สวัสดีค่ะเพื่อนๆ ทุกคน วันนี้ผู้เขียนจะมาอธิบายเกี่ยวกับการทำงานของ Principal Component Analysis หรือ PCA ทีละขั้นตอน เพื่อให้เพื่อนๆ สามารถเข้าใจและนำไปใช้ได้อย่างมั่นใจ คนไหนพื้นฐานคณิตศาสตร์ไม่ค่อยแน่น ไม่ต้องกังวลไปนะคะ ในบทความนี้จะอธิบายแบบละเอียดและเข้าใจง่าย คนที่ไม่ได้มี Strong Mathematical Background ก็สามารถทำความเข้าใจได้สบายค่ะ

ก่อนที่เราจะไปเริ่มทำความเข้าใจในแต่ละสเต็ปของการทำ PCA เรามาทำความรู้จักกันก่อนดีกว่า ว่า PCA คืออะไร และทำอะไรได้บ้าง

Overview

PCA เป็นวิธีการลด Dimension ของ Dataset ที่มีขนาดใหญ่ ด้วยการแปลง Variables ที่มีจำนวนมาก ให้มีจำนวนน้อยลงแต่ยัง Contains ข้อมูลส่วนใหญ่ของชุดข้อมูลไว้ได้

การลดจำนวน Variables ของชุดข้อมูลย่อมแลกมาด้วยการสูญเสียความแม่นยำเล็กน้อย อย่างไรก็ตามการลด Dimension ของข้อมูลจะช่วยให้การวิเคราะห์ง่ายและสะดวกมากขึ้น เนื่องจากชุดข้อมูลที่มีขนาดเล็กกว่านั้นง่ายต่อการ Explore และ Visualize การวิเคราะห์ข้อมูลจึงรวดเร็วมากขึ้นสำหรับ Machine Learning Algorithms โดยไม่ต้องประมวลผล Variables จำนวนมาก

A Step by Step Explanation of PCA

STEP 1: Standardization

จุดมุ่งหมายของขั้นตอนนี้คือการ Standardize Range ของ Continuous Variables ใน Dataset เพื่อให้แต่ละ Variables มีผลต่อการวิเคราะห์เท่าๆ กัน

กล่าวคือ หากมีความแตกต่างกันมากระหว่าง Range ของตัวแปรตั้งต้น ตัวแปรที่มี Range กว้างกว่าจะ Dominate ตัวแปรที่มี Range แคบกว่า (เช่น ตัวแปรที่มี Range ระหว่าง 0 ถึง 100 จะ Dominate ตัวแปรที่มี Range ระหว่าง 0 ถึง 1) นำไปสู่ผลลัพธ์ที่ Bias ดังนั้นการ Transform ข้อมูลให้เป็น Comparable Scales สามารถป้องกันปัญหานี้ได้

วิธีการทำ Standardization ไม่ยากอย่างที่คิด แค่นำค่าของแต่ละตัวแปร ลบด้วยค่าเฉลี่ย แล้วหารด้วย Standard Deviation ดังสมการด้านล่าง

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

ภาพที่ 1: The Formula for Standardized Values

หลังจากที่ทำการ Standardize เสร็จ ตัวแปรทั้งหมดจะถูกแปลงเป็น Scale เดียวกัน

STEP 2: Covariance Matrix Computation

จุดมุ่งหมายของขั้นตอนนี้คือการดูว่าแต่ละตัวแปรมีความสัมพันธ์กันหรือไม่ เนื่องจากบางครั้งตัวแปรมีความสัมพันธ์กันสูงในลักษณะที่มีข้อมูลซ้ำซ้อนกัน เพื่อระบุความสัมพันธ์เหล่านี้ เราจึงจำเป็นต้องคำนวณ Covariance Matrix ขึ้นมา

Covariance Matrix คือเมทริกซ์สมมาตร p × p (โดยที่ p คือจำนวน Dimension) ที่มี Covariance ของคู่ที่เป็นไปได้ทั้งหมดของตัวแปรในชุดข้อมูล ตัวอย่างเช่น สำหรับชุดข้อมูล 3 มิติ ที่มี 3 ตัวแปร (x, y, และ z) Covariance Matrix คือเมทริกซ์ 3 × 3 ดังนี้

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

ภาพที่ 2: Covariance Matrix for 3-Dimensional Data

เนื่องจาก Covariance ของตัวแปรใดตัวแปรหนึ่งกับตัวมันเอง เท่ากับ Variance ของตัวแปรนั้นๆ (Cov(a,a)=Var(a)) ดังนั้นในแนวทแยงหลัก (บนซ้ายไปขวาล่าง) ก็คือ Variance ของแต่ละตัวแปรนั่นเอง และเนื่องจาก Covariance เป็น Commutative (Cov(a,b)=Cov(b,a)) Covariance Matrix จึงมีความสมมาตรเมื่อเทียบกับเส้นทแยงมุมหลัก ซึ่งหมายความว่า Upper กับ Lower Triangle มีค่าเท่ากัน

ถึงตรงนี้เพื่อนๆ อาจจะสงสัยกันใช่ไหมคะ ว่าเจ้า Covariance นี่นำไปใช้ยังไง จริงๆ แล้วง่ายมากเลยค่ะ หลังจากที่เราคำนวณ Covariance Matrix เสร็จ สิ่งที่เราสนใจคือ Sign ของ Covariance แต่ละตัวใน Matrix นั่นเอง

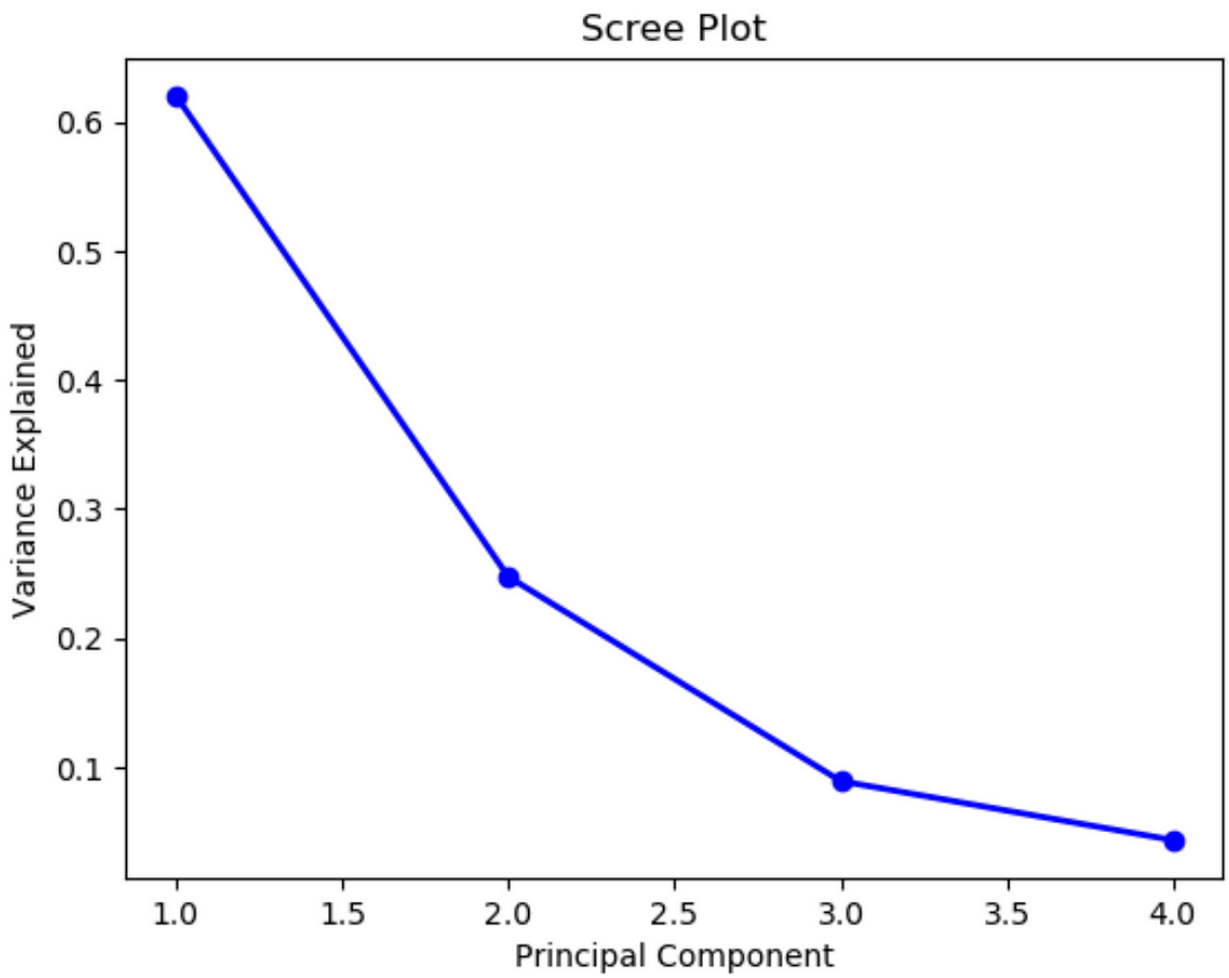
*ถ้าเป็น**ค่าบวก** ตัวแปรทั้งสองเพิ่มขึ้น หรือลดลงพร้อมกัน (Correlated)*

*ถ้าเป็น**ค่าลบ** ตัวแปรหนึ่งเพิ่มขึ้น เมื่ออีกตัวแปรลดลง (Inversely Correlated)*

STEP 3: Compute the Eigenvectors and Eigenvalues of the Covariance Matrix

Eigenvectors และ Eigenvalues เป็น Linear Algebra Concepts ที่เราจำเป็นต้องคำนวณจาก Covariance Matrix เพื่อกำหนด **Principal Components** ของข้อมูล แต่ก่อนที่จะอธิบายคอนเซ็ปต์เหล่านี้ เราต้องเข้าใจว่า Principal Components นั้นหมายถึงอะไร

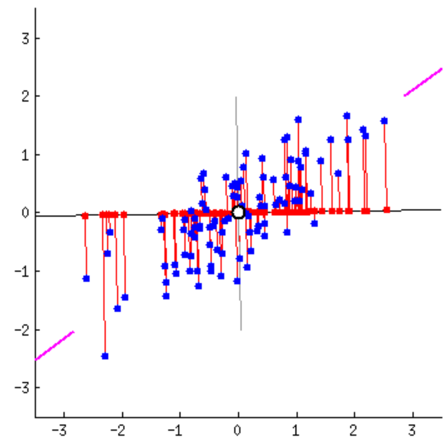
Principal Components คือตัวแปรใหม่ที่สร้างขึ้นเป็น Linear Combinations หรือ Mixtures ของตัวแปรตั้งต้นจากชุดข้อมูล โดย Combinations เหล่านี้จะไม่มีความสัมพันธ์กัน และข้อมูลส่วนใหญ่ภายในตัวแปรเริ่มต้นจะถูก Squeeze หรือ Compress ลงใน First Component แนวคิดคือ 10-Dimensional Data มี 10 Principal Components แต่ PCA จะพยายามใส่ข้อมูลที่เป็นไปได้สูงสุดใน First Component จากนั้นข้อมูลสูงสุดที่เหลืออยู่ใน Second Component และทำแบบนี้ต่อไปเรื่อยๆ ดังที่แสดงในแผนภาพด้านล่าง



ภาพที่ 3: Percentage of Variance for each Principal Component

สิ่งสำคัญอย่างหนึ่งที่ต้องตระหนักในที่นี้คือ Principal Components จะตีความได้น้อยกว่าตัวแปรตั้งต้นของชุดข้อมูล และไม่มีความหมายที่แท้จริง เนื่องจาก Components เหล่านี้สร้างเป็น Linear Combinations ของตัวแปรตั้งต้น

Wait!!! แล้ว Principal Components เหล่านี้ถูกสร้างขึ้นมาอย่างไรล่ะ?



ภาพที่ 4: How to Construct PC

ไม่ยากเลย! เนื่องจากมี Principal Components มากพอๆ กับที่มีตัวแปรในข้อมูล ดังนั้น Principal Components จึงถูกสร้างขึ้นในลักษณะที่ First Principal Component พิจารณาถึง Variance ที่เป็นไปได้มากที่สุดในชุดข้อมูล ยิ่ง Variance มากเท่าใด การกระจายตัวของจุดข้อมูลตามเส้นนั้นยิ่งมาก และยิ่งการกระจายไปตามเส้นมากเท่าไหร่ ยิ่งมีข้อมูลมากขึ้นเท่านั้น หรือถ้าพูดง่ายๆ ลองนึกถึง Principal Components เป็นแกนใหม่ที่มีมุมที่ดีที่สุดในการ Evaluate ข้อมูล เพื่อให้เห็นความแตกต่างระหว่าง Observations ได้ดีขึ้น ตัวอย่างเช่น สมมติว่า Scatter Plot ของชุดข้อมูลของเราเป็นดังที่แสดงตามภาพที่ 4 พอจะเดาได้ไหมว่า First Principal Component อยู่ตรงไหน? คำตอบคือเส้นสีดำที่หมุนไปตรงกับขีดสีม่วง เพราะมันผ่านจุดกำเนิดและเป็นเส้นที่ Projection ของแต่ละจุด (จุดสีแดง) แผ่ออกไปมากที่สุด หรือเส้นที่เพิ่ม Variance (ค่าเฉลี่ยของระยะทางกำลังสองจากจุดที่ทำการ Project (จุดสีแดง) ไปยังจุดกำเนิด)

The Second Principal Component คำนวณในลักษณะเดียวกัน โดยมีเงื่อนไขว่าจะต้องไม่สัมพันธ์กับ (เช่น ตั้งฉาก) First Principal Component และพิจารณาถึง Variance สูงสุดถัดไป และจะทำเหมือนเดิมต่อไปเรื่อยๆ จนกว่าจะมีการคำนวณ Principal Components ทั้งหมดเท่ากับจำนวนตัวแปรเดิม

ตอนนี้เราเข้าใจความหมายของ Principal Components แล้ว กลับมา

*ที่ **Eigenvectors** และ **Eigenvalues** สิ่งแรกที่เราต้องรู้เกี่ยวกับ 2 ค่านี้ คือทั้ง 2 ค่าจะมาเป็นคู่เสมอ Eigenvector ทุกตัวจึงมี Eigenvalue และจำนวนนั้นเท่ากับจำนวน Dimension ข้อมูล ตัวอย่างเช่น ชุดข้อมูล 3 Dimension มีตัวแปร 3 ตัว ดังนั้นจะมี Eigenvector 3 ค่าที่มี Eigenvalue ที่สอดคล้องกัน 3 ค่า*

บางคนอาจจะเริ่มสับสน ว่า Eigenvectors กับ Eigenvalues เกี่ยวกับ PCA อย่างไร ง่ายนิดเดียวเองค่ะ เพื่อนๆ จำภาพที่ 4 ได้ใช่ไหมคะ เบื้องหลังของภาพนั้นก็คือ Eigenvectors และ Eigenvalues นั่นเองค่ะ เนื่องจาก Eigenvectors ของ Covariance Matrix เป็น Directions ของ Axes ที่มี Variance มากที่สุด (Most Information) และเราเรียกว่า Principal Components และ Eigenvalues เป็นเพียงสัมประสิทธิ์ที่ติดอยู่กับ Eigenvectors ซึ่งบอก Variance ที่เกิดขึ้นในแต่ละ Principal Component

โดยการจัดลำดับ Eigenvectors จะจัดตามลำดับ Eigenvalues ของมันจากมากไปน้อย และเราก็จะได้ Principal Components ตามลำดับนัยสำคัญ

Example: สมมุติว่าชุดข้อมูลของเราเป็นแบบ 2 มิติ โดยมี 2 ตัวแปร x, y และ Eigenvectors และ Eigenvalues ของ Covariance Matrix มีดังนี้

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

ภาพที่ 5: Eigenvectors and Eigenvalues

หากเราจัดอันดับ Eigenvalues จากมากไปน้อย เราจะได้ $\lambda 1 > \lambda 2$ ซึ่งหมายความว่า Eigenvector ที่สอดคล้องกับ First Principal Component (PC1) คือ v1 และ Eigenvector ที่สอดคล้องกับ Second Component (PC2) คือ v2

หลังจากที่ได้ Principal Components แล้ว ในการคำนวณเปอร์เซ็นต์ของ Variance (Information) ที่คำนวณโดยแต่ละ Component สามารถทำได้โดยหาร Eigenvalue ของแต่ละ Component ด้วยผลรวมของ Eigenvalues ทั้งหมด ทีนี้เราลองมา Apply วิธีนี้กับตัวอย่างข้างต้นของเราบ้าง จะพบว่า PC1 และ PC2 มี Variance ของข้อมูลอยู่ที่ 96% และ 4% ตามลำดับ

STEP 4: Feature Vector

ดังที่เราเห็นในขั้นตอนที่แล้ว การคำนวณ Eigenvectors และเรียงลำดับตาม Eigenvalues ในลำดับจากมากไปน้อย ทำให้เราสามารถค้นหา Principal Components ตามลำดับความสำคัญได้ สิ่งที่เราทำในขั้นตอนนี้คือเลือกว่าจะเก็บ Components เหล่านี้ทั้งหมดหรือทิ้ง Components ที่มีนัยสำคัญน้อยกว่าไป (ที่มี Eigenvalues ต่ำ) และสร้างด้วย Components ที่เหลือเป็น Matrix ของเวกเตอร์ที่เราเรียกว่า Feature Vector

ดังนั้น Feature Vector เป็นเพียง Matrix ที่มีคอลัมน์ Eigenvectors ของ Components ที่เราตัดสินใจที่จะเก็บไว้ สิ่งนี้ทำให้เป็นก้าวแรกสู่การลดขนาดของข้อมูล เนื่องจากถ้าเราเลือกที่จะเก็บเฉพาะ p Eigenvectors (Components) ออกจาก n ชุดข้อมูล สุดท้ายเราจะมีข้อมูลเพียง p มิติเท่านั้น

Example: ต่อจากตัวอย่างของสเต็ปก่อนหน้า เราสามารถสร้าง Feature Vector ที่มีทั้ง Eigenvector v1 และ v2

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

ภาพที่ 6: Feature Vector

หรือทิ้ง Eigenvector v2 ซึ่งมีนัยสำคัญน้อยกว่า และสร้าง Feature Vector ที่มีเฉพาะ v1 เท่านั้น

$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

ภาพที่ 7: Feature Vector (v1)

การทิ้ง Eigenvector v2 จะทำให้ Dimension ของข้อมูลลดลงไป 1 Dimension และส่งผลให้ Information ในชุดข้อมูลสุดท้ายสูญหายไปบางส่วน แต่เนื่องจาก v2 มีข้อมูลเพียง 4% การสูญเสียจึงไม่สำคัญมากนัก และเรายังคงมีข้อมูล 96% ของ v1 อยู่

STEP 5: Recast the Data along the Principal Components Axes

ในสเต็ปก่อนหน้า นอกเหนือจากการทำ Standardization เรายังไม่ได้ทำการเปลี่ยนแปลงใดๆ กับข้อมูล เราเพียงแค่เลือก Principal Components และสร้าง Feature Vector แต่ชุดข้อมูลจะยังคงอยู่ใน Original Axes เหมือนเดิม

ในขั้นตอนสุดท้ายนี้ จุดมุ่งหมายคือการใช้ Feature Vector ที่สร้างขึ้นโดยใช้ Eigenvectors ของ Covariance Matrix เพื่อปรับทิศทางข้อมูลจากแกนเดิมไปยังแกนที่แสดงโดย Principal Components (ด้วยเหตุนี้จึงตั้งชื่อว่า Principal Components Analysis) ซึ่งทำได้โดยการคูณทรานสโพสของข้อมูลเดิมที่กำหนด ด้วยทรานสโพสของ Feature Vector นั่นเอง

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

ภาพที่ 8: Transformation Formula

เป็นอย่างไรกันบ้างคะสำหรับบทความอธิบาย PCA อย่างละเอียดและเข้าใจง่ายแบบนี้ หวังเป็นอย่างยิ่งเลยนะคะว่าบทความนี้พอจะเป็นประโยชน์ต่อเพื่อนๆ ไม่มากก็น้อย สุดท้ายนี้หากมีการใช้คำไม่เหมาะสมหรือมีข้อผิดพลาดประการใด ผู้เขียนต้องขออภัยไว้ ณ ที่นี้ด้วยนะคะ หากเพื่อนๆ มีข้อติชมสามารถคอมเม้นต์บอกกันมาได้เลยค่า ^^