**HW1**

id:110062209

name:簡晟棋

**Function Implement**

**problem 1:**

一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

自動產生的描述

1.用np.concatenate把train\_data跟test\_data串在一起變成all\_data

用pairwise\_distances算出all\_data中各兩點間的距離矩陣distance\_mat

2.對每個test\_data中的點取出

distance\_list = distance\_mat[len(train\_data)+id, :len(train\_data)]，即為該test data point與所有train data point的距離

將distance\_list從小到大sort後

把前k個最近的train\_data point的距離的平均放入k\_near\_distance\_list

3.用test\_label(已轉化成0,1的形式)與k\_near\_distance\_list計算roc\_auc\_score

­­**problem 2:**

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

1.隨機從train\_data取k個點當cluster center

2.用np.concatenate把cluster\_center跟train\_data串在一起變成train\_with\_cluster

用pairwise\_distances算出train\_with\_cluster中各兩點間的距離矩陣distance\_mat

對每個train data point找到離該point最近的cluster center

用cluster\_id紀錄每個point被分到哪個cluster

用cluster\_number紀錄每個cluster有幾個point

用cluster\_center\_tmp紀錄目前分類下的的cluster center(把該cluster每個point的值加起來取平均)

直到cluster center converge為止前持續進行2.

3. 用np.concatenate把cluster\_center跟test\_data串在一起變成test\_with\_cluster

用pairwise\_distances算出test\_with\_cluster中各兩點間的距離矩陣distance\_mat

4.把每個test point與離它最近的cluster之間的距離放入k\_cluster\_min\_distance\_list

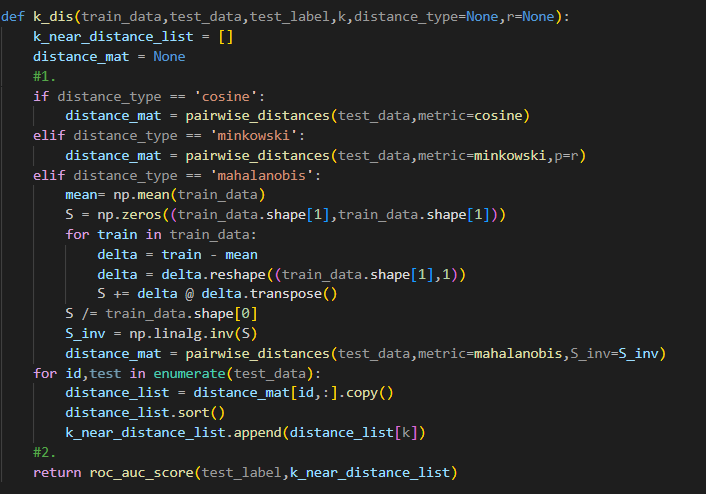
5.用test\_label(已轉化成0,1的形式)與k\_cluster\_min\_distance\_list計算roc\_auc\_score

**problem 3:**

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

先把cosine, minkowski和mahalanobis distance的function準備好



1.

如果distance\_type為’cosine’，用pairwise\_distances(test\_data,metric=cosine)算出每個test\_data中各兩點的cosine distance

如果distance\_type為'minkowski'，用pairwise\_distances(test\_data,metric= minkowski)算出每個test\_data中各兩點的minkowski distance，記得使p=r

如果distance\_type為’mahalanobis'，用pairwise\_distances(test\_data,metric= mahalanobis)加上train\_data的covariance matrix S的inverse把test\_data兩兩間的 mahalanobis distance算出來

用distance\_list取出每個test point與其他test data中的point的距離

將distance\_list從小到大sort後取出第k項，也就是離自己第k近的距離(第0項是自己)

把離自己第k近的距離放入k\_near\_distance\_list

2.用test\_label(已轉化成0,1的形式)與k\_near\_distance\_list計算roc\_auc\_score

**problem 4:**

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

k\_distance:用distance\_mat找出該point\_id與其他點的距離，由小到大sort後取出第k項為離自己第k近的距離(第0項是自己)

reachable\_distance:求出max(point\_o的k\_distance , point\_o和point\_p的距離)

lrd:把tuple(point\_p與該point id的距離,point id)放入distance\_list，從小到大sort後第1~k項的id就是離point\_p最近的k個點 (第0項是自己)

接著求出point\_p的local reachability distance(1/(離point\_p最近的k個點的reachable\_distance的平均))

lof: 把tuple(point\_p與該point id的距離,point id)放入distance\_list，從小到大sort後第1~k項的id就是離point\_p最近的k個點 (第0項是自己)

接著求出point\_p的local outlier factor(離point\_p最近的k個點的local reachability distance的平均/point\_p的local reachability distance)

一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

自動產生的描述

1用test\_data各兩點間距離矩陣distance\_mat與上面的lof求出每個test point的local outlier factor

把point\_id的local outlier factor放入distance\_list

把local outlier factor放入score\_for\_color\_tmp

2.如果drawing是true:

把globel variable test\_data\_draw、test\_label\_draw、score\_for\_color分別放入TSNE(n\_components=2).fit\_transform(test\_data)、test\_label、score\_for\_color\_tmp

3.用test\_label(已轉化成0,1的形式)與distance\_list計算roc\_auc\_score

**Calculate, Record, Drawing**

一張含有 文字, 螢幕擷取畫面, 字型, 設計 的圖片

自動產生的描述

knn\_score:紀錄 K Nearest Neighbor 在不同 k(對應 k1)的結果

kmeans\_score:紀錄 Cluster-based 在不同 k(對應 k2)的結果

Cosine\_dis\_score: 紀錄 Cosine Distance-based 的結果

Minkowski\_dis\_score = 紀錄 Minkowski Distance-based 在不同 r(對應 r)的結果

Mahalanobis\_dis\_score : 紀錄 Mahalanobis Distance-based 的結果

LOF\_score :紀錄 Density-based 的結果

test\_data\_draw: 在 Density-based 中畫圖用的 data

test\_label\_draw:test\_data\_draw 對應的 label

score\_for\_color: test\_data\_draw 對應的 score一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

先把test\_label轉換成0,1的形式

0:normal data(指定digit)

1:anomaly data(其他digit)

將每個digit的運算結果記錄在上面的list中

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

印出每個紀錄roc\_auc\_score的list的平均值

一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

自動產生的描述

分成兩張subplot

左邊用test\_data\_draw作圖，把對應的score\_for\_color當顏色塗上去，並附上colorbar sc

右邊把test\_data\_draw用test\_data\_draw區分成normal(blue)跟anomaly(orange)分別作圖

**Result**

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 螢幕擷取畫面 的圖片

自動產生的描述

**Observation**

**problem 5:**

1.在knn中，雖然不明顯，roc\_auc\_score會隨著k的上升而下降，因為太大的k會使normal point的anomaly score相對於anomaly point上升過多

2.在kmeans中，roc\_auc\_score會隨著k的上升而有明顯上升，因為多個cluster center能讓每個normal point都被分配到合適的cluster使anomaly score下降

3.k-distance除了Minkowski distance以外的roc\_auc\_score是4個方法中最好的，因為能使anomaly point與normal point的anomaly score有所區別

Mahalanobis distance有把不同維度的scale的影響考慮進去，比只考慮夾角大小，沒有考慮距離的Cosine distance表現好一點，而沒有把不同維度的scale的影響考慮進去Minkowski distance表現最差

在Minkowski distance中，r=2時表現最好，接著是r=inf，最後是r=1，因為歐式距離(r=2)比起切比雪夫距離(r=inf)與曼哈頓距離(r=1)較能確切表現兩點間的距離

4.local outlier factor的roc\_auc\_score是4個方法中最低的，因為從圖中可看到normal point與anomaly point的密度差距不夠明顯