- Q1. Assume there are 120 students in our B565 class, and each person has a 2% of chance of carrying coronavirus. We also know that the Omicron variant dominates and let's assume that it accounts for 90% of the new Covid cases. What's the probability that the entire class is free of coronavirus and Omicron, respectively? Show your work. [10 pts]
- i. The probability that the entire class is free of coronavirus:

If each student has a 2% chance of carrying coronavirus, then each student has a 98% chance of not carrying coronavirus.

1-0.02=0.98 is the probability that a student doesn't have the coronavirus.

Hence, if a class of 120 students doesn't have coronavirus, the probability of that happening is **0.98^120=0.08853787272** which is 8.8% approximately. (We multiply the probability of each student not having coronavirus).

ii. The probability that the entire class is free of Omicron:

If each student has a 2% chance of carrying coronavirus, then each student has a 0.02\*0.9=0.018 probability of carrying Omicron variant. The probability of a student not having Omicron is 1-0.018=0.982.

Hence, if a class of 120 students doesn't have coronavirus, the probability of that happening is **0.982^120=0.0.11307810828** which is 11.3% approximately. (We multiply the probability of each student not having Omicron).

Q2. Let  $\Omega$  be the space of possible outcomes of a fair die (with six sides) thrown twice. What is  $\Omega$ ? Let A be the event that a 4 is observed on either individual throw or the sum of both throws is at least 5. Let B be the event that the difference between the two throws is exactly two. Are A and B independent? What is the probability of B given A? What is the probability of A given B? Show your work. [15 pts]

```
i. \Omega={(1, 1),(1, 2),(1, 3),(1, 4),(1, 5),(1, 6),(2, 1),(2, 2),(2, 3),(2, 4),(2, 5),(2, 6),(3, 1),(3, 2),(3, 3), (3, 4),(3, 5),(3, 6),(4, 1),(4, 2),(4, 3),(4, 4),(4, 5),(4, 6),(5, 1),(5, 2),(5, 3),(5, 4),(5, 5),(5, 6), (6, 1),(6, 2),(6, 3),(6, 4),(6, 5),(6, 6)}. Size of \Omega is 36. n(\Omega)=36
```

ii. A=Event when a 4 is observed on either individual throw or the sum of both throws is at least 5

```
A = \{(1, 4), (1, 5), (1, 6), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\} 
n(A) = 36.
```

B=The event that the difference between the two throws is exactly two. B= $\{(1,3),(2,4),(3,5),(4,6),(3,1),(4,2),(5,3),(6,4)\}$ . n(B)=8

## A and B are independent.

Set  $A \cap B = \{(2,4),(3,5),(4,6),(4,2),(5,3),(6,4)\}$ .  $n(A \cap B) = 6$ .

 $P(A)=n(A)/n(\Omega)=30/36=5/6$ 

 $P(B)=n(B)/n(\Omega)=30/36=8/36=2/9$ 

 $P(A \cap B) = n(A \cap B)/n(Ω) = 6/36 = 1/6$ 

## iii. $P(B|A)=(P(A|B)P(B))/P(A)=P(A\cap B)/P(A)=(1/6)/(5/6)=1/5$

## iv. $P(A|B)=(P(B|A)P(A))/P(B)=P(A\cap B)/P(B)=(1/6)/(2/9)=3/4$

Q3. Recall that given two vectors x and y of n dimensions, their cosine similarity, Euclidean distance, and correlation can be computed as follows. [15 pts]

Cosine similarity,  $cos(x, y) = (x \cdot y)/||x||.||y||$ 

Euclidean distance,  $d(x, y) = (\sum n i=1(xi - yi)^2)^1/2$ 

Correlation, corr(x, y) = Sxy/(SxSy)

=(
$$\sum n i=1 (xi - x^-)(yi - y^-)$$
)/( $\sum n i=1 (xi - x^-)^2$ )^1/2.( $\sum n i=1 (yi - y^-)^2$ )^1/2

Sets can also be represented as vectors of zeros and ones, so for those vectors, Jaccard similarity (intersection over union) can be used. For the following vectors x and y, calculate the indicated similarity or distance measures. Show the steps.

- x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation and Euclidean.
- x = (0, 1, 0, 1), y = (1, 0, 1, 0), cosine, correlation, Euclidean, Jaccard
- x = (0, -1, 0, 1), y = (1, 0, -1, 0), cosine, correlation, Euclidean
- x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
- x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1) cosine, correlation

i. 
$$x = (1, 1, 1, 1), y = (2, 2, 2, 2)$$

a. Cosine Similarity:

$$x.y=1.2+1.2+1.2+1.2=8$$
  
 $||x||=(1^2+1^2+1^2+1^2)^{(1/2)}=2$   
 $||y||=(2^2+2^2+2^2+2^2)^{(1/2)}=4$   
 $\cos(x, y) = (x \cdot y)/||x||.||y||=8/(2*4)=1$ 

b. Correlation:

$$\begin{array}{l} x^-=(\sum n \ i=1(xi))/n=(1+1+1+1)/4=1 \\ y^-=(\sum n \ i=1(yi))/n=(2+2+2+2)/4=2 \\ (\sum n \ i=1 \ (xi-x^-)(yi-y^-))=(1-1)^*(2-2)+(1-1)^*(2-2)+(1-1)^*(2-2)+(1-1)^*(2-2)=0 \\ (\sum n \ i=1 \ (xi-x^-)^2)=(1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2=0 \\ (\sum n \ i=1 \ (xi-x^-)^2)^4/2=0 \\ (\sum n \ i=1 \ (yi-y^-)^2)=(2-2)^2+(2-2)^2+(2-2)^2+(2-2)^2=0 \\ (\sum n \ i=1 \ (yi-y^-)^2)^4/2=0 \end{array}$$

Since, both the numerator and the denominator are 0, the correlation for vectors **x** and **y** are not defined.

c. Euclidean distance:

$$\sum n i=1(xi-yi)^2=(1-2)^2+(1-2)^2+(1-2)^2+(1-2)^2=4*(-1)^2=4$$
  
( $\sum n i=1(xi-yi)^2)^1/2=4^(1/2)=2$ 

ii. 
$$x = (0, 1, 0, 1), y = (1, 0, 1, 0)$$

a. Cosine Similarity:

$$x.y=0.1+1.0+0.1+1.0=0$$
  
 $||x||=(0^2+1^2+0^2+1^2)^{(1/2)}=2^{(1/2)}$   
 $||y||=(1^2+0^2+1^2+0^2)^{(1/2)}=2^{(1/2)}$   
 $\cos(x, y) = (x \cdot y)/||x||.||y||=0$ 

b. Correlation:

$$x' = (\sum n i = 1(xi))/n = (0+1+0+1)/4 = 1/2$$
  
 $y' = (\sum n i = 1(yi))/n = (1+0+1+0)/4 = 1/2$ 

$$(\sum n i=1 (xi - x^{-})(yi - y^{-})) = (0-1/2)(1-1/2)+(1-1/2)(0-1/2)+(0-1/2)(1-1/2)+(1-1/2)(0-1/2)=(-1/4)^{*}4=-1$$

(
$$\sum$$
 n i=1 (xi - x^)^2)=(0-1/2)^2+(1-1/2)^2+(0-1/2)^2+(1-1/2)^2=4\*1/4=1 ( $\sum$  n i=1 (xi - x^)^2)^1/2=1

$$\sum_{i=1}^{\infty} n_i = 1 (y_i - y^-)^2 = (1-1/2)^2 + (0-1/2)^2 + (1-1/2)^2 + (0-1/2)^2 = 4*1/4 = 1$$
  
 $(\sum_{i=1}^{\infty} n_i = 1 (y_i - y^-)^2)^1/2 = 1$ 

Hence,

$$corr(x, y) = Sxy/(SxSy) = -1/1*1 = -1$$

c. Euclidean distance:

$$\sum n i=1(xi-yi)^2=(0-1)^2+(1-0)^2+(0-1)^2+(1-0)^2=4$$
  
( $\sum n i=1(xi-yi)^2)^1/2=4^4(1/2)=2$ 

d. Jaccard Coefficient:

iii. 
$$x = (0, -1, 0, 1), y = (1, 0, -1, 0)$$

a. Cosine Similarity:

$$x.y=0.1+-1.0+0.-1+1.0=0$$
  
 $||x||=(0^2+(-1)^2+0^2+1^2)^(1/2)=2^(1/2)$   
 $||y||=(1^2+0^2+(-1)^2+0^2)^(1/2)=2^(1/2)$   
 $\cos(x, y) = (x \cdot y)/||x||.||y||=0$ 

b. Correlation:

$$x^{-}=(\sum n i=1(xi))/n=(0+-1+0+1)/4=0$$

$$\begin{array}{l} y^-=(\sum n\ i=1(yi))/n=(1+0+-1+0)/4=0\\ (\sum n\ i=1\ (xi-x^-)(yi-y^-))\\ =(0)(1)+(-1)(0)+(0)(-1)+(1)(0)=0\\ (\sum n\ i=1\ (xi-x^-)^2)=(0)^2+(-1)^2+(0)^2+(1)^2=2\\ (\sum n\ i=1\ (xi-x^-)^2)=(1)^2+2^4/2\\ (\sum n\ i=1\ (yi-y^-)^2)=(1)^2+(0)^2+(-1)^2+(0)^2=2\\ (\sum n\ i=1\ (yi-y^-)^2)^2+(1/2=2^4/2\\ (\sum n\ i=1\ (yi-y^-)^2)^2+(1/2=2^4/2\\ (\sum n\ i=1\ (yi-y^-)^2)^2+(1/2=2^4/2\\ (\sum n\ i=1\ (yi-y^-)^2)^2+(-1-0)^2+(0-(-1))^2+(1-0)^2=4\\ (\sum n\ i=1(xi-yi)^2)^2+(1/2)=2\\ (\sum n\ i=1(xi))/n=(1+1+0+1+0+1)/6=4/6=2/3\\ (\sum n\ i=1\ (xi-x^-)(yi-y^-))\\ =(1-2/3)(1-2/3)+(1-2/3)(1-2/3)+(1-2/3)+(1-2/3)(0-2/3)+(0-2/3)+(1-2/3)(1-2/3)\\ (\sum n\ i=1\ (xi-x^-)(yi-y^-))\\ =(1-2/3)(1-2/3)^2+(1-2/3)^2+(0-2/3)^2+(1-2/3)^2+(1-2/3)^2+(1-2/3)^2=1/9+1/9+4/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (xi-x^-)^2)^2+(1-2/3)^2+(1-2/3)^2+(0-1/2)^2+(0-1/2)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (yi-y^-)^2)=(1-2/3)^2+(1-2/3)^2+(1-2/3)^2+(0-1/2)^2+(0-1/2)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (yi-y^-)^2)=(1-2/3)^2+(1-2/3)^2+(1-2/3)^2+(0-1/2)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (yi-y^-)^2)=(1-2/3)^2+(1-2/3)^2+(1-2/3)^2+(0-1/2)^2+(0-1/2)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (yi-y^-)^2)=(1-2/3)^2+(1-2/3)^2+(1-2/3)^2+(0-1/2)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+1/9\\ =4/3\\ (\sum n\ i=1\ (yi-y^-)^2+(1-2/3)^2+(1-2/3)^2+(1-2/3)^2+(1-2/3)^2=1/9+1/9+1/9+4/9+1/9\\ =4$$

 $(\sum n i=1 (yi - y^{-})^{2})^{1/2}=(4/3)^{1/2}$ 

 $corr(x, y) = Sxy/(SxSy) = 1/3/(4/3)^1/2*(4/3)^1/2=0.25$ 

Hence,

c. Jaccard Coefficient:

$$JC = (f11)/(f01+f10+f11)=3/(1+1+3)=0.6$$

v. 
$$x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1)$$

a. Cosine Similarity:

$$x.y=2.-1+-1.1+0.-1+2.0+0.0+-3.-1=-2-1+3=0$$
  
 $||x||=(2^2+-1^2+0^2+2^2+0^2+-3^2)^{(1/2)}=18^{(1/2)}$   
 $||y||=(-1^2+1^2+-1^2+0^2+0^2+-1^2)^{(1/2)}=4^{(1/2)}=2$   
 $\cos(x, y) = (x \cdot y)/||x||.||y||=0/(2^{*}(18)^{*}1/2)=0.0$ 

b. Correlation:

$$\begin{array}{l} x^-=(\sum n \ i=1(xi))/n=(2-1+0+2+0-3)/6=0 \\ y^-=(\sum n \ i=1(yi))/n=(-1+1-1+0+0-1)/6=-2/6 \\ \\ (\sum n \ i=1 \ (xi-x^-)(yi-y^-)) \\ =(2)(-1+2/6)+(-1)(1+2/6)+(0)(1+2/6)+(2)(-2/6)+(0)(+2/6)+(-3)(-1+2/6)=0 \\ \\ (\sum n \ i=1 \ (xi-x^-)^2)=(2)^2+(-1)^2+(0)^2+(2)^2+0^2+(-3)^2=18 \\ (\sum n \ i=1 \ (xi-x^-)^2)^4/2=18^4/2 \\ (\sum n \ i=1 \ (yi-y^-)^2)=(-1+2/6)^2+(1+2/6)^2+(1+2/6)^2+(-2/6)^2+(-2/6)^2+(-1+2/6)^2=10/3 \\ (\sum n \ i=1 \ (yi-y^-)^2)^4/2=(10/3)^4/2 \\ \\ \text{Hence,} \end{array}$$

 $corr(x, y) = Sxy/(SxSy) = 0/((18^{1/2})*((10/3)^{1/2})) = 0$ 

con-001 concon-003 concon-005 5 rows × 573 columns In [2]: microbiome.shape Out[2]: (344, 573) microbiome.describe() In [3]: Out[3]:  $k\_Archaea|p\_Euryarchaeota|c\_Methanobacteria|o\_Methanobacteriales|f\_Methanobacteriacea|$ count mean std min 25% 50% **75**% 8 rows × 572 columns In [4]: class\_column = microbiome['Class'].tolist() In [5]: | microbiome = microbiome.drop('Class', 1) In [6]: from sklearn.decomposition import PCA pca = PCA(n\_components=2) pca.fit(microbiome) #the results will show that PC1 dominates for the original data print("variance explained", pca.explained\_variance\_ratio\_, "singular\_val ues", pca.singular\_values\_) print(pca.explained\_variance\_ratio\_.sum()) #transform the data according to the PCA results microbiome\_transformed = pca.transform(microbiome) variance explained [0.29922075 0.06936047] singular\_values [337.41686876 162.45270412] 0.3685812133629984 Using PCA, the data matrix is projected to its first two principal components. The projected values of the original data are stored in a pandas DataFrame object named projected. In [7]: projected = pd.DataFrame(microbiome\_transformed, columns=['pc1', 'pc2'], in dex=range(1,345))projected['Class'] =class\_column projected Out[7]: pc2 Class pc1 **1** -2.943049 -2.248436 n **2** -7.215752 1.836456 **3** 24.051681 -4.850619 n **4** -8.684207 6.431572 n **5** 2.565706 -3.969367 n **340** -7.294597 1.556832 n **341** -2.864927 18.011063 **342** -8.629940 -7.939053 **343** -6.603719 -2.611686 **344** -8.060710 -9.187658 344 rows × 3 columns In [8]: import matplotlib.pyplot as plt colors = {'n':'b', 't2d':'r'} markerTypes = {'n':'+', 't2d':'x'} for classType in markerTypes: d = projected[projected['Class']==classType] plt.scatter(d['pc1'],d['pc2'],c=colors[classType],s=60,marker=marker Types[classType]) plt.xlabel("PC1") plt.ylabel("PC2") Out[8]: Text(0, 0.5, 'PC2') 40 30 20 2 10 -10-20 -30 Ó 20 60 40 PC1 Finally, we draw a scatter plot to display the projected values. We can see that the class 't2d' values overlap with the 'n' class values, in many cases. Hence, there is not much variability captured using PCA with 2 components. From the given output, the variance is very low. Hence, PCA is a good approach to reduce this dataset. Now, we will perform t-SNE on this dataset. In [9]: #t-sne: from sklearn.manifold import TSNE from numpy import reshape import seaborn as sns microbiome1 = pd.read\_csv("T2D\_abundance.csv", delim\_whitespace=True) microbiome1.head() Out[9]: k\_\_Archaea|p\_\_Euryarchaeota|c\_\_Methanobacteria|o\_\_Methanobacteriales|f\_\_Methanobacteriaceae con-001 con-002 con-003 con-004 con-5 rows × 573 columns In [10]: class\_column1 = microbiome1['Class'].tolist() class\_column1 Out[10]: ['n', 'n', 'n', 'n' 'n', 'n' 'n', 't2d', 'n', 't2d', 't2d' 't2d' 't2d' 't2d' 't2d', 'n', 'n', 'n', 'n', 'n', 'n' 'n', 'n' 'n', 'n', 'n', 'n', 'n', 'n' 'n' 'n', 'n', 'n', 'n', 'n', 'n'] In [11]: microbiome1 = microbiome1.drop('Class', 1) In [12]: x = microbiome1y = class\_column1 Then, we'll define the model by using the TSNE class, here the n components parameter defines the number of target dimensions. In [13]: tsne = TSNE(n\_components=2, verbose=1, random\_state=123)  $z = tsne.fit_transform(x)$ [t-SNE] Computing 91 nearest neighbors... [t-SNE] Indexed 344 samples in 0.069s... [t-SNE] Computed neighbors for 344 samples in 0.241s... [t-SNE] Computed conditional probabilities for sample 344 / 344 [t-SNE] Mean sigma: 9.567743 [t-SNE] KL divergence after 250 iterations with early exaggeration: 67.8 [t-SNE] KL divergence after 1000 iterations: 0.865274 Now, we will visualize the data. In [14]: df = pd.DataFrame() df["y"] = ydf["comp-1"] = z[:,0]df["comp-2"] = z[:,1]sns.scatterplot(x="comp-1", y="comp-2", hue=df.y.tolist(),palette=sns.co lor\_palette("hls", 2), data=df).set(title="Microbiome Data T-SNE projecti Out[14]: [Text(0.5, 1.0, 'Microbiome Data T-SNE projection')] Microbiome Data T-SNE projection 20 10 -10-20 n t2d 10 15 comp-1 t-SNE offers slightly better separation of the classes and there is not much overlap between the data points. It offers good dimensionality reduction too. Q5. The plot below was used to demonstrate the Curse of Dimensionality. Implement a code to simulate your own data, and generate your special plot of curse of dimensionality. Try dimensions from 2 to 50 with a step size of 1. And for each dimension, randomly generate 500 data points. Use Euclidean distance. [30 pts] We are assuming A is the set of points given and B is the reference point, if you compute the distance from B to each of the points in B, the difference between the maximum distance and the minimum distance goes to 0. In [21]: import numpy as np import matplotlib.pyplot as plt import math log\_values = [] for step in range(2,51): # Generate 500 random points in N dimensions.  $A = [np.random.randint(-50, 50, step) for _ in range(500)]$ B = np.random.randint(-50, 50, step)# Computing distances as per the given problem #https://www.delftstack.com/howto/numpy/calculate-euclidean-distanc diffs = [np.linalg.norm(a-B) for a in A] mxd = max(diffs)mnd = min(diffs) #taking log of whole equation log\_value = math.log10((mxd-mnd)/mnd) log\_values.append( log\_value ) # Plotting the curve for simulated data using matplotlib plt.plot(range(2,51),log\_values) plt.xlabel('Number of dimensions') plt.ylabel('log10( (MAX\_DIST - MIN\_DIST )/MIN\_DIST )') plt.show() 1.75 /MIN\_DIST) 1.50 1.25 1/( LSIQ NIW - LSIQ XWW) 0.75 0.50 0.25 0.00 -0.25 10 20 30 Number of dimensions If we take log of only the numerator in the function for the y axis, we get a better and much smoother curve. In [22]: import numpy as np import matplotlib.pyplot as plt import math log\_values = [] for step in range(2,51): # Generate 500 random points in N dimensions.  $A = [np.random.randint(-50, 50, step) for _ in range(500)]$ B = np.random.randint(-50,50,step) # Computing distances as per the given problem #https://www.delftstack.com/howto/numpy/calculate-euclidean-distanc diffs = [np.linalg.norm(a-B) for a in A] mxd = max(diffs)mnd = min(diffs) #taking log of whole equation log\_value = math.log10(mxd-mnd)/mnd log\_values.append( log\_value ) # Plotting the curve for simulated data using matplotlib plt.plot(range(2,51),log\_values) plt.xlabel('Number of dimensions') plt.ylabel('log10( (MAX\_DIST - MIN\_DIST )/MIN\_DIST )') plt.show() ( 2.00 · 1.75 · 1.75 · 1.50 · 1.25 · ₹ 1.00 0.75 (MAX\_DIST\_ 0.50 (0.50 0.00 20 30 Number of dimensions References: [1]. https://stackoverflow.com/questions/69247773/generating-a-curse-of-dimensionality-chart [2]. https://www.datatechnotes.com/2020/11/tsne-visualization-example-in-python.html [3]. <a href="https://www.kaggle.com/morecoding/scalingpca-demo/">https://www.kaggle.com/morecoding/scalingpca-demo/</a> [4]. <a href="https://www.kaggle.com/morecoding/datapreprocessing">https://www.kaggle.com/morecoding/datapreprocessing</a> [5]. https://www.delftstack.com/howto/numpy/calculate-euclidean-distance/

In [ ]:

Q4. Analyze a microbiome dataset. [30 pts]

First, we will perform PCA on this dataset

• This dataset includes the microbiome profiles of 344 people, some with type 2 diabetes, and others without. The microbiome profile for a person stores the relative abundance of different bacterial specie found in the stool sample collected from that person. The last column shows the

• Perform PCA and t-SNE on the dataset and visualize the data in 2D space. In the plots, each

• Report what you learn from the PCA analyses. How much variability of the data is captured by using only two dimensions? Is PCA a good approach for dimensionality reduction for this dataset?

• Does t-SNE result in a good dimensionality reduction of this dataset? Why or why not

microbiome = pd.read\_csv("T2D\_abundance.csv", delim\_whitespace=True)

 $k\_Archaea|p\_Euryarchaeota|c\_Methanobacteria|o\_Methanobacteriales|f\_Methanobacteriaceae$ 

class (with diabetes or not), and the other columns are for the relative abundances.

Do you see clusters of people according to their disease status?

• The dataset is available here.

data point is a user.

In [1]: import pandas as pd

Out[1]:

import numpy as np

microbiome.head()