L10 Dimensionality Reduction การลดมิติข้อมูล

การลดมิติข้อมูล เป็นขั้นตอนที่ถูกนำเสนอเพื่อแก้ปัญหาของมิติข้อมูลที่สูง ไม่มีการจัดการข้อมูลในเบื้องต้น บางข้อมูลไม่มีประโยชน์ในการวิเคราะห์ ส่งผลกระทบต่อความถูกต้องของโมเดล และทำให้สิ้นเปลืองทรัพยากรในการประมวลผลอีกด้วย

* Review of eigenvalues and eigenvectors

Graphical user interface, text, application

Description automatically generated

A picture containing table

Description automatically generated

* PCA and Kernel PCA
  + การวิเคราะห์หาองค์ประกอบหลักของข้อมูลภายใต้เงื่อนไขความแปรปรวน (Variance) สูงสุด หรือเป็นกระบรวนการแปลงข้อมูล X บนเมทริกซ์การแปลง P เพื่อเป็นข้อมูลเชิงเส้น Y
  + Y = PX
* Multi-dimensional distance scaling (MDS)
* Non-linear manifold learning
  + Locally Linear Embedding (LLE)

Local-Linear Embedding (LLE) [21]ถูกนำเสนอในเวลาเดียวกันกับ Isomap โดยประมาณ มีข้อดีหลายประการเหนือ Isomap รวมถึงการเพิ่มประสิทธิภาพที่เร็วขึ้นเมื่อนำไปใช้เพื่อใช้ประโยชน์จากอัลกอริทึมเมทริกซ์แบบกระจัดกระจายและผลลัพธ์ที่ดีกว่าสำหรับปัญหามากมาย LLE เริ่มต้นด้วยการหาเพื่อนบ้านที่ใกล้ที่สุดของแต่ละจุด จากนั้นจะคำนวณชุดของน้ำหนักสำหรับแต่ละจุดที่อธิบายจุดได้ดีที่สุดเป็นการรวมกันเชิงเส้นของเพื่อนบ้าน ในที่สุดก็ใช้เทคนิคการเพิ่มประสิทธิภาพตาม eigenvector เพื่อค้นหาการฝังจุดในมิติต่ำเพื่อให้แต่ละจุดยังคงอธิบายด้วยการรวมกันเชิงเส้นของเพื่อนบ้าน LLE มีแนวโน้มที่จะจัดการกับความหนาแน่นของตัวอย่างที่ไม่สม่ำเสมอได้ไม่ดีเนื่องจากไม่มีหน่วยคงที่เพื่อป้องกันไม่ให้น้ำหนักลอยเนื่องจากภูมิภาคต่างๆมีความหนาแน่นของตัวอย่างแตกต่างกัน LLE ไม่มีโมเดลภายใน

LLE คำนวณพิกัด Barycentric จุดX ฉันขึ้นอยู่กับประเทศเพื่อนบ้านXเจ จุดเดิมถูกสร้างขึ้นใหม่โดยการรวมเชิงเส้นซึ่งกำหนดโดยเมทริกซ์น้ำหนักW ijของเพื่อนบ้าน ข้อผิดพลาดในการสร้างใหม่ถูกกำหนดโดยฟังก์ชันต้นทุนE ( W )

A picture containing text, watch, clock

Description automatically generated

น้ำหนักW ijหมายถึงจำนวนของการมีส่วนร่วมของจุดX jในขณะที่สร้างจุดX iขึ้นใหม่ ฟังก์ชันต้นทุนจะลดลงภายใต้ข้อ จำกัด สองข้อ: (a) จุดข้อมูลแต่ละจุดX iถูกสร้างขึ้นใหม่จากเพื่อนบ้านเท่านั้นดังนั้นจึงบังคับให้W ijเป็นศูนย์ถ้าจุดX jไม่ใช่เพื่อนบ้านของจุดX iและ (b) ผลรวม ของทุกแถวของเมทริกซ์น้ำหนักเท่ากับ 1

A picture containing text, clock, watch

Description automatically generated

จุดข้อมูลเดิมจะถูกเก็บรวบรวมในDพื้นที่มิติและเป้าหมายของขั้นตอนวิธีคือการลดมิติเพื่อdดังกล่าวว่าD >> d น้ำหนักเดียวกันW IJที่ reconstructs ฉัน TH จุดข้อมูลในDมิติจะถูกนำมาใช้เพื่อสร้างจุดเดียวกันในที่ต่ำกว่าdมิติ แผนที่การอนุรักษ์พื้นที่ใกล้เคียงถูกสร้างขึ้นตามแนวคิดนี้ แต่ละจุด X iในพื้นที่มิติDจะถูกจับคู่กับจุด Y iในพื้นที่มิติdโดยการย่อฟังก์ชันต้นทุนให้น้อยที่สุด

A picture containing text, watch, clock

Description automatically generated

ในฟังก์ชันต้นทุนนี้ไม่เหมือนกับฟังก์ชันก่อนหน้านี้น้ำหนัก W ijจะถูกคงไว้และการย่อขนาดจะทำในจุด Y iเพื่อปรับพิกัดให้เหมาะสม ปัญหาการลดนี้สามารถแก้ไขได้โดยการแก้เบาบางN X N Eigen ปัญหาค่า ( Nเป็นจำนวนจุดข้อมูล) ซึ่งด้านล่างdภัณฑ์ไอเกนเวกเตอร์ให้ชุดมุมฉากของพิกัด โดยทั่วไปจุดข้อมูลที่มีการสร้างขึ้นใหม่จากKเพื่อนบ้านที่ใกล้ที่สุดที่วัดจากระยะทางยุคลิด สำหรับการนำไปใช้ดังกล่าวอัลกอริทึมจะมีพารามิเตอร์Kฟรีเพียงตัวเดียวซึ่งสามารถเลือกได้โดยการตรวจสอบความถูกต้องข้าม

* + Isometric Feature Map (ISOMAP)

Isomap [20]คือการรวมกันของอัลกอริทึมฟลอยด์-Warshallกับคลาสสิกหลายมิติขูดหินปูน Classic Multidimensional Scaling (MDS) ใช้เมทริกซ์ของระยะทางแบบคู่ระหว่างจุดทั้งหมดและคำนวณตำแหน่งสำหรับแต่ละจุด Isomap อนุมานว่าระยะทางแบบคู่จะรู้กันระหว่างจุดใกล้เคียงเท่านั้นและใช้อัลกอริทึม Floyd – Warshall เพื่อคำนวณระยะทางแบบจับคู่ระหว่างจุดอื่น ๆ ทั้งหมด สิ่งนี้จะประมาณเมทริกซ์เต็มรูปแบบของระยะทาง geodesic แบบคู่ระหว่างจุดทั้งหมดได้อย่างมีประสิทธิภาพ จากนั้น Isomap จะใช้ MDS แบบคลาสสิกเพื่อคำนวณตำแหน่งมิติที่ลดลงของจุดทั้งหมด Landmark-Isomap เป็นตัวแปรหนึ่งของอัลกอริทึมนี้ที่ใช้จุดสังเกตเพื่อเพิ่มความเร็วโดยต้องเสียค่าใช้จ่ายบางส่วน

ในการเรียนรู้ที่หลากหลายข้อมูลอินพุตจะถูกสุ่มตัวอย่างจากท่อร่วมมิติที่มีมิติต่ำซึ่งฝังอยู่ภายในช่องว่างเวกเตอร์มิติที่สูงขึ้น สัญชาตญาณหลักที่อยู่เบื้องหลัง MVU คือการใช้ประโยชน์จากความเป็นเส้นตรงของท่อร่วมในท้องถิ่นและสร้างแผนที่ที่รักษาพื้นที่ใกล้เคียงในทุกจุดของท่อร่วมที่อยู่เบื้องหลัง

L11 Clustering Basics

* Dissimilarity measures and proximity matrix
* Basic K-means algorithm
* Choosing number of clusters
* Kernel K-means

L12 Clustering Advanced

* Gaussian mixture model
* Hierarchical clustering
* Density-based clustering

L13 Factor Analysis

* Factor model and fitting
* Number of factors and variance explained
* Factor rotation
* Factor scores