2 Machine Learning

1. Data Analysis (Karin Kelley, 2566)

Data Analysis คือ การนำ Data หรือข้อมูลจากผลประกอบการในอดีตมาวิเคราะห์ และหา

สาเหตุของปัญหา หรือต้นตอที่มาที่ไปของข้อมูลอย่างแท้จริง ‘เพื่อให้เห็นภาพขั้นตอนของ

ธุรกิจได้ชัดเจนขึ้น’ ต้องบอกว่า Data Analysis นั้นเป็นศาสตร์ที่มีความยืดหยุ่นสูง ต้องใช้

ความเข้าใจและอัปเดตข้อมูลอยู่ตลอดเวลา

โดย Data Analysis จะยึดหลักการของ ‘การวิเคราะห์และหาจุดเชื่อมโยงของข้อมูล’ ซึ่งสิ่ง

เหล่านี้จะส่งผลให้แผนธุรกิจของคุณถูกปรับปรุงไปในทางที่ดีขึ้น เพราะข้อมูลจะช่วยให้คุณ

มองเห็นภาพรวมและสาเหตุของปัญหาที่มีส่วนทำให้ธุรกิจไม่เกิดผลลัพธ์เท่าที่ควร

กระบวนการวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลคือ กระบวนการวิเคราะห์ข้อมูลหรืออีกวิธีหนึ่งคือขั้นตอนการวิเคราะห์ข้อมูล เกี่ยวข้องกับการรวบรวมข้อมูลทั้งหมด ประมวลผล สำรวจข้อมูล และใช้เพื่อค้นหารูปแบบและข้อมูลเชิงลึกอื่นๆ

* 1. Data Cleaning (Ronald Van Loon, 2566)

ข้อมูลบางส่วนที่คุณรวบรวมไว้อาจไม่มีประโยชน์ ดังนั้นจึงถึงเวลาที่จะต้องล้างข้อมูลออก กระบวนการนี้เป็นที่ที่คุณลบช่องว่าง บันทึกที่ซ้ำกัน และข้อผิดพลาดพื้นฐาน จำเป็นต้องล้างข้อมูลก่อนส่งข้อมูลเพื่อทำการวิเคราะห์

* 1. Missing Data

ข้อมูลหายไปหรือค่าหายไป เกิดขึ้นเมื่อคุณไม่มีข้อมูลที่เก็บไว้สำหรับตัวแปรหรือผู้เข้าร่วมบางตัว ข้อมูลอาจหายไปเนื่องจากการป้อนข้อมูลที่ไม่สมบูรณ์ อุปกรณ์ทำงานผิดปกติ ไฟล์สูญหาย และสาเหตุอื่นๆ อีกมากมาย

ในชุดข้อมูลใดๆ มักจะมีข้อมูลที่ขาดหายไป ในการวิจัยเชิงปริมาณค่าที่หายไปจะปรากฏเป็นเซลล์ว่าง

ประเภทของข้อมูลที่ขาดหายไป

ข้อมูลที่ขาดหายไปถือเป็นข้อผิดพลาดเนื่องจากข้อมูลของคุณไม่ได้แสดงถึงมูลค่าที่แท้จริงของสิ่งที่คุณกำหนดไว้เพื่อวัด

เหตุผลของข้อมูลที่หายไปเป็นสิ่งสำคัญที่ต้องพิจารณา เนื่องจากจะช่วยให้คุณระบุประเภทของข้อมูลที่หายไปและสิ่งที่คุณต้องทำเกี่ยวกับข้อมูลนั้น

ข้อมูลที่ขาดหายไปมีสามประเภทหลัก

1. หายไปโดยสิ้นเชิงโดยไม่ได้ตั้งใจ

เมื่อข้อมูลหายไปโดยสิ้นเชิงโดยการสุ่ม (MCAR) ความน่าจะเป็นที่ค่าใดค่าหนึ่งจะหายไปจากชุดข้อมูลของคุณจะไม่เกี่ยวข้องกับสิ่งอื่นใด

ค่าที่หายไปจะถูกกระจายแบบสุ่ม จึงสามารถมาจากที่ใดก็ได้ในการแจกแจงค่าทั้งหมดของคุณ ข้อมูล MCAR เหล่านี้ยังไม่เกี่ยวข้องกับตัวแปรอื่นๆ ที่ตรวจไม่พบอีกด้วย

ข้อมูลมักถูกพิจารณาว่าเป็น MCAR หากดูเหมือนไม่เกี่ยวข้องกับค่าเฉพาะหรือตัวแปรอื่นๆ ในทางปฏิบัติ เป็นเรื่องยากที่จะบรรลุสมมติฐานนี้ เนื่องจาก "ความสุ่มที่แท้จริง" นั้นหาได้ยาก

เมื่อข้อมูลสูญหายเนื่องจากอุปกรณ์ทำงานผิดปกติหรือสูญเสียตัวอย่าง จะถือเป็น MCAR

1. หายไปโดยบังเอิญ

ข้อมูลที่ขาดหายไปโดยการสุ่ม (MAR) ไม่ ได้ หายไปจากการสุ่ม คำนี้เป็นคำ เรียกชื่อผิดเล็กน้อย

ข้อมูลที่หายไปประเภทนี้แตกต่างอย่างเป็นระบบจากข้อมูลที่คุณรวบรวม แต่ตัวแปรอื่นๆ ที่สังเกตพบสามารถนำมาพิจารณาได้ทั้งหมด

ความน่าจะเป็นที่จุดข้อมูลจะหายไปนั้นสัมพันธ์กับตัวแปรที่สังเกตได้อื่น แต่ไม่เกี่ยวข้องกับค่าเฉพาะของจุดข้อมูลนั้นเอง

1. ขาดไม่ได้โดยบังเอิญ

ข้อมูลที่ขาดหายไปไม่ได้สุ่ม (MNAR) หายไปด้วยเหตุผลที่เกี่ยวข้องกับค่าเหล่านั้น

ข้อมูลที่ขาดหายไปประเภทนี้เป็นสิ่งสำคัญที่ต้องค้นหา เนื่องจากคุณอาจขาดข้อมูลจากกลุ่มย่อยที่สำคัญภายในตัวอย่างของคุณ ตัวอย่างของคุณอาจไม่ได้เป็นตัวแทนของประชากร ของคุณ

* 1. Imbalanced Data (Saikat Mazumder, 2566)

Imbalanced data หมายถึงชุดข้อมูลประเภทต่างๆที่คลาสเป้าหมายมีการกระจายข้อมูลที่ไม่สม่ำเสมอ กล่าวคือ ป้ายกำกับคลาสหนึ่งมีจำนวนข้อมูลที่สูงมาก และอีกป้ายหนึ่งมีจำนวนข้อมูลที่ต่ำมาก

1.3.1 Resampling

เทคนิคนี้ใช้ในการสุ่มตัวอย่างหรือลดตัวอย่างคลาสส่วนน้อยหรือส่วนใหญ่ เมื่อเรา

ใช้ชุดข้อมูลที่ไม่สมดุล เราสามารถสุ่มตัวอย่างคลาสส่วนน้อยโดยใช้การแทนที่ได้ เทคนิคนี้เรียกว่าการสุ่มตัวอย่างมากเกินไป ในทำนองเดียวกัน เราสามารถสุ่มลบแถวออกจากคลาสส่วนใหญ่เพื่อจับคู่กับคลาสของชนกลุ่มน้อยซึ่งเรียกว่าการสุ่มตัวอย่างต่ำ หลังจากการสุ่มตัวอย่างข้อมูลแล้ว เราจะได้ชุดข้อมูลที่สมดุลสำหรับทั้งคลาสส่วนใหญ่และคลาสส่วนน้อย

1.3.1.1 Oversampling

รูปภาพประกอบด้วย ข้อความ, ภาพหน้าจอ, สี่เหลี่ยมผืนผ้า, ออกแบบ

คำอธิบายที่สร้างโดยอัตโนมัติ

1.3.1.2 Undersampling

รูปภาพประกอบด้วย ข้อความ, ภาพหน้าจอ, สี่เหลี่ยมผืนผ้า, ออกแบบ

คำอธิบายที่สร้างโดยอัตโนมัติ

1. Data Visualization (IBM)

การแสดงข้อมูลผ่านการใช้กราฟิกทั่วไป เช่น แผนภูมิ แผนภูมิ อินโฟกราฟิก และแม้แต่ภาพเคลื่อนไหว การแสดงข้อมูลด้วยภาพเหล่านี้จะสื่อสารความสัมพันธ์ของข้อมูลที่ซับซ้อนและข้อมูลเชิงลึกที่ขับเคลื่อนด้วยข้อมูลในลักษณะที่เข้าใจง่าย การแสดงภาพข้อมูลเป็นขั้นตอนสำคัญในกระบวนการวิทยาศาสตร์ข้อมูล ช่วยให้ทีมและบุคคลสามารถถ่ายทอดข้อมูลไปยังเพื่อนร่วมงานและผู้มีอำนาจตัดสินใจได้อย่างมีประสิทธิภาพมากขึ้น

2.1 Types of data visualizations

2.1.1 heatmap : การแสดงภาพกราฟิกเหล่านี้มีประโยชน์ในการแสดงภาพข้อมูลพฤติกรรมตามสถานที่ นี่อาจเป็นตำแหน่งบนแผนที่หรือแม้แต่หน้าเว็บก็ได้

รูปภาพ

2.1.2 boxplot(**Mike Yi**) : พล็อตกล่อง (หรือที่เรียกว่า พล็อตกล่องและมัสสุ) ใช้กล่องและเส้นเพื่อแสดงการแจกแจงของกลุ่มข้อมูลตัวเลขตั้งแต่หนึ่งกลุ่มขึ้นไป ขีดจำกัดของกล่องระบุช่วงของส่วนกลาง 50% ของข้อมูล โดยมีเส้นกึ่งกลางทำเครื่องหมายค่ามัธยฐาน เส้นจะขยายออกจากแต่ละกล่องเพื่อจับช่วงของข้อมูลที่เหลืออยู่ โดยมีจุดวางอยู่เหนือขอบเส้นเพื่อระบุค่าผิดปกติ

รูปภาพประกอบด้วย แผนภาพ, ไลน์, วางแผน, พล็อต

คำอธิบายที่สร้างโดยอัตโนมัติ

2.1.3 pairplot(Sarath SL, 2562) : pairplot มีประโยชน์เมื่อคุณต้องการใช้การวิเคราะห์ข้อมูลเชิงสำรวจ (“EDA”) มันจะแสดงภาพข้อมูลที่กำหนดเพื่อค้นหาความสัมพันธ์ระหว่างข้อมูลเหล่านั้น โดยที่ตัวแปรสามารถเป็นแบบต่อเนื่องหรือแบบแบ่งหมวดหมู่ได้

รูปภาพประกอบด้วย ข้อความ, ออกแบบ, ภาพประกอบ

คำอธิบายที่สร้างโดยอัตโนมัติ

1. Feature Engineering(Apipol Sukgler, 2560)

Feature Engineering เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างข้อมูล เพื่อหาว่า feature ใดควรนำมาใช้ในการวิเคราะห์

ข้อมูลเยอะมากที่สามารถนำมาเป็น ตัวแปร (input feature) ที่จะนำมาใช้ predict มี 2 หลักการคือ

* มีค่าของ feature นั้นในขณะที่กำลังทำการ predict
* โดยที่ค่าที่นำมาใช้ต้องเป็นตัวเลขหรือมีการจัดกลุ่มไว้แล้ว

โดยเราต้องดูว่า feature ไหนไม่มีประโยชน์แล้วควรจะเอาออก เพราะทำให้เกิดความ

ผิดพลาด (noise) ในการหาความสัมพันธ์ที่แท้จริง (Signal) ยิ่งเรามีอัตรา signal / noise น้อยเท่าไหร่ ก็จะยิ่งมีความแม่นยำมากขึ้น

รูปภาพประกอบด้วย ข้อความ, แผนภาพ, ไลน์, พล็อต

คำอธิบายที่สร้างโดยอัตโนมัติ

แต่ในการเลือกใช้ feature มากหรือน้อยมันก็มี การแลกเปลี่ยนที่เท่าเทียม (trade-off)

1. ใช้ทุก feature ที่คุณคิดว่ามีผลต่อการ predict ใน model ถ้าแม่นยำเพียงพอ
2. นำเอา feature ที่คิดว่าไม่มีความเกี่ยวข้องโดยตรงมาใช้ร่วมกัน ถ้าแม่นยำเพียงพอก
3. เริ่มจากใช้ feature selection algorithm ในการหา feature ที่ดีที่สุด

หาก feature ที่เราสนใจมีไม่ครบแต่ถ้ามีข้อมูลมากพอ เราก็มีวิธีจัดการกับข้อมูลที่หายไป

4.Model Development

4.1 Logistic Regression การถดถอยโลจิสติกมักใช้สำหรับการทำนายและการจำแนกปัญหา กรณีการใช้งานบางส่วน ได้แก่

* การตรวจจับการฉ้อโกง:แบบจำลองการถดถอยโลจิสติกสามารถช่วยทีมระบุความผิดปกติของข้อมูล ซึ่งเป็นการคาดการณ์การฉ้อโกง พฤติกรรมหรือลักษณะบางอย่างอาจมีความเกี่ยวข้องกับกิจกรรมการฉ้อโกงสูงกว่า ซึ่งเป็นประโยชน์อย่างยิ่งต่อการธนาคารและสถาบันการเงินอื่น ๆ ในการปกป้องลูกค้าของตน บริษัทที่ใช้ SaaS ได้เริ่มนำแนวทางปฏิบัติเหล่านี้มาใช้เพื่อกำจัดบัญชีผู้ใช้ปลอมออกจากชุดข้อมูลของตนเมื่อทำการวิเคราะห์ข้อมูลเกี่ยวกับประสิทธิภาพทางธุรกิจ
* การทำนายโรค: ในทางการแพทย์ วิธีการวิเคราะห์นี้สามารถใช้เพื่อทำนายความเป็นไปได้ในการเกิดโรคหรือการเจ็บป่วยสำหรับประชากรที่กำหนด องค์กรด้านการดูแลสุขภาพสามารถตั้งค่าการดูแลเชิงป้องกันสำหรับบุคคลที่มีแนวโน้มสูงขึ้นสำหรับความเจ็บป่วยที่เฉพาะเจาะจง
* การคาดคะเนการเลิกใช้งาน : พฤติกรรมเฉพาะอาจบ่งบอกถึงการเลิกใช้งานในหน้าที่ต่างๆ ขององค์กร ตัวอย่างเช่น ฝ่ายทรัพยากรบุคคลและทีมผู้บริหารอาจต้องการทราบว่ามีพนักงานที่มีประสิทธิภาพสูงภายในบริษัทที่มีความเสี่ยงที่จะลาออกจากองค์กรหรือไม่ ข้อมูลเชิงลึกประเภทนี้สามารถกระตุ้นการสนทนาเพื่อทำความเข้าใจปัญหาภายในบริษัท เช่น วัฒนธรรมหรือค่าตอบแทน หรืออีกทางหนึ่ง องค์กรการขายอาจต้องการเรียนรู้ว่าลูกค้ารายใดที่มีความเสี่ยงที่จะพาธุรกิจของตนไปที่อื่น สิ่งนี้สามารถกระตุ้นให้ทีมตั้งค่ากลยุทธ์การรักษาเพื่อหลีกเลี่ยงการสูญเสียรายได้

4.2 ซัพพอร์ตเวกเตอร์ (Support Vector Machine : SVM)

เป็นเทคนิคหนึ่งที่ได้รับความนิยมอย่างแพร่หลายในงานที่เกี่ยวข้องกับการจดจำรูปแบบตลอดจนการแก้ปัญหาการจัดกลุ่ม โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (optimal separating hyperplane) เมื่อเราพิจารณาข้อมูลที่ประกอบด้วยข้อมูล 2 กลุ่ม ดังสมการที่ 1

………(1)

เมื่อ

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ดังสมการที่ 2

………(2)

……….(3)

กลุ่มข้อมูล x จากสมการที่ 3 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงแต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (kernel function) ดังสมการที่ 4

………(4)

เมื่อ แทน ฟังก์ชันสาหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่

อยู่ในรูปเชิงเส้นสามารถแบ่งแยกได้

แทน ค่าน้ำหนักที่เชื่อมโยงจาก feature space ไปสู่ output space

**b** แทน ค่าโน้มเอียง (bias)

แทน ซัพพอร์ตเวกเตอร์ โดย

แทน จำนวนซัพพอร์ตเวกเตอร์

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าข้อมูลใน feature space ที่ใกล้ที่สุดดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุดและเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ต เวกเตอร์” (support vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์

ดังนั้นจึงต้องมีการก าหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร (slack variable) ดังสมการที่ 5 และ 6 ดังนี้

เมื่อกำหนดให้ y = 1 ………(5)

เมื่อกำหนดให้ y = -1 ………(6)

จากการก าหนดค่า > 0 ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมทชีนบรรลุวัตถุประสงค์ใน2 ส่วนคือการเพิ่มระยะแบ่งแยกให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังสมการที่ 7

Minimize ………(7)

โดยที่ :

0,i = 1,2,…,N

และมีเคอร์เนลฟังก์ชัน ที่นิยมใช้อยู่ 3 ชนิดด้วยกันคือ

* โพลิโนเมียล (polynomial) :

* เรเดียลเบสิสฟังก์ชัน (Radial Basis Function-RBF) :

* ซิกมอยด์ (sigmoid) :

4.3 Random Forest

Random Forest คืออัลกอริธึมการเรียนรู้ของเครื่องที่ใช้กันทั่วไปซึ่งรวมเอาต์พุตของแผนผังการตัดสินใจหลาย ๆ อันเพื่อให้ได้ผลลัพธ์เดียว ความสะดวกในการใช้งานและความยืดหยุ่นได้กระตุ้นให้เกิดการนำไปใช้ เนื่องจากสามารถจัดการกับปัญหาการจัดหมวดหมู่และการถดถอยได้

เนื่องจากแบบจำลองฟอเรสต์แบบสุ่มประกอบด้วยต้นไม้การตัดสินใจหลายต้น จึงควรเริ่มต้นด้วยการอธิบายอัลกอริทึมแผนผังการตัดสินใจโดยสังเขป โครงสร้างการตัดสินใจเริ่มต้นด้วยคำถามพื้นฐาน สามารถถามคำถามหลายชุดเพื่อหาคำตอบคำถามเหล่านี้ประกอบขึ้นเป็นโหนดการตัดสินใจในแผนผัง ซึ่งทำหน้าที่เป็นวิธีการแยกข้อมูลคำถามแต่ละข้อช่วยให้แต่ละคนตัดสินใจขั้นสุดท้ายได้ ซึ่งจะแสดงด้วยโหนดปลายใบ การสังเกตที่เหมาะกับเกณฑ์จะเป็นไปตามสาขา "ใช่" และที่ไม่เป็นไปตามเกณฑ์อื่น แผนผังการตัดสินใจพยายามค้นหาการแยกส่วนที่ดีที่สุดเพื่อจัดชุดข้อมูลย่อย และโดยทั่วไปแล้วต้นไม้เหล่านี้จะได้รับการฝึกอบรมผ่านอัลกอริทึมการจัดหมวดหมู่และแผนผังการถดถอย (CART) เมตริก

แม้ว่าแผนผังการตัดสินใจจะเป็นอัลกอริทึมการเรียนรู้แบบมีผู้สอนทั่วไป แต่ก็มีแนวโน้มที่จะเกิดปัญหาได้ เช่น ความลำเอียงและการเกินพอดี อย่างไรก็ตาม เมื่อต้นไม้การตัดสินใจหลายต้นประกอบกันเป็นชุดในอัลกอริทึมของฟอเรสต์แบบสุ่ม พวกมันทำนายผลลัพธ์ที่แม่นยำยิ่งขึ้น โดยเฉพาะอย่างยิ่งเมื่อต้นไม้แต่ละต้นไม่มีความสัมพันธ์กัน

วิธีการเรียนรู้ทั้งมวลประกอบด้วยชุดตัวแยกประเภทเช่น ต้นไม้การตัดสินใจและการคาดคะเนจะถูกรวมเข้าด้วยกันเพื่อระบุผลลัพธ์ที่ได้รับความนิยมสูงสุด วิธีการรวมที่รู้จักกันดีที่สุดคือการบรรจุถุงหรือที่เรียกว่าการรวมบูตสแตรปและการเพิ่มได้แนะนำวิธีการบรรจุถุง ในวิธีนี้ จะมีการเลือกตัวอย่างข้อมูลแบบสุ่มในชุดการฝึกด้วยการแทนที่ ซึ่งหมายความว่าสามารถเลือกจุดข้อมูลแต่ละจุดได้มากกว่าหนึ่งครั้ง หลังจากสร้างตัวอย่างข้อมูลหลายตัวอย่างแล้ว โมเดลเหล่านี้จะได้รับการฝึกอบรมโดยอิสระ และขึ้นอยู่กับประเภทของงาน เช่น การถดถอยหรือการจำแนกประเภท ค่าเฉลี่ยหรือส่วนใหญ่ของการคาดคะเนเหล่านั้นจะให้ค่าประมาณที่แม่นยำยิ่งขึ้น วิธีการนี้มักใช้เพื่อลดความแปรปรวนภายในชุดข้อมูลที่มีสัญญาณรบกวน

อัลกอริทึมของฟอเรสต์แบบสุ่มเป็นส่วนเสริมของวิธีการบรรจุถุงเนื่องจากใช้ทั้งการบรรจุถุงและการสุ่มลักษณะพิเศษเพื่อสร้างฟอเรสต์ของการตัดสินใจที่ไม่สัมพันธ์กัน การสุ่มคุณลักษณะ หรือที่เรียกว่าคุณลักษณะบรรจุถุงหรือ “ วิธีการสุ่มพื้นที่ย่อย ” สร้างชุดย่อยของคุณลักษณะแบบสุ่ม ซึ่งรับประกันความสัมพันธ์ที่ต่ำระหว่างแผนผังการตัดสินใจ นี่เป็นข้อแตกต่างที่สำคัญระหว่างต้นไม้ตัดสินใจและฟอเรสต์แบบสุ่ม ในขณะที่แผนผังการตัดสินใจพิจารณาการแบ่งคุณลักษณะที่เป็นไปได้ทั้งหมด ฟอเรสต์แบบสุ่มจะเลือกเฉพาะส่วนย่อยของคุณลักษณะเหล่านั้นเท่านั้น

1. มันทำงานอย่างไร

อัลกอริทึมฟอเรสต์แบบสุ่มมีไฮเปอร์พารามิเตอร์หลักสามตัว ซึ่งจำเป็นต้องตั้งค่าก่อนการฝึก ซึ่งรวมถึงขนาดโหนด จำนวนทรี และจำนวนคุณลักษณะที่สุ่มตัวอย่าง จากตรงนั้น สามารถใช้ตัวจำแนกฟอเรสต์แบบสุ่มเพื่อแก้ปัญหาการถดถอยหรือการจำแนกประเภท

อัลกอริทึมของฟอเรสต์แบบสุ่มประกอบด้วยคอลเล็กชันของต้นไม้ตัดสินใจ และต้นไม้แต่ละต้นในชุดประกอบด้วยตัวอย่างข้อมูลที่ดึงมาจากชุดการฝึกที่มีการแทนที่ ซึ่งเรียกว่าตัวอย่าง Bootstrap จากตัวอย่างการฝึกนั้น หนึ่งในสามของตัวอย่างนั้นถูกกันไว้เป็นข้อมูลทดสอบ ซึ่งเรียกว่าตัวอย่างที่แกะออกจากถุง (oob) ซึ่งเราจะกลับมาพูดถึงในภายหลัง อีกตัวอย่างหนึ่งของการสุ่มจะถูกแทรกผ่านคุณสมบัติการบรรจุถุง เพิ่มความหลากหลายให้กับชุดข้อมูลและลดความสัมพันธ์ระหว่างแผนผังการตัดสินใจ การกำหนดคำทำนายจะแตกต่างกันไปขึ้นอยู่กับประเภทของปัญหา สำหรับงานการถดถอย ต้นไม้การตัดสินใจแต่ละรายการจะถูกหาค่าเฉลี่ย และสำหรับงานการจำแนกประเภท การลงคะแนนเสียงส่วนใหญ่—เช่น ตัวแปรตามหมวดหมู่ที่พบบ่อยที่สุด จะให้คลาสที่คาดการณ์ไว้ สุดท้าย ตัวอย่าง oob จะใช้สำหรับการตรวจสอบข้าม

ภาพที่ 2.9 (ที่มา : https://www.ibm.com/topics/ramdom-forest)

2. ประโยชน์และความท้าทายของ Random Forest

มีข้อดีและความท้าทายที่สำคัญหลายประการที่ random forest นำเสนอเมื่อใช้สำหรับการจำแนกประเภทหรือปัญหาการถดถอย

ประโยชน์หลัก

• ลดความเสี่ยงของการ overfitting: แผนผังการตัดสินใจจะเสี่ยงต่อการ overfitting เนื่องจากพวกมันมีแนวโน้มที่จะพอดีกับตัวอย่างทั้งหมดภายในข้อมูลการฝึกอบรม อย่างไรก็ตาม เมื่อมีจำนวนต้นไม้การตัดสินใจที่แข็งแกร่งในฟอเรสต์แบบสุ่ม ตัวแยกประเภทจะไม่ทำให้โมเดลมากเกินไป เนื่องจากค่าเฉลี่ยของต้นไม้ที่ไม่สัมพันธ์กันจะลดความแปรปรวนโดยรวมและข้อผิดพลาดในการทำนาย

• ให้ความยืดหยุ่น: เนื่องจากฟอเรสต์แบบสุ่มสามารถจัดการทั้งงานการถดถอยและการจำแนกประเภทด้วยความแม่นยำสูง จึงเป็นวิธีที่นิยมในหมู่นักวิทยาศาสตร์ข้อมูล การรวมคุณสมบัติยังทำให้ตัวจำแนกประเภทฟอเรสต์แบบสุ่มเป็นเครื่องมือที่มีประสิทธิภาพในการประมาณค่าที่ขาดหายไป เนื่องจากจะรักษาความถูกต้องเมื่อข้อมูลบางส่วนขาดหายไป

• ง่ายต่อการกำหนดความสำคัญของคุณลักษณะ: Random Forest ทำให้ง่ายต่อการประเมินความสำคัญของตัวแปรหรือการมีส่วนร่วมกับแบบจำลอง มีสองสามวิธีในการประเมินความสำคัญของฟีเจอร์ ความสำคัญของ Gini และค่าเฉลี่ยการลดลงของสิ่งเจือปน (MDI) มักจะใช้เพื่อวัดว่าความแม่นยำของแบบจำลองลดลงมากเพียงใดเมื่อไม่รวมตัวแปรที่กำหนด อย่างไรก็ตาม ความสำคัญของการเปลี่ยนรูปหรือที่เรียกว่าค่าเฉลี่ยความแม่นยำในการลดลง (MDA) เป็นอีกหนึ่งมาตรการที่สำคัญ MDA ระบุค่าความแม่นยำที่ลดลงโดยเฉลี่ยโดยการสุ่มเปลี่ยนค่าคุณลักษณะในตัวอย่าง oob

3. ความท้าทายที่สำคัญ

• กระบวนการที่ใช้เวลานาน: เนื่องจากอัลกอริทึมของฟอเรสต์แบบสุ่มสามารถจัดการกับชุดข้อมูลขนาดใหญ่ได้ จึงสามารถคาดการณ์ได้แม่นยำมากขึ้น แต่อาจประมวลผลข้อมูลได้ช้า เนื่องจากเป็นการประมวลผลข้อมูลสำหรับแผนผังการตัดสินใจแต่ละรายการ

• ต้องการทรัพยากรมากขึ้น: เนื่องจากฟอเรสต์สุ่มประมวลผลชุดข้อมูลที่มีขนาดใหญ่ขึ้น พวกเขาจึงต้องใช้ทรัพยากรมากขึ้นในการจัดเก็บข้อมูลนั้น

• ซับซ้อนกว่า: การคาดคะเนของแผนผังการตัดสินใจแบบเดี่ยวจะตีความได้ง่ายกว่าเมื่อเทียบกับแบบฟอเรสต์

4.4 ต้นไม้แห่งการตัดสินใจ (Decision Trees : DT)

แผนผังการตัดสินใจเป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนแบบไม่มีพารามิเตอร์ ซึ่งใช้สำหรับทั้งงานการจัดหมวดหมู่และการถดถอย มีโครงสร้างแบบลำดับชั้น

รูปภาพประกอบด้วย แผนภาพ

คำอธิบายที่สร้างโดยอัตโนมัติ

**ภาพที่ 2.10** (ที่มา : <https://www.ibm.com/topics/decision-trees>)

จากแผนภาพด้านบน ต้นไม้การตัดสินใจเริ่มต้นด้วยโหนดรูท ซึ่งไม่มีสาขาใดๆ ที่เข้ามา สาขาขาออกจากรูทโหนดจากนั้นฟีดไปยังโหนดภายในหรือที่เรียกว่าโหนดการตัดสินใจ ตามคุณสมบัติที่มีอยู่ โหนดทั้งสองประเภทจะทำการประเมินเพื่อสร้างชุดย่อยที่เป็นเนื้อเดียวกัน ซึ่งแสดงโดยโหนดลีฟหรือโหนดเทอร์มินัล โหนดปลายสุดแสดงถึงผลลัพธ์ที่เป็นไปได้ทั้งหมดภายในชุดข้อมูล ตัวอย่างเช่น สมมติว่าคุณกำลังพยายามประเมินว่าคุณควรไปเล่นเซิร์ฟหรือไม่ คุณอาจใช้กฎการตัดสินใจต่อไปนี้เพื่อตัดสินใจเลือก

รูปภาพประกอบด้วย แผนภาพ

คำอธิบายที่สร้างโดยอัตโนมัติ

**ภาพที่ 2.11** (ที่มา : https://www.ibm.com/topics/decision-trees)

โครงสร้างผังงานประเภทนี้ยังสร้างการแสดงการตัดสินใจที่ย่อยง่าย ช่วยให้กลุ่มต่างๆ ทั่วทั้งองค์กรเข้าใจได้ดีขึ้นว่าเหตุใดการตัดสินใจจึงเกิดขึ้น

การเรียนรู้แผนผังการตัดสินใจใช้กลยุทธ์การแบ่งแยกและพิชิตโดยการค้นหาอย่างละโมบเพื่อระบุจุดแยกที่เหมาะสมที่สุดภายในต้นไม้ จากนั้นกระบวนการแยกนี้จะถูกทำซ้ำในลักษณะจากบนลงล่าง วนซ้ำจนกว่าเรคคอร์ดทั้งหมดหรือส่วนใหญ่จะถูกจัดประเภทภายใต้ป้ายกำกับระดับเฉพาะ ไม่ว่าจุดข้อมูลทั้งหมดจะถูกจัดประเภทเป็นชุดที่เป็นเนื้อเดียวกันหรือไม่นั้นขึ้นอยู่กับความซับซ้อนของแผนผังการตัดสินใจเป็นส่วนใหญ่ ต้นไม้ขนาดเล็กสามารถรับโหนดใบบริสุทธิ์ได้ง่ายกว่า นั่นคือจุดข้อมูลในคลาสเดียว อย่างไรก็ตาม เมื่อต้นไม้มีขนาดโตขึ้น การรักษาความบริสุทธิ์นี้ก็ยิ่งยากขึ้นเรื่อยๆ และมักจะส่งผลให้ข้อมูลตกหล่นภายในทรีย่อยที่กำหนดน้อยเกินไป เมื่อสิ่งนี้เกิดขึ้น จะเรียกว่าการกระจายตัวของข้อมูล และมักจะนำไปสู่การใช้มากเกินไปเป็นผลให้ต้นไม้ตัดสินใจชอบต้นไม้ขนาดเล็ก ซึ่งสอดคล้องกับหลักการของ parsimony ใน Occam's Razor คือ "ไม่ควรเพิ่มจำนวนเอนทิตีเกินความจำเป็น" ต้นไม้การตัดสินใจควรเพิ่มความซับซ้อนในกรณีที่จำเป็นเท่านั้น เนื่องจากคำอธิบายที่ง่ายที่สุดมักจะดีที่สุด เพื่อลดความซับซ้อนและป้องกันการใส่มากเกินไป มักใช้การตัดแต่งกิ่ง นี่เป็นกระบวนการที่จะลบสาขาที่แบ่งคุณลักษณะที่มีความสำคัญต่ำออก ความพอดีของแบบจำลองสามารถประเมินผ่านกระบวนการตรวจสอบข้ามได้ อีกวิธีหนึ่งที่แผนผังการตัดสินใจสามารถรักษาความถูกต้องได้คือการสร้างกลุ่มผ่านอั ลกอริธึมฟ อเรสต์แบบสุ่ม ลักษณนามนี้ทำนายผลลัพธ์ได้แม่นยำกว่า โดยเฉพาะอย่างยิ่งเมื่อต้นไม้แต่ละต้นไม่มีความสัมพันธ์กัน

1. ประเภทของต้นไม้การตัดสินใจ

อัลกอริทึมของ Hunt ซึ่งพัฒนาขึ้นในทศวรรษที่ 1960 เพื่อจำลองการเรียนรู้ของมนุษย์ในด้านจิตวิทยา เป็นรากฐานของอัลกอริทึมแผนผังต้นไม้การตัดสินใจที่เป็นที่นิยมมากมาย เช่น

* **ID3**: Ross Quinlan ได้รับเครดิตในการพัฒนา ID3 ซึ่งเป็นชวเลขของ "Iterative Dichotomiser 3" อัลกอริทึมนี้ใช้ประโยชน์จากเอนโทรปีและการรับข้อมูลเป็นเมตริกเพื่อประเมินการแยกผู้สมัคร งานวิจัยบางส่วนของ Quinlan เกี่ยวกับอัลกอริทึมนี้จากปี 1986
* **C4.5**: อัลกอริทึมนี้ถือเป็นการทำซ้ำในภายหลังของ ID3 ซึ่งพัฒนาโดย Quinlan สามารถใช้การรับข้อมูลหรือเพิ่มอัตราส่วนเพื่อประเมินจุดแยกภายในแผนผังการตัดสินใจ
* **CART**: คำว่า CART เป็นคำย่อของ "ต้นไม้การจำแนกประเภทและการถดถอย" และได้รับการแนะนำโดย Leo Breiman อัลกอริทึมนี้มักใช้สิ่งเจือปน Gini เพื่อระบุคุณลักษณะในอุดมคติที่จะแยกออก สิ่งเจือปน Gini วัดว่าแอตทริบิวต์ที่สุ่มเลือกนั้นจัดประเภทผิดบ่อยเพียงใด เมื่อประเมินโดยใช้สิ่งเจือปน Gini ค่าที่ต่ำกว่าจะเหมาะสมกว่า

1. วิธีเลือกแอตทริบิวต์ที่ดีที่สุดในแต่ละโหนด

แม้ว่าจะมีหลายวิธีในการเลือกแอตทริบิวต์ที่ดีที่สุดในแต่ละโหนด แต่สองวิธีคือ การรับข้อมูลและการไม่บริสุทธิ์ของ Gini ทำหน้าที่เป็นเกณฑ์การแยกที่เป็นที่นิยมสำหรับแบบจำลองแผนผังการตัดสินใจ ช่วยในการประเมินคุณภาพของแต่ละเงื่อนไขการทดสอบและความสามารถในการจำแนกกลุ่มตัวอย่างในชั้นเรียนได้ดีเพียงใด

* + - เอนโทรปีและการรับข้อมูล

เป็นการยากที่จะอธิบายการรับข้อมูลโดยไม่พูดถึงเรื่องเอนโทรปีก่อน เอนโทรปีเป็นแนวคิดที่เกิดจากทฤษฎีสารสนเทศซึ่งใช้วัดความไม่บริสุทธิ์ของค่าตัวอย่าง ถูกกำหนดโดยสูตรต่อไปนี้ โดยที่

สูตรเอนโทรปี:

S แทน ชุดข้อมูลที่คำนวณเอนโทรปี

C แทน คลาสในเซต S

p(c) แสดงถึงสัดส่วนของจุดข้อมูลที่อยู่ในคลาส c ต่อจำนวนจุดข้อมูลทั้งหมดในชุด S

ค่าเอนโทรปีสามารถอยู่ระหว่าง 0 ถึง 1 หากตัวอย่างทั้งหมดในชุดข้อมูล S อยู่ในคลาสเดียวกัน ค่าเอนโทรปีจะเท่ากับศูนย์ ถ้าครึ่งหนึ่งของตัวอย่างถูกจำแนกเป็นคลาสหนึ่งและอีกครึ่งหนึ่งอยู่ในอีกคลาสหนึ่ง เอนโทรปีจะสูงสุดที่ 1 เพื่อเลือกคุณลักษณะที่ดีที่สุดเพื่อแยกและค้นหาแผนผังการตัดสินใจที่เหมาะสมที่สุด คุณลักษณะที่มีขนาดเล็กที่สุด ควรใช้จำนวนเอนโทรปี ข้อมูลที่ได้รับแสดงถึงความแตกต่างของเอนโทรปีก่อนและหลังการแยก

แอตทริบิวต์ที่กำหนด แอตทริบิวต์ที่ได้รับข้อมูลสูงสุดจะสร้างการแยกที่ดีที่สุดเนื่องจากทำหน้าที่ได้ดีที่สุดในการจำแนกข้อมูลการฝึกอบรมตามการจัดประเภทเป้าหมาย การรับข้อมูลมักจะแสดงด้วยสูตรต่อไปนี้ โดยที่

สูตรการรับข้อมูล :

a แสดงถึงแอตทริบิวต์หรือป้ายกำกับคลาสเฉพาะ

Entropy(S) คือเอนโทรปีของชุดข้อมูล S

|Sv|/ |S| แสดงถึงสัดส่วนของค่าใน ต่อจำนวนของค่าในชุดข้อมูล S

Entropyคือเอนโทรปีของชุดข้อมูล

* + - สิ่งเจือปน Gin

สิ่งเจือปน Gini คือความน่าจะเป็นของการจำแนกจุดข้อมูลแบบสุ่มในชุดข้อมูลอย่างไม่ถูกต้อง หากได้รับการติดป้ายตามการแจกแจงคลาสของชุดข้อมูล เช่นเดียวกับเอนโทรปี ถ้าตั้งค่า S คือบริสุทธิ์ ดังนั้น สิ่งเจือปนจะเป็นศูนย์ นี่แสดงโดยสูตรต่อไปนี้

สูตรสิ่งเจือปน Gini:

1. ข้อดีและข้อเสียของ Decision Tree

ในขณะที่แผนผังการตัดสินใจสามารถใช้ได้ในกรณีการใช้งานที่หลากหลายขั้นตอนวิธีอื่นๆ มักจะมีประสิทธิภาพดีกว่าขั้นตอนวิธี แผนผังการตัดสินใจ ที่กล่าวว่า ต้นไม้การตัดสินใจมีประโยชน์อย่างยิ่งสำหรับ การทำเหมืองข้อมูล และงานค้นหาความรู้ เรามาสำรวจประโยชน์และความท้าทายที่สำคัญของการใช้แผนผังการตัดสินใจเพิ่มเติมด้านล่าง

**ข้อดี**

* + ง่ายต่อการตีความ: ตรรกะบูลีนและการแสดงภาพต้นไม้การตัดสินใจทำให้เข้าใจและบริโภคได้ง่ายขึ้น ลักษณะลำดับชั้นของแผนผังการตัดสินใจยังทำให้ง่ายต่อการดูว่าแอตทริบิวต์ใดที่สำคัญที่สุด ซึ่งอัลกอริทึมอื่นๆ อาจไม่ชัดเจนเสมอไป
  + ไม่จำเป็นต้องเตรียมข้อมูลเพียงเล็กน้อยหรือไม่มีเลย: แผนผังการตัดสินใจมีลักษณะเฉพาะหลายอย่าง ซึ่งทำให้มีความยืดหยุ่นมากกว่าตัวแยกประเภทอื่นๆ สามารถจัดการประเภทข้อมูลต่างๆ เช่น ค่าที่ไม่ต่อเนื่องหรือค่าต่อเนื่อง และค่าต่อเนื่องสามารถแปลงเป็นค่าหมวดหมู่ได้โดยใช้เกณฑ์ นอกจากนี้ยังสามารถจัดการค่าที่มีค่าขาดหายไป ซึ่งอาจเป็นปัญหาสำหรับตัวแยกประเภทอื่นๆ
  + ยืดหยุ่นมากขึ้น: ต้นไม้การตัดสินใจสามารถใช้ประโยชน์จากทั้งการจัดหมวดหมู่และงานการถดถอย ทำให้มีความยืดหยุ่นมากกว่าอัลกอริทึมอื่นๆ นอกจากนี้ยังไม่ไวต่อความสัมพันธ์พื้นฐานระหว่างแอตทริบิวต์ ซึ่งหมายความว่าหากตัวแปรสองตัวมีความสัมพันธ์กันสูง อัลกอริทึมจะเลือกเพียงหนึ่งในคุณสมบัติที่จะแยกออก

**ข้อเสีย**

* + มีแนวโน้มที่จะ overfitting: โครงสร้างการตัดสินใจที่ซับซ้อนมีแนวโน้มที่จะ overfit และไม่สามารถสรุปได้ดีกับข้อมูลใหม่ สถานการณ์นี้สามารถหลีกเลี่ยงได้ผ่านกระบวนการก่อนการตัดแต่งกิ่งหรือหลังการตัดแต่งกิ่ง การตัดแต่งก่อนการตัดแต่งจะหยุดการเจริญเติบโตของต้นไม้เมื่อมีข้อมูลไม่เพียงพอ ในขณะที่การตัดแต่งภายหลังการตัดแต่งจะลบทรีย่อยที่มีข้อมูลไม่เพียงพอหลังจากการสร้างต้นไม้
  + ตัวประมาณค่าความแปรปรวนสูง: ความแปรผันเล็กน้อยภายในข้อมูลสามารถสร้างแผนผังการตัดสินใจที่แตกต่างกันมาก การใส่ถุงหรือการหาค่าเฉลี่ยของค่าประมาณ สามารถเป็นวิธีการลดความแปรปรวนของแผนผังการตัดสินใจ อย่างไรก็ตาม วิธีการนี้มีข้อจำกัดเนื่องจากอาจนำไปสู่ตัวทำนายที่สัมพันธ์กันสูง
  + มีราคาแพงกว่า: เนื่องจากแผนผังการตัดสินใจใช้วิธีค้นหาแบบละโมบในระหว่างการก่อสร้าง การฝึกอบรมจึงมีราคาแพงกว่าเมื่อเทียบกับอัลกอริทึมอื่นๆ
  + ไม่รองรับอย่างสมบูรณ์ใน scikit-learn: Scikit-learn เป็นไลบรารีการเรียนรู้ของเครื่องยอดนิยมที่ใช้ Python แม้ว่าไลบรารีนี้จะมี โมดูล Decision Tree) การใช้งานในปัจจุบันไม่สนับสนุนตัวแปรที่เป็นหมวดหมู่

4.5 K-Nearest

K-Nearest หรือที่รู้จักในชื่อ KNN หรือ k-NN เป็นตัวแยกประเภทการเรียนรู้แบบไม่มีพารามิเตอร์และมีผู้ดูแล ซึ่งใช้ความใกล้ชิดเพื่อจัดหมวดหมู่หรือคาดการณ์เกี่ยวกับการจัดกลุ่มจุดข้อมูลแต่ละจุด แม้ว่าจะใช้ได้กับปัญหาการถดถอยหรือปัญหาการจำแนกประเภท แต่โดยทั่วไปจะใช้เป็นอัลกอริทึมการจำแนกประเภท โดยทำงานบนสมมติฐานที่ว่าจุดที่คล้ายกันสามารถพบได้ใกล้กัน

รูปภาพประกอบด้วย ข้อความ, แผนภาพ, ภาพหน้าจอ, ไลน์

คำอธิบายที่สร้างโดยอัตโนมัติ

ปัญหาการถดถอยใช้แนวคิดที่คล้ายกันกับปัญหาการจำแนกประเภท แต่ในกรณีนี้ ค่าเฉลี่ยของ K- Nearest

จะถูกนำมาทำนายเกี่ยวกับการจำแนกประเภท ความแตกต่างหลักที่นี่คือการจัดหมวดหมู่จะใช้สำหรับค่าที่ไม่ต่อเนื่อง ในขณะที่การถดถอยจะใช้กับค่าต่อเนื่อง อย่างไรก็ตาม ก่อนที่จะทำการจำแนกประเภทได้ จะต้องกำหนดระยะทางเสียก่อน

1. การวัดระยาทาง

โดยสรุป เป้าหมายของอัลกอริธึม5 K-Nearest คือการระบุ5 K-Nearest ที่ใกล้ที่สุดของจุดสืบค้นที่กำหนด เพื่อให้เราสามารถกำหนดป้ายกำกับคลาสให้กับจุดนั้นได้ เพื่อดำเนินการดังกล่าว KNN มีข้อกำหนดบางประการ

กำหนดหน่วยวัดระยะทาง

ในการพิจารณาว่าจุดข้อมูลใดอยู่ใกล้จุดสืบค้นมากที่สุด จะต้องคำนวณระยะห่างระหว่างจุดสืบค้นและจุดข้อมูลอื่นๆ หน่วยวัดระยะทางเหล่านี้ช่วยสร้างขอบเขตการตัดสินใจ ซึ่งแบ่งพาร์ติชันจุดสืบค้นไปยังจุดต่างๆ โดยทั่วไปคุณจะเห็นขอบเขตการตัดสินใจแสดงเป็นภาพด้วยแผนภาพ

ระยะทาง Euclidean (p=2): นี่เป็นการวัดระยะทางที่ใช้บ่อยที่สุด และจำกัดไว้เพียงเวกเตอร์ที่มีมูลค่าจริงเท่านั้น เมื่อใช้สูตรด้านล่างนี้ จะวัดเส้นตรงระหว่างจุดสืบค้นกับจุดอื่นที่กำลังวัด

d(x,y) =

ระยะทาง Manhattan (p=1) : นี่เป็นการวัดระยะทางยอดนิยมอีกรูปแบบหนึ่ง ซึ่งวัดค่าสัมบูรณ์ระหว่างจุดสองจุด นอกจากนี้ยังเรียกอีกอย่างว่าระยะทางรถแท็กซี่หรือระยะทางบล็อกเมือง เนื่องจากมักแสดงเป็นภาพด้วยตาราง ซึ่งแสดงให้เห็นว่าบุคคลหนึ่งอาจนำทางจากที่อยู่หนึ่งไปยังอีกที่หนึ่งผ่านถนนในเมืองได้อย่างไร

ระยะทาง Minkowski : การวัดระยะทางนี้เป็นรูปแบบทั่วไปของการวัดระยะทางแบบยุคลิดและแมนฮัตตัน พารามิเตอร์ p ในสูตรด้านล่างช่วยให้สามารถสร้างหน่วยวัดระยะทางอื่นๆ ได้ ระยะทางแบบยุคลิดแทนด้วยสูตรนี้ เมื่อ p เท่ากับ 2 และระยะทางแมนฮัตตันแทนด้วย p เท่ากับ 1

ระยะ Hamming : เทคนิคนี้มักใช้กับเวกเตอร์บูลีนหรือสตริง เพื่อระบุจุดที่เวกเตอร์ไม่ตรงกัน ด้วยเหตุนี้ จึงเรียกอีกอย่างว่าเมตริกที่ทับซ้อนกัน สามารถแสดงได้ด้วยสูตรต่อไปนี้:

1. การกำหนด k

ค่า k ในอัลกอริธึม k-NN กำหนดจำนวนจุดที่จะถูกตรวจสอบ เพื่อกำหนดการจำแนกประเภทของจุดสืบค้นเฉพาะ ตัวอย่างเช่น หาก k=1 อินสแตนซ์จะถูกกำหนดให้กับคลาสเดียวกันกับเพื่อนบ้านที่ใกล้ที่สุด การกำหนด k อาจเป็นการกระทำที่สมดุลเนื่องจากค่าที่แตกต่างกันสามารถนำไปสู่การใส่มากเกินไปหรือน้อยเกินไปได้ ค่า k ที่ต่ำกว่าสามารถมีความแปรปรวนสูงได้ แต่มีอคติต่ำ และค่า k ที่มากขึ้นอาจทำให้เกิดอคติสูงและความแปรปรวนที่ต่ำกว่า การเลือก k จะขึ้นอยู่กับข้อมูลอินพุตเป็นส่วนใหญ่ เนื่องจากข้อมูลที่มีค่าผิดปกติหรือสัญญาณรบกวนมากกว่ามีแนวโน้มที่จะทำงานได้ดีกว่าหากค่า k สูงกว่า โดยรวมแล้ว ขอแนะนำให้ใช้ตัวเลขคี่สำหรับ k เพื่อหลีกเลี่ยงความสัมพันธ์ในการจำแนกประเภท และกลยุทธ์การตรวจสอบข้ามสามารถช่วยให้คุณเลือก k ที่เหมาะสมที่สุดสำหรับชุดข้อมูลของคุณได้

1. ข้อดี ข้อเสีย

ข้อดี

- ใช้งานง่าย: เนื่องจากอัลกอริทึมมีความเรียบง่ายและแม่นยำ จึงเป็นหนึ่งในตัวแยกประเภทแรกที่นักวิทยาศาสตร์ข้อมูลหน้าใหม่จะได้เรียนรู้

- ปรับเปลี่ยนได้อย่างง่ายดาย : เมื่อมีการเพิ่มตัวอย่างการฝึกอบรมใหม่ อัลกอริธึมจะปรับตามข้อมูลใหม่ เนื่องจากข้อมูลการฝึกอบรมทั้งหมดจะถูกเก็บไว้ในหน่วยความจำ

- ไฮเปอร์พารามิเตอร์ไม่กี่ตัว: KNN ต้องการเพียงค่า ak และการวัดระยะทาง ซึ่งต่ำเมื่อเปรียบเทียบกับอัลกอริธึมการเรียนรู้ของเครื่องอื่น ๆ

ข้อเสีย

- ปรับขนาดได้ไม่ดี : เนื่องจาก KNN เป็นอัลกอริธึมแบบขี้เกียจ จึงใช้หน่วยความจำและการจัดเก็บข้อมูลมากกว่าเมื่อเทียบกับตัวแยกประเภทอื่น ซึ่งอาจมีค่าใช้จ่ายสูงทั้งในแง่เวลาและเงิน หน่วยความจำและพื้นที่เก็บข้อมูลที่มากขึ้นจะทำให้ค่าใช้จ่ายทางธุรกิจเพิ่มขึ้น และข้อมูลที่มากขึ้นอาจใช้เวลาในการคำนวณนานขึ้น แม้ว่าโครงสร้างข้อมูลที่แตกต่างกัน เช่น Ball-Tree จะถูกสร้างขึ้นเพื่อจัดการกับความไร้ประสิทธิภาพในการคำนวณ แต่ตัวแยกประเภทที่แตกต่างกันอาจเหมาะสมที่สุด ขึ้นอยู่กับปัญหาทางธุรกิจ

- คำสาปแห่งมิติ : อัลกอริธึม KNN มีแนวโน้มที่จะตกเป็นเหยื่อของคำสาปแห่งมิติ ซึ่งหมายความว่าอัลกอริทึมนี้จะทำงานได้ไม่ดีกับอินพุตข้อมูลมิติสูง บางครั้งเรียกอีกอย่างว่าปรากฏการณ์พีค (ลิงก์อยู่นอก ibm.com) โดยที่หลังจากที่อัลกอริธึมได้รับคุณลักษณะที่เหมาะสมที่สุด คุณลักษณะเพิ่มเติมจะเพิ่มจำนวนข้อผิดพลาดในการจัดหมวดหมู่ โดยเฉพาะอย่างยิ่งเมื่อขนาดตัวอย่างเล็กลง

- มีแนวโน้มที่จะมีการติดตั้งมากเกินไป : เนื่องจาก "คำสาปแห่งมิติ" KNN จึงมีแนวโน้มที่จะมีการติดตั้งมากเกินไป แม้ว่าเทคนิคการเลือกคุณลักษณะและการลดขนาดจะถูกนำมาใช้เพื่อป้องกันไม่ให้สิ่งนี้เกิดขึ้น ค่า k ก็สามารถส่งผลกระทบต่อพฤติกรรมของแบบจำลองได้เช่นกัน ค่า k ที่ต่ำกว่าสามารถทำให้ข้อมูลพอดีกับข้อมูลได้ ในขณะที่ค่า k ที่สูงกว่ามีแนวโน้มที่จะ "ปรับให้เรียบ" ค่าทำนาย เนื่องจากเป็นการหาค่าเฉลี่ยของค่าในพื้นที่ที่ใหญ่กว่าหรือบริเวณใกล้เคียง อย่างไรก็ตาม หากค่า k สูงเกินไป ก็อาจทำให้ข้อมูลไม่พอดีได้

4.6 MLP

โครงข่ายประสาทเทียมหลายชั้นที่เชื่อมต่อโดยสมบูรณ์เรียกว่า Multilayer Perceptron (MLP)

รูปภาพประกอบด้วย แผนภาพ, ข้อความ, ไลน์, วงกลม

คำอธิบายที่สร้างโดยอัตโนมัติ

มี 3 ชั้น รวมชั้นซ่อน 1 ชั้น หากมีเลเยอร์ที่ซ่อนอยู่มากกว่า 1 เลเยอร์ จะเรียกว่า Deep ANN MLP เป็นตัวอย่างทั่วไปของโครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า ในรูปนี้ หน่วยกระตุ้นที่ i ในชั้น lth แสดงเป็น ai(l)

จำนวนเลเยอร์และจำนวนเซลล์ประสาทเรียกว่าไฮเปอร์พารามิเตอร์ของโครงข่ายประสาทเทียม และสิ่งเหล่านี้จำเป็นต้องมีการปรับแต่ง ต้องใช้เทคนิคการตรวจสอบความถูกต้องข้ามเพื่อค้นหาค่าที่เหมาะสมที่สุดสำหรับสิ่งเหล่านี้

ขั้นตอนการเรียนรู้ MLP มีดังต่อไปนี้

* เริ่มต้นด้วยเลเยอร์อินพุต เผยแพร่ข้อมูลส่งต่อไปยังเลเยอร์เอาต์พุต ขั้นตอนนี้เป็นการขยายพันธุ์ไปข้างหน้า
* ขึ้นอยู่กับผลลัพธ์ ให้คำนวณข้อผิดพลาด (ความแตกต่างระหว่างผลลัพธ์ที่คาดการณ์ไว้และผลลัพธ์ที่ทราบ) ข้อผิดพลาดจะต้องลดลง
* เผยแพร่ข้อผิดพลาดอีกครั้ง ค้นหาอนุพันธ์ของมันตามแต่ละน้ำหนักในเครือข่าย และอัปเดตแบบจำลอง

ทำซ้ำสามขั้นตอนข้างต้นในหลายยุคเพื่อเรียนรู้ตุ้มน้ำหนักในอุดมคติ

สุดท้าย ผลลัพธ์จะถูกใช้ผ่านฟังก์ชันขีดจำกัดเพื่อรับเลเบลคลาสที่คาดการณ์ไว้

Forward Propagation in MLP

ในขั้นตอนแรก ให้คำนวณหน่วยการเปิดใช้งาน al(h) ของเลเยอร์ที่ซ่อนอยู่

หน่วยการเปิดใช้งานเป็นผลมาจากการใช้ฟังก์ชันการเปิดใช้งาน φ กับค่า z จะต้องสร้างความแตกต่างได้จึงจะสามารถเรียนรู้น้ำหนักโดยใช้การไล่ระดับสีแบบไล่ระดับ ฟังก์ชันการเปิดใช้งาน φ มักเป็นฟังก์ชัน sigmoid (logistic)

ช่วยให้เกิดความไม่เชิงเส้นที่จำเป็นในการแก้ปัญหาที่ซับซ้อน เช่น การประมวลผลภาพ

เส้นโค้งซิกมอยด์

เส้นโค้งซิกมอยด์เป็นเส้นโค้งรูปตัว S

รูปภาพประกอบด้วย ไลน์, พล็อต, ข้อความ, แผนภาพ

คำอธิบายที่สร้างโดยอัตโนมัติ

การเปิดใช้งานเลเยอร์ที่ซ่อนอยู่จะแสดงเป็น:

z(h) = a(in) W(h)

a(h) =

สำหรับ output:

Z(out) = A(h) W(out)

A(out) =

5. การประเมิน Model