**（同學你好！雖然你做了PPT，但是我們還是希望使用Canva來匯報。因此，我把你的內容整合到了Canva的投影片中，你可以打開這個鏈接檢視並熟悉自己的內容，如果你有想法需要修改一些，也請直接在Canva上修改。）**

**（以下第一頁的內容，我會在匯報的一開始就提到，不需要同學再做介紹了。）**

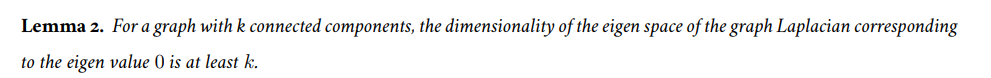
**第一頁**:引言 在人工智能、資訊檢索和數據挖掘領域中，通常會遇到低維數據嵌入高維空間的情況。因此，複雜數據的適當表示是研究人員感興趣的話題。以生物感知為例：刺激信號通常具有高維表示和低維內在結構。如果能將感知到的高維信號恢復為低維信號，將大大有利於生物感知的研究。 關於降維的論文和方法有很多，但大多數方法並未利用數據所在流形的結構。Belkin和Niyogi(2022)提出了拉普拉斯特徵映射方法以及一個新的分析框架，用於基於幾何動機的數據維度降低演算法。 該結果能夠反映流形的內在幾何結構，因為最佳嵌入是由流形上的拉普拉斯貝爾特拉米算子提供的。圖鄰接矩陣的拉普拉斯算子可以看作是對流形上的拉普拉斯貝爾特拉米算子的近似，而數據點的嵌入映射（離散的）來自於對整個流形上定義的特徵映射/自然映射（連續的）的近似。 拉普拉斯特徵映射方法（以下簡稱LE）的幾個特點如下：

1. 在建立映射後，所需的計算相對簡單。建立映射的過程中需要搜索最近鄰域，這一步並不容易。
2. 它保留了局部性，因此對於異常值和噪聲相對不敏感。
3. LE方法降維的副產品是其聚類屬性：當數據在二維平面上表示時，原始的聚類關係很容易顯示出來。這一原理類似於在機器學習和計算機視覺中開發的譜聚類方法。相比之下，全局方法不顯示聚類傾向。然而，並非所有數據集都有意義的聚類，應謹慎使用。
4. 當N較大時，t的選擇變得更加重要。這是因為如果一個點與圖上的大多數點相連，則應使用熱核的參數t來創建“差異模式”，從而區分真正接近和遠距離的點，並達到“保持局部性”的效果。

**第二頁**: 圖中是一個IRIS DATASET 3D散點圖，展示了使用前三個變量的數據分佈。也可觀察出setosa品種的特徵幾乎與另兩種花朵分離，而versicolor和virginica之間則較相近。

（同學，在這裡，因為數據集原有四個維度，請說明清楚三維散點圖選用了哪三個維度。請查字典了解”Sepal”和”Petal”的意思。此外，圖片頁面較空，我又給了一張從另外一個角度觀察的圖片。）

**第三頁，第四頁**: 圖像顯示了**k=10**, **k=30**, **k=50,k=75**, **k=100** 和 **k=150**的KNN方法的結果。

在k=10時，setosa品種幾乎完全分離，而versicolor和virginica之間有一定的重疊。這反映了較小的k值強調了數據的局部結構，這是因為在Laplacian Eigenmaps（LE）方法中，數據點是通過保持小範圍內鄰近點的局部連接性來進行映射的。當設定k=10時，基於鄰接圖的特性，我們可以看到數據點按其地理或特徵相似度被有效地分組。至於為何當k=10時為何能達到顯著的“二分”效果，下下頁的LEMMA會說明。

在**k=30**時，三個品種的分布都開始彼此分離，尤其是Versicolor和Virginica品種，它們在嵌入空間中表現出更清晰的群聚邊界。

在**k=50**時，~~品種setosa仍能夠保持清晰的分離，這表明其在二維空間中的特徵相對獨特且一致~~。然而，versicolor和virginica的點開始在空間中分散

在**k=75**時，這張圖中三個品種的點分佈更加分散。Setosa品種仍然清晰地與其他兩個品種分離，而Versicolor和Virginica品種之間的界限也更為明確。(請明確說出，這是knn分離效果最好的一張圖)

隨著k值增加到k=100和k=150時， Versicolor和Virginica這兩個品種的分類效果顯著變差。~~這一現象可能由於在二維空間中的嵌入丟失了一部分高維數據的結構信息。~~在數據總量為150的情況下，將k設為100~~或150~~意味著（點之間的聯繫非常稠密，而點之間的差異變得相對較小）~~幾乎每個點都與其他所有點相連~~，這對於保持局部特征是不利的。由於未使用熱核，而是採用了均等的連接權重，這進一步弱化了數據點間真實的相似性差異，導致全局連接過度，影響了分類效果的準確性。（當將k設為150時，所有的點都彼此相連，鄰接圖沒有保留任何有意義的信息）

**第五頁**: **引理 2：** 當圖形中存在k個互不相交的連通子圖時，圖拉普拉斯算子的特征值0的維度至少為k，

**證明：** 現在考慮一個具有 k 個連通子圖，通過調整行和列，我們可以得到一個分塊對角拉普拉斯矩陣。一張含有 文字, 圖表, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

根據引理 1，一個完全連通圖的拉普拉斯算子的特徵空間是一維的，且唯一的單位特徵向量是全一向量然後，對於，我們有特徵向量使得 

然後，這些可以被構造成原始拉普拉斯算子對應於特徵值 0 的特徵空間的一個正交基：一張含有 文字, 圖表, 行, 平行 的圖片

自動產生的描述

介紹完了LEMMA之後，我們可以~~深入~~理解當k=10時為何能達到顯著的“二分”效果，根據上面的lemma中說明了當圖形中存在k個互不相交的連通子圖時，圖拉普拉斯算子的特征值0的維度至少為k。這意味著當圖形被分成k個連通子圖時，每個連通分量在低維表徵中可以獨立存在，因此當k設置為10時，可能導致數據在首要的維度上呈現明顯的分割。

**第六頁**: 在IRIS DATASET的PCA分析中，我們僅在第一主成分上觀察到有效的區分。在第二主成分上，三種品類的區分效果不佳。從PCA圖中可見，Setosa品種（綠色點）與其他兩個品種（Versicolor和Virginica）在第一主成分上有很好的分離，而Versicolor和Virginica在第二主成分上有一定的重疊。而和我前述的KNN方法比較，KNN方法的效果依賴於選擇的k值，但k值選的好，像是前述的k=75，其在將原始數據從高維空間還原回其嵌入的低維流形（即原始數據的基本結構或本質特徵）的過程中，（LE）方法表現得更好。