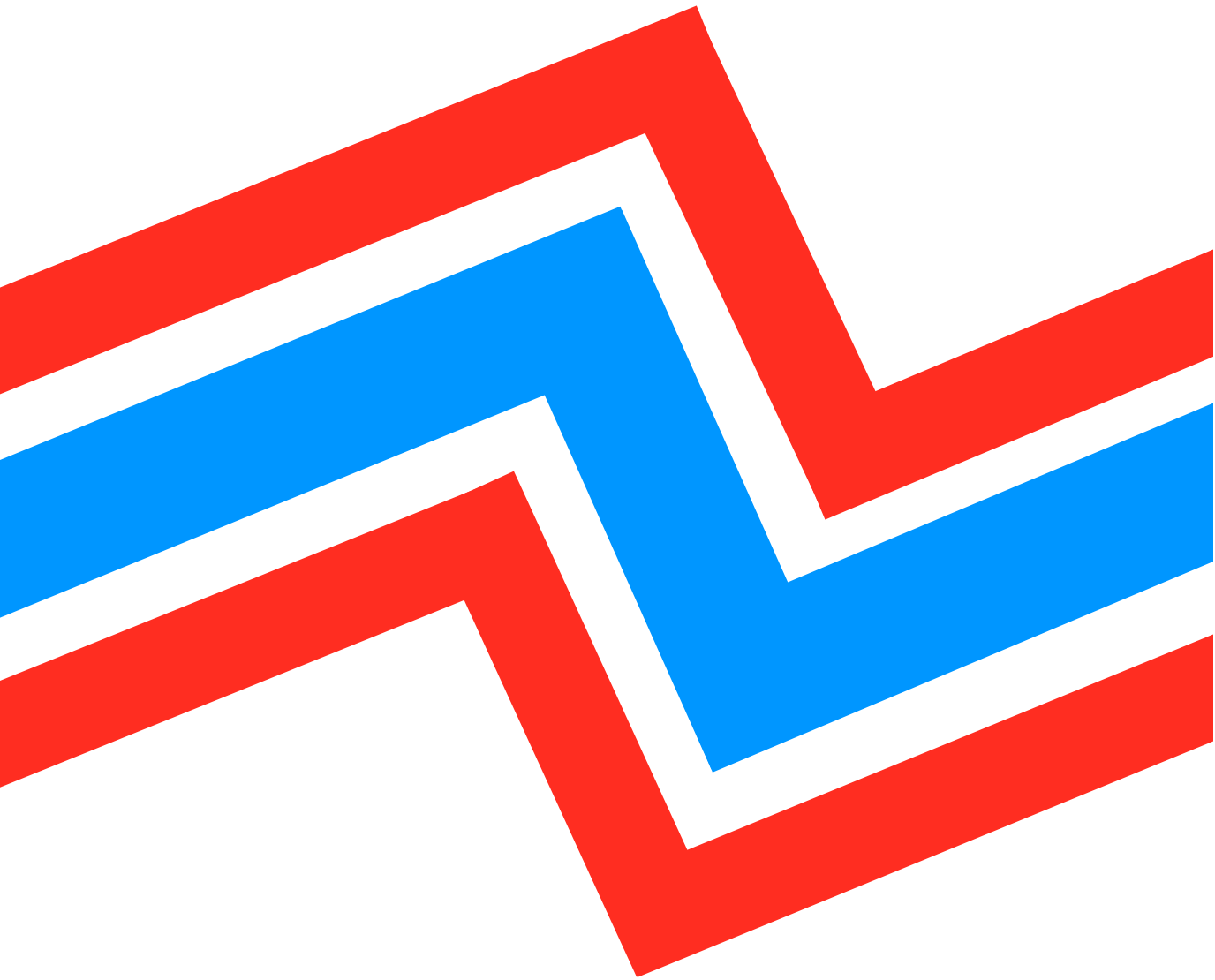
An Introduction to Data Mining Techniques

Thai version



Eakasit Pacharawongsakda, Ph.D.
Certified RapidMiner Analyst
Data Cube: http://facebook.com/datacube.th

การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น

โดย ดร. เอกสิทธิ์ พัชรวงศ์ศักดา

หสม. ดาต้า คิวบ์
http://facebook.com/datacube.th
http://www.dataminingtrend.com

ชื่อผู้แต่ง ดร. เอกสิทธิ์ พัชรวงศ์ศักดา

ชื่อหนังสือ การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น

จำนวนหน้า 124 หน้า

พิมพ์ครั้งที่ 1

เดือนปีที่พิมพ์ สิงหาคม 2557

ชื่อโรงพิมพ์ บริษัท เอเซีย ดิจิตอลการพิมพ์ จำกัด
21/19-20 ถ.งามวงศ์วาน แขวงลาดยาว
เขตจตุจักร กรุงเทพมหานคร 10900

จัดจำหน่ายโดย หสม. ดาต้า คิวบ์
53/486 หมู่ 13 นวนคร ต.คลองหนึ่ง
อ.คลองหลวง จ.ปทุมธานี

ออกแบบปก นางสาว กมนนัทธ์ บางแวก

ราคา 499 บาท

แด่ พ่อ แม่ ผู้ให้กำเนิด
และอาจารย์ผู้ประสิทธิ ประศาสตร์วิชาความรู้ต่างๆ

คำนำ

ย้อนหลังไปเมื่อ 12 ปีก่อน การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิง (Data Mining) ยังรู้จักกันในวงแคบส่วนใหญ่เป็นนักศึกษาปริญญาโทและเอกที่สนใจทำงานวิจัยทางด้านนี้ ผมเองเริ่มต้นรู้จักกับดาต้า ไมน์นิงเมื่อประมาณ 12 ปีก่อนเช่นกัน ในสมัยที่เป็นนักศึกษาปริญญาตรีตัวเล็กๆ ในห้องปฏิบัติการวิจัยการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery Laboratory) ในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ในช่วงเวลาที่ผ่านมาผมได้เห็นการเปลี่ยนแปลงเกี่ยวกับความสนใจของผู้คนในเรื่องดาต้า ไมน์นิงอย่างมากมาย ตั้งแต่ตอนแรกที่ความสนใจอยู่ในวงแคบดังที่ได้กล่าวมาแล้วจนมาถึงปัจจุบันที่มีผู้สนใจเพิ่มขึ้นเป็นวงกว้าง เช่น บริษัทเอกชนหรือธนาคารต่างเริ่มให้ความสนใจนำการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงไปใช้งานกันมากขึ้นหรือมหาวิทยาลัยบางแห่งเริ่มจัดให้มีการเรียนการสอนเกี่ยวกับเรื่องนี้ในระดับชั้นปริญญาตรี จากความนิยมที่เพิ่มขึ้นและการเก็บเกี่ยวประสบการณ์ในการทำงานและการอบรมการวิเคราะห์ข้อมูลทางด้านดาต้า ไมน์นิง ทำให้ผมคิดอยากจะเขียนหนังสือสักเล่มซึ่งทำการแนะนำเทคนิคการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงเบื้องต้นสำหรับนักศึกษาและผู้สนใจขึ้นมาและนั่นเองคือที่มาของหนังสือเล่มนี้ที่ชื่อว่า An Introduction to Data Mining Techniques โดยในหนังสือเล่มนี้ผมจะแสดงหลักการทำงานของวิธีการทางด้านดาต้า ไมน์นิงไม่ว่าจะเป็น การหากฎความสัมพันธ์ (association rules discovery) การแบ่งกลุ่มข้อมูล (clustering) และ การจำแนกประเภทข้อมูล (classification) พร้อมทั้งตัวอย่างการทำงานของวิธีการเหล่านี้เพื่อให้ผู้อ่านเข้าใจได้ง่ายโดยที่ไม่ต้องมีความรู้พื้นฐานทางด้านคณิตศาสตร์ขั้นสูง โดยการจัดพิมพ์ครั้งที่ 2 นี้ผมตั้งใจเผยแพร่ความรู้ทางด้าน ดาต้า ไมน์นิงให้กับผู้สนใจโดยทั่วไป ผมอยากให้คนรู้จักดาต้า ไมน์นิงว่าแท้จริงแล้วเป็นอย่างไร ไม่ใช่เป็นเพียงแค่คิดเอาเองว่าบิ๊ก ดาต้า (Big Data) คือ ดาต้า ไมน์นิง ผมขออธิบายแบบนี้ครับว่าบิ๊ก ดาต้า แบ่งเป็น 2 ส่วน ส่วนแรกคือการจัดเก็บข้อมูลที่มีจำนวนมหาศาลมาเก็บไว้ ซึ่งในปัจจุบันประเทศไทยเราคงเริ่มอยู่ที่ระดับนี้ ในส่วนที่ 2 คือ การนำข้อมูลที่มีจำนวนมากมาทำการวิเคราะห์ สำหรับผมคิดว่าการวิเคราะห์ข้อมูลคงต้องใช้เวลาอีกสักพักสำหรับบิ๊ก ดาต้าและเมื่อถึงเวลานั้นผมคิดว่าท่านผู้อ่านจะมีความสามารถในการวิเคราะห์ข้อมูลเหล่านี้ได้เป็นอย่างดีแล้ว

สุดท้ายนี้หนังสือเล่มนี้คงไม่สามารถเกิดขึ้นได้ถ้าผมไม่ได้มีโอกาสเรียนรู้เรื่องการวิเคราะห์ข้อมูลด้วยดาต้า ไมน์นิงจาก รศ. ดร. กฤษณะ ไวยมัย จากมหาวิทยาลัยเกษตรศาสตร์ ดร. สุภาวดี อิงศรีสว่าง จากศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ศ. ดร. ธนารักษ์ ธีระมั่นคง จากสถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทยาลัยธรรมศาสตร์ และ ศ. ดร. นิค เซอร์โคน (Nick Cercone) จากมหาวิทยาลัย ยอร์ค (York University) ประเทศแคนาดา

ดร. เอกสิทธิ์ พัชรวงศ์ศักดา
28 สิงหาคม 2557

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

บทที่ 1 การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง (Data Mining)

- แนะนำการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง (Data Mining)
  - ความหมายของดาต้า ไมน์นิง
  - การประยุกต์ใช้งานดาต้า ไมน์นิง
- ข้อมูลในรูปแบบต่างๆ
  - ข้อมูลแบบที่มีโครงสร้าง (structured data)
  - ข้อมูลแบบที่ไม่มีโครงสร้าง (unstructured data)
- เทคนิคในการวิเคราะห์ข้อมูลด้วยดาต้า ไมน์นิง
  - เทคนิคที่การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)
  - เทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning)

1.1 แนะนำการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง

เคยลองสังเกตุดูไหมครับว่ารอบๆ ตัวเราเต็มไปด้วย ... **ข้อมูล** ... ซึ่งตัวเราเองเป็นส่วนหนึ่งที่ทำให้ข้อมูลเหล่านี้เกิดขึ้น ท่านผู้อ่านลองจินตนาการตามเรื่องที่ผมจะเล่าต่อไปนี้ดูนะครับ ...

"เช้าวันทำงานที่เร่งรีบวันหนึ่ง ปิติตื่นนอนด้วยความสดชื่น แต่แล้วความหิวก็บังเกิดขึ้นตามมาเขาจึงเดินเข้าไปซื้อขนมจีบและซาลาเปาที่ร้านสะดวกซื้อซึ่งเปิดตลอด 24 ชั่วโมงที่ตั้งอยู่หน้าปากซอย เมื่อท้องอิ่มก็ได้เวลาที่ต้องออกไปทำงานเสียที ปิติจึงได้ขับรถออกไปและสักระยะหนึ่งเขาก็พบว่าน้ำมันรถนั้นใกล้จะหมดจึงจำเป็นต้องเลี้ยวรถเข้าไปเติมน้ำมันในปั๊มเสียก่อน เมื่อน้ำมันถูกเติมจนเต็มแล้วพนักงานปั๊มจึงเดินมาเพื่อขอเก็บเงิน ปิติหยิบบัตรเครดิตออกมาเพื่อจ่ายค่าน้ำมัน หลังจากนั้นจึงเดินทางต่อไปทำงาน หน้าที่ของปิติ คือ ผู้ให้บริการข้อมูล (call center) ของบริษัทเครือข่ายโทรคมนาคมรายใหญ่แห่งหนึ่ง ซึ่งในแต่ละวันเขาจะต้องให้บริการลูกค้าที่มีปัญหาในด้านต่างๆ และโดยเฉลี่ยในหนึ่งวันนั้นปิติต้องให้บริการลูกค้าหลายสิบราย ในช่วงบ่ายของวันนั้น ชูใจซึ่งเป็นเพื่อนสนิทของปิติคิดอยากจะดูภาพยนตร์จึงได้โทรศัพท์มาชวนปิติเพื่อนัดทานข้าวแล้วก็ดูภาพยนตร์ในตอนเย็นหลังเลิกงาน ปิติเห็นด้วยกับชูใจ เธอจึงทำการซื้อบัตรภาพยนตร์ผ่านทางระบบอินเตอร์เน็ตและชำระเงินด้วยบัตรเครดิตของเธอเองก่อน หลังจากนั้นเธอจึงนัดเขาให้ไปเจอกันที่ห้างสรรพสินค้าที่เปิดใหม่ใจกลางกรุง แต่ปิติลองคิดดูแล้วว่าถ้าเขาขับรถไปก็คงไม่สะดวกจึงได้ตัดสินใจจะเดินทางไปด้วยรถไฟฟ้า เมื่อถึงตอนเลิกงานปิติจึงได้เดินทางออกจากที่ทำงานไปยังสถานีรถไฟฟ้า และโดยสารรถไฟฟ้าไป เมื่อไปถึงห้างสรรพสินค้าและพบชูใจแล้ว ทั้งสองยังมีเวลาเหลือพอที่จะไปหาอาหารรับประทาน ปิติและชูใจจึงเดินไปเดินมาพบว่ามีร้านบะหมี่ญี่ปุ่นชื่อน่าอร่อยแห่งหนึ่งมีโปรโมชันใหม่ที่น่าสนใจที่

สามารถให้ลูกค้าสามารถสร้างเมนูใหม่ขึ้นมาเองได้ ปิติและชูใจจึงได้ตัดสินใจรับประทานอาหารที่ร้านนี้ เมื่อทั้งสองได้รับอาหารแล้วต่างก็ถ่ายรูปอาหารของตัวเองและโพสต์ลงไปในเว็บไซต์ Facebook และ Instagram เพื่อแชร์ภาพเหล่านี้ให้กับเพื่อนของตนเองได้เห็น หลังจากท้องอิ่มแล้วจึงได้เวลาไปชมภาพยนตร์ ทั้งสองใช้เวลาในการเพลิดเพลินกับการชมภาพยนตร์ไปร่วม 2 ชั่วโมง เมื่อออกมาจึงพบว่าดึกมากแล้วปิติและชูใจจึงได้แยกย้ายกันกลับบ้าน"

จากตัวอย่างที่ยกขึ้นมานี้ก็จะพอสรุปได้ว่าปิติและชูใจสร้างข้อมูลขึ้นมามากมาย ดังนี้

- ข้อมูลการซื้อสินค้าในร้านสะดวกซื้อ
- ข้อมูลการใช้บริการผ่านบัตรเครดิตในการเติมน้ำมันและซื้อบัตรชมภาพยนตร์
- ข้อมูลลูกค้าที่มาใช้บริการสอบถามข้อมูลทางโทรศัพท์ (call center)
- ข้อมูลการใช้โทรศัพท์ในการติดต่อสื่อสาร
- ข้อมูลอาหารที่เลือกรับประทาน
- ข้อมูลรูปภาพต่างๆ ที่แสดงบนเว็บไซต์เครือข่ายสังคม (online social network) ต่างๆ

จะเห็นได้ว่าข้อมูลส่วนใหญ่นั้นเกิดจากการทำกิจวัตรประจำวันของเราทั้งสิ้นไม่ว่าจะเป็นแบบออฟไลน์ (offline) เช่น การซื้อสินค้าจากร้านค้าต่างๆ หรือข้อมูลในรูปแบบออนไลน์ (online) เช่น การ โพสต์ข้อความหรือรูปภาพต่างๆ ขึ้นไปบนเว็บไซต์เครือข่ายทางสังคม เป็นต้น ข้อมูลเหล่านี้มักจะถูกเก็บไว้ในรูปแบบดิจิตอล (digital) เพื่อให้สะดวกต่อการค้นหา โดยในแต่ละวันข้อมูลเหล่านี้ก็จะมีปริมาณเพิ่มมากขึ้นเรื่อยๆ จนทำให้เราก้าวเข้าสู่ยุคที่เรียกว่าบิ๊ก ดาต้า (Big Data) หรือยุคที่มีข้อมูลขนาดมหาศาลเมื่อข้อมูลมีจำนวนมากขึ้นย่อมทำให้เกิดความต้องการนำข้อมูลเหล่านี้มาใช้เพื่อก่อให้เกิดประโยชน์มากที่สุด วิธีการหนึ่งที่นิยมใช้กันมากในปัจจุบันคือการวิเคราะห์หาความสัมพันธ์ที่ซ่อนอยู่ในข้อมูล วิธีการนี้เรียกว่า "การขุดเหมืองข้อมูล" (Data Mining) หรือเรียกทับศัพท์ว่า ดาต้า ไมน์นิง (ซึ่งในหนังสือเล่มนี้ผมขอเรียกชื่อทับศัพท์เพื่อให้เป็นสากลและเข้าใจได้ง่ายกว่าครับ) ดังนั้น ในหนังสือเล่มนี้เราจะได้มาเรียนรู้ถึงวิธีการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เพื่อค้นหา "**สิ่งที่มีประโยชน์ที่แอบซ่อนอยู่ภายในข้อมูลเหล่านี้"** กันครับ

โดยเนื้อหาในบทนี้จะอธิบาย 2 เรื่องครับ คือ (1) ดาต้า ไมน์นิงคืออะไร และ (2) ควรจะทำดาต้า ไมน์นิงเมื่อไรและจะได้ประโยชน์อะไร ซึ่งรายละเอียดจะมีดังต่อไปนี้

(1) ความหมายของดาต้า ไมน์นิง

ดาต้า ไมน์นิงเป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า "เหมืองข้อมูล" ซึ่งเป็นคำศัพท์ที่ใช้เปรียบกับการขุดเหมืองแร่ทั่วๆ ไป โดยในการขุดเหมืองแร่นั้นสิ่งที่ต้องการก็คือแร่ที่มีค่า เช่น เพชร พลอย ต่างๆ ในขั้นตอนการทำเหมืองแร่นั้นจะต้องระเบิดภูเขาใหญ่หลายๆ ลูกเพื่อค้นหาแร่ที่ต้องการ ซึ่งแร่ที่พบนั้น

ก็ได้ออกมาน้อยมากเมื่อเทียบกับหินที่โดนระเบิดจากภูเขา เช่นเดียวกันเมื่อในองค์กรหรือบริษัทมีภูเขาของข้อมูลที่มีขนาดมหาศาล บริษัทจึงต้องการขุดค้นหาลงไปในข้อมูลเหล่านี้เพื่อให้ได้สิ่งที่มีค่าซึ่งอยู่ในข้อมูลเหล่านี้ ทว่านี่เป็นการเปรียบเปรยให้เห็นลักษณะที่คล้ายกันระหว่างการขุดเหมืองแร่และการขุดเหมืองข้อมูลนะครับ ถ้าพูดกันตามหลักวิชาการแล้วมีผู้เชี่ยวชาญหลายท่านได้ให้ความหมายของดาต้า ไมน์นิงไว้ดังนี้ครับ

- "*The exploration and analysis of large quantities of data in order to discover meaningful patterns and rules*"
  "เป็นการวิเคราะห์ข้อมูลเพื่อค้นหารูปแบบ (patterns) หรือ กฎ (rules) ที่เกิดขึ้นในฐานข้อมูลขนาดใหญ่" (จากหนังสือ Data Mining Techniques for Marketing, Sales and Customer Relationship Management 3rd Edition)

- "*Extraction of interesting (non-trivial, previously, unknown and potential useful) information from data in large databases*"
  "เป็นกระบวนการดึงข่าวสารที่น่าสนใจ และมีประโยชน์แต่ ไม่เคยรู้มาก่อนจากฐานข้อมูลขนาดใหญ่" (จากหนังสือ Data Mining Concept and Techniques 2nd Edition)

จากหนังสือทั้งสองเล่มเราสามารถสรุปได้ว่า **ดาต้า ไมน์นิง** คือ "การค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่" อาทิเช่น ข้อมูลการซื้อขายสินค้าในซุเปอร์มาร์เก็ตต่างๆ ซึ่งข้อมูลนี้จะเก็บรายการสินค้าที่ลูกค้าซื้อในแต่ละครั้งโดยเมื่อทำการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง แล้วจะได้สิ่งที่มีประโยชน์เช่น "ลูกค้าส่วนใหญ่ที่ซื้อเบียร์มักจะซื้อผ้าอ้อมด้วย" จะเห็นได้ว่าข้อมูลนี้เป็นข้อมูลที่ไม่เคยคิดว่ามีความสัมพันธ์กันและไม่เคยรู้มาก่อนเลย เมื่อได้ความรู้แบบนี้ออกมาแล้วอาจจะนำไปออกโปรโมชันหรือช่วยในการจัดวางชั้นสินค้าในซุปเปอร์มาร์เก็ตต่อไปได้

(2) การประยุกต์ใช้งานดาต้า ไมน์นิง

การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงนี้มีตัวอย่างความสำเร็จให้เห็นอยู่เยอะครับ ผมขอยกตัวอย่างแรกซึ่งเป็นตัวอย่างคลาสสิกให้ท่านผู้อ่านได้ทราบก่อนครับ นั่นก็คือการที่ห้างวอล์มาร์ท (Walmart) ได้ทำการค้นพบพฤติกรรมการซื้อสินค้าของลูกค้าที่เป็นเพศชายว่า ในช่วงเย็นของวันศุกร์มักจะมีลูกค้ากลุ่มหนึ่งมาซื้อสินค้าบางอย่างควบคู่กันไป นั่นก็คือ "เบียร์และผ้าอ้อม" โดยจากการวิเคราะห์เจาะลึกลงไปก็พบเหตุผลว่าการที่สินค้าสองอย่างนี้มีการซื้อร่วมกันบ่อยๆ เพราะว่า พ่อบ้านส่วนใหญ่มักจะซื้อผ้าอ้อมให้ลูกน้อยและไม่ได้ออกไปผับมากนักหลังจากจากมีลูกจึงได้ซื้อเบียร์ไปดื่มในช่วงสุดสัปดาห์ [1] หลังจากที่ห้างวอล์มาร์ทรู้ถึงพฤติกรรมแบบนี้ทางห้างก็สามารถที่จะจัดวางสินค้าสองชนิดนี้ให้สามารถค้นหาได้ง่ายๆ หรือ

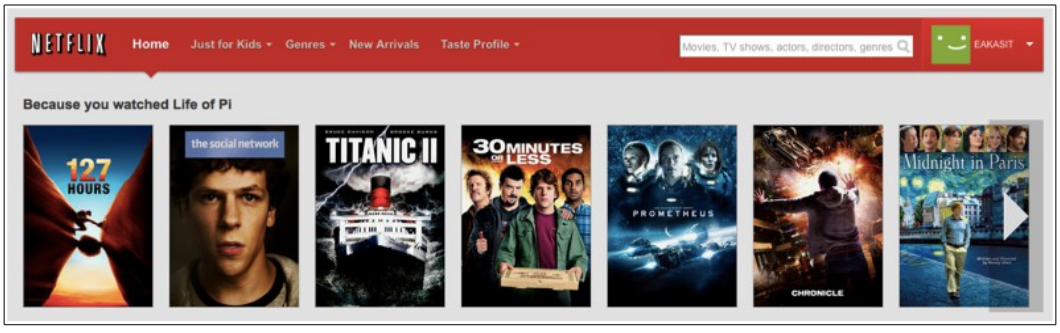
มองเห็นได้ง่ายเพื่อเพิ่มโอกาสที่ลูกค้าจะได้ซื้อติดไม้ติดมือกันไปด้วยครับ ส่วนตัวอย่างที่สองก็ยังคงมาจากห้างสรรพสินค้าเหมือนกันครับ นั่นคือห้างทาร์เก็ท (Target) ห้างนี้เป็นห้างที่เกิดขึ้นมาทีหลังทำให้การจะแข่งขันกับห้างวอล์มาร์ทที่มีอยู่ก่อนแล้วก็คงไม่ใช่เรื่องง่าย ดังนั้นทางห้างจึงพยายามหาวิธีที่จะดึงดูดให้ลูกค้ามาซื้อสินค้ากับทางห้างให้มากขึ้นและรักษาฐานลูกค้าที่มีอยู่ให้เชื่อใจและอยากกลับมาซื้อสินค้าที่ห้างของตนเองให้ได้มากที่สุด จากการวิจัยทางการตลาดของห้างทาร์เก็ทพบว่า เมื่อครอบครัวมีสมาชิกใหม่เกิดขึ้นคนในครอบครัวก็จะเริ่มมีการจับจ่ายใช้สอยมากขึ้นเพื่อรองรับการขยายขนาดของครอบครัว ดังนั้นเมื่อทราบเช่นนี้แล้วทางห้างทาร์เก็ทจึงได้ทำการวิเคราะห์พฤติกรรมของลูกค้าผู้หญิงที่มาซื้อสินค้าและพบว่าเมื่อลูกค้าเหล่านี้เริ่มตั้งครรภ์ ลูกค้าจะมีพฤติกรรมการซื้อสินค้าที่เปลี่ยนไป เช่น เริ่มมีการซื้อวิตามินบำรุงมากขึ้น เปลี่ยนไปกินอาหารที่มีประโยชน์ หรือแม้กระทั่งซื้อตู้เสื้อผ้าเพิ่ม จากรูปแบบพฤติกรรมลักษณะนี้ทำให้ทางห้างสามารถส่งโปรโมชันที่เกี่ยวกับการตั้งครรภ์หรือสินค้าสำหรับเด็กให้กับลูกค้ากลุ่มนี้ได้ นอกจากนี้ห้างทาร์เก็ทยังมีความมั่นใจว่าถ้าลูกค้าเชื่อใจที่จะซื้อสินค้าให้กับบุตรที่เกิดขึ้นใหม่แล้วลูกค้าเหล่านี้ก็จะเชื่อใจซื้อสินค้าชนิดอื่นๆ ของทางห้างไปอีกเรื่อยๆ [2]

เมื่อย้อนกลับมาดูห้างสรรพสินค้าในประเทศไทยเราจะพบว่าในบ้านเราก็เริ่มมีการเก็บข้อมูลเพื่อทำการวิเคราะห์กันมากขึ้นโดยท็อปส์ ซุปเปอร์มาร์เก็ต (TOPS supermarket) ในเครือบริษัทเซ็นทรัล (Central) เป็นห้างสรรพสินค้าแรกที่มีการวิเคราะห์ข้อมูลการซื้อสินค้าโดยใช้ข้อมูลจากบัตร สปอท (SPOT) ซึ่งเป็นบัตรสมาชิก (loyalty card) ของท็อปส์เอง หลังจากทำการวิเคราะห์ข้อมูลแล้วท็อปส์จะนำเสนอโปรโมชัน (promotion) ที่แตกต่างกันให้กับลูกค้าแต่ละคนผ่านทางระบบที่เรียกว่า เพอร์ซันนัล ช็อปปิง ลิสต์ (Personal Shopping List) ซึ่งสมาชิกของท็อปส์สามารถดูโปรโมชันที่ถูกจัดเตรียมไว้ให้ตัวเองได้โดยการนำบัตร สปอท ไปสอดที่เครื่องอ่านที่อยู่หน้าทางเข้าของซุเปอร์มาร์เก็ต [3] เราลองมาดูห้างสรรพสินค้าอื่นๆ กันบ้างนะครับ ผมขอยกตัวอย่างของห้างเทสโก้ โลตัส (Tesco Lotus) ซึ่งเป็นห้างค้าส่งของประเทศอังกฤษ ต้องขอบอกก่อนครับว่าห้างเทสโก้ที่ประเทศอังกฤษนั้นได้ว่าจ้างบริษัท ดันน์ฮัมบี้ (Dunnhumby) ทำการวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิงมานานมากแล้วครับ แต่ในประเทศไทยเราเริ่มมีขึ้นเมื่อประมาณ 3-4 ปีที่ผ่านมาโดยการเริ่มจากให้ลูกค้าสมัครบัตร คลับ การ์ด (Club card) หลังจากนั้นเมื่อลูกค้ามาซื้อสินค้าแต่ละครั้งทางห้างเทสโก้ โลตัสก็จะทราบพฤติกรรมการซื้อสินค้าของลูกค้าแต่ละรายเพื่อทำการวิเคราะห์ข้อมูลและส่งโปรโมชันออกไปให้ลูกค้าแต่ละคน นอกจากนี้ยังมีตัวอย่างกรณีศึกษาของเทสโก้ โลตัส ที่พยายามจะทำการยกเลิกการขายอาหารสุนัขถุงใหญ่ขนาด 10-20 กิโลกรัมเนื่องจากมีความเชื่อว่าลูกค้าไม่ชอบที่จะแบกอาหารสุนัขถุงใหญ่ขนาดนี้ไปที่รถ แต่จากการวิเคราะห์พฤติกรรมการบริโภคโดยบริษัท ดันน์ฮัมบี้ ประเทศไทย พบว่ามีลูกค้ากลุ่มหนึ่งที่ชอบซื้ออาหารสุนัขถุงใหญ่ขนาด 10-20 กิโลกรัมอยู่เป็นจำนวนหนึ่งซึ่งคนกลุ่มนี้ก็จะไม่ซื้ออาหารสุนัขถุงเล็กเช่นกัน จากการทดสอบของเทสโก้ โลตัสโดยการนำสินค้าชนิดนี้ออกในบางสาขาพบว่ายอดขายลดลง [4] นี่คือตัวอย่างของการเข้าใจพฤติกรรมของผู้บริโภคจากการวิเคราะห์ข้อมูลที่มีอยู่

ที่ผ่านมาผมได้แนะนำตัวอย่างการนำเทคนิคดาต้า ไมน์นิงไปประยุกต์ใช้กับข้อมูลในรูปแบบออฟไลน์ (offline) ซึ่งหมายถึงไม่ได้ใช้ในอินเตอร์เน็ตกันไปบ้างแล้ว ในยุคปัจจุบันนี้เราได้เข้าสู่ยุคที่อินเตอร์เน็ตก้าวหน้าไปมากครับ เรามาลองดูตัวอย่างการนำเทคนิคดาต้า ไมน์นิงไปประยุกต์ใช้ในเว็บไซต์ต่างๆ กันดูบ้างครับ เริ่มแรกผมขอแนะนำตัวอย่างที่เห็นได้ชัดเจนที่สุดคือเว็บไซต์ อเมซอน (Amazon.com) ซึ่งเป็นเว็บไซต์ที่ขายหนังสือผ่านทางอินเตอร์เน็ต ผมคิดว่าผู้อ่านหลายๆ ท่านน่าจะเคยเข้าไปดูหรือใช้บริการของเว็บไซต์นี้กันมาบ้างแล้ว เมื่อเราเข้าไปค้นหาหนังสือเราจะพบว่าอเมซอนจะมีระบบแนะนำหนังสือ (recommendation system) ให้โดยเป็นการวิเคราะห์ข้อมูลจากพฤติกรรมการซื้อหนังสือของลูกค้าก่อนหน้านี้ ตัวอย่างของระบบแนะนำหนังสือของเว็บไซต์อเมซอนแสดงในรูปที่ 1-1 ซึ่งผมทำการเลือกค้นหาหนังสือที่เกี่ยวข้องกับการใช้ซอฟต์แวร์ RapidMiner ซึ่งเป็นซอฟต์แวร์ที่ช่วยในการวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิง ระบบก็จะทำการแนะนำหนังสือที่เกี่ยวข้องมาให้

รูปที่ 1-1 ระบบแนะนำสินค้าของเว็บไซต์อเมซอน

นอกจากเว็บไซต์อเมซอนที่เป็นผู้ริเริ่มใช้ระบบแนะนำสินค้าเป็นแห่งแรกแล้วยังมีเว็บไซต์อื่นๆ ที่ใช้หลักการนี้เช่นกัน ตัวอย่างเช่น เว็บไซต์ เน็ตฟลิกซ์ (Netflix) ซึ่งเป็นผู้ให้บริการดูภาพยนต์ผ่านทางอินเตอร์เน็ตก็มีระบบแนะนำภาพยนต์ที่ลูกค้าส่วนใหญ่มักจะดูพร้อมกันไปด้วยดังเช่นรูปที่ 1-2



รูปที่ 1-2 ระบบแนะนำสินค้าของเว็บไซต์เน็ตฟลิกซ์

นอกจากนี้ยังมีตัวอย่างเว็บไซต์หรือระบบอื่นๆ เช่น ระบบซื้อแอพพลิเคชันบนโทรศัพท์มือถือที่มีระบบปฏิบัติการ แอนดรอยด์ (Android) ที่เรียกว่า เพลย์ สโตร์ (Play Store) ก็มีระบบแนะนำแอพพลิเคชันที่มีผู้ใช้มักจะดาวน์โหลดร่วมกันบ่อยๆ อยู่

นอกจากนี้ยังมีการนำเทคนิคดาต้า ไมน์นิงไปใช้ค้นหาข้อมูลที่สำคัญจากข้อมูลในเครือข่ายสังคมออนไลน์ต่างๆ เช่น การวิเคราะห์ทัศนคติของเรื่องต่างๆ ที่ผู้คนกำลังสนใจอยู่ ตัวอย่างเช่น เว็บไซต์ sentiment140.com ได้ทำการดึงข้อความทวีต (tweet) จากเว็บไซต์ทวิตเตอร์ (twitter.com) ที่เกี่ยวข้องกับสินค้าหรือคำศัพท์ (keyword) ที่ผู้ใช้สนใจออกมาแล้วจึงทำการวิเคราะห์ว่าข้อความใดมีทัศนคติเชิงบวกหรือเชิงลบกับสินค้าที่พิจารณาอยู่ ตัวอย่างของเว็บไซต์ sentiment140.com แสดงในรูปที่ 1-3 โดยข้อความที่เป็นสีแดง คือ ทัศนคติเชิงลบ ข้อความที่เป็นสีเขียน คือ ทัศนคติเชิงบวก และข้อความที่เป็นสีขาว คือ ทัศนคติที่เป็นกลาง การวิเคราะห์ข้อมูลลักษณะนี้จะช่วยองค์กรหรือบริษัทได้เข้าใจความคิดเห็นของลูกค้าได้มากขึ้น

รูปที่ 1-3 ผลการวิเคราะห์ข้อความในเว็บไซต์ทวิตเตอร์ซึ่งแสดงทัศนคติเชิงบวก ลบ หรือเป็นกลาง

1.2 ข้อมูลในรูปแบบต่าง ๆ

หลังจากที่เราได้เห็นตัวอย่างการนำเทคนิคดาต้า ไมน์นิงไปใช้งานกันบ้างแล้ว ถัดมาผมขอแนะนำเรื่องเกี่ยวกับข้อมูลและคำศัพท์ที่เกี่ยวข้องในการวิเคราะห์ข้อมูลที่จะได้พบในบทถัดไป โดยปกติแล้วข้อมูลที่สามารถนำมาวิเคราะห์ได้จะต้องเป็นแบบมีโครงสร้าง (structured data) หรืออยู่ในรูปแบบตาราง เช่น ข้อมูลของสมาชิก หรือ ข้อมูลการซื้อขายสินค้าต่างๆ แต่ในปัจจุบันเราก้าวเข้าสู่ยุคของบิ๊ก ดาต้า ซึ่งมีข้อมูลขนาดมหาศาลแต่ข้อมูลเหล่านี้มักจะไม่มีโครงสร้าง (unstructured data) เช่น ข้อความต่างๆ ไม่ว่าจะอยู่ในอีเมล์ (e-mail) หรือเครือข่ายสังคมออนไลน์ (social network) ต่างๆ ถ้าเราต้องการนำข้อมูลที่ไม่มีโครงสร้างนี้มาทำการวิเคราะห์ จำเป็นจะต้องทำการแปลงรูปแบบข้อมูลเหล่านี้ให้อยู่ในรูปแบบที่มีโครงสร้าง หรือ รูปแบบตารางเสียก่อน ในหนังสือเล่มนี้ผมจะเน้นไปที่ข้อมูลที่มีโครงสร้างแต่ขอแนะนำวิธีการแปลงข้อมูลที่ไม่มีโครงสร้างสักเล็กน้อย เรามาลองดูคำศัพท์ต่างๆ ที่เกี่ยวข้องกับข้อมูลทั้งสองแบบกันก่อนครับ

(1) ข้อมูลแบบมีโครงสร้าง (structured data)

ข้อมูลแบบมีโครงสร้างเป็นข้อมูลทั่วๆ ไปที่เรามักจะพบเห็นกันในรูปแบบของตาราง เช่น ข้อมูลที่เก็บรายละเอียดของสมาชิก หรือ การซื้อขายสินค้า โดยปกติแล้วข้อมูลเหล่านี้มักจะเก็บอยู่ในไฟล์ประเภท Excel หรือในฐานข้อมูลต่างๆ ตัวอย่างเช่น ข้อมูลสมาชิกที่แสดงในตารางที่ 1-1 ซึ่งประกอบด้วย 5 แถวและ 5 คอลัมน์ ในหนังสือที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงส่วนใหญ่จะเรียกข้อมูลแต่ละ "**แถว**" ว่า "**ตัวอย่าง** (example)" หรือ "**อินสแตนซ์** (instance)" และข้อมูลแต่ละ "**คอลัมน์**" ว่า "**แอตทริบิวต์** (attribute)" หรือ "**ฟีเจอร์** (feature)" ซึ่งในหนังสือเล่มนี้ผมขอเรียกข้อมูลในแต่ละแถวว่า **ตัวอย่าง** และแต่ละคอลัมน์ว่า **แอตทริบิวต์** ครับ และบรรทัดแรกในตารางที่ 1-1 คือชื่อของแต่ละแอตทริบิวต์ ดังนั้นเราสามารถอ่านได้ว่าลูกค้าที่มีหมายเลขสมาชิกเป็น 2014-00001 ชื่อว่า สมชาย เป็นเพศชาย อายุ 33 ปีและมีร้ายได้ 33,000 บาท เป็นต้น

ตารางที่ 1-1 แสดงข้อมูลรายละเอียดของลูกค้า

หมายเลขสมาชิก	ชื่อ	เพศ	อายุ	รายได้
2014-00001	สมชาย	ชาย	33	33,000
2014-00002	สมหญิง	หญิง	35	35,000
2014-00003	ปิติ	ชาย	23	23,000
2014-00004	มานี	หญิง	22	22,000
2014-00005	ชูใจ	หญิง	23	23,000

(2) **ข้อมูลแบบไม่มีโครงสร้าง** (unstructured data)

ดังที่ได้กล่าวไปแล้วว่าข้อมูลส่วนใหญ่จะเป็นข้อมูลแบบที่ไม่มีโครงสร้าง เช่น ข้อความ หรือ รูปภาพต่างๆ แต่ข้อมูลเหล่านี้ก็มีความสำคัญ เช่น ตัวอย่างของการนำไปประยุกต์เพื่อหาทัศนคติต่างๆ ของลูกค้าที่เกิดขึ้น ในหัวข้อนี้ผมขอยกตัวอย่างการแปลงข้อมูลที่เป็นรูปแบบข้อความให้เป็นข้อมูลในรูปแบบตารางโดยใช้วิธีการแปลงข้อมูล (data transformation) เช่น ข้อความข่าวในตารางที่ 1-2

ตารางที่ 1-2 แสดงข้อความข่าวต่างๆ

เอกสาร	ข้อความในเอกสาร
1	ชาวนาหลายจังหวัดอีสานปลื้มทหาร หลังได้รับเงินจำนำข้าว
2	ชาวนาโคราช แห่รับเงินจำนำข้าวจาก ทหาร-ธ.ก.ส. พร้อมขอบคุณ ค.ส.ช.
3	เกษตรกร เฮ ผู้ว่าฯปราจีนบุรี นำเงินจำนำข้าวคืนชาวนาครบทุกราย

จากข้อความในตารางที่ 1-2 เราสามารถตัดข้อความออกมาเป็นคำต่างๆ ได้ เช่น ชาวนา หลายจังหวัด อีสาน ปลื้ม ทหาร หลัง เป็นต้น หลังจากนั้นคัดเลือกคำที่เป็นคำสำคัญและนำมาเป็นชื่อของแต่ละแอตทริบิวต์ และพิจารณาว่าในแต่ละเอกสารมีคำใดปรากฏขึ้นบ้าง ซึ่งถ้าเอกสารมีคำนั้นปรากฏขึ้นจะมีค่าเป็น Y และไม่มีปรากฏขึ้นจะมีค่าเป็น N ดังแสดงในตารางที่ 1-3 ซึ่งแปลความหมายได้ว่า เอกสารที่ 1 มีคำว่า "ชาวนา", "ปลื้ม", "รับเงิน", "จำนำข้าว", "ทหาร" เกิดขึ้นแต่ไม่มีคำว่า "ขอบคุณ" ปรากฏอยู่เลย

ตารางที่ 1-3 แสดงข้อความที่ได้ทำการแปลงมาเป็นข้อมูลในรูปแบบตาราง

เอกสาร	ชาวนา	ปลื้ม	รับเงิน	จำนำข้าว	ทหาร	ขอบคุณ
1	Y	Y	Y	Y	Y	N
2	Y	N	Y	Y	Y	Y
3	Y	N	N	Y	N	N

จากตัวอย่างนี้เป็นการแปลงข้อมูลในรูปแบบที่ไม่มีโครงสร้าง คือ ข้อความให้เป็นข้อมูลในรูปแบบที่มีโครงสร้าง คือ ตาราง ถัดจากนี้ผมจะแนะนำให้ท่านผู้อ่านรู้จักกับเทคนิคในการวิเคราะห์ข้อมูลด้วยดาต้าไมน์นิงซึ่งมี 2 ประเภทหลักดังจะได้อธิบายในหัวข้อถัดไปครับ

1.3 เทคนิคการวิเคราะห์ข้อมูลด้วยดาต้า ไมน์นิง

เทคนิคในการวิเคราะห์ข้อมูลด้วยดาต้า ไมน์นิงนั้นสามารถได้เป็น 2 ประเภทหลักๆ คือ

- เทคนิคการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)
- เทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning)

(1) เทคนิคการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)

เทคนิคประเภทแรกนี้จะเน้นที่การพิจารณาข้อมูลเป็นหลัก เช่น พิจารณาว่าข้อมูลมีความสัมพันธ์กันในลักษณะใดบ้าง เทคนิคในประเภทนี้จะแบ่งย่อยได้อีก คือ เทคนิคการค้นหากฎความสัมพันธ์ (association rule) และ การแบ่งกลุ่มข้อมูล (clustering) ซึ่งจะได้อธิบายรายละเอียดในบทที่ 2 และ 3 ต่อไปตามลำดับ

(2) เทคนิคการเรียนรู้แบบมีผู้สอน (supervised learning)

เทคนิคในประเภทนี้จะเน้นการเรียนรู้จากข้อมูลที่มีอยู่ในอดีตเพื่อนำมาสร้างโมเดลสำหรับทำนายหรือคาดการณ์สิ่งที่เกิดขึ้นในอนาคต โมเดลในที่นี้อาจจะเป็นสมการทางคณิตศาสตร์ หรือ กฎต่างๆ ก็เป็นได้ เทคนิคการเรียนรู้แบบมีผู้สอนนี้สามารถแบ่งย่อยได้อีก คือ การจำแนกประเภทข้อมูล (classification) และ การประมาณค่าข้อมูล (regression) ซึ่งทั้งสองเทคนิคจะมีลักษณะที่คล้ายกันมากแต่แตกต่างกันที่คำตอบที่ต้องการทำนาย ซึ่งการจำแนกประเภทข้อมูลจะทำนายข้อมูลที่มีค่าเป็น นอมินอล (nominal) เช่น เพศชาย หญิง หรือค่าที่ไม่ใช่ตัวเลขนั่นเอง ส่วนการประมาณค่าข้อมูลจะใช้กับข้อมูลคำตอบที่เป็นตัวเลขเท่านั้น ผมจะอธิบายรายละเอียดของทั้ง 2 วิธีนี้ในบทที่ 4 ครับ

ในหนังสือเล่มนี้ผมเน้นไปที่การแนะนำให้รู้จักเทคนิคในการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงต่างๆ และแสดงตัวอย่างการทำงานเบื้องต้นให้ทราบ แต่ขออนุญาตไม่ลงในรายละเอียดเรื่องทฤษฎีและการนำไปใช้งานในด้านหนึ่งด้านใดโดยเฉพาะ สำหรับผู้อ่านที่สนใจรายละเอียดของเทคนิคต่างๆ ในเชิงลึก ผมขอแนะนำให้อ่านหนังสือ Introduction to Concepts and Techniques in Data Mining and Application to Text Mining ซึ่งเขียนโดย ศ. ดร. ธนารักษ์ ธีระมั่นคง จากสถาบันเทคโนโลยีนานาชาติ สิรินธร มหาวิทยาลัยธรรมศาสตร์

บทที่ 4 การจำแนกประเภทข้อมูล (Classification)

- การจำแนกประเภทข้อมูลและการประยุกต์ใช้งาน
- ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล
- การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล
- เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Decision Tree
- เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Naive Bayes
- เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี K-Nearest Neighbors (K-NN)
- เทคนิคการจำแนกประเภทข้อมูลด้วยวิธี Neural Network

4.1 การจำแนกประเภทข้อมูลและการประยุกต์ใช้งาน

ในสองบทที่ผ่านมาผมได้แนะนำให้รู้จักเทคนิคทางดาต้า ไมน์นิงที่เป็นแบบไม่มีผู้สอน (unsupervised learning) ซึ่งได้แก่ การหากฎความสัมพันธ์ในบทที่ 2 และการแบ่งกลุ่มข้อมูลในบทที่ 3 บทนี้เราจะมาทำความรู้จักกับเทคนิคแบบที่มีผู้สอน (supervised learning) กันครับนั่นคือ เทคนิคการจำแนกประเภทข้อมูล (classification) ในชีวิตเราได้พบกับการประยุกต์ใช้งานเทคนิคการจำแนกประเภทข้อมูลนี้กันอยู่มากมายครับ แต่บางครั้งเราอาจจะไม่ทันรู้ตัว เช่น การทำนายสภาพอากาศในวันถัดไปว่าฝนจะตกมากน้อยแค่ไหน หรือมีอุณหภูมิเท่าไร หรือ การจำแนกอีเมล์ (e-mail) ออกเป็นประเภทสแปม (spam) หรือแบบปกติ (normal) นอกจากนี้ในแอพพลิเคชันหลายตัว เช่น Facebook ก็สามารถค้นหาใบหน้าของคนที่ปรากฏในรูปภาพได้ว่ามีใครบ้าง สิ่งเหล่านี้เกิดจากนำเทคนิคการจำแนกประเภทข้อมูลไปประยุกต์ใช้ครับ

เทคนิคการจำแนกประเภทข้อมูลนี้จะนำข้อมูลที่มีในอดีตมาสอนระบบเพื่อให้เรียนรู้รูปแบบที่เกิดขึ้นในข้อมูลแล้วจึงสร้างเป็นสมการหรือโมเดล (model) ขึ้นมาเพื่อหาคำตอบให้สำหรับข้อมูลใหม่ เช่น การจำแนกอีเมล์ออกเป็นสแปมหรือแบบปกติต้องมีการนำข้อมูลของอีเมล์ประเภทสแปมและประเภทปกติมาให้คอมพิวเตอร์ทำการเรียนรู้เสียก่อน หลังจากนั้นจึงสร้างโมเดลการจำแนกประเภทของอีเมล์และใช้จำแนกอีเมล์ที่เข้ามาใหม่ว่าเป็นแบบสแปมหรือแบบปกติ ตัวอย่างอื่น เช่น การคาดการณ์ว่าพรุ่งนี้จะมีอุณหภูมิเท่าไร ก็ต้องอาศัยข้อมูลวันก่อนหน้าที่ได้ทำการเก็บมา เป็นต้น จากสองตัวอย่างที่ได้ยกมาจะเห็นว่าคำตอบที่สนใจนั้นต่างกัน คือ ในตัวอย่างแรกคำตอบที่สนใจคือประเภทของอีเมล์ ซึ่งมีผลที่เป็นไปได้ 2 แบบ คือ สแปม และ ปกติ ซึ่งคำตอบที่เป็นประเภทของค่าต่างๆ เหล่านี้ในการวิเคราะห์ข้อมูลจะเรียกว่า “**คลาส**” (class) หรือ “**ลาเบล**” (label) และเทคนิคที่หาคำตอบเหล่านี้เรียกว่า **เทคนิคการจำแนกประเภทข้อมูล** หรือ classification ส่วนตัวอย่างที่ 2 สนใจที่จะหาคำตอบที่แตกต่างจากไปจากในตัวอย่างแรกเพราะเน้นหาคำตอบที่เป็นเชิงปริมาณ หรือ จำนวนตัวเลข เป็นหลัก ดังนั้นเทคนิคในลักษณะนี้จึงเรียกว่า **เทคนิคการ**

ประมาณค่า หรือ regression ต่อจากนี้ไปผมจะแนะนำให้รู้จักกับขั้นตอนในการจำแนกประเภทข้อมูลซึ่งโดยปกติแล้วจะแบ่งเป็น 3 ขั้นตอนใหญ่ๆ คือ

ตารางที่ 4-1 แสดงข้อมูลสภาพอากาศย้อนหลัง 14 วัน

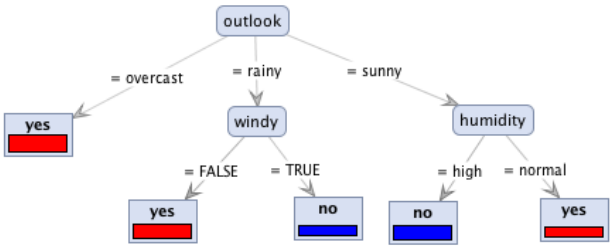
No.	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	mild	normal	false	Yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

1. ขั้นตอนการสร้างโมเดล

ขั้นตอนนี้เป็นขั้นตอนแรกในการจำแนกประเภทข้อมูล โดยการนำข้อมูลเทรนนิง ดาต้า (training data) หรือข้อมูลที่ใช้ในการเรียนรู้ มาสร้างเป็นโมเดลขึ้นมาด้วยเทคนิคการจำแนกประเภทข้อมูลแบบต่างๆ เช่น วิธี Decision Tree (หัวข้อที่ 4.4) วิธี Naive Bayes (หัวข้อที่ 4.5) วิธี K-Nearest Neighbors (หัวข้อที่ 4.6) และ วิธี Neural Network (หัวข้อที่ 4.7)

ในการจำแนกประเภทข้อมูลนั้นเราจำเป็นอย่างมากที่จะต้องมีข้อมูลเทรนนิง ดาต้า เพื่อให้คอมพิวเตอร์ได้เรียนรู้จากตัวอย่างของข้อมูล โดยข้อมูลเทรนนิง ดาต้านี้ คือ ข้อมูลที่มีในอดีตโดยจะประกอบด้วย 2 ส่วนคือ แอตทริบิวต์ทั่วไปและแอตทริบิวต์ที่เป็นคลาสคำตอบที่สนใจ ตัวอย่างเช่น ข้อมูลในตารางที่ 4-1 ซึ่งเป็นข้อมูลที่เก็บสภาพอากาศย้อนหลัง 14 วันเพื่อดูว่าในแต่ละวันจะมีการจัดการแข่งขันเบสบอลขึ้นหรือไม่ [9] ซึ่งแอตทริบิวต์ **Outlook**, **Temperature**, **Humidity** และ **Windy** จะเป็นแอตทริบิวต์ประเภททั่วไปที่ใช้ในการพิจารณาว่าถ้าค่าในแอตทริบิวต์เหล่านี้เป็นลักษณะไหนแล้วจึงมีการจัดแข่งเบสบอล ส่วนแอตทริบิวต์ที่เป็นคลาสคำตอบที่เราสนใจคือ แอตทริบิวต์ **Play** สำหรับข้อมูลของท่านผู้อ่านเองก็จะต้องพิจารณาว่าแอตทริบิวต์ไหนที่จะเป็นคลาสคำตอบ นอกจากนี้สิ่งที่ควรคำนึงเป็นอย่างมากก็คือข้อมูลแอตทริบิวต์ทั่วไปจะต้องมีความสัมพันธ์กับแอตทริบิวต์ที่เป็นคลาสคำตอบจึงจะทำให้โมเดลที่เราสร้างได้มีประสิทธิภาพและน่าเชื่อถือมาก

หลังจากนั้นเทคนิคในการจำแนกประเภทข้อมูลต่างๆ จะเรียนรู้จากข้อมูลเทรนนิง ดาต้าและสร้างเป็นโมเดลในรูปแบบต่างๆ ออกมา เช่น โมเดล Decision Tree ในรูปที่ 4-1



รูปที่ 4-1 โมเดล Decision Tree ที่สร้างได้จากข้อมูลในตารางที่ 4-1
(สร้างจากซอฟต์แวร์ RapidMiner Studio 6)

จากโมเดล Decision Tree ที่สร้างได้สามารถสรุปเป็นกฎได้ เช่น

1. IF **Outlook** = overcast THEN **play** = yes
2. IF **Outlook** = sunny AND **Humidity** = high THEN **play** = no

จากกฎข้อที่ 1 แปลความหมายได้ว่าถ้าสภาพอากาศเป็นแบบ overcast จะมีการจัดแข่งขันเบสบอล แต่จากกฎข้อที่ 2 แปลความหมายได้ว่าถ้าสภาพอากาศเป็นแบบ sunny และความชื้นมากแล้วจะไม่สามารถจัดแข่งเบสบอลได้

2. **ขั้นตอนการทดสอบประสิทธิภาพของโมเดล**

หลังจากที่สร้างโมเดลขึ้นมาได้แล้ว ขั้นตอนถัดมาจะต้องทำการวัดประสิทธิภาพของโมเดลที่สร้างได้ ซึ่งผมขอแยกส่วนนี้มาอธิบายในรายละเอียดในหัวข้อ**ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล** (หัวข้อที่ 4.2) และ **การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล** (หัวข้อที่ 4.3)

3. **ขั้นตอนนำโมเดลไปใช้งานเพื่อทำนายข้อมูลใหม่**

หลังจากที่วัดประสิทธิภาพของโมเดลได้ผลเป็นที่น่าเชื่อถือหรือพอใจแล้ว เราสามารถนำโมเดลที่สร้างได้นี้ไปใช้ในการทำนายข้อมูลที่เข้ามาใหม่ เช่น ถ้าพบว่าสภาพอากาศในวันปัจจุบันมีค่าดังตารางที่ 4-2 จะทำนายได้ว่าจะมีการจัดแข่งเบสบอลขึ้น เนื่องจากแอตทริบิวต์ Outlook มีค่าเป็น sunny และ Humidity เป็น normal แล้ว Play จะเป็น yes (พิจารณาจากโมเดลในรูปที่ 4-1)

ตารางที่ 4-2 แสดงข้อมูลสภาพอากาศในวันปัจจุบันซึ่งยังไม่ทราบคลาสคำตอบ

No.	Outlook	Temperature	Humidity	Windy	Play
15	sunny	hot	normal	false	?

ในหัวข้อถัดไปผมจะแนะนำให้รู้จักกับตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูลต่างๆ ซึ่งประกอบด้วย ตาราง confusion matrix, ค่า precision, ค่า recall, ค่า f-measure และ ค่า accuracy เป็นต้น

4.2 ตัววัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

ดังที่ได้กล่าวไปแล้วว่าการนำโมเดลไปใช้งานจริงได้นั้นเราจำเป็นจะต้องทราบประสิทธิภาพของโมเดลเสียก่อน โดยทั่วไปแล้วจะมีตัววัดที่นิยมใช้กันในงานวิจัยและการทำงานต่างๆ อยู่ 5 ค่า คือ

- **Precision** เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส
- **Recall** เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส
- **F-measure** เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

- Accuracy เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส

ก่อนที่จะไปรู้จักกับตัววัดประสิทธิภาพของโมเดลตัวต่างๆ ผมขอแนะนำให้รู้จักกับตาราง confusion matrix ก่อนครับ confusion matrix คือ ตารางแบบจัตุรัสโดยมีจำนวนแถวเท่ากับจำนวนคอลัมน์และเท่ากับจำนวนคลาส เช่น ในตารางที่ 4-1 มีคลาสคำตอบอยู่ 2 ค่า คือ yes และ no ฉะนั้นตาราง confusion matrix นี้จะสร้างได้เป็นตารางขนาด 2x2 ดังในตารางที่ 4-3 โดยข้อมูลด้านคอลัมน์คือ คลาสที่อยู่ในข้อมูลเทรนนิงดาต้า (actual) และข้อมูลในแนวแถว คือ คลาสที่โมเดลทำนายมาได้ (predicted)

ตารางที่ 4-3 แสดงตาราง confusion matrix ของข้อมูล weather ซึ่งมี 2 คลาส

predicted / actual	yes	no
yes	TP	FP
no	FN	TN

จากในตารางที่ 4-3 ค่าที่แสดงในช่องต่างๆ ของตารางประกอบด้วย

- True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่
- True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่
- False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่
- False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

ตารางที่ 4-4 แสดงข้อมูลแอตทริบิวต์ Play จากเทรนนิง ดาต้า 10 ตัวแรกและค่าที่ทำนายได้

No.	Actual	Predicted
1	*no*	*no*
2	*no*	*no*
3	yes	no
4	**yes**	**yes**
5	yes	no
6	no	yes

ตารางที่ 4-4 (ต่อ) แสดงข้อมูลแอตทริบิวต์ Play จากเทรนนิง ดาต้า 10 ตัวแรกและค่าที่ทำนายได้

No.	Actual	Predicted
7	yes	yes
8	*no*	*no*
9	yes	no
10	yes	yes

เพื่อความเข้าใจที่มากขึ้นผมขอยกตัวอย่างข้อมูลที่อยู่ในเทรนนิง ดาต้าและที่ทำนายออกมาได้มาแสดงให้ดูในตารางที่ 4-4 โดยที่กำลังพิจารณาคลาส Play = yes ดังนั้นจะสรุปได้ว่า

- True Positive (TP) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = yes
  - มีจำนวน 3 ตัว (แถวที่เป็น**ตัวหนา** คือ แถวที่ 4, 7 และ 10)
- True Negative (TN) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Play = no
  - มีจำนวน 3 ตัว (แถวที่เป็น*ตัวเอียง* คือ แถวที่ 1, 2 และ 8)
- False Positive (FP) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = yes
  - มีจำนวน 1 ตัว (แถวที่ขีดเส้นใต้ คือ แถวที่ 6)
- False Negative (FN) คือ จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาส Play = no
  - มีจำนวน 3 ตัว (แถวที่ตัวอักษรปกติ คือ แถวที่ 3, 5 และ 9)

ดังนั้นจึงสร้างตาราง confusion matrix ได้ดังตารางที่ 4-5

ตารางที่ 4-5 แสดงตาราง confusion matrix ของข้อมูล weather ซึ่งมี 2 คลาส

predicted / actual	yes	no
yes	3	1
no	3	3

หลังจากที่เราสร้างตาราง confusion matrix ได้ดังตารางที่ 4-5 แล้วเรามาดูวิธีคำนวณค่า Precision, Recall, F-measure และ Accuracy กันต่อเลยดีกว่าครับ

- **Precision** เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาส

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

ดังนั้น Precision ของคลาส yes คือ

Precision (**Play**=yes) = 3/4 = 75%

- **Recall** เป็นการวัดความถูกต้องของโมเดล โดยพิจารณาแยกทีละคลาส

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

ดังนั้น Recall ของคลาส yes คือ

Recall (**Play**=yes) = 3/6 = 50%

- **F-measure** เป็นการวัดค่า Precision และ Recall พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

F-measure (**Play**=yes) = 2 x 75% x 50% / (75% + 50%) = 60%

- **Accuracy** เป็นการวัดความถูกต้องของโมเดล โดยพิจารณารวมทุกคลาส คือ จำนวน True Positive ของทุกคลาสรวมกันได้เท่ากับ 6/10 = 60%

4.3 การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล

ในหัวข้อที่ผ่านมาผมแนะนำให้รู้จักกับตัววัดประสิทธิภาพของโมเดลไปแล้ว การวัดประสิทธิภาพแบบนี้ได้จำเป็นต้องแบ่งข้อมูลออกเป็น 2 ส่วน โดยที่ส่วนที่ 1 ใช้เพื่อสร้างโมเดลและส่วนที่ 2 ให้โมเดลทำนายค่าคลาสคำตอบออกมา การแบ่งข้อมูลเพื่อทำการทดสอบนี้มี 3 วิธีการใหญ่ ดังนี้

(1) วิธี Self Consistency Test

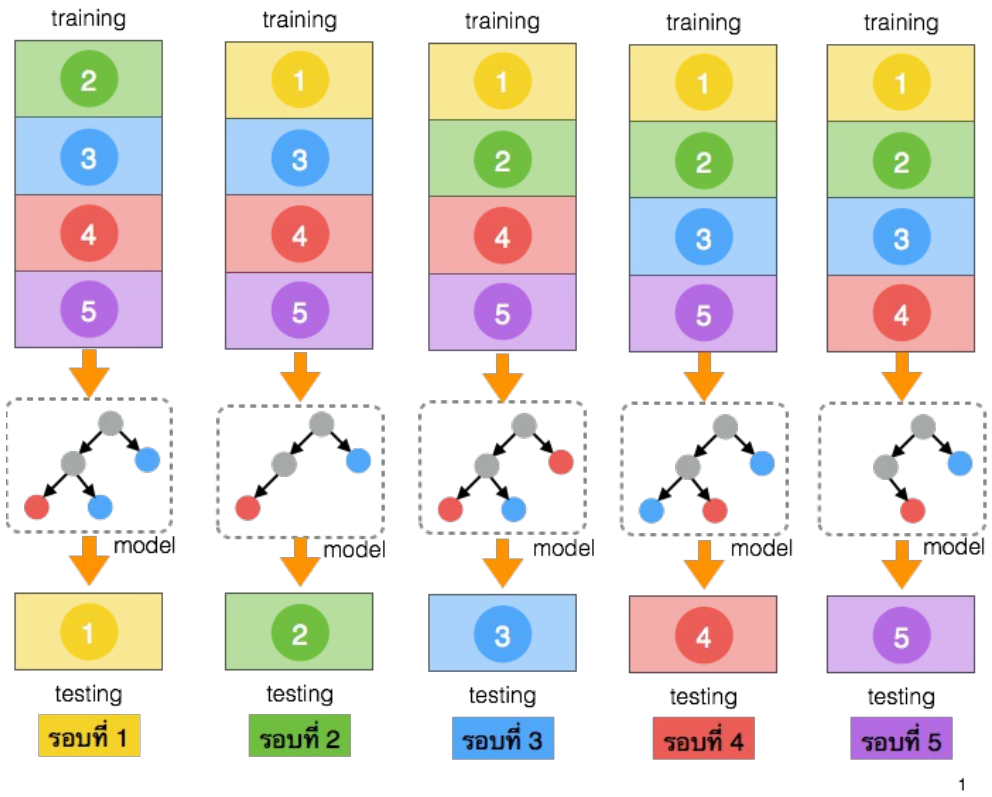
วิธี Self Consistency Test หรือบางครั้งเรียกว่า Use Training Set นี้เป็นวิธีการที่ง่ายที่สุด นั่นคือ ข้อมูลที่ใช้ในการสร้างโมเดลและข้อมูลที่ใช้ในการทดสอบโมเดลเป็นข้อมูลชุดเดียวกัน กระบวนการนี้เริ่มจากสร้างโมเดลด้วยข้อมูลเทรนนิง ดาต้า หลังจากนั้นนำโมเดลที่สร้างได้มาทำนายข้อมูลเทรนนิง ดาต้า ชุดเดิม ตัวอย่างเช่น นำข้อมูลเทรนนิง ดาต้า ในตารางที่ 4-1 มาสร้างโมเดลและทดสอบโมเดลเป็นต้น การวัดประสิทธิภาพด้วยวิธีนี้จะให้ผลการวัดประสิทธิภาพที่มีค่าสูงมาก (อาจจะเข้าใกล้ 100%) เนื่องจากเป็นข้อมูลชุดเดิมที่ระบบได้ทำการเรียนรู้มาแล้ว แต่ผลการวัดที่ได้ไม่เหมาะที่จะนำไปรายงานในงานวิจัยต่างๆ ซึ่งวิธีการนี้เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพเพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น ถ้าได้ผลการวัดที่น้อยแสดงว่าโมเดลไม่เหมาะสมกับข้อมูล จึงไม่ควรจะนำไปทดสอบด้วยวิธีการแบ่งข้อมูลแบบต่างๆ

(2) วิธี Split Test

วิธี Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70% ต่อ 30% หรือ 80% ต่อ 20% โดยข้อมูลส่วนที่หนึ่ง (70% หรือ 80%) ใช้ในการสร้างโมเดลและข้อมูลส่วนที่สอง (30% หรือ 20%) ใช้ในการทดสอบประสิทธิภาพของโมเดล ตัวอย่างเช่น แบ่งข้อมูลเทรนนิง ดาต้า ในตารางที่ 4-1 แบ่งข้อมูล 10 ตัวอย่าง (14 x 0.7 = 9.8) ในการสร้างโมเดลและข้อมูล 4 ตัวอย่าง (14 x 0.3 = 4.2) ใช้ในการทดสอบประสิทธิภาพของโมเดล เป็นต้น แต่การทดสอบแบบ Split Test นี้ทำการสุ่มข้อมูลเพียงครั้งเดียวซึ่งในบางครั้งถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้สร้างโมเดลทำให้ผลการวัดประสิทธิภาพได้ออกมาดี ในทางตรงข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สร้างโมเดลมากทำให้ผลการวัดประสิทธิภาพได้ออกมาแย่ ดังนั้นจึงควรใช้วิธี Split Test นี้หรือทำการสุ่มหลายๆ ครั้ง แต่ข้อดีของวิธีการนี้คือใช้เวลาในการสร้างโมเดลน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่มาก

(3) วิธี Cross-validation Test

วิธีนี้เป็นวิธีที่นิยมใช้ในการทดสอบประสิทธิภาพของโมเดลเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน (มักจะแสดงด้วยค่า k) เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล ทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ เช่น การทดสอบด้วยวิธี 5-fold cross-validation ในรูปที่ 4-2



รูปที่ 4-2 ตัวอย่างการแบ่งข้อมูลแบบ 5-fold cross-validation

จากรูปที่ 4-2 แบ่งข้อมูลเทรนนิง ดาต้าออกเป็น 5 ส่วนที่มีจำนวนเท่ากัน หลังจากนั้นทำการทดสอบประสิทธิภาพของโมเดล 5 ครั้ง ดังนี้

- **รอบที่** 1 ใช้ข้อมูลส่วนที่ 2,3,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 1
- **รอบที่** 2 ใช้ข้อมูลส่วนที่ 1,3,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 2
- **รอบที่** 3 ใช้ข้อมูลส่วนที่ 1,2,4 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 3
- **รอบที่** 4 ใช้ข้อมูลส่วนที่ 1,2,3 และ 5 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 4
- **รอบที่** 5 ใช้ข้อมูลส่วนที่ 1,2,3 และ 4 สร้างโมเดล และใช้โมเดลทำนายข้อมูลส่วนที่ 5

จะเห็นได้ว่าข้อมูลทุกชุด (ข้อมูลทุกตัวอย่าง) จะได้เป็นตัวทดสอบประสิทธิภาพของโมเดล โดยในการทดสอบประสิทธิภาพของโมเดลแต่ละรอบจะได้จำนวน TP, TN, FP, FN ใส่ลงไปในตาราง confusion matrix และบวกเพิ่มเข้าไป สุดท้ายจะได้ตาราง confusion matrix ที่เป็นค่ารวมทั้งหมดหลังจากนั้นจึงทำการ

ประวัติผู้เขียน

ชื่อ เอกสิทธิ์ พัชรวงศ์ศักดา (EAKASIT PACHARAWONGSAKDA)

การศึกษา

- ปริญญาเอก วิทยาการคอมพิวเตอร์ สถาบันเทคโนโลยีนานาชาติสิรินธร (SIIT) มหาวิทยาลัยธรรมศาสตร์ (ทุนโครงการปริญญาเอกกาญจนาภิเษก)
- ปริญญาโทวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ (บางเขน)
- ปริญญาตรีวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์ (เกียรตินิยมอันดับ 2) มหาวิทยาลัยเกษตรศาสตร์ (บางเขน)

ประสบการณ์

- Data Scientist ที่บริษัท GloriSys Cloud Solutions Co., Ltd.
- ผู้ร่วมก่อตั้ง ห้างหุ้นส่วนสามัญ ดาต้า คิวบ์ (data cube)
- ได้รับทุน visiting PhD Student ที่มหาวิทยาลัยยอร์ค (York) เมืองโตรอนโต ประเทศแคนาดา
- ได้รับ statement of completion หลักสูตร Data Mining with Weka จากมหาวิทยาลัย Waikato
- อดีตที่ปรึกษาด้านการวิเคราะห์ข้อมูลให้กับสมาคมส่งเสริมเทคโนโลยี (ไทย-ญี่ปุ่น)
- ผู้ก่อตั้งหลักสูตร An Introduction to Data Mining (Workshop with WEKA)
- วิทยากรประจำหลักสูตร Practical Data Mining with RapidMiner Studio 6
- วิทยากรประจำหลักสูตร Web Application Development using Weka and PHP
- วิทยากรประจำหลักสูตร Basic Data Mining with WEKA
- วิทยากรรับเชิญอบรม Practical Data Mining with RapidMiner Studio 6 ที่มหาวิทยาลัยเทคโนโลยีราชมงคงสุวรรณภูมิ
- วิทยากรรับเชิญอบรม Practical Data Mining with RapidMiner Studio 6 ที่มหาวิทยาลัยราชภัฏสวนสุนันทา
- วิทยากรรับเชิญอบรม Practical Data Mining with RapidMiner Studio 6 ที่มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ
- วิทยากรรับเชิญอบรม Basic Data Mining for Marketing using RapidMiner Studio 6 ที่ธนาคารเพื่อการเกษตรและสหกรณ์ (ธ.ก.ส.)

ประสบการณ์ (ต่อ)

- วิทยากรรับเชิญอบรม Data Mining with WEKA ที่ บริษัท ระยองวิศวกรรมและซ่อมบำรุง จำกัด (REPCO)
- วิทยากรรับเชิญอบรม Data Mining with WEKA ที่คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา

- วิทยากรรับเชิญอบรมหลักสูตร An Introduction to Data Mining ที่ บริษัท โทเทิ่ล แอ็คเซ็ส คอมมูนิเคชั่น จำกัด (มหาชน) – DTAC
- วิทยากรรับเชิญอบรม Data Mining with WEKA ที่ บริษัท โมโนเทคโนโลยี จำกัด
- วิทยากรรับเชิญอบรม Data Mining with WEKA ที่คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ
- วิทยากรรับเชิญหลักสูตร ขุดเหมืองข้อมูลด้วย Weka รุ่นที่ 1 – 7 สถาบัน Open Source Education Center
- วิทยากรหัวข้อ “Training Data Mining with Weka” จัดโดยมูลนิธิศักดิ์พรทรัพย์
- ผู้เขียนบทความ “ขุดเหมืองข้อมูลด้วย Weka” นิตยสาร OpenSource2Day
- ประสบการณ์พัฒนาระบบงาน Help Desk ด้วยเทคนิค data mining
- ประสบการณ์พัฒนาระบบพยากรณ์น้ำฝนด้วยเทคนิค data mining ฯลฯ
- อดีตผู้ร่วมก่อตั้งและวิทยากรของ หสม. โอเพน ไมเนอร์
- อดีตผู้ช่วยนักวิจัย ห้องปฏิบัติการ Information Systems ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) งานวิจัยด้าน Data Mining, Bioinformatics และ Chemoinformatics
- อดีตที่ปรึกษาโครงงานปริญญาตรี ห้องปฏิบัติการ Data Analysis and Knowledge Discovery Laboratory (DAKDL) คณะวิศวกรรมศาสตร์ (คอมพิวเตอร์) มหาวิทยาลัยเกษตรศาสตร์

ผลงานวิจัยที่มีส่วนร่วม

- **วารสารวิชาการระดับนานาชาติ** (International Journals)

1. Eakasit Pacharawongsakda and Thanaruk Theeramunkong (2013), “**Predict Subcellular Locations of Singleplex and Multiplex Proteins by Semi-Supervised Learning and Dimension-Reducing General Mode of Chou’s PseAAC**”, NanoBioscience, IEEE Transactions on, vol.12, no.4, pp.311-320, Dec. 2013 (Impact Factor (2013): 1.768)
2. Eakasit Pacharawongsakda and Thanaruk Theeramunkong (2013), “**Multi-Label Classification Using Dependent and Independent Dual Space Reduction**”, The Computer Journal, Oxford University Press, vol.56, no.9, pp.1113-1135, Feb. 2013 (Impact Factor (2013): 0.888)
3. Eakasit Pacharawongsakda, Sunai Yokwai and Supawadee Ingsriswang (2009), “**Potential natural product discovery from microbes through a diversity-guided computational framework**”, Applied Microbiology and Biotechnology, 82: 579 (Impact Factor (2012): 3.689)

4. Wasna Viratyosin, Supawadee Ingsriswang, Eakasit Pacharawongsakda and Prasit Palittapongarnpim (2008), "**Genome-wide subcellular localization of putative outer membrane and extracellular proteins in Leptospira interrogans serovar Lai genome using bioinformatics approaches**", BMC Genomics, 9 (Impact Factor (2012): 4.400)
5. Supawadee Ingsriswang and Eakasit Pacharawongsakda (2007), "**sMOL Explorer: an open source, web-enabled database and exploration tool for Small MOLecules datasets**", Bioinformatics, 18: pp. 2498-2500 (Impact Factor (2012): 5.323)

- **การประชุมวิชาการระดับนานาชาติ** (International Conferences)

1. Eakasit Pacharawongsakda and Thanaruk Theeramunkong (2013), "**A Two-Stage Dual Space Reduction Framework for Multi-label Classification**", Proceedings of the the third Quality issues, measures of interestingness and evaluation of data mining models workshop, Gold Coast, Australia, April 14-17, 2013
2. Samatcha Thanangthanakij, Eakasit Pacharawongsakda, Nattapong Tongtep, Pakinee Aimmanee and Thanaruk Theeramunkong (2012), "**An Empirical Study on Multi-dimensional Sentiment Analysis from User Service Reviews**". Proceedings of the seventh International Conference on Knowledge, Information and Creativity Support Systems, Melbourne, Australia, November 8 – 10, 2012 (Cited by 2)
3. Eakasit Pacharawongsakda, Cholwich Nattee and Thanaruk Theeramunkong (2012), "**Improving Multi-label Classification Using Semi-supervised Learning and Dimensionality Reduction**", Proceedings of the twelfth Pacific Rim International Conference on Artificial Intelligence (PRICAI), Kuching Sarawak, Malaysia, September 3-7, 2012.
4. Eakasit Pacharawongsakda and Thanaruk Theeramunkong (2012), "**Towards More Ecient Multi-Label Classification using Dependent and Independent Dual Space Reduction**", Proceedings of the sixteenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Kuala Lumpur, Malaysia, May 29- June 1, 2012. (Cited by 4)
5. Eakasit Pacharawongsakda and Thanaruk Theeramunkong (2011), "**Improving Classifier Chains for Multi-label Classification using Dual Space Reduction**", Proceedings of The 6th International Conference on Knowledge, Information and Creativity Support Systems, Beijing, China, October, 22-24, 2011. (Best Paper Award)

6. Sunai Yokwai, Boonyarat Phadermrod, Eakasit Pacharawongsakda and Supawadee Ingsriswang (2008), "**Using Molecular Systematics and GIS-based Modeling Approaches for Selection of Potential Sites to Explore the Desirable Microbial Products**", Proceedings of GeoInformatics 2008, Guangzhou, China, June 28-29 2008
7. Eakasit Pacharawongsakda, Sunai Yokwai, Nitsara Karoonuthaisiri, Duangdao Wichadakul and Supawadee Ingsriswang (2008), "**ESTplus: An Integrative System for Comprehensive and Customized EST Analysis and Proteomic Data Matching**", Proceedings of The 2nd International Conference on Bioinformatics and Biomedical Engineering (iCBBE2008), Shanghai, China. May 16-18, 2008 (Cited by 3)
8. Duangdao Wichadakul, Supawadee Ingsriswang, Eakasit Pacharawongsakda, Boonyarat Phadermrod and Sunai Yokwai (2008), "**ATGC-Dom: Alignment, Tree, and Graph for Comparative proteomes by DOMain architecture**", Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), Singapore, Mar 30-Apr 2 2008 (Co-Winner of Poster Special Commendation Award)
9. Wasna Viratyosin, Supawadee Ingsriswang, Eakasit Pacharawongsakda and Prasit Palittapongarnpim (2005), "**Computational Framework Analysis for Secreted Proteins in Leptospira interrogans Genome**", Poster Presentation at of the Third Asia-Pacific Bioinformatics Conference (APBC2005), Singapore, 17-21 January 2005
10. Supawadee Ingsriswang, Wasna Viratyosin, Eakasit Pacharawongsakda and Prasit Palittapongarnpim (2005), "**Post-Prediction through Hierarchical Partitioning and Discriminant Analysis for Decision Support of Putative Secreted Protein Identification**", Poster Presentation at the Third Asia-Pacific Bioinformatics Conference (APBC2005), Singapore, 17-21 January 2005
11. Eakasit Pacharawongsakda, Supawadee Ingsriswang, Anuttara Nathalang and Warren Brockelman (2004), "**Mo-Singto: A Mobile Integrated Data Logging and Mapping System for Study of Forest Regeneration**", Proceedings of the IUFRO4.11 Conference on Applications of Statistics, Information Systems and Computers in Natural Resources Monitoring and Management, Taiwan, June 7-11, 2004.

- **การประชุมวิชาการระดับชาติ** (National Conferences)

1. Boonyarat Phadermrod, Sunai Yokwai, Eakasit Pacharawongsakda and Supawadee Ingsriswang (2008), "**An Opensource GIS-based Application to Improve the Molecular Systematic Study of Microorganisms in National Parks of Thailand**", Proceedings of the 12th Annual Symposium on Computational Science and Engineering (ANSCSE 12 International Symposium), Ubon Rajathanee University, Thailand, Mar 27-29 2008
2. Eakasit Pacharawongsakda, Thanawin Rakthanmanon, Thanapat Kangkachit and Kritsana Waiyamai (2007), "**Prunning Strategies for Improving Sequential Pattern Mining of Protein Sequence Dataset**", Proceedings of the 11th National Computational Science and Engineering Conference (NCSEC 2007), Bangkok, Thailand.
3. Eakasit Pacharawongsakda and Supawadee Ingsriswang (2007), "**Discovering Discriminative Dipeptides in Subcellular Localization of Gram-Negative Bacterial Proteins**", Proceedings of the 4th International Joint Conference on Computer Science and Software Engineering, (JCSSE2007), Khonkaen, Thailand, 2-4 May 2007, pp. 343-348
4. Boonyarat Phadermrod ,Sunai Yokwai, Eakasit Pacharawongsakda, Supawadee Ingsriswang, (2007) "**An Open-source Three-Tier Architecture for Mapping Microbial Diversity**", Proceedings of NSTDA Annual Conference, (NAC 2007), Pathumthani, Thailand, 28-30 March 2007
5. Eakasit Pacharawongsakda and Supawadee Ingsriswang (2006), "**NAT : An Open Source Platform for Searchable Structure and Property Database of Natural Bioactive Compounds**", Proceedings of the 10th National Computational Science and Engineering Conference (NCSEC 2006), Khonkaen, Thailand, 25-27 October 2006, pp.553-560
6. ! Siriwon Taewijit, Eakasit Pacharawongsakda and Supawadee Ingsriswang (2006), "**Improved Subcellular Localization Prediction of Gram-Positive Bacterial Proteins Using Feature Selection and DDAG-SVM**", Proceedings of the 10th National Computer Science and Engineering Conference (NCSEC 2006), 25-27 October 2006, Khonkaen, Thailand, pp.535-543
7. Eakasit Pacharawongsakda and Supawadee Ingsriswang (2006), "**AAA: Automatic AIDS Antiviral Screening System Using SVM with Maximal Frequent Molecular Fragments**", Proceeding of the 10th Annual National Symposium on Computational Science and Engineering (ANSCSE10), Chiangmai, Thailand, March 22-24, pp. 171-

175, ISBN: 974-656-924-4

8. Eakasit Pacharawongsakda and Supawadee Ingsriswang (2005), "**Subcellular Localization Prediction of Gram-negative Bacterial Proteins Using Decision Directed Acyclic Graph Support Vector Machines**", Proceeding of the 9th National Computer Science and Engineering Conference, Bangkok (NCSEC 2005), Thailand, 27-28 October 2005, pp.161-170, ISBN-974-677-541-3

หลักสูตรการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining โดยซอฟต์แวร์ RapidMiner Studio 6 (ขั้นพื้นฐานและปานกลาง)

ภาพรวมของหลักสูตร

โลกในยุคปัจจุบันได้ก้าวเข้าไปสู่ยุคที่เรียกว่า “Big Data” หรือ “ข้อมูลอภิมหาศาล” เนื่องจากในแต่ละวันมีข้อมูลเกิดขึ้นมากมาย อาทิเช่น ข้อมูลสมาชิกของ Facebook ข้อมูลการซื้อสินค้าจากในซุปเปอร์มาร์เกตต่างๆ และเพื่อให้เกิดประโยชน์มากที่สุดเราจำเป็นต้องนำข้อมูลอภิมหาศาลเหล่านี้มาทำการวิเคราะห์ (analyze) ซึ่งเทคนิคหนึ่งที่ได้รับการนิยมอย่างสูงในปัจจุบัน คือ เทคนิค Data Mining ซึ่งเป็นเทคนิคที่ค้นหาความสัมพันธ์ในข้อมูล เช่น ถ้าลูกค้าซื้อเบียร์แล้วลูกค้าจะซื้อผ้าอ้อมร่วมไปด้วย หรือถ้าเรากด Like หน้า Facebook page เราจะเห็นว่า Facebook มีระบบแนะนำpage อื่นๆ ที่เกี่ยวข้องมาให้ด้วย หรือ การสร้างโมเดลเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต เช่น ทำนายยอดขายในไตรมาสถัดไป หรือ การทำนายว่าพนักงานคนไหนที่จะลาออกจากบริษัทในช่วง 3 เดือนข้างหน้า ตัวอย่างเหล่านี้ล้วนเป็นผลมาจากการวิเคราะห์ข้อมูลทางด้าน Data Mining

การวิเคราะห์ข้อมูลด้วย Data Mining นี้กำลังเป็นที่นิยมไปทั่วโลกด้วยแรงขับเคลื่อนอย่างหนึ่งคือ การมีซอฟต์แวร์ที่ช่วยให้ทำการวิเคราะห์ได้ง่ายขึ้น แต่ซอฟต์แวร์ส่วนใหญ่จะเป็นซอฟต์แวร์เชิงพาณิชย์ (commercial software) เช่น SAS Enterprise Miner หรือ IBM Intelligent Miner ทว่าการลงทุนซื้อซอฟต์แวร์เชิงธุรกิจเหล่านี้มาใช้งานอาจจะไม่คุ้มค่าในการลงทุนสำหรับผู้ประกอบการวิสาหกิจขนาดกลางและขนาดย่อม (SMEs) หรืออาจารย์ นักวิจัย และ นักศึกษาระดับปริญญาโทและเอก ในมหาวิทยาลัยต่างๆ ดังนั้นวิธีการหนึ่งที่จะทำให้เราสามารถวิเคราะห์ข้อมูลเหล่านี้ได้คือการใช้ open source software ที่สามารถดาวน์โหลดมาใช้งานได้โดยไม่เสียค่าใช้จ่าย (ฟรี !!!) เช่น ซอฟต์แวร์ Weka ผมคลุกคลีกับ Weka มาเป็นเวลาหลายปี เคยเขียนคู่มือการใช้งาน Weka Explorer ลงในนิตยสาร OpenSource2Day สร้างหลักสูตรการอบรมการใช้งาน Weka Explorer และอบรมการใช้งานซอฟต์แวร์ตัวนี้มาเป็นจำนวนเกือบยี่สิบรุ่น แม้ว่าซอฟต์แวร์นี้จะใช้งานได้ง่ายสำหรับผู้เริ่มต้นและสะดวกที่จะนำไปใช้ในการพัฒนา Web Application แต่ในหลายๆ ครั้งผมมักจะพบข้อจำกัดหรือความยากในการแสดงผลจากซอฟต์แวร์ตัวนี้ ดังนั้นผมจึงหันมาสนใจซอฟต์แวร์ตัวอื่นที่สามารถทดแทนหรือดีกว่าซอฟต์แวร์ Weka Explorer และผมก็พบกับซอฟต์แวร์ RapidMiner Studio 6 ซึ่งเป็นซอฟต์แวร์ทาง Data Mining ที่ได้รับการโหวตว่ามีผู้ใช้งานมากที่สุดจากเว็บไซต์ KDnuggets.com เมื่อปี 2013 ในหลักสูตรนี้ผมจะแนะนำให้คุณรู้จักการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining ตั้งแต่ระดับต้นจน (basic) จนถึงระดับกลาง (intermediate) ด้วยการใช้ซอฟต์แวร์ RapidMiner Studio 6 ซึ่งเป็นเวอร์ชันล่าสุด ถ้าคุณยังลังเลว่าคุณควรจะมาเข้าร่วมอบรมหลักสูตรนี้กับผมหรือไม่ ผมขอถาม 8 คำถามสั้นๆ ดังนี้ครับ

- สนใจการวิเคราะห์ข้อมูลด้วย Data Mining แต่ไม่รู้จะเริ่มยังไงดี
- อยากรู้ว่าลูกค้าซื้อสินค้าอะไรเป็นส่วนใหญ่
- อยากเข้าใจพฤติกรรมการบริโภคของลูกค้า
- อยากทำงานวิจัยทางด้าน text mining
- อยากทำงานวิจัยทางด้าน image processing
- ไม่ชอบการเขียนโปรแกรมแต่อยากวิเคราะห์ข้อมูลที่ซับซ้อนได้
- เคยเข้าร่วมการอบรมการใช้งาน Weka Explorer มาแล้วและอยาก update ความรู้ทาง Data Mining ใหม่ๆ ด้วยซอฟต์แวร์ใหม่ๆ

ถ้าคุณตอบว่า “ใช่” ในคำถามข้อใดข้อหนึ่ง ผมขอแนะนำว่าคุณควรจะมาเข้าร่วมอบรมกับผมครับ และคุณจะรู้ว่าทำไมผมถึงเปลี่ยนใจจาก Weka Explorer มาตกหลุมรักซอฟต์แวร์ที่ชื่อว่า RapidMiner Studio 6 ครับ

<u>เนื้อหาการอบรม</u>

วันที่ 1

- แนะนำการวิเคราะห์ข้อมูลด้วยเทคนิค Data Mining และการใช้ประโยชน์ในงานวิจัย
- แนะนำกระบวนการ CRISP-DM เบื้องต้นสำหรับการวิเคราะห์ข้อมูล
- แนะนำส่วนต่างๆ ของซอฟต์แวร์ RapidMiner Studio 6
- การนำข้อมูลไฟล์ Excel, CSV เข้ามาใช้ใน RapidMiner Studio 6
- ลักษณะของแอตทริบิวต์ (attribute) ต่างๆ ในชุดข้อมูล
- การเขียนไฟล์ให้อยู่ในรูปแบบของ Excel และ CSV
- การแสดงข้อมูลในกราฟแบบต่างๆ เช่น scatter plot, time series
- การค้นหา Outlier ซึ่งเป็นข้อมูลที่แตกต่างจากข้อมูลอื่นๆ
- การค้นหาข้อมูลที่ผิดพลาด (missing value) และแทนที่ด้วยค่าที่กำหนดเองหรือค่าทางสถิติ
- การแปลงข้อมูลด้วยเทคนิค discretization แบบกำหนดช่วงเองหรือแบบอัตโนมัติ
- การลดจำนวนข้อมูลด้วยการ sampling แบบต่างๆ
- การเลือกแอตทริบิวต์เพื่อใช้ในการวิเคราะห์ข้อมูล
- แนะนำการหากฎความสัมพันธ์ (association rules) และการประยุกต์ใช้งานด้านต่างๆ
- แนะนำเทคนิคการหากฎความสัมพันธ์ด้วยเทคนิค Apriori และ FP Growth
- การแปลงข้อมูลจากฐานข้อมูล relation database ให้เป็นฐานข้อมูล transaction database
- การหากฎความสัมพันธ์ด้วยเทคนิค FP Growth ซึ่งเป็นวิธีที่มีประสิทธิภาพมากที่สุด
- Workshop การหากฎความสัมพันธ์จากข้อมูลการซื้อสินค้าจำนวนมากกว่า 100,000 transactions ด้วย RapidMiner Studio

- แนะนำการแบ่งกลุ่มข้อมูล (clustering) และการประยุกต์ใช้งานด้านต่างๆ
- แนะนำตัววัดประสิทธิภาพของการแบ่งกลุ่มข้อมูล
- แนะนำการแบ่งกลุ่มข้อมูลด้วยเทคนิค K-Means และ DBScan
- Workshop การแบ่งกลุ่มข้อมูลทางด้านการศึกษาและการแพทย์ด้วย RapidMiner Studio 6

วันที่ 2

- แนะนำการจำแนกประเภทข้อมูล (classification)
- การวัดประสิทธิภาพของการจำแนกประเภทข้อมูล
- แนะนำเทคนิค Linear Regression และการประยุกต์ใช้งาน
- การใช้งาน Linear Regression ใน RapidMiner Studio 6
- แนะนำเทคนิค Naive Bayes และการประยุกต์ใช้งาน
- การใช้งาน Naive Bayes ใน RapidMiner Studio 6
- แนะนำเทคนิค Decision Tree และการประยุกต์ใช้งาน
- การใช้งาน Decision Tree ใน RapidMiner Studio 6
- แนะนำเทคนิค K-Nearest Neighbours (KNN) และการประยุกต์ใช้งาน
- การใช้งาน KNN ใน RapidMiner Studio 6
- แนะนำเทคนิค Neural Networks และการประยุกต์ใช้งาน
- การใช้งาน Neural Networks ใน RapidMiner Studio 6
- แนะนำเทคนิค Support Vector Machines (SVM) และการประยุกต์ใช้งาน
- การใช้งาน SVM ใน RapidMiner Studio 6
- Workshop การจำแนกประเภทข้อมูลในงานด้านต่างๆ
    - ด้านธุรกิจ
    - ด้านการศึกษา
    - ด้านการแพทย์
- การคัดเลือกแอตทริบิวต์ (attribute selection) และการประยุกต์ใช้ในการจำแนกประเภทข้อมูล
- Workshop การคัดเลือกแอตทริบิวต์และการจำแนกประเภทข้อมูลในงานด้านต่างๆ

วันที่ 3

- แนะนำการทำ Text Mining ด้วย RapidMiner Studio 6
- Workshop การจำแนกข้อความที่เป็น spam จาก SMS
- Workshop การแบ่งกลุ่มข้อมูลจากข้อความรีวิว (Review)
- Workshop การหากฎความสัมพันธ์จากข้อความรีวิว

- แนะนำการทำ Image Mining ด้วย RapidMiner Studio 6
- Workshop การจำแนกรูปภาพออกเป็นประเภทต่างๆ

<u>วิทยากร</u>

- ดร. เอกสิทธิ์ พัชรวงศ์ศักดา วิทยาศาสตร์คอมพิวเตอร์ สถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทยาลัยธรรมศาสตร์

วันเวลาและสถานที่อบรม

- ระยะเวลาการอบรม 3 วัน
- ณ โรงแรม เค. ยู. โฮม มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน

ค่าลงทะเบียน

- ราคา 6,500 บาท (ลดพิเศษจากราคาเต็ม 6,900 บาท)
- ผู้เข้าร่วมอบรมจะได้รับหนังสือประกอบการอบรม flash drive

ติดตามรายละเอียดเพิ่มเติมได้ที่ htpp://www.dataminingtrend.com หรือ
http://facebook.com/datacube.th

การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิง เบื้องต้น

ย้อนหลังไปเมื่อ 12 ปีก่อน การวิเคราะห์ข้อมูลด้วยเทคนิค ดาต้า ไมน์นิง (Data Mining) ยังรู้จักกันในวงแคบ ส่วนใหญ่เป็นนักศึกษาปริญญาโทและเอกที่สนใจทำงานวิจัยทางด้านนี้ ผมเองเริ่มต้นรู้จักกับดาต้า ไมน์นิง เมื่อประมาณ 12 ปีก่อนเช่นกัน ในสมัยที่เป็นนักศึกษาปริญญาตรีตัวเล็กๆ ในห้องปฏิบัติการวิจัยการค้นหาความรู้จากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery Laboratory) ในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ในช่วงเวลาที่ผ่านมาผมได้เห็นการเปลี่ยนแปลงเกี่ยวกับความสนใจของผู้คนในเรื่องดาต้า ไมน์นิงอย่างมากมาย ตั้งแต่ตอนแรกที่ความสนใจอยู่ในวงแคบดังที่ได้กล่าวมาแล้วจนมาถึงปัจจุบันที่มีผู้สนใจเพิ่มขึ้นเป็นวงกว้าง เช่น บริษัทเอกชนหรือธนาคารต่างเริ่มให้ความสนใจนำการวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิงไปใช้งานกันมากขึ้นหรือมหาวิทยาลัยบางแห่งเริ่มจัดให้มีการเรียนการสอนเกี่ยวกับเรื่องนี้ในระดับชั้นปริญญาตรี จากความนิยมที่เพิ่มขึ้นและการเก็บเกี่ยวประสบการณ์ในการทำงานและการอบรมการวิเคราะห์ข้อมูลทางด้านดาต้า ไมน์นิง ทำให้ผมคิดอยากจะเขียนหนังสือสักเล่มซึ่งทำการแนะนำเทคนิคการวิเคราะห์ข้อมูลทางดาต้า ไมน์นิงเบื้องต้น สำหรับนักศึกษาและผู้สนใจขึ้นมาและนั่นเองคือที่มาของหนังสือเล่มนี้ที่ชื่อว่า An Introduction to Data Mining Techniques โดยในหนังสือเล่มนี้ผมจะแสดงหลักการทำงานของวิธีการทางด้านดาต้า ไมน์นิง ประกอบด้วย

- การหากฎความสัมพันธ์ (association rules discovery)
  - การหากฎความสัมพันธ์ด้วยวิธี Apriori
- การแบ่งกลุ่มข้อมูล (clustering)
  - การแบ่งกลุ่มข้อมูลด้วยวิธี K-Means
  - การแบ่งกลุ่มข้อมูลด้วยวิธี Agglomerative Clustering
- การจำแนกประเภทข้อมูล (classification)
  - การจำแนกประเภทข้อมูลด้วยวิธี Decision Tree
  - การจำแนกประเภทข้อมูลด้วยวิธี Naive Bayes
  - การจำแนกประเภทข้อมูลด้วยวิธี K-Nearest Neighbors
  - การจำแนกประเภทข้อมูลด้วยวิธี Neural Network

พร้อมทั้งตัวอย่างการทำงานของวิธีการเหล่านี้เพื่อให้ผู้อ่านเข้าใจได้ง่ายโดยที่ไม่ต้องมีความรู้พื้นฐานทางด้านคณิตศาสตร์ขั้นสูง