การคำนวณ Likelihood ของลำดับผลการสังเกต คือ

$$P(O|Q,\lambda) = \prod_{t=1}^{T} P(O_t|q_t,\lambda) \tag{2.15}$$

การคำนวณความน่าจะเป็นแบบไปข้างหน้าสำหรับผลการสังเกตการณ์ทานไอศกรีม [3 1 3] จากลำดับสถานะซ่อนที่เป็นไปได้ทั้งหมด [hot hot cold] ตามสมการ

$$P([3\ 1\ 3]|[hot\ hot\ cold]) = P(3|hot) \times P(1|hot) \times P(3|cold) \tag{2.16}$$

และแสดงในแบบรูปภาพของการคำนวณ



รูปที่ 2.10 การคำนวณ Likelihood ของผลการสังเกตสำหรับเหตุการณ์

จากรูปที่ 2.10 เป็นการคำนวณ Likelihood ของผลการสังเกตสำหรับเหตุการณ์การทานไอศกรีม [3 1 3] กำหนดให้มีลำดับสถานะซ่อนคือ [hot hot cold]

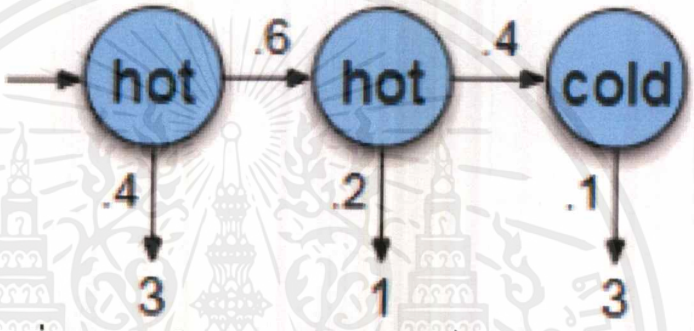
แต่ว่าในความเป็นจริงแล้ว ไม่มีทางที่จะรู้ได้อย่างแน่นอน ว่าอะไรคือลำดับสถานะซ่อนที่แน่นอน จึงต้องการคำนวณความน่าจะเป็นร่วมของเหตุการณ์การทานไอศกรีม [3 1 3] โดยสรุปจากสภาพอากาศที่เป็นไปได้ทั้งหมดแทน ด้วยการให้น้ำหนักความน่าจะเป็นของสถานะซ่อนเหล่านั้น ในลำดับแรกให้ทำการคำนวณความน่าจะเป็นร่วมของการอยู่ในลำดับสภาพอากาศ Q และทำให้เกิดลำดับ O ของการทานไอศกรีม ซึ่งรูปทั่วไป คือ

$$\begin{aligned} P(O,Q|\lambda) &= P(O|Q,\lambda) \times P(Q,\lambda) \\ &= \prod_{t=1}^{T} P(O_t|q_t,\lambda) \times \prod_{t=1}^{T} P(q_t|q_{t-1},\lambda) \end{aligned} \tag{2.17}$$

การคำนวณของความน่าจะเป็นร่วมของผลการสังเกตการทานไอศกรีม [3 1 3] และสถานะซ่อนที่เป็นไปได้ [hot hot cold] แสดงให้เห็นในสมการ

$$P([3\ 1\ 3], [hot\ hot\ cold]) = P(hot|start) \times P(hot|hot) \times P(cold|hot) \times P(3|cold) \times P(1|hot) \times P(3|cold) \quad (2.18)$$

และแสดงให้เห็นรูปภาพของการคำนวณ



รูปที่ 2.11 การคำนวณของความน่าจะเป็นร่วมของเหตุการณ์

จากรูปที่ 2.11 เป็นการคำนวณของความน่าจะเป็นร่วมของเหตุการณ์การทานไอศกรีม [3 1 3] และมีลำดับสถานะ [hot hot cold]

ในตอนนี้รู้วิธีคำนวณความน่าจะเป็นร่วมของผลการสังเกตกับแต่ละลำดับของสถานะที่ซ่อนอยู่ สามารถคำนวณความน่าจะเป็นทั้งหมดของลำดับผลการสังเกตด้วยการรวมความน่าจะเป็นร่วมของลำดับผลการสังเกตกับแต่ละลำดับสถานะซ่อนที่เป็นไปได้ทั้งหมด

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) = \sum_{Q} P(O|Q, \lambda)P(Q|\lambda) \quad (2.19)$$

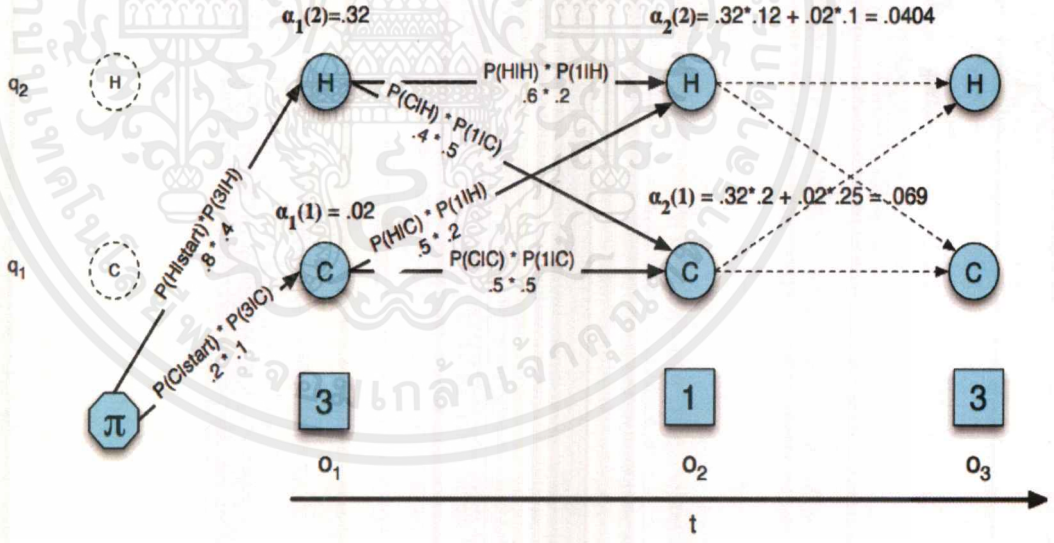
สำหรับในตัวอย่างนี้ขอยกตัวอย่างเพียงบางส่วนของผลรวมลำดับสถานะซ่อนที่เป็นไปได้ทั้งหมด 8 เหตุการณ์

$$\begin{aligned} P([3\ 1\ 3]) = {} & P([3\ 1\ 3], \text{cold cold cold}) + P([3\ 1\ 3], \text{cold cold hot}) \\ & + P([3\ 1\ 3], \text{cold hot cold}) + P([3\ 1\ 3], \text{cold hot hot}) \\ & + P([3\ 1\ 3], \text{hot cold cold}) + P([3\ 1\ 3], \text{hot cold hot}) \\ & + P([3\ 1\ 3], \text{hot hot cold}) + P([3\ 1\ 3], \text{hot hot hot}) \end{aligned}$$

สำหรับแบบจำลองฮิดเดนมาร์คอฟมี N สถานะซ่อนและลำดับผลการสังเกตของ T ผลการสังเกตก็จะมี N^T ลำดับสถานะซ่อนที่เป็นไปได้ สำหรับในงานจริง ๆ จะมี N และ T ทั้งคู่ที่มีค่าใหญ่มาก และแน่นอนว่า N^T ก็จะมีค่าที่ใหญ่มาก ๆ ดังนั้นไม่สามารถที่จะคำนวณ Likelihood ของผลการสังเกตทั้งหมดได้ด้วยการแยกคิด Likelihood ของผลการสังเกตสำหรับแต่ละลำดับของสถานะซ่อนแล้วนำมารวมกันได้

ใช้ขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) ที่มี $O(N^2T)$ ซึ่งจะมีประสิทธิภาพกว่าวิธีการคำนวณแบบนั้น ซึ่งขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) เป็นประเภทของขั้นตอนวิธีแบบหลักการพลวัต (Dynamic Programming Algorithm) นั่นคืออัลกอริทึมจะใช้ตารางสำหรับการเก็บค่าสื่อกลางที่มันจะสร้างความน่าจะเป็นของลำดับผลการสังเกต ขั้นตอนวิธีแบบไปข้างหน้า คำนวณความน่าจะเป็นของผลการสังเกตด้วยผลรวมของความน่าจะเป็นของทุกเส้นทางสถานะซ่อนที่เป็นไปได้ทั้งหมดที่สามารถสร้างลำดับผลการสังเกตได้ แต่มันจะมีประสิทธิภาพโดยการรวบแต่ละเส้นทางเหล่านั้นให้อยู่ใน Forward Trellis อันเดียว

ในรูปที่ 2.12 แสดงตัวอย่างของ Forward Trellis สำหรับคำนวณ Likelihood ของ [3 1 3] เมื่อกำหนดลำดับสถานะซ่อน [hot hot cold]



รูปที่ 2.12 Forward Trellis สำหรับการคำนวณ Likelihood

จากรูปที่ 2.12 Forward Trellis สำหรับการคำนวณ Likelihood ของผลการสังเกตทั้งหมดสำหรับเหตุการณ์การทานไอศกรีม [3 1 3] สถานะซ่อนแทนด้วยวงกลม , ผลการสังเกตแทนด้วยสี่เหลี่ยม ในรูปแสดงให้เห็นถึงการคำนวณของ $\alpha_t(j)$ สำหรับสองสถานะซ่อน ณ สองช่วงเวลาต่อกัน

ในการคำนวณในแต่ละโหนดเป็นไปตาม (2.21) ผลจากการคำนวณความน่าจะเป็นแสดงให้เห็นในแต่ละโหนด เป็นไปตามสมการ (2.20)

แต่ละโหนดของขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) $\alpha_t(j)$ คือความน่าจะเป็นของสถานะที่ j ณ เวลาที่ t เมื่อกำหนดค่า λ ค่าของแต่ละโหนด $\alpha_t(j)$ จะถูกคำนวณมาจากความน่าจะเป็นรวมของทุก ๆ ทางเดินที่มายังโหนด แต่ละโหนดจะมีความน่าจะเป็นดังนี้

$$\alpha_t(j) = P(O_1, O_2, \dots, O_t, q_t = S_j|\lambda) \tag{2.20}$$

ในที่นี้ $q_t = S_j$ หมายความว่า สถานะ ณ เวลาที่ t ของลำดับคือสถานะ j คำนวณความน่าจะเป็น α นี้ด้วยผลรวมความน่าจะเป็นของเส้นทางทั้งหมดที่นำมาสู่โหนดปัจจุบัน สำหรับการกำหนดให้ สถานะ q_i ณ เวลา t ค่าความน่าจะเป็น α จะถูกคำนวณดังนี้

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(O_t) \tag{2.21}$$

มี 3 ปัจจัยที่จะถูกนำมาคูณในสมการ (2.24) ในการคำนวณความน่าจะเป็นแบบไปข้างหน้า ณ เวลาที่ t คือ

$\alpha_{t-1}(i)$ ทางเดินของความน่าจะเป็นแบบไปข้างหน้า ณ เวลาก่อนหน้า

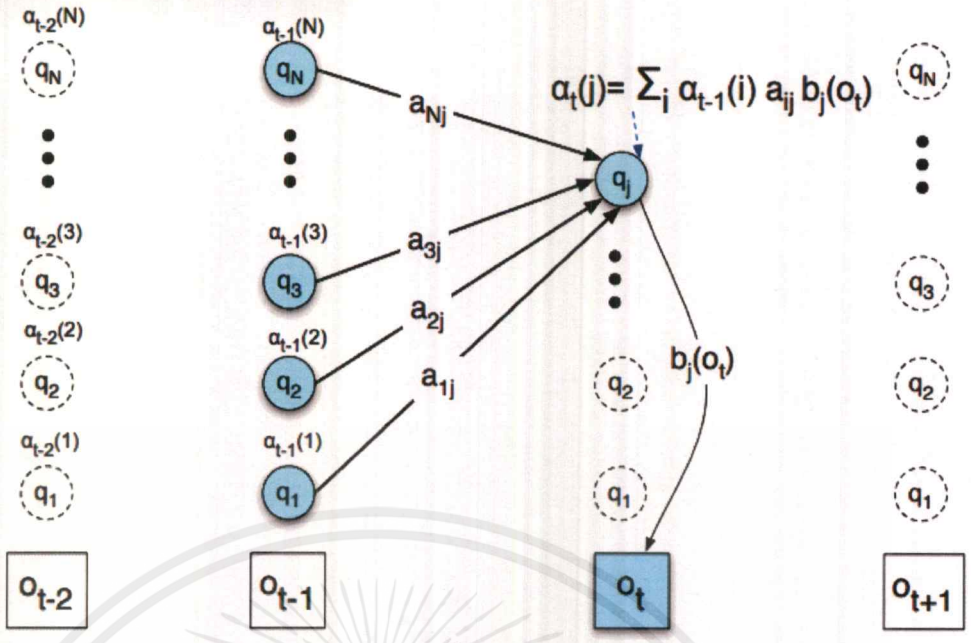
a_{ij} ค่าความน่าจะเป็นเปลี่ยนสถานะจากสถานะก่อนหน้า q_i มาถึงสถานะปัจจุบัน q_j

$b_j(O_t)$ ค่าความน่าจะเป็นของผลการสังเกต O_t ณ เวลาที่ t เมื่อกำหนดให้สถานะปัจจุบันคือ j

พิจารณาการคำนวณในรูปที่ 2.12 ของ $\alpha_2(2)$ ความน่าจะเป็นแบบไปข้างหน้า ณ เวลาที่ 2 ในสถานะที่ 2 ได้สร้างผลการสังเกตบางส่วน คือ [3 1] คำนวณจากความน่าจะเป็น α ณ เวลาที่ 1 มาจาก 2 ทาง ซึ่งแต่ละเส้นทางจะประกอบไปด้วย 3 ปัจจัยด้านบน : $\alpha_1(1) \times P(H|C) \times P(1|H)$ และ $\alpha_1(2) \times P(H|H) \times P(1|H)$

รูปที่ 2.13 แสดงให้เห็นถึงอีกมุมมองหนึ่งของแต่ละขั้นตอนในการคำนวณค่าในแต่ละโหนดของไดอะแกรม

ให้ 2 คำจำกัดความของขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) คือ Pseudocode ในรูปที่ 2.14 และคำสั่งของการเรียกซ้ำแบบมีเงื่อนไข (Recursive) ไว้ตรงนี้



รูปที่ 2.13 แสดงให้เห็นถึงขั้นตอนของ Forward Algorithm

จากรูปที่ 2.13 การแสดงผลการคำนวณของ $\alpha_t(i)$ เพียงหนึ่งค่าในไดอะแกรม ด้วยผลรวมของทุกค่า α_{t-1} ก่อนหน้า และให้น้ำหนักด้วยความน่าจะเป็นแบบทรานเซียนท์ a และคูณด้วยความน่าจะเป็นของผลการสังเกต $b_i(O_{t+1})$ สำหรับแบบจำลองฮิดเดนมาร์คอฟโดยส่วนมากที่มีค่าความน่าจะเป็นในการเปลี่ยนสถานะเท่ากับศูนย์ ดังนั้นสถานะที่เป็นศูนย์นี้จะมีมีส่วนเกี่ยวข้องกับความน่าจะเป็นแบบไปข้างหน้าของสถานะปัจจุบัน สถานะซ่อนแทนด้วยวงกลม , สถานะผลการสังเกตแทนด้วยสี่เหลี่ยม , โหนดที่แรเงาจะถูกรวมอยู่ในการคำนวณความน่าจะเป็นของ $\alpha_t(i)$

function FORWARD(*observations* of len T, *state-graph* of len N) **returns** *forward-prob*

create a probability matrix *forward*$[N,T]$
for each state s **from** 1 **to** N **do** ; initialization step
  $forward[s,1] \leftarrow \pi_s * b_s(o_1)$
for each time step t **from** 2 **to** T **do** ; recursion step
  for each state s **from** 1 **to** N **do**

$$forward[s,t] \leftarrow \sum_{s'=1}^{N} forward[s',t-1] * a_{s',s} * b_s(o_t)$$

$$forwardprob \leftarrow \sum_{s=1}^{N} forward[s,T] \quad \text{; termination step}$$

return *forwardprob*

รูปที่ 2.14 Forward Algorithm

ขั้นตอนการทำงาน Forward Algorithm

1. ค่าเริ่มต้น

$$\alpha_1(j) = \pi_j b_j(O_1) \qquad 1 \le j \le N$$

2. ทำซ้ำ

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(O_t) \qquad 1 \le j \le N, 1 < t \le T$$

3. จุดสิ้นสุด

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

2.8 การคำนวณ Likelihood โดยใช้ขั้นตอนวิธีแบบไปข้างหลัง (Backward Algorithm)

การเรียนรู้ : หากมีลำดับผลการสังเกต และเซตของสถานะที่เป็นไปได้ในแบบจำลองฮิดเดนมาร์คอฟ จะสามารถคำนวณหาค่าพารามิเตอร์เมทริกซ์ A และ B

การป้อนข้อมูลไปยังอัลกอริทึม การเรียนรู้ดังกล่าวจะเป็นลำดับที่ไม่บ่งบอก(สถานะซ่อน)กำกับของค่าสังเกต O ที่จะส่งผลให้เกิดสถานะซ่อน Q

ตัวอย่าง

สำหรับแบบจำลองไอศกรีม ลำดับของค่าสังเกต $O = \{1,3,2, \dots\}$ และเซตของสถานะซ่อน H และ C

อัลกอริทึมพื้นฐานสำหรับการกระบวนการเรียนรู้ของแบบจำลองฮิดเดนมาร์คอฟ คือ ขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) และขั้นตอนวิธีแบบไปข้างหลัง (Backward Algorithm) หรือ Baum-Wel Algorithm ซึ่งใช้ Expectation-Maximization หรือ EM Algorithm ในการหาค่าประมาณ Likelihood ที่มากที่สุดของพารามิเตอร์ของแบบจำลองฮิดเดนมาร์คอฟ อัลกอริทึมดังกล่าวจะช่วยในการปรับค่าความน่าจะเป็นในการเปลี่ยนสถานะ A และความน่าจะเป็นของสัญลักษณ์การสังเกต B ในแบบจำลองฮิดเดนมาร์คอฟ

Expectation-Maximization หรือ EM เป็นอัลกอริทึมแบบทำซ้ำในการคำนวณเพื่อประมาณความน่าจะเป็นเริ่มต้น จากนั้นใช้ค่าประมาณเหล่านั้นเพื่อประมาณค่าให้ดีขึ้นไปอีก และปรับปรุงความน่าจะเป็นในทุก ๆ ครั้งที่ทำซ้ำ

เริ่มต้นด้วยการพิจารณากรณีง่ายๆของการเรียนรู้ ซึ่งทราบทั้งอุณหภูมิและจำนวนไอศกรีมในแต่ละวัน ลองคิดดูว่าชุดผลการสังเกต การป้อนข้อมูลต่อไปนี้และรู้ไปถึงลำดับสถานะซ่อน

3	3	2
LS	LS	HS
1	1	2
HS	HS	HS
1	2	3
HS	LS	LS

รูปที่ 2.15 ตัวอย่างชุดข้อมูล

สิ่งนี้จะช่วยให้สามารถคำนวณพารามิเตอร์ของแบบจำลองฮิดเดนมาร์คอฟได้อย่างง่าย เพียงแค่ประมาณความเป็นไปได้สูงสุดจากข้อมูลฝึกสอน โดยเริ่มจากคำนวณหา $\boldsymbol{\pi}$ จากการนับ 3 สถานะซ่อนเริ่มต้น

$$\pi_h = 1/3 \, , \pi_c = 2/3$$

ต่อไปสามารถคำนวณหาเมทริกซ์ A ได้จากการเปลี่ยนแปลงสถานะโดยละเว้นสถานะซ่อนตัวสุดท้าย

$$P(hot|hot) = 2/3 \quad P(cold|hot) = 1/3$$
$$P(cold|cold) = 1/2 \quad P(hot|cold) = 1/2$$

และเมทริกซ์ B

$$P(1|hot) = 0/4 = 0 \qquad P(1|cold) = 3/5 = 0.6$$
$$P(2|hot) = 1/4 = 0.25 \qquad P(2|cold) = 2/5 = 0.4$$
$$P(3|hot) = 3/4 = 0.75 \qquad P(3|cold) = 0$$

หากแต่ความเป็นจริงแบบจำลองฮิดเดนมาร์คอฟไม่สามารถคำนวณค่าเหล่านี้ได้ตรงๆจากลำดับของค่าสังเกตนั้นเป็นเพราะไม่อาจทราบได้ว่าจริง ๆ แล้วเส้นทางของสถานะที่ผ่านเครื่องมือเพื่อรับข้อมูลที่กำหนด

ตัวอย่าง

สมมติว่าไม่ได้บอกอุณหภูมิวันที่ 2 ให้ จึงต้องเดาเอาเอง แต่เมื่อมีความน่าจะเป็นดังกล่าวและอุณหภูมิของวันอื่น ๆ ทำให้สามารถคำนวณทางทฤษฎีของเบย์กับความน่าเป็นทั้งหมดเพื่อที่จะได้

ค่าประมาณความน่าจะเป็นของอุณหภูมิที่หายไปและใช้ค่าเหล่านั้นเพื่อคำนวณอุณหภูมิสำหรับวันที่ 2 ได้

แต่ปัญหาที่แท้จริงนั้นยากกว่านั้น เพราะไม่อาจทราบการคำนวณของการอยู่ในสถานะซ่อนเร้นใด ๆ Baum-Welch Algorithm แก้ไขปัญหานี้ด้วยการประมาณค่าแบบทำซ้ำ เริ่มจากการประมาณค่าความน่าจะเป็นในการเปลี่ยนสถานะและความน่าจะเป็นของสัญลักษณ์การสังเกต โดยการคำนวณความน่าจะเป็นไปแบบข้างหน้า (Forward Probability) ที่สอดคล้องกับค่าสังเกตที่ถูกกำหนดมาแล้วหารด้วยความน่าจะเป็นโดยรวมทุกเส้นทางที่ต่างกันทั้งหมดที่มีส่วนทำให้เกิดความน่าจะเป็นแบบไปข้างหน้านี้

เพื่อให้เข้าใจอัลกอริทึม จึงจำเป็นต้องกำหนดความน่าจะเป็นที่เป็นประโยชน์ที่เกี่ยวข้องกับความน่าจะเป็นไปข้างหน้า โดยเรียกว่า ความน่าจะเป็นแบบย้อนกลับ (Backward Probability) โดยกำหนดสัญลักษณ์เป็น β คือความน่าจะเป็นของการมองเห็นค่าสังเกตจากเวลาที่ $t+1$ จนถึงเวลาสุดท้าย โดยกำหนดให้สถานะปัจจุบันอยู่ในสถานะที่ i ณ เวลาที่ t และเมื่อกำหนดพารามิเตอร์ λ

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T | q_t = S_i, \lambda) \tag{2.22}$$

ขั้นตอนการทำงานมีลักษณะคล้ายกับขั้นวิธีแบบไปข้างหน้า (Forward Algorithm)

1. ค่าเริ่มต้น

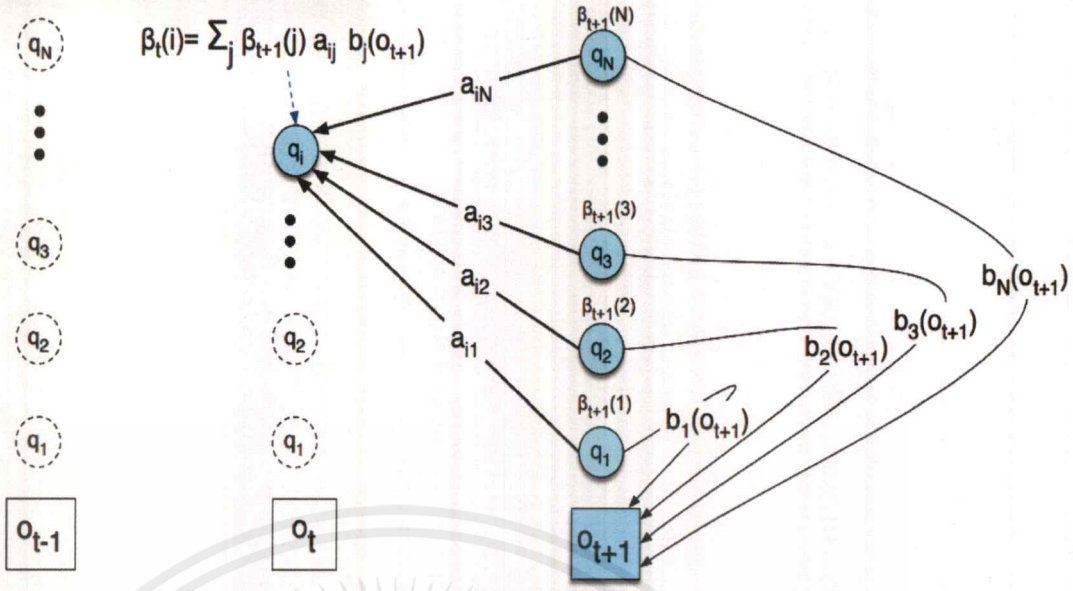
$$\beta_T(i) = 1 \qquad 1 \le i \le N$$

2. ทำซ้ำ

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \qquad 1 \le i \le N, 1 < t \le T$$

3. จุดสิ้นสุด

$$P(O|\lambda) = \sum_{j=1}^{N} \pi_j b_j(o_1) \beta_1(j)$$



รูปที่ 2.16 แสดงให้เห็นถึงขั้นตอนของ Backward Algorithm

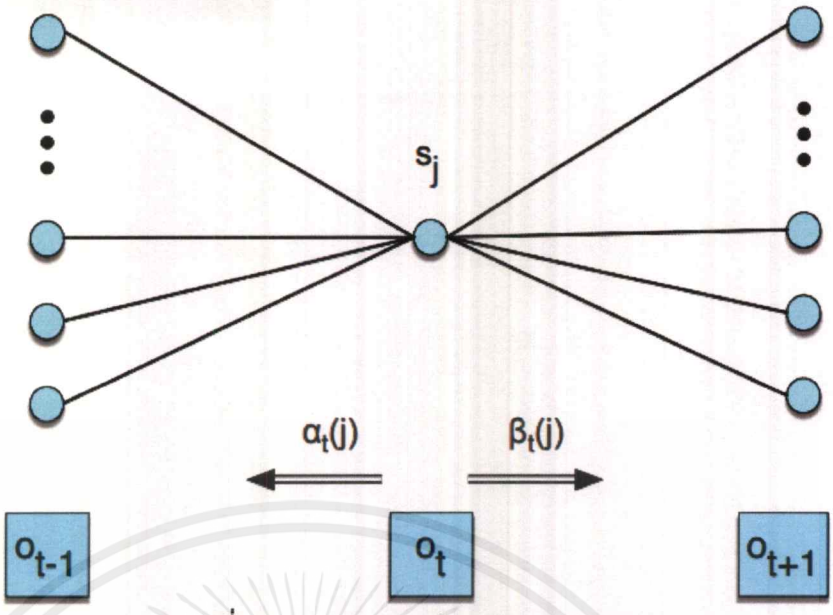
จากรูปที่ 2.16 เป็นการคำนวณของ $\beta_t(i)$ ด้วยผลรวมทั้งหมดของค่า $\beta_{t+1}(j)$ ถูกให้น้ำหนักด้วยความน่าจะเป็นในการเปลี่ยนสถานะ a_{ij} และความน่าจะเป็นของสัญลักษณ์การสังเกต $b_j(o_{t+1})$ ส่วนสถานะเริ่มต้นและสถานะสุดท้ายจะไม่ถูกยกขึ้นมาให้เห็น

สำหรับการคำนวณความน่าจะเป็นที่จะอยู่ในสถานะ j ณ เวลา t เมื่อสถานะ i ซ่อนทับกับสถานะ j ซึ่งแทนด้วย $\gamma_t(j)$

$$\gamma_t(j) = P(q_t = S_j|O,\lambda) \tag{2.23a}$$

$$\gamma_t(j) = \frac{P(O_1, O_2, \dots, O_t, q_t = S_j|\lambda) \cdot P(O_{t+1}, O_{t+2}, \dots, O_T|q_t = S_j, \lambda)}{P(O|\lambda)} \tag{2.23b}$$

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{P(O|\lambda)} \tag{2.23c}$$



รูปที่ 2.17 แสดงการคำนวณของ $\gamma_t(j)$
ความน่าจะเป็นของการอยู่ในสถานะ j ณ เวลาที่ t

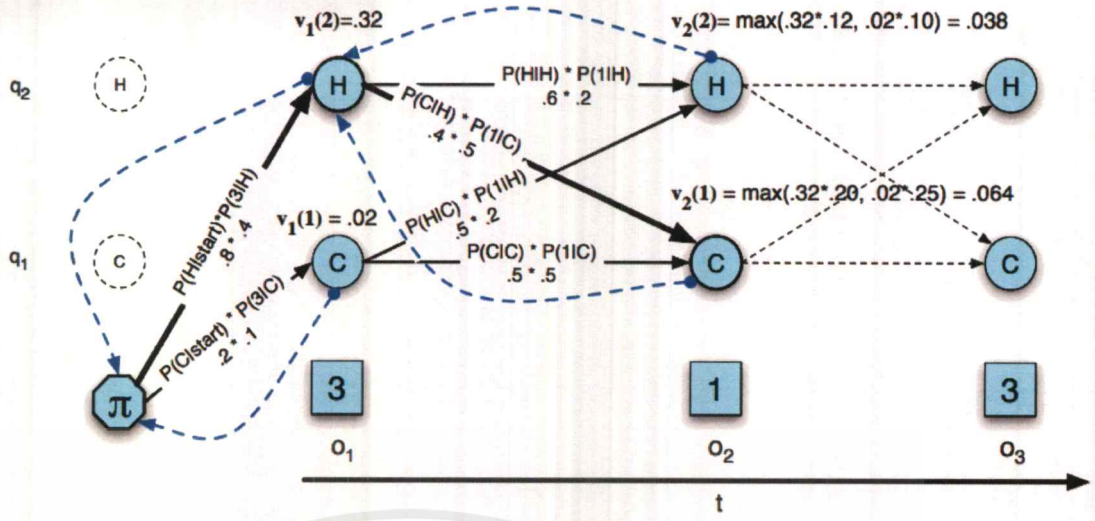
2.9 การถอดรหัสวิเทอร์บิ

สำหรับทุกแบบจำลองฮิดเดนมาร์คอฟจะประกอบไปด้วยตัวแปรซ่อน ซึ่งการพิจารณาว่าลำดับของตัวแปรซ่อนนั้นคืออะไร จะเป็นการกำหนดแหล่งที่มาของลำดับผลการสังเกตได้ซึ่งเรียกว่า การถอดรหัส (Decoding) สำหรับตัวอย่างของแบบจำลองไอศกรีม กำหนดให้ลำดับของผลการสังเกตไอศกรีมคือ [3 1 3] และการถอดรหัสนั้นคือการหาว่าอะไรคือลำดับของสภาพอากาศที่เป็นไปได้มากที่สุด

"การถอดรหัส : กำหนดให้ $\lambda = (A, B, \pi)$ และลำดับของผลการสังเกต $O = O_1, O_2, \ldots, O_T$ เพื่อหาความเป็นไปได้มากที่สุดของลำดับสถานะซ่อน $Q = q_1, q_2, \ldots, q_T$"

ที่เสนอในการหาลำดับที่ดีที่สุดนั้น จะดูจากแต่ละลำดับสถานะซ่อนที่เป็นได้ (HHH, HHC, HCH, etc) สามารถใช้ขั้นตอนวิธีแบบไปข้างหน้า ในการคำนวณ Likelihood ของลำดับผลการสังเกตเมื่อกำหนดลำดับสถานะซ่อนได้ จากนั้นควรจะเลือกลำดับสถานะซ่อนที่มี Likelihood ของผลการสังเกตที่มากที่สุด ซึ่งคล้ายกับเนื้อหาส่วนที่แล้วที่ไม่สามารถคำนวณได้เนื่องจากมีจำนวนของลำดับสถานะซ่อนมากเกินไป

โดยทั่วไปแล้วขั้นตอนวิธี Decoding สำหรับแบบจำลองฮิดเดนมาร์คอฟ คือวิเทอร์บิ (Viterbi Algorithm) ซึ่งคล้ายกับขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) แต่วิเทอร์บิจะใช้ค่าที่มากที่สุดของทุกเส้นทางสถานะซ่อนที่เป็นไปได้ทั้งหมดแทน



รูปที่ 2.18 ตัวอย่างการแสดงแผนภาพวิเทอร์บิ

รูปที่ 2.18 วิเทอร์บิ ใช้สำหรับการคำนวณหาเส้นทางผ่านสถานะซ่อนที่ดีที่สุดสำหรับเหตุการณ์ของการกินไอศกรีม [3 1 3] สถานะซ่อนแทนด้วยวงกลม , ผลการสังเกตแทนด้วยสี่เหลี่ยม ขณะที่วงกลมเส้นประบ่งบอกถึงการเปลี่ยนเส้นทางที่ผิด ในรูปแสดงให้เห็นถึงการคำนวณของ $v_t(j)$ สำหรับ 2 สถานะ ณ 2 ช่วงเวลาต่อกัน ในการคำนวณค่าความน่าจะเป็นแต่ละโหนดเป็นไปตามสมการ (2.25) ผลลัพธ์ของความน่าจะเป็นจะแสดงให้เห็นในแต่ละโหนด ตามสมการ (2.24)

รูปที่ 2.18 แสดงให้เห็นตัวอย่างของแผนภาพวิเทอร์บิ สำหรับการคำนวณลำดับสถานะซ่อนที่ดีที่สุดสำหรับลำดับของผลการสังเกต [3, 1, 3] ซึ่งจะทำการประมวลผลลำดับผลการสังเกตจากซ้ายไปขวา แต่ละโหนด $v_t(j)$ แทนด้วยความน่าจะเป็นที่ดีที่สุดของหนึ่งเส้นทางในสถานะที่ j ณ เวลาที่ t และผ่านลำดับสถานะซ่อนที่เป็นไปได้มากที่สุด $q_1, q_2, \ldots, q_{t-1}$ เมื่อกำหนดพารามิเตอร์ λ

ซึ่งค่าในแต่ละโหนดจะถูกคำนวณโดยการทำซ้ำเส้นทางที่เป็นไปได้มากที่สุดที่นำมาสู่โหนดปัจจุบันแต่ละโหนดแสดงความน่าจะเป็นดังนี้

$$v_t(j) = max_{q_1,\ldots,q_{t-1}} P(q_1, \ldots, q_{t-1}, O_1, O_2, \ldots, O_t, q_t = S_j|\lambda) \qquad (2.24)$$

ได้แสดงให้เห็นถึงทางเดินที่เป็นไปได้มากที่สุด ด้วยการหาค่าที่มากที่สุดของทุก ๆ สถานะซ่อนก่อนหน้าที่เป็นไปได้ทั้งหมด $max_{q_1,\ldots,q_{t-1}}$ เหมือน ๆ กับอัลกอริทึมหลักการพลวัต (Dynamic Programming Algorithm) เนื่องจากได้คำนวณความน่าจะเป็นที่จะอยู่ในทุกสถานะ ณ เวลา $t-1$ คำนวณความน่าจะเป็นวิเทอร์บิ โดยการใช้ความน่าจะเป็นที่มากที่สุดของทางเดินที่นำไปสู่โหนดปัจจุบัน สำหรับสถานะที่กำหนด q_j ณ เวลา t ค่า $v_t(j)$ จะถูกคำนวณเป็น

$$v_t(j) = max_{1 \le i \le N-1}\ v_{t-1}(i)\ a_{ij} b_j(O_t) \qquad (2.25)$$

ปัจจัยสามประการที่ถูกนำมาคูณในสมการ (2.25) นี้เพื่อคำนวณความน่าจะเป็นวิเทอร์บิ ณ เวลา t คือ

$v_{t-1}(i)$ ความน่าจะเป็นวิเทอร์บิของทางเดินก่อนหน้าจาก ณ เวลาก่อนหน้า

a_{ij} ค่าความน่าจะเป็นแบบการส่งจากสถานะก่อนหน้า q_i มาถึง สถานะปัจจุบัน q_j

$b_j(O_t)$ ค่าความน่าจะเป็นของผลการสังเกต O ณ เวลาที่ t เมื่อกำหนดให้สถานะปัจจุบันคือ j

function VITERBI(*observations* of len *T*,*state-graph* of len *N*) **returns** *best-path*, *path-prob*

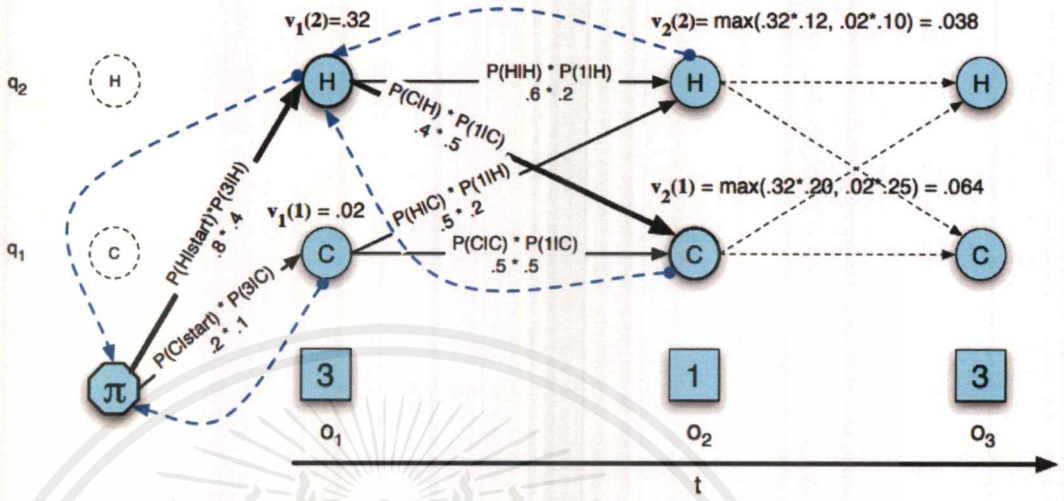
create a path probability matrix *viterbi*[*N*,*T*]
for each state *s* **from** 1 **to** *N* **do** ; initialization step
  viterbi[*s*,1] $\leftarrow \pi_s * b_s(o_1)$
  backpointer[s,1] $\leftarrow 0$
for each time step *t* **from** 2 **to** *T* **do** ; recursion step
  for each state *s* **from** 1 **to** *N* **do**
    viterbi[s,t] $\leftarrow \max_{s'=1}^{N}$ *viterbi*[$s', t-1$] $* a_{s',s} * b_s(o_t)$
    backpointer[s,t] $\leftarrow \operatorname{argmax}_{s'=1}^{N}$ *viterbi*[$s', t-1$] $* a_{s',s} * b_s(o_t)$
bestpathprob $\leftarrow \max_{s=1}^{N}$ *viterbi*[s, T] ; termination step
bestpathpointer $\leftarrow \operatorname{argmax}_{s=1}^{N}$ *viterbi*[s, T] ; termination step
bestpath $\leftarrow$ the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

รูปที่ 2.19 Viterbi algorithm

สำหรับการหาลำดับของสถานะซ่อนที่เหมาะสมที่สุด ให้ลำดับของผลการสังเกตและพารามิเตอร์ฮิดเดนมาร์คอฟ $\lambda = (A, B, \pi)$ อัลกอริทึมส่งกลับเส้นทางผ่านของสถานะซ่อนที่ถูกกำหนดค่า Likelihood มากที่สุดของลำดับของผลการสังเกต

แสดงให้เห็นถึง Pseudocode สำหรับขั้นตอนวิธีวิเทอร์บิ (Viterbi Algorithm) คล้าย ๆ กับ ขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) เพียงแต่แทนที่จะใช้ผลรวมของค่าความน่าจะเป็นของเวลาก่อนหน้าวิเทอร์บิ เลือกที่จะใช้ค่าสูงสุดจากสถานะ i ของเวลาก่อนหน้าแทน และวิธีวิเทอร์บิมีส่วนประกอบหนึ่งส่วนที่ขั้นตอนวิธีแบบไปข้างหน้าไม่มี นั่นคือ จุดย้อนกลับ (Backpointers) นั่นเป็นเพราะว่าขั้นตอนวิธีแบบไปข้างหน้าต้องการที่จะคำนวณ Likelihood ของผลการสังเกต ในขณะที่วิเทอร์บิต้องการสร้างความน่าจะเป็นและลำดับของสถานะซ่อนที่เป็นไปได้มากที่สุด โดยการคำนวณลำดับสถานะซ่อนที่ดีที่สุดและติดตามสถานะซ่อนแต่ละสถานะที่จะนำมาสู่สถานะซ่อน

ปัจจุบัน เป็นดังรูปที่ 2.19 และในตอนสุดท้ายวิเทอร์บิ อัลกอริทึมจะหาเส้นทางย้อนกลับที่ดีที่สุดไปยังจุดเริ่มต้น[15]



รูปที่ 2.20 ตัวอย่างการแสดงแผนภาพการย้อนกลับของวิเทอร์บิ

จากรูปที่ 2.20 การย้อนกลับของวิเทอร์บิในขณะที่มีการคำนวณในแต่ละเส้นทางไปยังสถานะใหม่สำหรับผลการสังเกตถัดไปจะเก็บจุดย้อนกลับ (แสดงด้วยเส้นประ) ที่เป็นเส้นทางที่ดีที่สุดที่นำไปสู่สถานะนี้

สุดท้ายนี้ สามารถกำหนดการทำซ้ำของวิเทอร์บิ ได้ดังนี้

1. ค่าเริ่มต้น

$$v_1(j) = \pi_j b_j(o_1) \qquad 1 \le j \le N$$

$$bt_1(j) = 0 \qquad 1 \le j \le N$$

2. ทำซ้ำ

$$v_t(j) = max_{1 \le i \le N-1}\, v_{t-1}(i)\, a_{ij} b_j(o_t) \qquad 1 \le j \le N, 1 \le t \le T$$

$$bt_t(j) = argmax_{1 \le i \le N-1}\, v_{t-1}(i)\, a_{ij} b_j(o_t) \qquad 1 \le j \le N, 1 \le t \le T$$

3. จุดสิ้นสุด

The Best score: $P* = max_{1 \le i \le N-1}\, v_T(i)$

The start of backtrace: $q_T* = argmax_{1 \le i \le N-1}\, v_T(i)$

ในบทที่ 3 จะอธิบายถึงสิ่งที่ได้ดำเนินการไปแล้วและสิ่งที่ต้องดำเนินการต่อไปในอนาคตของงานวิจัยเรื่องการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ

บทที่ 3
วิธีการดำเนินงานวิจัย

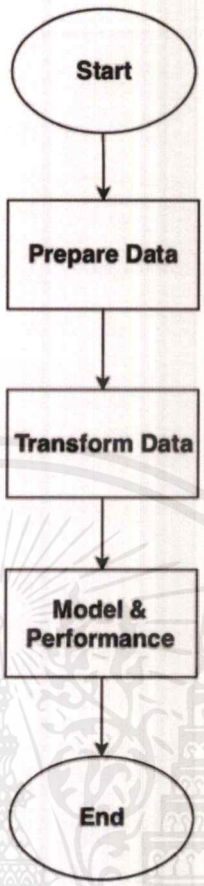
ในงานวิจัยการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองนี้ ข้อมูลที่ใช้จะเป็นข้อมูลจากร้านค้าออนไลน์ร้านหนึ่ง โดยการใช้ Web Scraping ในการดึงข้อมูล ข้อมูลที่ดึงออกมานั้นเป็นข้อมูลที่อยู่บนหน้าเว็บไซต์ของเจ้าของร้านนั้น ซึ่งข้อมูลเหล่านั้นเป็นข้อมูลที่เปิดเผยให้ทุกคน หรือสาธารณะรับรู้ได้ ดังนั้นจึงทำให้ได้ข้อมูลจริงเพื่อมาวิจัยศึกษาการสร้างแบบจำลองการให้คะแนนสินเชื่อแก่กลุ่มคนประเภทพ่อค้าแม่ค้าออนไลน์ในการประกอบการตัดสินใจในการให้สินเชื่อแก่ผู้กู้

ในส่วนของขั้นตอนการดำเนินงานวิจัย จะแบ่งออกเป็น 4 ขั้นตอนหลัก คือ

1. ขั้นตอนการออกแบบผังงาน
2. ขั้นตอนการเตรียมข้อมูล
3. ขั้นตอนการแปลงข้อมูล
4. ขั้นตอนการเข้าแบบจำลองและการวัดประสิทธิภาพแบบจำลอง

3.1 ขั้นตอนการออกแบบผังงาน

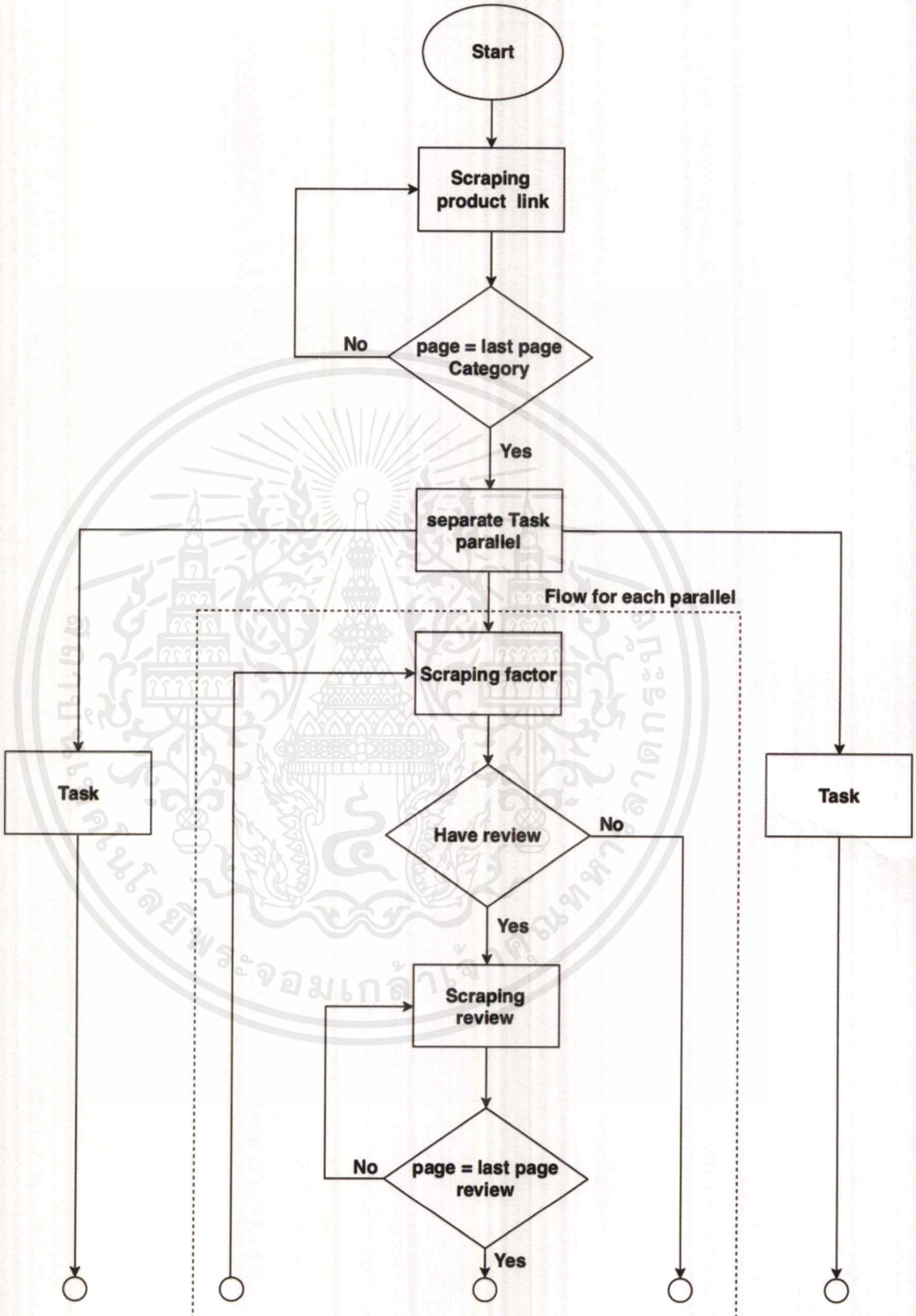
แผนภาพแสดงลำดับขั้นตอนการทำงาน (Flowchart Diagram) เป็นเครื่องมือที่ใช้เพื่อรวบรวมจัดลำดับความคิดเห็น แสดงขั้นตอนการทำงานที่ชัดเจนและใช้วางแผนการทำงานขั้นแรก โดยสัญลักษณ์ Flowchart แสดงถึงการทำงานลักษณะต่าง ๆ ที่เชื่อมต่อกัน Flowchart ถูกใช้ในการออกแบบ เพื่อช่วยให้เห็นภาพสิ่งที่เกิดขึ้นและช่วยให้เข้าใจกระบวนการทำงานและอาจช่วยหาข้อบกพร่องภายในงานอีกด้วย เช่น ปัญหาคอขวด (ปัญหาที่มีงานไปกองที่ส่วนใดส่วนหนึ่งและส่วนอื่นเกิดการรอ) เป็นต้น

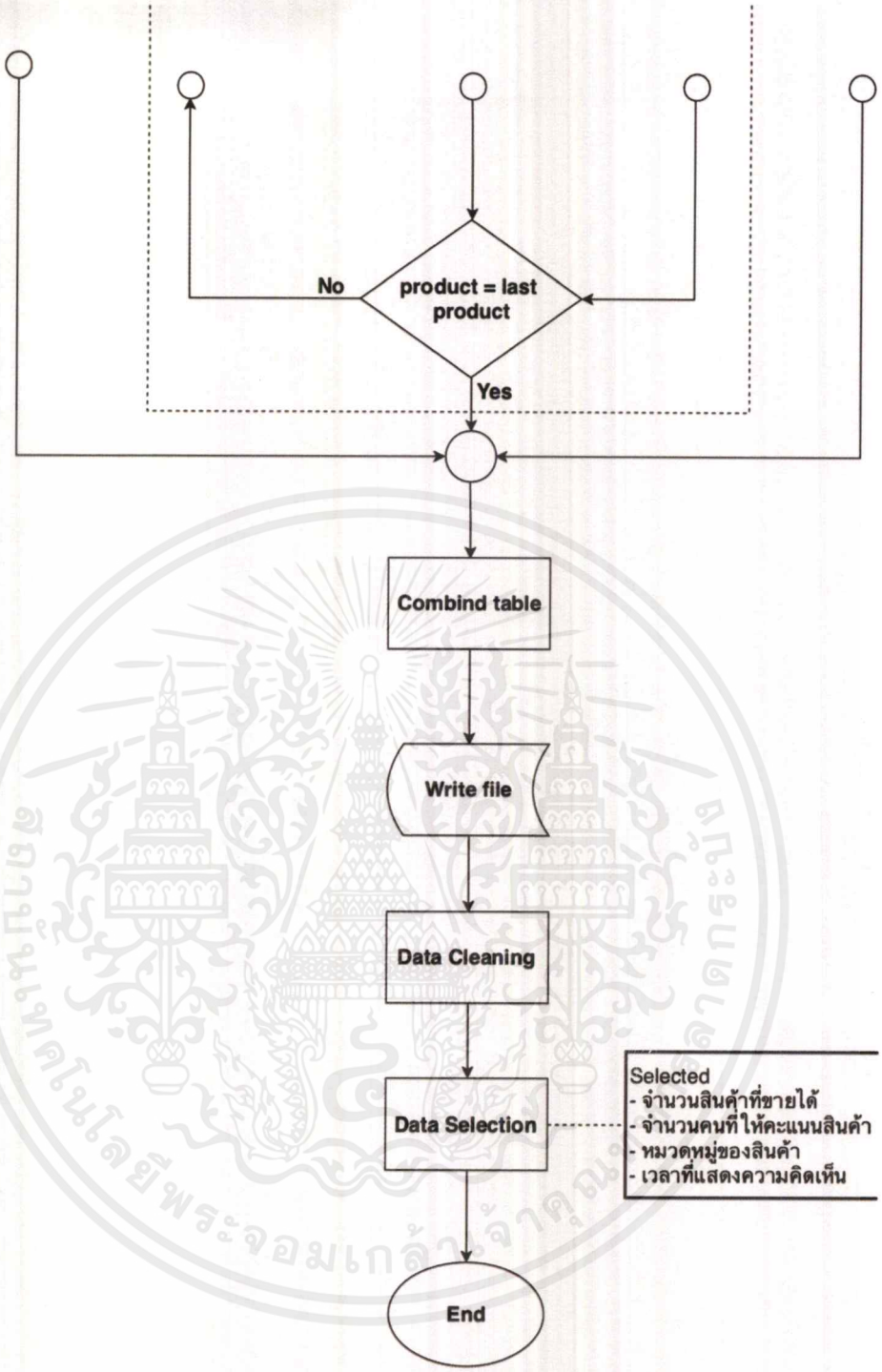


รูปที่ 3.1 แสดงส่วนการทำงานหลัก

3.2 ขั้นตอนการเตรียมข้อมูล

ในส่วนการทำงานนี้จะเป็นส่วนที่จัดเตรียมข้อมูลร้านค้าออนไลน์ โดยจะจัดเตรียมข้อมูลร้านค้าจากเว็บไซต์เพื่อนำไปใช้ในส่วนการทดลองวิจัยการรู้จำการให้คะแนนสินเชื่อร้านค้าต่อไปได้ ซึ่งจะประกอบไปด้วย 3 ขั้นตอน ได้แก่ ขั้นตอนการดึงข้อมูล (Web Scraping) ขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning) และ ขั้นตอนการคัดเลือกข้อมูล (Data Selection) โดยขั้นตอนการทำงานมีกระบวนการตามรูปที่ 3.2





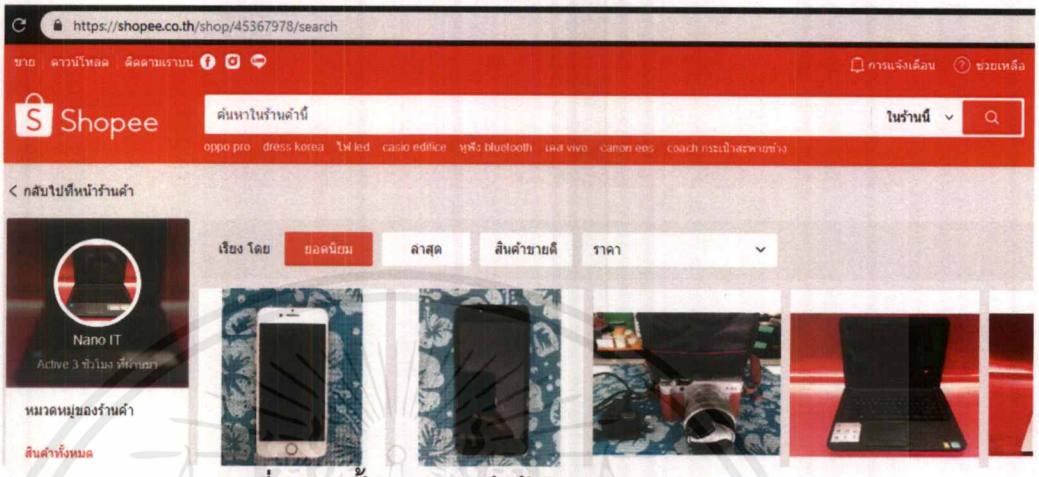
รูปที่ 3.2 แสดงการทำงานในขั้นตอนการเตรียมข้อมูล

3.2.1 ขั้นตอนการดึงข้อมูล

ข้อมูลที่จะใช้ในงานวิจัยนี้ เป็นข้อมูลที่อยู่บนหน้าเว็บไซต์ Shopee นั้นถูกออกแบบมาให้แค่ผู้ใช้งาน (User) ดู ซึ่งหมายความว่าสำหรับผู้ใช้งานเท่านั้น ไม่สามารถดึงข้อมูลออกมาวิเคราะห์ได้โดยตรง ต้องทำการดึงข้อมูล (Web Scraping) ในข้อมูลที่สนใจหลังจากนั้นจึงทำ

การแปลงข้อมูล (Transform) ให้อยู่ในรูปแบบทั่วไป (Format) จึงจะสามารถนำมาวิเคราะห์ใช้ได้ โดยขั้นตอนการทำ Web Scraping สามารถทำได้ดังนี้

1. หา Base URL ที่เป็นเป้าหมาย



รูปที่ 3.3a ขั้นตอนการดึงข้อมูล (Web Scraping)

2. คลิกขวาบนหน้าเว็บไซต์แล้วเลือก Inspect > Elements tab เพื่อรู้ชื่อ Element ที่สนใจและจำเป็นต่อการเขียนโปรแกรม

รูปที่ 3.3b ขั้นตอนการดึงข้อมูล (Web Scraping)

3. ทำการ Test Crawling ใช้ R packages คือ xml2 rvest และ tidyverse โดย Core คำสั่งหลักที่จะใช้ดึงข้อมูลมี 3 คำสั่ง คือ

read_html() = ดึง HTML target URL

html_node() = ใช้ Access CSS nodes หรือ XML path ต่าง ๆ ที่อยู่บนหน้า Web page นั้น โดยการเลือกใช้ CSS

html_text() = ใช้ดึง Text value ที่อยู่ใน CSS หรือ XML path ที่เลือก

4. บันทึกข้อมูล

write.csv(o2c_listings,"/your_path_name/rename_your_file.csv", row.name = FALSE)

ซึ่งคณะผู้จัดทำนั้นจะดึงข้อมูลออกมา 2 แบบ คือ ข้อมูลจากร้านค้าที่นำมาวิจัยศึกษา และข้อมูลจากสินค้าในแต่ละหมวดหมู่ที่ร้านค้านั้นมีสินค้าอยู่

ข้อมูลจากร้านค้าที่นำมาวิจัยศึกษา เป็นหน้าร้านออนไลน์ของร้านค้าร้านหนึ่ง ซึ่งข้อมูลที่ดึงมาประกอบไปด้วยข้อมูลดังต่อไปนี้

1. ชื่อสินค้า
2. จำนวนสินค้าที่ขายได้
3. จำนวนสินค้าคงเหลือ
4. จำนวนคนที่ให้คะแนนสินค้า
5. คะแนนของสินค้าชิ้นนั้น (0-5 คะแนน)
6. ราคาสินค้า
7. ยอด favorite ของสินค้า
8. ชื่อร้านค้า
9. หมวดหมู่ของสินค้า
10. ความคิดเห็น
    - 10.1 ชื่อคนแสดงความคิดเห็น
    - 10.2 เวลาที่แสดงความคิดเห็น
    - 10.3 ความคิดเห็น
    - 10.4 ดาวที่ให้

สำหรับข้อมูลจากสินค้าในแต่ละหมวดหมู่ที่ร้านค้านั้นมีสินค้าอยู่ เนื่องจากร้านค้าออนไลน์ร้านหนึ่งมีสินค้าที่หลากหลายและอาจจะมีสินค้าต่างหมวดหมู่ เพื่อเป็นการดำเนินการศึกษางานวิจัยจึงต้องตรวจสอบว่าร้านค้าที่สนใจ มีหมวดหมู่สินค้าอะไรบ้าง และทำการดึงสินค้าทุกชนิดในหมวดหมู่ที่ร้านค้านั้น ๆ มีทั้งหมด โดยรายละเอียดข้อมูลที่ดึงนั้นเป็นไปตามข้อมูลจากร้านค้าที่นำมาวิจัยศึกษา

3.2.2 ขั้นตอนการทำความสะอาดข้อมูล

เนื่องจากข้อมูลที่ดึงมาได้นั้นเป็นข้อมูลจริงทั้งหมดบนหน้าเว็บไซต์ มีโอกาสที่ข้อมูลนั้นจะไม่สมบูรณ์จากการพิมพ์ผิด พิมพ์ตก เครื่องมือเกิดเสีย (error) ข้อมูลตัวเลขมีความเป็นไปได้น้อยมาก เช่น สั่งซื้อสินค้าจำนวน 1000000000000 ชิ้น เป็นต้น ซึ่งในทางเทคนิคจะเรียกว่าข้อมูลที่อยู่นอกกลุ่มว่า “Outlier” หรือบางครั้งเป็นข้อมูลที่ตกหล่นหายไป (Missing value) และเนื่องด้วยเหตุที่ข้อมูลมีขนาดใหญ่ (Big Data) มนุษย์ไม่สามารถจัดการได้ไหว จึงจำเป็นต้องใช้เครื่องมือเข้ามาช่วย

3.2.3 ขั้นตอนการคัดเลือกข้อมูล

หลังจากขั้นตอนการทำความสะอาดข้อมูล จำเป็นต้องคัดเลือกเฉพาะหัวข้อของข้อมูลที่นำไปวิเคราะห์ต่อได้ นั่นก็คือ ชื่อสินค้า จำนวนสินค้าที่ขายได้ ราคาสินค้า ชื่อร้านค้า หมวดหมู่สินค้า ชื่อคนแสดงความคิดเห็น และเวลาที่แสดงความคิดเห็น จากร้านค้าที่นำมาวิจัยศึกษา และจากสินค้าในแต่ละหมวดหมู่ที่ร้านค้านั้น ๆ มีสินค้าอยู่

3.3 ขั้นตอนการแปลงข้อมูล

ในส่วนนี้จะเป็นการนำข้อมูลเข้าสู่กระบวนการแปลงสภาพข้อมูลเพื่อนำข้อมูลที่ต้องใช้เข้าสู่แบบจำลองในขั้นตอนต่อไป

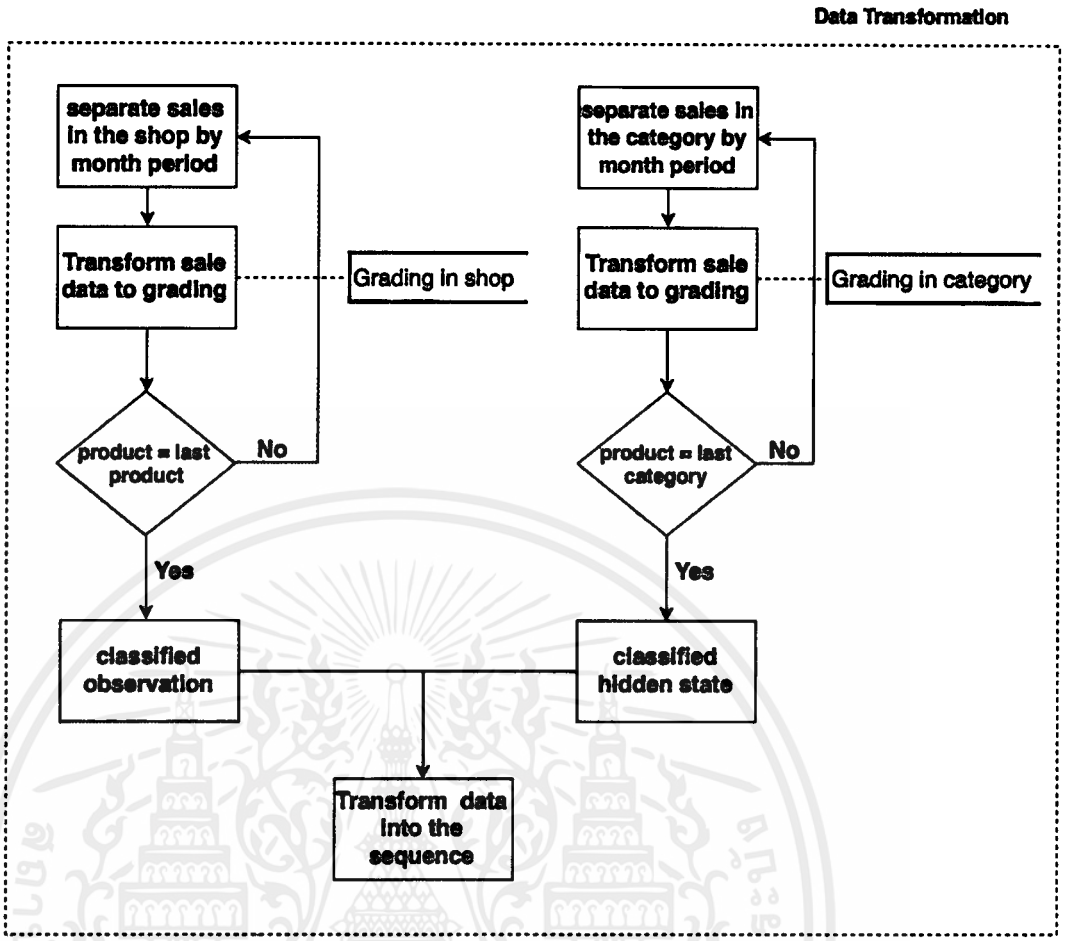
3.3.1 ขั้นตอนการเทียบอัตราส่วน

เนื่องจากข้อมูลจำนวนสินค้าที่ขายได้จากร้านค้าที่นำมาวิจัยศึกษาเป็นยอดขายทั้งหมดของสินค้าชิ้นนั้น ทำให้ไม่สามารถทราบจำนวนสินค้าที่ขายได้ในแต่ละเดือน ทางคณะผู้จัดทำจึงจำเป็นต้องเขียนโปรแกรมนับจำนวนผู้ที่เข้ามาแสดงความคิดเห็นสินค้านั้น ๆ ในแต่ละเดือนแล้วทำการอัตราส่วนกับจำนวนผู้ที่เข้ามาแสดงความคิดเห็นสินค้านั้น ๆ ทั้งหมดทุกเดือน คูณกับยอดขายทั้งหมดที่ขายได้ เพื่อให้ได้ยอดขายสินค้าในแต่ละเดือน

$$\text{ยอดขายสินค้าแต่ละเดือน} = \frac{\text{จำนวนผู้แสดงความคิดเห็นสินค้าแต่ละเดือน}}{\text{จำนวนผู้แสดงความคิดเห็นสินค้าทั้งหมด}} \times \text{ยอดขายสินค้าทั้งหมด}$$

3.3.2 ขั้นตอนการแบ่งกลุ่มข้อมูล

เป็นการนำข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล ซึ่งก็คือชื่อสินค้า จำนวนสินค้าที่ขายได้ ราคาสินค้า ชื่อร้านค้า หมวดหมู่สินค้า ชื่อคนแสดงความคิดเห็น และเวลาที่แสดงความคิดเห็นจากร้านค้าที่นำมาวิจัยศึกษา โดยจะแบ่งการทำงานออกเป็น 2 ส่วน ตามรูปที่ 3.4



รูปที่ 3.4 Flowchart แสดงการแบ่งข้อมูล

ในส่วนการแบ่งกลุ่มข้อมูลจากร้านค้าที่นำมาวิจัยศึกษา จะใช้กระบวนการวิธีการแจกแจงปกติ (Normal Distribution) การกำหนดระดับ (Grading) และการแปลงข้อมูลโดยใช้ลอการิทึม (Logarithm Transformation) จากทฤษฎีดังกล่าวในบทที่ 2

โดยจะนำยอดขายสินค้าแต่ละเดือนของทุกสินค้าภายในร้านมาจัดระดับยอดขายของสินค้าแต่ละชนิด แบ่งกลุ่มข้อมูลภายในร้านค้าออกเป็น 3 ประเภท คือ

- 1 (ขายไม่ดี) หากเปอร์เซ็นไทล์ของยอดขายของสินค้าในเดือนนั้น ๆ ตกอยู่ในช่วงเปอร์เซนไทล์ที่ 0-33
- 2 (ขายดี) หากเปอร์เซ็นไทล์ของยอดขายของสินค้าในเดือนนั้น ๆ ตกอยู่ในช่วงเปอร์เซนไทล์ที่ 33-66
- 3 (ขายดีมาก) หากเปอร์เซ็นไทล์ของยอดขายของสินค้าในเดือนนั้น ๆ ตกอยู่ในช่วงเปอร์เซนไทล์ที่ 66-100

ซึ่งการแบ่งกลุ่มข้อมูลภายในร้านค้าทั้ง 3 ประเภทนี้ เป็นการแบ่งหาผลการสังเกต (Classified Observation)

และนำยอดขายสินค้าแต่ละเดือนของทุกหมวดหมู่ที่ร้านมีมาจัดระดับยอดขายของสินค้าแต่ละชนิดในหมวดนั้น ๆ แบ่งกลุ่มข้อมูลภายในหมวดหมู่ออกเป็น 2 ประเภท คือ

- LS (Low Sold) หากเปอร์เซ็นไทล์ของยอดขายของสินค้าในเดือนนั้น ๆ เมื่อเทียบกับหมวดหมู่สินค้าตกอยู่ในช่วงค่าเฉลี่ยน้อยกว่าสองเท่าของส่วนเบี่ยงเบนมาตรฐาน
- HS (High Sold) หากเปอร์เซ็นไทล์ของยอดขายของสินค้าในเดือนนั้น ๆ เมื่อเทียบกับหมวดหมู่สินค้าตกอยู่ในช่วงค่าเฉลี่ยมากกว่าหรือเท่ากับสองเท่าของส่วนเบี่ยงเบนมาตรฐาน

ซึ่งการแบ่งกลุ่มข้อมูลภายในหมวดหมู่ทั้ง 2 ประเภทนี้เป็นการแบ่งหาสถานะซ่อน (Classified Hidden State)

ขั้นตอนนี้มีจุดประสงค์เพื่อเป็นเกณฑ์ใช้สำหรับการจัดระดับของยอดขายสินค้าภายในหมวดหมู่และภายในร้านค้า

1. จัดระดับยอดขายสินค้าภายในหมวดหมู่ หมายถึง เป็นการจัดระดับว่ายอดขายสินค้าแต่ละชิ้นอยู่ในระดับไหนเมื่อเทียบทั้งหมวดหมู่

2. จัดระดับยอดขายสินค้าภายในร้าน หมายถึง เป็นการจัดระดับว่ายอดขายสินค้าแต่ละชิ้นภายในร้านอยู่ในระดับไหนเมื่อเทียบทั้งร้าน

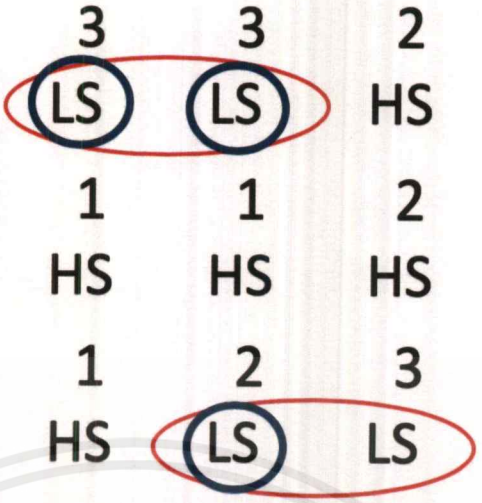
3.4 ขั้นตอนการเข้าแบบจำลองและการวัดประสิทธิภาพแบบจำลอง

ในส่วนนี้จะเป็นการนำข้อมูลที่ผ่านกระบวนการตรวจสอบ แก้ไข ปรับปรุง และแบ่งกลุ่มข้อมูลเป็นที่เรียบร้อยมาเข้าแบบจำลองและวัดประสิทธิภาพเพื่อหาผลลัพธ์ที่ต้องการ

โดยกำหนดองค์ประกอบสำคัญในการเข้าแบบจำลอง ดังนี้

1. สถานะ (State) มี 2 สถานะ คือ LS (Low Sold) และ HS (High Sold)
   สัญลักษณ์ที่ใช้ $S = \{S_1 = LS, S_2 = HS\}$ โดยที่ q_t คือสถานะ ณ เวลาที่ t
2. สัญลักษณ์ของผลการสังเกต (Observation) มี 3 แบบ คือ 1(ขายไม่ดี) 2(ขายดี) และ 3(ขายดีมาก)
   สัญลักษณ์ที่ใช้ $V = \{V_1 = 1, V_2 = 2, V_3 = 3\}$
3. การแจกแจงค่าความน่าจะเป็นในการเปลี่ยนสถานะ (Transition Matrix) a_{ij}

ตัวอย่างการคำนวณ



รูปที่ 3.5 แสดงความสัมพันธ์การคำนวณของ a_{ij}

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad 1 \le i, j \le N$$

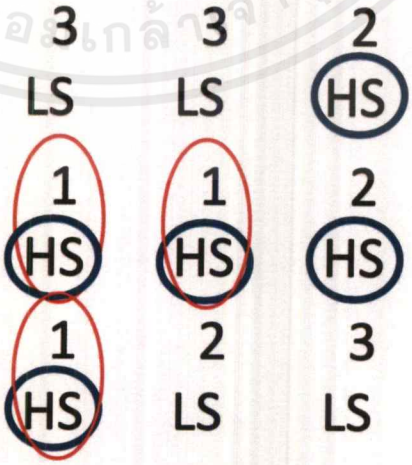
$$a_{11} = P(q_{t+1} = LS | q_t = LS)$$

$$= \frac{P(q_{t+1} = LS, q_t = LS)}{P(q_t = LS)}$$

$$= \frac{2}{3}$$

4. การแจกแจงของความน่าจะเป็นของสัญลักษณ์การสังเกต (Observation Symbol Probability Distribution) $b_j(k)$

ตัวอย่างการคำนวณ



รูปที่ 3.6 แสดงความสัมพันธ์การคำนวณของ $b_j(k)$

$$b_j(k) = P(V_k \ at \ t | q_t = S_j) \quad 1 \le j \le N, 1 \le k \le M$$

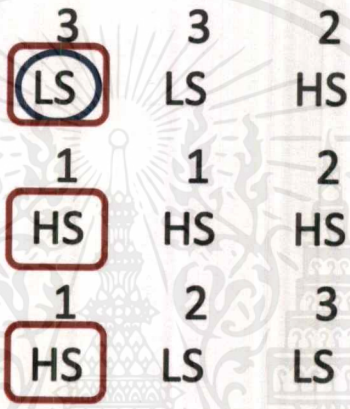
$$b_2(1) = P(V_1 \ at \ t | q_t = HS)$$

$$= \frac{P(V_1 \ at \ t, q_t = HS)}{P(q_t = HS)}$$

$$= \frac{3}{5}$$

5. การแจกแจงของสถานะเริ่มต้นของแต่ละสถานะ (Initial Probability Distribution) π_i

ตัวอย่างการคำนวณ

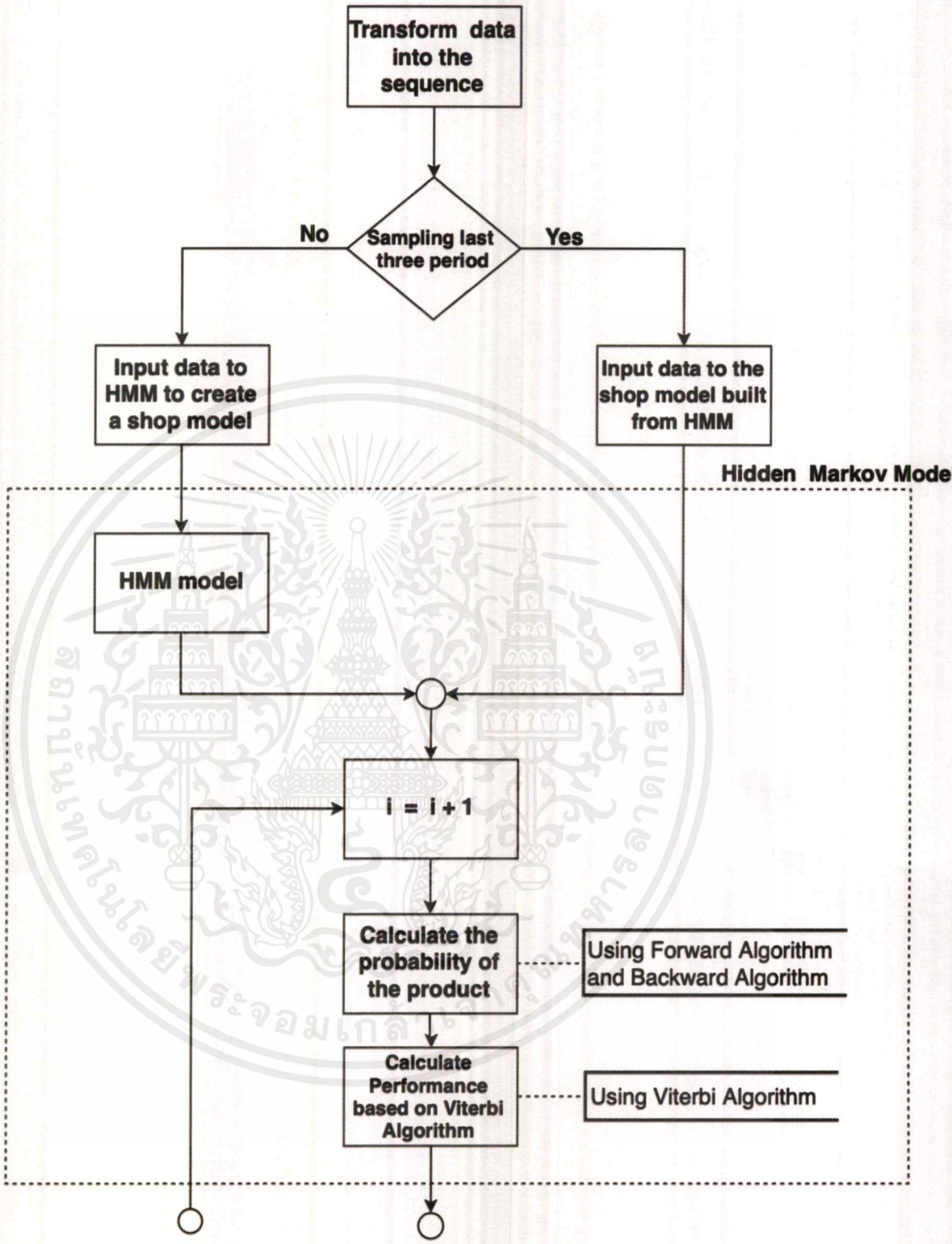


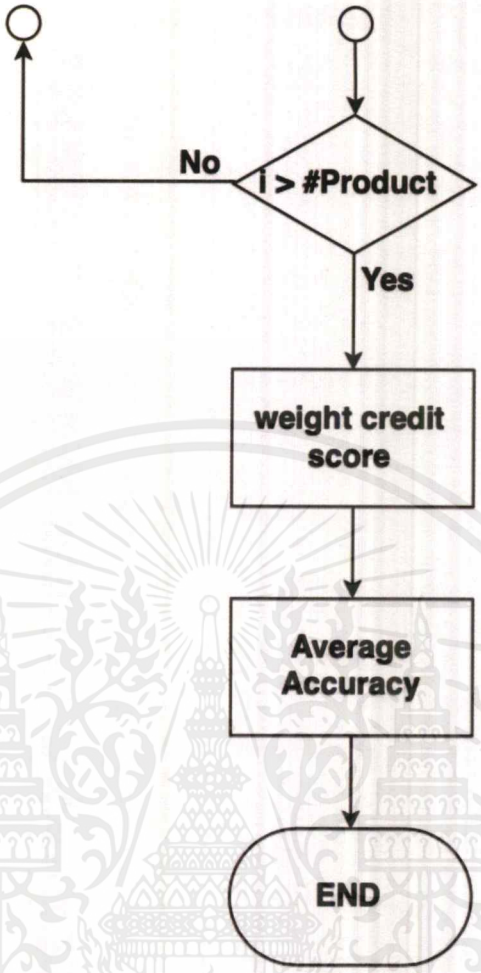
รูปที่ 3.7 แสดงความสัมพันธ์การคำนวณของ π_i

$$\pi_i = P(q_1 = S_i) \quad 1 \le i \le N$$

$$\pi_1 = P(q_1 = LS)$$

$$= \frac{1}{3}$$





รูปที่ 3.8 Flowchart แสดงการเข้าแบบจำลองและการวัดประสิทธิภาพแบบจำลอง

แรกเริ่มทำการสร้างแบบจำลองของร้านค้า โดยการนำระดับยอดขายสินค้าแต่ละเดือนภายในร้านค้าและยอดขายสินค้าแต่ละเดือนภายในหมวดหมู่ที่ร้านค้ามีตลอดระยะเวลาที่สินค้าเริ่มขายจนถึงเดือนปัจจุบันเข้าแบบจำลองฮิดเดนมาร์คอฟตามทฤษฎีดังกล่าว

คำนวณความน่าจะเป็นที่จะขายดีของแต่ละสินค้าและแต่ละเดือนด้วยวิธีการคำนวณ Likelihood ในบทที่ 2 โดยใช้ขั้นตอนวิธีแบบไปข้างหน้า (Forward Algorithm) นำมาคูณกับแบบไปข้างหลัง (Backward Algorithm) รวมทั้งทำนายผลการสังเกตของแต่ละสินค้าและแต่ละเดือนด้วยวิธีการถอดรหัสวิเทอร์บิในบทที่ 2 หาทั้งสองอย่างทำจนครบสินค้าทั้งหมดของร้านค้า

หลังจากที่ได้ผลการทำนายผลการสังเกตของแต่ละสินค้าและแต่ละเดือนแล้ว ในขั้นตอนต่อไปจะเป็นการวัดความแม่นยำของผลการสังเกตของแต่ละสินค้าและแต่ละเดือน ซึ่งหมายความว่าทำการ

ทดสอบแบบจำลองผลการทำนายเมื่อเทียบกับข้อมูลจริงมีความแม่นยำมากน้อยเพียงใด โดยใช้ขั้นตอนวิธีการ Confusion Matrix ในบทที่ 2

เมื่อหาความน่าจะเป็นที่จะขายดีของแต่ละสินค้าและแต่ละเดือน และการทำนายผลการสังเกตแต่ละสินค้าและแต่ละเดือนครบแล้วเรียบร้อยจึงกลายมาเป็นแบบจำลองของร้านค้า

หาคะแนนของสินค้าแต่ละชนิด ด้วยการนำยอดขายภายในร้านแต่ละสินค้าของสามเดือนล่าสุดเข้าแบบจำลองร้านค้า ผลลัพธ์ที่ได้จะเป็นความน่าจะเป็นที่จะขายดีแต่ละสินค้าและแต่ละเดือนของสามเดือนล่าสุด นำความน่าจะเป็นแต่ละสินค้าทั้งสามเดือนนั้นมาหาค่าเฉลี่ย และคำนวณหาคะแนนสินเชื่อของร้านค้า โดยการนำคะแนนของสินค้าแต่ละชนิดมาหาค่าเฉลี่ยถ่วงน้ำหนักถ่วงโดยอัตราส่วนของยอดขายชนิดนี้กับยอดขายทั้งหมดในร้าน

หาความแม่นยำการทำนายของสินค้าแต่ละชนิด ด้วยการนำผลการสังเกตแต่ละสินค้าของสามเดือนล่าสุดเข้าแบบจำลองร้าน ผลลัพธ์ที่ได้จะเป็นค่าความแม่นยำแต่ละสินค้าและแต่ละเดือนของสามเดือนล่าสุด นำค่าความแม่นยำแต่ละสินค้าทั้งสามเดือนนั้นมาหาค่าเฉลี่ย และคำนวณการวัดประสิทธิ์ภาพของร้านค้า โดยการนำค่าความแม่นยำของสินค้าแต่ละชนิดมาหาค่าเฉลี่ยถ่วงน้ำหนักถ่วงโดยเวลาทั้งหมดนับตั้งแต่สินค้าชนิดนั้นเริ่มขายจนถึงเดือนปัจจุบัน

ในบทที่ 4 จะอธิบายถึงผลการดำเนินงานของงานวิจัยเรื่องการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ

บทที่ 4

ผลการดำเนินงานและอภิปรายผล

ในการจัดทำงานวิจัยเรื่องการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ มีวัตถุประสงค์เพื่อเป็นทางเลือกหนึ่งในการประกอบการพิจารณาการให้สินเชื่อ ในปัจจุบันพ่อค้าแม่ค้าออนไลน์มีจำนวนมากขึ้น การขอสินเชื่อสำหรับกลุ่มคนประเภทนี้ยังคงเป็นเรื่องยากในการเตรียมหลักฐานทางการเงิน เนื่องจากข้อมูลหรือประวัติการค้าขายของพ่อค้าแม่ค้ากลุ่มนี้ส่วนใหญ่ถูกจัดเก็บบนหน้าเว็บไซต์ออนไลน์ จึงเป็นสาเหตุสำคัญที่ทางผู้จัดทำได้นำข้อมูลเหล่านี้มาสร้างแบบจำลองสำหรับร้านค้านั้น ๆ เพื่อวิเคราะห์และสร้างคะแนนสินเชื่อส่วนบุคคลนั่นเอง ซึ่งมีผลการดำเนินงาน ดังนี้

4.1 ผลจากการเตรียมข้อมูล

หลังจากได้ข้อมูลจากขั้นตอนการดึงข้อมูลแล้วจะต้องมีการทำความสะอาดข้อมูล เนื่องจากข้อมูลดิบที่ได้มานั้นยังไม่อยู่ในรูปแบบที่ใช้งานได้ และหลังจากทำความสะอาดข้อมูลแล้วนั้น ต้องเข้ากระบวนการคัดเลือก และแปลงข้อมูลเพื่อให้ได้ข้อมูลที่ต้องการไปใช้ในขั้นตอนต่อไป

	item_nam	item_sold	item_remain	item_rat	item_star	item_price	item_fav	item_	item_	item_review	id
1	ถูกมาก!!!	0	12	0	0	950	1	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
2	ไฟ LED ขน	6	297	2	5	279	6	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
3	กล้องติดรถ	0	500	0	0	691	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
4	กล้องติดรถ	0	1	0	0	1299	3	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
5	กล้องมองห	24	77	9	4.4	270	26	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
6	ที่ชาร์จโทร	0	94	0	0	129	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
7	กล้องติดรถ	0	100	0	0	849	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
8	T 658 กล้อ	0	16000	0	0	1062	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
9	กล้อง Bor	1	5099	1	3	154	0	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
#	<U+2764>	0	3	0	0	492	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	เครื่องชาร์	2	652	1	5	78	1	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
#	กล้องบันที	0	22	0	0	177	2	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	กล้องวิดีโอ	85	9913	33	4	468	333	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
#	B5GR <U+0	0	100	0	0	779	3	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	proof pf7	0	2	0	0	3590	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	อะแดปเตอ	1	599	1	5	81	0	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
#	Smart HD	0	500	0	0	1890	1	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	(ใส่โค้ด[N	0	2000	0	0	950	3	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	กล้องมองห	0	500	0	0	440	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam
#	ร้านแนะนำ	8	2	3	5	1790	21	กล้องส	0	[{"name":"someone can't loading review that have error"}]	cate_cam
#	กล้องมองห	0	5	0	0	213	0	กล้องส	0	["ยังไม่มีคะแนน"]	cate_cam

รูปที่ 4.1 ตารางข้อมูลดิบ

	item_nam	item_sold	item_rem	item_ratir	item_star	item_prio	item_favc	item_cate	item_sho	id	name	time	star	comment
1	NEW!! หูฟ	3	117	2	5	2330	23	True wirel	unicca268	cate_1	ajichanta	5/3/2019	5	NA
2	NEW!! หูฟ	3	117	2	5	2330	23	True wirel	unicca268	cate_1	luckylahm	4/19/2019	5	NA
3	หูฟังไร้สา	106	94	54	4.8	1650	252	In-Ear	unicca268	cate_2	monchai1	5/6/2019	5	
4	หูฟังไร้สา	106	94	54	4.8	1650	252	In-Ear	unicca268	cate_2	m_may_m	4/5/2019	5	
5	หูฟังไร้สา	106	94	54	4.8	1650	252	In-Ear	unicca268	cate_2	tawatchai	4/3/2019	5	
6	หูฟังไร้สา	106	94	54	4.8	1650	252	In-Ear	unicca268	cate_2	jibja123	3/30/2019	5	ส่งของไวมากค่ะ สินค้า
7	หูฟัง Huav	69	197	25	4.8	449	76	In-Ear	unicca268	cate_3	worradorr	4/29/2019	5	
8	หูฟัง Huav	69	197	25	4.8	449	76	In-Ear	unicca268	cate_3	khunchai1	4/27/2019	5	ส่งรวดเร็วทันใจดีครับช
9	หูฟัง Huav	69	197	25	4.8	449	76	In-Ear	unicca268	cate_3	tordatace	4/16/2019	5	
10	หูฟัง Huav	69	197	25	4.8	449	76	In-Ear	unicca268	cate_3	ywandg	4/10/2019	5	
11	OPPO R11	10	190	1	5	199	40	In-Ear	unicca268	cate_4	tine17y	11/7/2018	5	การให้บริการจากร้านค้
12	หูฟัง Earpl	0	25	NA	NA	3950	6	In-Ear	unicca268	cate_5	ยังไม่มีคะ	NA	NA	NA
13	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	aditepeie	5/10/2019	5	เสียงดีราคาถูก แบรตร์
14	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	chupview	5/10/2019	5	
15	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	seksan486	5/10/2019	5	
16	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	prachayas	5/9/2019	5	
17	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	bell0847	5/9/2019	5	
18	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	lethej	5/9/2019	5	ดีมาก ส่งเร็ว ถูกและดี
19	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	600112.3	5/9/2019	5	
20	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	pepepopa	5/8/2019	5	
21	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	mindlll	5/8/2019	5	
22	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	manaswe	5/4/2019	5	
23	ร้านแนะน	474	49	253	4.7	329	403	In-Ear	jjsuperche	cate_6	oattemasi	5/3/2019	5	

รูปที่ 4.2 ตารางที่ผ่านกระบวนการเตรียมข้อมูลเป็นที่เรียบร้อย

4.2 ผลจากการแปลงข้อมูล

ขั้นตอนนี้มีจุดประสงค์เพื่อนำข้อมูลเข้าสู่กระบวนการแปลงสภาพข้อมูลเพื่อนำข้อมูลที่ต้องใช้เข้าสู่แบบจำลองและเป็นเกณฑ์ใช้สำหรับการจัดระดับของยอดขายสินค้าภายในหมวดหมู่ LS (Low Sold) และ HS (High Sold) และภายในร้านค้า 1 (ขายไม่ดี) 2 (ขายดี) และ 3 (ขายดีมาก)

							q_t	O_t	t ของสินค้าแต่ละชนิด
	id	item_cate	total_review	total_sold	date	n_reviewer	sold_period	state	observe	time_index
1	product_1	แบตเตอรี่สำรอง	681	1958	28/5/2018	4	11.50073421	LS	2	1
2	product_1	แบตเตอรี่สำรอง	681	1958	28/6/2018	2	5.750367107	LS	1	2
3	product_1	แบตเตอรี่สำรอง	681	1958	28/7/2018	8	23.00146843	HS	2	3
4	product_1	แบตเตอรี่สำรอง	681	1958	28/8/2018	26	74.75477239	HS	3	4
5	product_1	แบตเตอรี่สำรอง	681	1958	28/9/2018	37	106.3817915	HS	3	5
6	product_1	แบตเตอรี่สำรอง	681	1958	28/10/2018	47	135.133627	HS	3	6
7	product_1	แบตเตอรี่สำรอง	681	1958	28/11/2018	151	434.1527166	HS	3	7
8	product_1	แบตเตอรี่สำรอง	681	1958	28/12/2018	92	264.5168869	HS	3	8
9	product_1	แบตเตอรี่สำรอง	681	1958	28/1/2019	84	241.5154185	HS	3	9
10	product_1	แบตเตอรี่สำรอง	681	1958	28/2/2019	104	299.0190896	HS	3	10
11	product_1	แบตเตอรี่สำรอง	681	1958	28/3/2019	51	146.6343612	HS	3	11
12	product_1	แบตเตอรี่สำรอง	681	1958	28/4/2019	47	135.133627	HS	3	12

รูปที่ 4.3 แสดงข้อมูลที่ถูกแบ่งกลุ่มข้อมูล

4.3 ผลจากการเข้าแบบจำลองและการวัดประสิทธิภาพแบบจำลอง

ในการสร้างแบบจำลองเพื่อวัดคะแนนสินค้าร้านค้าออนไลน์ร้านหนึ่ง ผู้จัดทำได้นำข้อมูลมาเข้าแบบจำลองฮิดเดนมาร์คอฟโดยใช้โปรแกรม R Studio โดยความน่าจะเป็นที่จะขายดีของแต่ละสินค้า

ภายในร้านที่ได้จากการคำนวณในแบบจำลองร้านค้าและผ่านการใช้ค่าเฉลี่ยถ่วงน้ำหนักกับยอดขายสินค้าสามเดือนล่าสุดจะแสดงถึงคะแนนสินเชื่อของผู้กู้รายนั้น

จากภาพที่ 4.3 จะเห็นว่าข้อมูลในตารางประกอบด้วยข้อมูลต่าง ๆ ที่ปรากฏบนหน้าเว็บไซต์ ซึ่งผ่านกระบวนการขั้นตอนการเตรียมข้อมูลและขั้นตอนการแบ่งกลุ่มข้อมูลเป็นที่เรียบร้อย โดยที่ LS และ HS คือข้อมูลที่ต้องนำเข้าแบบจำลองฮิดเดนมาร์คอฟเพื่อสร้างแบบจำลองของร้านค้า ซึ่ง LS คือ สถานะ Low Sold และ HS คือ สถานะ Hight Sold ของระดับยอดขายในหมวดหมู่สินค้าเดียวกัน

	LS	HS
LS	0.7423995	0.2576005
HS	0.1104349	0.8895651

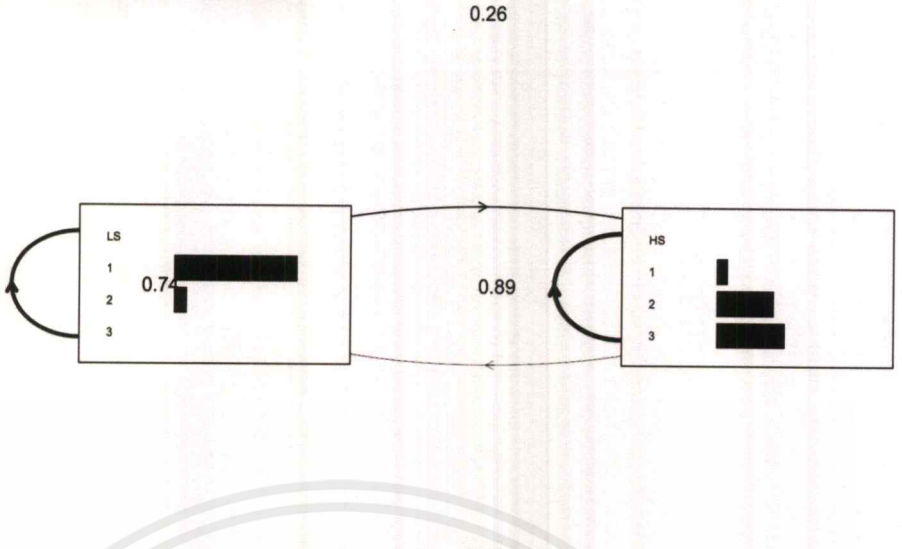
ตารางที่ 4.1 แสดงความน่าเป็นในการเปลี่ยนสถานะ

	1	2	3
LS	0.90767721	0.09227069	0.000052096
HS	0.07920288	0.41934672	0.501450400

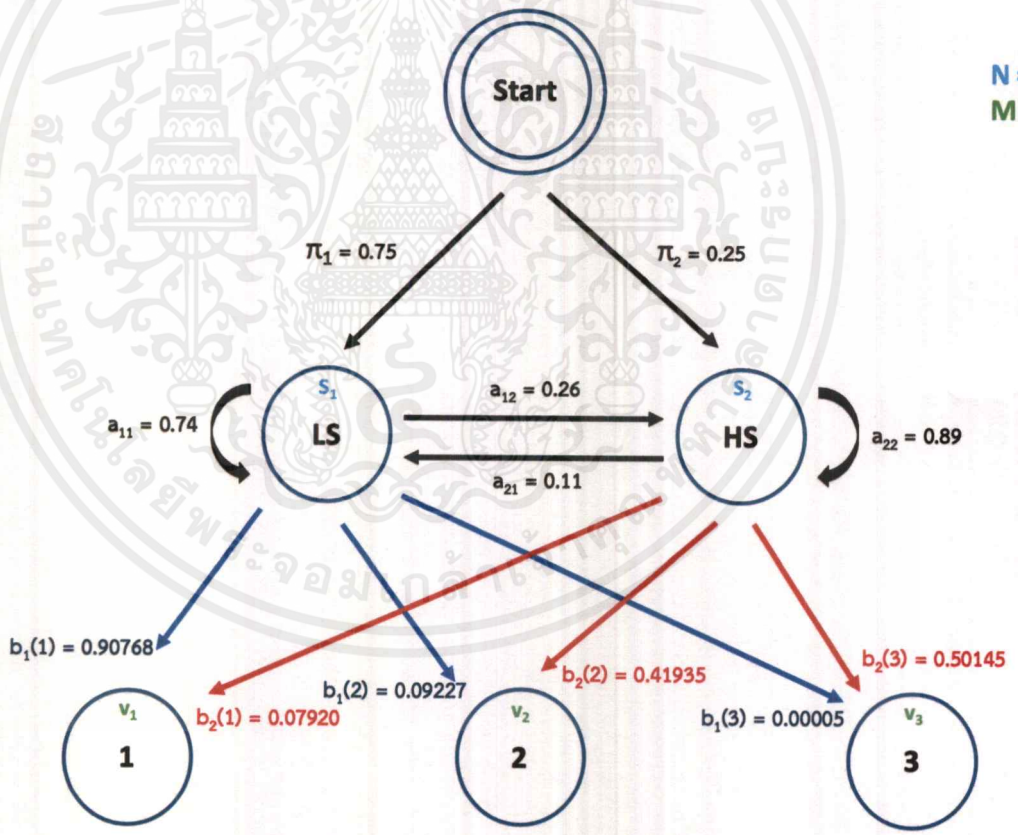
ตารางที่ 4.2 แสดงความน่าจะเป็นของสัญลักษณ์การสังเกต

	LS	HS
Begin	0.7498423	0.2501577

ตารางที่ 4.3 แสดงความน่าจะเป็นสถานะเริ่มต้นของแต่ละสถานะ



รูปที่ 4.4 แสดงแบบจำลองโดยการคำนวณจากโปรแกรม



รูปที่ 4.5 แสดงแบบจำลองของร้านค้า

จากรูปที่ 4.5 แสดงให้เห็นถึงความน่าจะเป็นของสถานะต่าง ๆ ในแบบจำลองโดยการคำนวณจากโปรแกรมจะได้ผลลัพธ์ ดังนี้

$a_{11} = 0.74$ ซึ่งหมายความว่า ความน่าจะเป็นในการเปลี่ยนสถานะจาก LS เป็น LS เท่ากับ 0.74

$a_{12} = 0.26$ ซึ่งหมายความว่า ความน่าจะเป็นในการเปลี่ยนสถานะจาก LS เป็น HS เท่ากับ 0.26

$a_{21} = 0.11$ ซึ่งหมายความว่า ความน่าจะเป็นในการเปลี่ยนสถานะจาก HS เป็น LS เท่ากับ 0.11

$a_{22} = 0.89$ ซึ่งหมายความว่า ความน่าจะเป็นในการเปลี่ยนสถานะจาก HS เป็น HS เท่ากับ 0.89

$b_1(1) = 0.90768$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 1 ในสถานะ LS เท่ากับ 0.90768

$b_1(2) = 0.09227$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 2 ในสถานะ LS เท่ากับ 0.09227

$b_1(3) = 0.00005$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 3 ในสถานะ LS เท่ากับ 0.00005

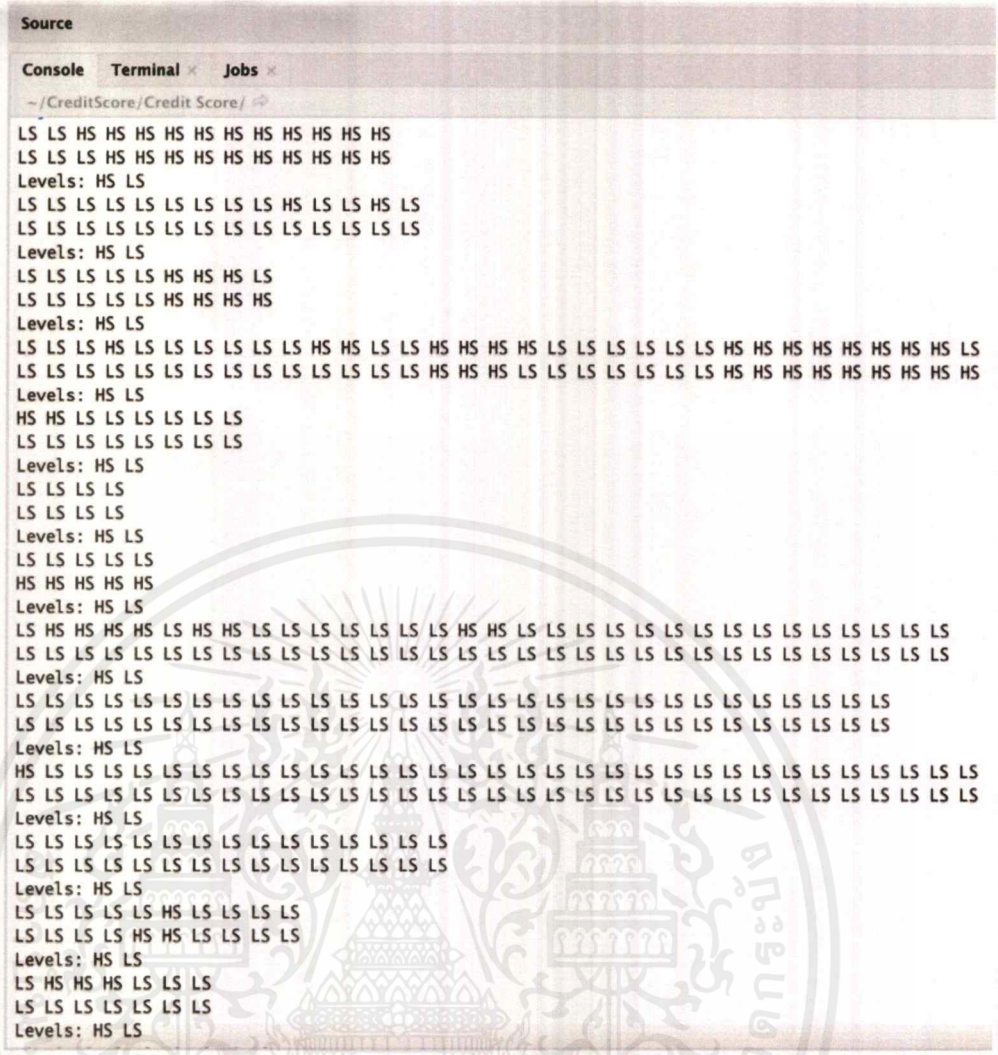
$b_2(1) = 0.07920$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 1 ในสถานะ HS เท่ากับ 0.07920

$b_2(2) = 0.41935$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 2 ในสถานะ HS เท่ากับ 0.41935

$b_2(3) = 0.50145$ ซึ่งหมายความว่า ความน่าจะเป็นที่ร้านค้ามีระดับของยอดขายในร้านเป็น 3 ในสถานะ HS เท่ากับ 0.50145

$\pi_1 = 0.75$ ซึ่งหมายความว่า ความน่าจะเป็นเริ่มต้นของสถานะ LS เท่ากับ 0.75

$\pi_2 = 0.25$ ซึ่งหมายความว่า ความน่าจะเป็นเริ่มต้นของสถานะ HS เท่ากับ 0.25



รูปที่ 4.6 แสดงผลการทำนายการเปลี่ยนสถานะยอดขายในหมวดหมู่

คะแนนสินเชื่อส่วนบุคคลที่ถูกสร้างจากแบบจำลองฮิดเดนมาร์คอฟนั้น สินค้าแต่ละชิ้นจะมีการให้คะแนนของค่าน้ำหนักที่ต่างกัน ซึ่งค่าน้ำหนักที่ว่านั่นคือจำนวนสินค้าที่ขายได้ ผลลัพธ์จึงเป็นค่าเฉลี่ยถ่วงน้ำหนัก ดังรูปที่ 4.7 คะแนนเต็มของแบบจำลองนี้คือ 10 คะแนน จะเห็นว่าคะแนนของร้านนี้อยู่ในระดับที่ค่อนข้างสูง และเพื่อสร้างความเชื่อมั่นให้กับผลลัพธ์ที่ได้จากแบบจำลอง การวัดประสิทธิภาพในหัวข้อถัดไปจึงเป็นอีกหนึ่งเครื่องมือที่จะช่วยยืนยันว่าคะแนนสินเชื่อส่วนบุคคลที่ได้มานี้มีความน่าเชื่อถือเพียงใด

```
[1] "Credit score : 9.5"
```

รูปที่ 4.7 แสดงคะแนนสินเชื่อส่วนบุคคล

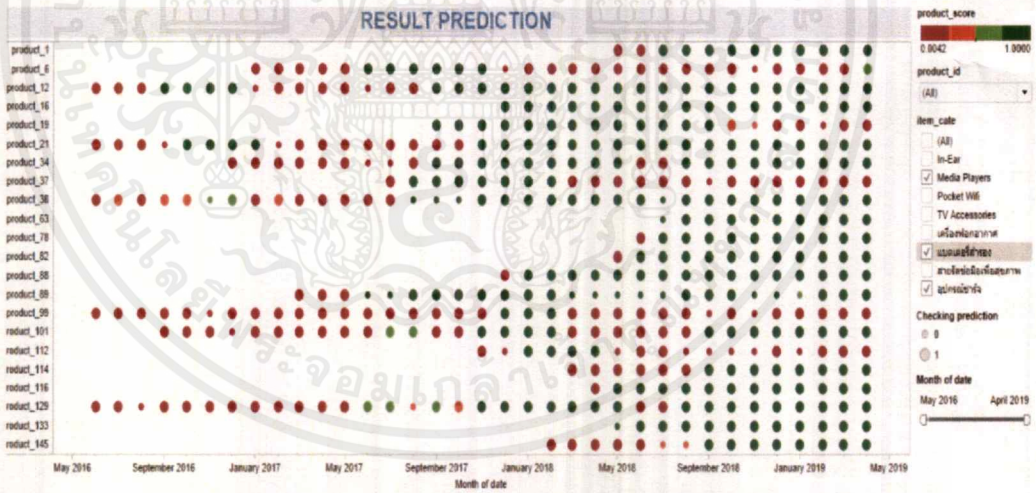
การวัดประสิทธิภาพของแบบจำลองนั้นมีความสำคัญมากอีกขั้นตอนหนึ่ง เนื่องจากเป็นสิ่งที่ช่วยยืนยันว่าแบบจำลองที่สร้างมานั้นมีความน่าเชื่อถือมากน้อยเพียงใดหากถูกนำออกมาใช้งานจริง โดยกระบวนนี้จะใช้การวัดประสิทธิภาพข้อมูลของตาราง Confusion Matrix แต่สินค้าแต่ละตัวนั้นจะมีการให้คะแนนของน้ำค่าหนักที่ต่างกัน ผลลัพธ์จึงเป็นค่าเฉลี่ยถ่วงน้ำหนัก ดังรูปที่ 4.8 ดังนั้นผลคะแนนสินเชื่อของร้านค้าร้านนี้ซึ่งผ่านแบบจำลองที่สร้างขึ้นจากแบบจำลองฮิดเดนมาร์คอฟ มีความน่าเชื่อถือ

```
[1] "Avg.Accuracy of test data : 93.2%"
```

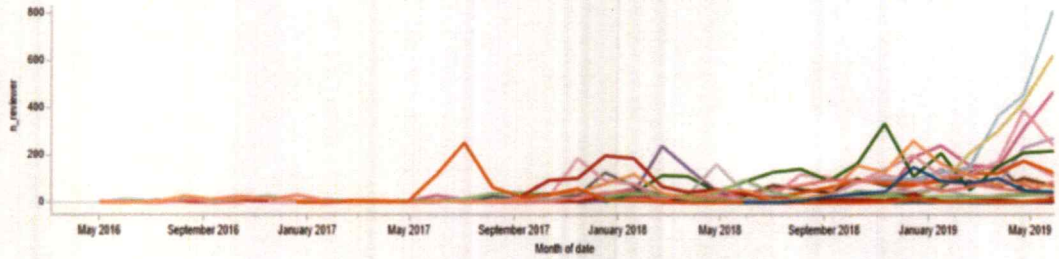
รูปที่ 4.8 แสดงผลการวัดประสิทธิภาพแบบจำลอง

4.4 การแสดงผลลัพธ์ด้วย Dashboard

เนื่องจากการนำเสนอข้อมูลมีเพียงตัวเลขมากมาย เพื่อให้สามารถสรุปข้อมูลได้ง่าย ๆ ใช้เวลาในการตีความสั้น ๆ และสามารถตอบโจทย์ในทางธุรกิจได้ ทางคณะผู้จัดทำจึงนำข้อมูลที่ได้จากแบบจำลองมานำเสนอในมุมมองใหม่ๆ



รูปที่ 4.9a แสดงผลลัพธ์จากแบบจำลองของร้านค้าในรูปแบบ Dashboard



รูปที่ 4.9b แสดงจำนวนผู้มารีวิวสินค้า

ในรูปที่ 4.9a แกนด้านซ้ายมือเป็นสินค้าทั้งหมดภายในหมวดหมู่ของร้านแห่งหนึ่ง แกนนอนเป็นเดือนนับตั้งแต่สินค้าชิ้นนั้นเริ่มขายจนถึงปัจจุบัน แกนด้านขวามือ ใช้เลือกหมวดของสินค้าที่สนใจ (ต้องการแสดงผล) สีต่าง ๆ บ่งบอกระดับคะแนนของสินค้าชิ้นนั้น ๆ ในเดือนนั้น ได้แก่ แดง ส้ม เขียวอ่อน และเขียวแก่ โดยเรียงลำดับจากน้อยไปมาก ขนาดของวงกลมบ่งบอกการวัดประสิทธิภาพของผลลัพธ์ของตัวสินค้าในเดือน ๆ นั้น ยิ่งเล็กยิ่งไม่ดี

ในรูปที่ 4.9b แกนด้านซ้ายมือเป็นจำนวนผู้ที่มารีวิวให้กับสินค้าชนิดนั้น ๆ ตั้งแต่เริ่มมีการขายจนถึงปัจจุบัน โดยสินค้าแต่ละชนิดสามารถแยกได้จากเส้นแต่ละสี

ในบทที่ 5 จะเป็นการอธิยายถึงข้อสรุปและข้อเสนอแนะของงานวิจัยเรื่องการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ

บทที่ 5
สรุปผลและข้อเสนอแนะ

ในการทำปัญหาพิเศษครั้งนี้ได้สรุปผลของการดำเนินงาน ข้อเสนอแนะและอุปสรรคของการทำงานเพื่อที่จะเป็นประโยชน์ต่อผู้เข้ามาศึกษา สามารถนำไปปรับปรุง และพัฒนาให้มีการสร้างแบบจำลองที่มีประสิทธิภาพและความแม่นยำมากยิ่งขึ้น

5.1 สรุปผลการดำเนินงาน

ในการจัดทำงานวิจัยเรื่องการใช้ข้อมูลจากร้านค้าออนไลน์ในการสร้างแบบจำลองสินเชื่อส่วนบุคคลเพื่อพิจารณาการให้คะแนนสินเชื่อ มีวัตถุประสงค์เพื่อศึกษาการใช้ข้อมูลบนโซเชียลมีเดียมาพิจารณาให้คะแนนสินเชื่อส่วนบุคคลเพื่อเป็นแนวทางในการเพิ่มช่องทางและวิธีการให้สินเชื่อแก่ผู้กู้ที่เป็นกลุ่มคนประเภทพ่อค้าแม่ค้าออนไลน์โดยมีเกณฑ์พื้นฐานการตัดสินใจของธนาคารในการพิจารณาอนุมัติสินเชื่อ

ข้อมูลที่ดึงจากเว็บไซต์ร้านค้าออนไลน์ นำมาผ่านกระบวนการแบ่งกลุ่ม และนำเข้าแบบจำลองร้านค้าที่สร้างขึ้นโดยแบบจำลองฮิดเดนมาร์คอฟสามารถคำนวณคะแนนสินเชื่อส่วนบุคคลของร้านค้าที่สนใจ สรุปได้ดังนี้

1. คะแนนสินเชื่อส่วนบุคคลของร้านค้าที่สนใจ คือ 9.5 เมื่อเทียบกับข้อมูลยอดขายของร้านค้าร้านนี้

2. ประสิทธิภาพแบบจำลองของร้านค้า คือ 93.2%

5.2 ปัญหาและอุปสรรค

1. ต้องศึกษาและหาแบบจำลองที่เหมาะสมกับข้อมูลจากร้านค้าและความสัมพันธ์ของพ่อค้าแม่ค้าออนไลน์กับลูกค้าของร้านค้าออนไลน์ที่สนใจ

2. ข้อมูลของสินค้าภายในร้านและข้อมูลสินค้าภายในหมวดหมู่มีจำนวนมาก จึงจำเป็นต้องใช้เวลาเยอะในการดึงข้อมูลและจำเป็นต้องใช้คอมพิวเตอร์ที่มีความเร็วค่อนข้างมากในการทำงาน

3. ราคาสินค้าเวลาตั้งแต่แรกเริ่มจนกระทั่งเวลาที่ทำการดึงข้อมูล อาจเป็นราคาไม่คงที่เนื่องจากอาจจะมีโปรโมชั่นลดราคาตามเทศกาลบนเว็บไซต์ออนไลน์ หรือเป็นพ่อค้าแม่ค้าทำการลดราคาด้วยตัวเอง ทำให้การคำนวณยอดขายจากร้านค้าไม่ใช่ยอดขายที่แท้จริงเพราะใช้ราคาสินค้าที่ติดป้ายไว้

4. เทรนด์การซื้อสินค้าของลูกค้า บางเวลาการสั่งซื้อสินค้าของลูกค้ามีจำนวนมากซึ่งขึ้นอยู่กับสถานการณ์ หรือเหตุการณ์ ณ ตอนนั้น ทำให้ไม่สามารถคำนวณยอดขายสินค้าต่อเดือนโดยมีสถานการณ์เป็นตัวแปรหลักได้อย่างแน่นอน

5.3 ข้อเสนอแนะและแนวทางในการพัฒนา

จากการศึกษาการใช้ข้อมูลในโซเชียลมีเดียเพื่อให้คะแนนสินเชื่อส่วนบุคคล พบว่าการศึกษาครั้งนี้ยังมีความครอบคลุมไม่มากพอและมีข้อจำกัดในเรื่องต่าง ๆ เพื่อปรับปรุงและเพิ่มประสิทธิภาพแบบจำลองทางผู้จัดทำมีข้อเสนอแนะ คือ ควรทำแบบจำลองร้านค้าให้มีความครอบคลุมในเรื่องราคาสินค้า ณ เวลานั้น ๆ เพื่อให้การคำนวณยอดขายมีความเที่ยงตรงมากที่สุด รวมถึงการนำรีวิวของลูกค้ามาประกอบการวัดคุณภาพของร้านค้าเพื่อพิจารณาคะแนนสินเชื่อ

เอกสารอ้างอิง

[1] *"Shopee ออนไลน์" แพลตฟอร์มช็อปปิ้งออนไลน์ที่โตเร็วสุดในอาเซียน*. (2017, November 3). Retrieved from sanook: https://www.sanook.com/money/524013/

[2] ภูริยากร, พ. (2007). *การประมาณทอพอโลยีของแบบจำลองฮิดเดนมาร์คอฟสำหรับการรู้จำหน่วยเสียงภาษาไทยโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม*. Retrieved from CUIR at Chulalongkorn University: http://cuir.car.chula.ac.th/handle/123456789/15408

[3] รดิศพงศ์, ท. (2012). การแปลงข้อมูลผลการวิจัยโดยวิธีทางสถิติ. *วารสารกรมวิทยาศาสตร์บริการ ปีที่ 60 ฉบับที่ 189*, 16-19.

[4] *ค่ามาตรฐาน (Standard Normal Score)*. (n.d.). Retrieved from opendurian: https://www.opendurian.com/learn/standard_normal_score/

[5] บวรธำรงชัย, ฉ. (2010). *การรู้จำจังหวัดในป้ายทะเบียนรถไทยด้วยแบบจำลองความน่าจะเป็น*. Retrieved from CUIR at Chulalongkorn University: http://cuir.car.chula.ac.th/handle/123456789/21151

[6] เพียสุพรรณ, น. (2014, 11 28). *การแจกแจงปกติและเส้นโค้งปกติ (Normal Curve)*. Retrieved from Normal Distribution' s Site: http://mathdistribution.weebly.com/358536343619364935923585364935923591361135853605363636493621363236483626365736093650358836573591361358536053636.html

[7] ภิรมย์รื่น, อ. (2018, April 5). *สรุปเรื่องเกี่ยวกับการวัดประสิทธิภาพ Model ของ Machine Learning แบบ Classification กันนิดหน่อย*. Retrieved from Blog@Auoychai: https://www.auoychai.com/big-datadata-analytic/สรุปเรื่องเกี่ยวกับการ/

[8] ศรประสิทธิ์, พ. (2010). *แผนการสอน วิชา ศร 232สถิติ เศรษฐศาสตร์ 1 ECS 232 Economics Statistics 1*. Retrieved from SWU - Economics - Pattranuch - Yola: http://pattranuch.yolasite.com/resources/ECS232/6.%20stat.pdf

[9] ชาติวัฒนานนท์, น. (1997). *ขนาดตัวอย่างสำหรับตัวสถิติทดสอบที ในกรณีที่ประชากรมีการแจกแจงไม่เป็นปกติ*. Retrieved from CUIR at Chulalongkorn University: http://cuir.car.chula.ac.th/dspace/handle/123456789/7768?src=%2Fdspace%2Fsimple-search%3Fhistory%3D%26location%3D%252F%26query%3D%26rpp%3D10%2

6sort_by%3Dscore%26order%3Ddesc%26filter_field_1%3Dauthor%26filter_type_1%3Dcontains%26filter_value_1%3D%25E0%25B

[10] Draxl, V. (2018, February 4). *BACHELOR PAPER Web Scraping Data Extraction from websites.* Retrieved from ACADEMIA: https://www.academia.edu/35901535/BACHELOR_PAPER_Web_Scraping_Data_Extraction_from_websites

[11] Hamilton, H. J. (n.d.). *Confusion Matrix.* Retrieved from Howard J. Hamilton: http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html

[12] Jaibun, D. (2011, 10 14). *5. การวัดตำแหน่งที่ของข้อมูล.* Retrieved from **โรงเรียนมหิดลวิทยานุสรณ์** : http://www.mwit.ac.th/~math/E_Learning/MATH30203/sources/Statistics_05.pdf

[13] Nanehkaran, Y. A. (2013). An Introduction To Electronic Commerce. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 2*, 190-193.

[14] Plagad. (2010, August 26). *Confusion Matrix.* Retrieved from Plagad's Blog: https://plagad.wordpress.com/2010/08/26/confusion-matrix/

[15] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 257-286.

[16] *StatEasyUse (สถิติง่ายง่าย).* (2015, October 11). Retrieved from Blogger: http://stateasyuse.blogspot.com/2015/10/blog-post_11.html

[17] Wikipedia. (2019, May 7). *Sensitivity and specificity.* Retrieved from wikipedia: https://en.wikipedia.org/wiki/Sensitivity_and_specificity