**ทำความเข้าใจเกี่ยวกับ K-หมายถึงการจัดกลุ่ม**

เป้าหมาย

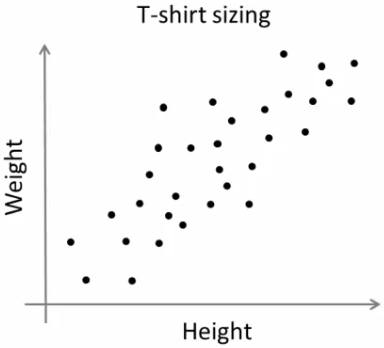
ในบทนี้เราจะเข้าใจแนวคิดของ K-Means Clustering วิธีการทำงาน ฯลฯ

ทฤษฎี

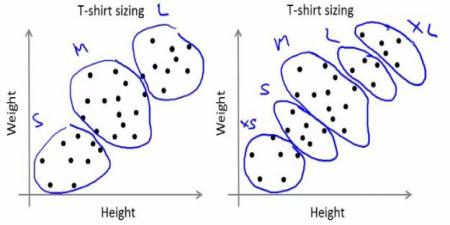
เราจะจัดการกับตัวอย่างที่ใช้ทั่วไป

ปัญหาเสื้อยืด

พิจารณา บริษัท ซึ่งจะปล่อยเสื้อยืดรุ่นใหม่ออกสู่ตลาด เห็นได้ชัดว่าพวกเขาจะต้องผลิตแบบจำลองในขนาดที่แตกต่างกันเพื่อตอบสนองคนทุกขนาด ดังนั้น บริษัท จึงให้ข้อมูลเกี่ยวกับความสูงและน้ำหนักของคนและนำเสนอกราฟดังต่อไปนี้



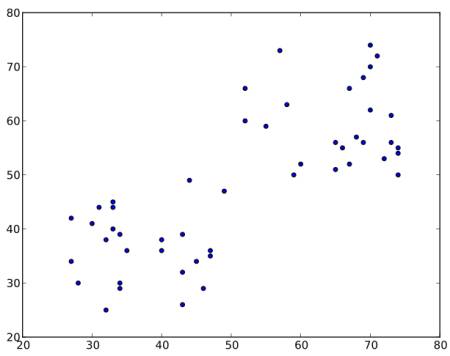
บริษัท ไม่สามารถสร้างเสื้อยืดได้ทุกขนาด แทนพวกเขาแบ่งคนที่มีขนาดเล็กขนาดกลางและขนาดใหญ่และการผลิตเพียง 3 รุ่นนี้ซึ่งจะเหมาะกับทุกคน การจัดกลุ่มคนเหล่านี้ออกเป็นสามกลุ่มสามารถทำได้โดยการจัดกลุ่มแบบ k-means และอัลกอริทึมให้เรามี 3 ขนาดที่ดีที่สุดซึ่งจะตอบสนองทุกคน ถ้าไม่เช่นนั้น บริษัท สามารถแบ่งคนออกเป็นกลุ่มได้มากขึ้นอาจเป็น 5 คนเป็นต้น ตรวจสอบภาพด้านล่าง:



มันทำงานอย่างไร?

อัลกอริทึมนี้เป็นกระบวนการซ้ำ เราจะอธิบายทีละขั้นตอนด้วยความช่วยเหลือของภาพ

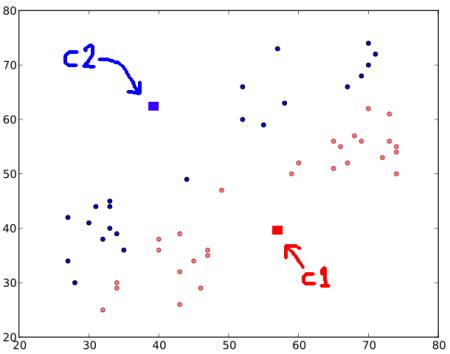
พิจารณาชุดของข้อมูลดังนี้ (คุณสามารถพิจารณาได้ว่าเป็นปัญหาเสื้อยืด) เราจำเป็นต้องจัดกลุ่มข้อมูลนี้ออกเป็นสองกลุ่ม



ขั้นตอน: 1 - อัลกอริทึมสุ่มเลือกเซนทรอยด์สองตัวC1และC2(บางครั้งข้อมูลสองตัวจะถูกนำมาเป็น centroids)

ขั้นตอนที่ 2 - คำนวณระยะทางจากแต่ละจุดไปยังเซนทรอยด์ หากข้อมูลการทดสอบใกล้ชิดมากขึ้นC1ข้อมูลนั้นจะมีข้อความว่า '0' ถ้าอยู่ใกล้ชิดC2แล้วให้ทำเครื่องหมายว่า '1' (หากมีเซนทรอยด์อยู่มากขึ้นจะมีป้ายกำกับว่า '2', '3' ฯลฯ )

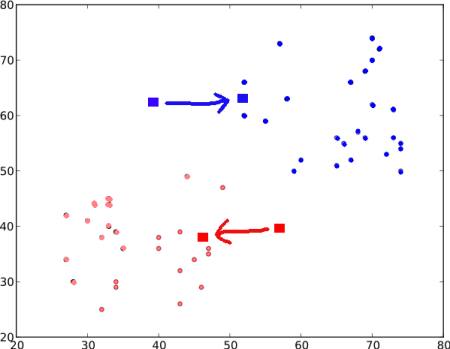
ในกรณีของเราเราจะกำหนดสี '0' ทั้งหมดที่มีป้ายกำกับเป็นสีแดงและ '1' มีข้อความกำกับว่าเป็นสีน้ำเงิน ดังนั้นเราจึงได้ภาพต่อไปนี้หลังจากการดำเนินการข้างต้น



ขั้นตอนที่ 3 - ถัดไปเราจะคำนวณค่าเฉลี่ยของจุดสีน้ำเงินและจุดสีแดงทั้งหมดแยกกันซึ่งจะเป็นเซนทรอยด์ใหม่ของเรา นั่นคือC1และC2เปลี่ยนเป็น centroids ที่คำนวณใหม่ (โปรดจำไว้ว่าภาพที่แสดงไม่ใช่ค่าที่แท้จริงและไม่ใช่ขนาดที่แท้จริง แต่เป็นเพียงการสาธิตเท่านั้น)

และอีกครั้งให้ทำขั้นตอนที่ 2 ด้วย centroids ใหม่และข้อมูลป้ายกำกับเป็น '0' และ '1'

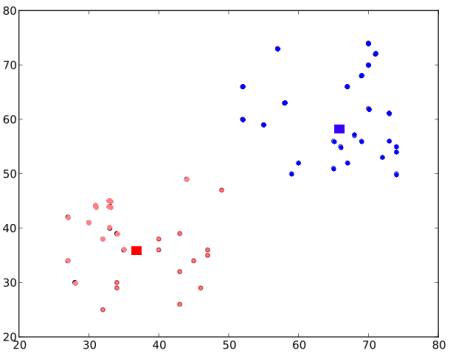
เราจึงได้ผลลัพธ์ดังนี้:



ตอนนี้ขั้นตอนที่ 2และขั้นตอนที่ 3จะถูกทำซ้ำจนกระทั่งทั้งสอง centroids ถูก converged ไปยังจุดคงที่ (หรืออาจจะหยุดขึ้นอยู่กับเกณฑ์ที่เรามีให้เช่นจำนวนสูงสุดของการทำซ้ำหรือความถูกต้องเฉพาะเจาะจงถึง ฯลฯ ) จุดเหล่านี้เป็นเช่นว่าผลรวมของระยะทางระหว่างข้อมูลการทดสอบและ centroids ที่สอดคล้องกันของพวกเขามีขั้นต่ำ หรือเพียงแค่รวมระยะทางระหว่างC1 \ leftrightarrow Red \ _PointsและC2 \ leftrightarrow Blue \ _Pointsต่ำสุด

ลดระยะห่าง (C1, Red \ _Point) + \ sum_ (All \: Blue \ _Points} ระยะทาง (C2, Blue \ _Point) \ bigg \ n \

ผลสุดท้ายเกือบจะเป็นดังนี้:



ดังนั้นนี่เป็นเพียงความเข้าใจที่เข้าใจง่ายของ K-Means Clustering สำหรับรายละเอียดเพิ่มเติมและคำอธิบายทางคณิตศาสตร์โปรดอ่านคู่มือการเรียนรู้เรื่องเครื่องมาตรฐานหรือตรวจสอบลิงก์ในแหล่งข้อมูลเพิ่มเติม มันเป็นเพียงชั้นบนสุดของ K-Means clustering มีจำนวนมากของการปรับเปลี่ยนอัลกอริทึมนี้เช่นวิธีการเลือก centroids เริ่มต้นวิธีการเร่งกระบวนการย้ำเป็นต้น