HOMEWORK 2

Username: Ameya Dalvi email: abdalvi@iu.edu

1) 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?

Answer:

The probability of a person having Coronavirus P(C) is **0.02.**

Probability of a student not having Coronavirus would be

= 1 - P(C)
= 1 – 0.02
= **0.98**

For all 120 students to be free of Coronavirus, **all 120 students** should have probability of 0.98.
Hence, the probability that entire class is free of coronavirus is $\mathbf{(0.98)^{120}}$

We know that 90% of new covid cases are Omicron. Hence, for a person infected with coronavirus the probability of that variant being Omicron **P(O)** would be

P(O) = (0.9) * P(C)

Thus, P(O) = (0.9) * (0.02)
= **0.018**

Now, probability of that variant not being Omicron would be

= 1 – P(O)
= 1 – 0.018
= **0.982**

Hence, the probability of entire class being free of Omicron would be $\mathbf{(0.982)^{120}}$

2) Let Ω be the space of possible outcomes of a fair die (with six sides) thrown twice. What is Ω? 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?

Answer:

Sample space Ω for all possible outcomes of a fair die can be given as:

(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)

All possible outcomes of a fair die with six sides thrown twice = 36

Now, the event that 4 is observed on either individual throw **(Event1):**

{(1,4), (2,4), (3,4), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6),(5,4), (6,4)}

N(event that 4 is observed on either individual throw) = 11

P(event that 4 is observed on either individual throw) = 11/36

And, the sum of both throws is at least 5 **(Event2):**:

{(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(sum of both throws is at least 5) = 30

P(sum of both throws is at least 5) = 30/36

N(Event1 ∩ Event2) = 11
P(Event1 ∩ Event2) = 11/36

Let A be the event that a 4 is observed on either individual throw or the sum of both throws is at least 5:

$$P(A) = P(\text{Event1} \cup \text{Event2}) = P(\text{Event1}) + P(\text{Event2}) - P(\text{Event1} \cap \text{Event2})$$
$$= 11/36 + 30/36 - 11/36$$
$$= 30/36$$

Let B be the event that the difference between the two throws is exactly two:

$$B = \{(1,3), (2,4), (3,1), (3,5), (4,2), (4,6), (5,3), (6,4)\}$$

$N(B) = 8$, $P(B) = 8/36$

To check: Are A and B independent?

A and B will be independent if $\mathbf{P(A \cap B) \neq P(A) * P(B)}$

$\mathbf{P(A)*P(B)} = 30/36 * 8/36$
$= 5/6 * 2/9$
$= 10/54$

$\mathbf{P(A \cap B)} = 11/36$
Since**,** $\mathbf{P(A \cap B) \neq P(A) * P(B),}$ events A and B are **not independent.**

Now, $(A \cap B) = \{(2,4), (3,5), (4,2), (4,6), (5,3), (6,4)\}$
$N(A \cap B) = 6$, $P(A \cap B) = 6/36 = 1/36$

Probability of B given A, $P(B|A) = P(A \cap B) / P(A) = (1/6) / (30/36)$
$= 1/6 * 36/30$
$= 1/5 = 0.2$

Probability of A given B, $P(A|B) = P(A \cap B) / P(B) = (1/6) / (8/36)$
$= 1/6 * 36/8$
$= 3/4 = 0.75$

3) 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

Cosine Similarity: ***cos(x,y)*** **=** $\frac{x}{||x||} * \frac{y}{||y||}$

Correlation: ***corr(x,y)*** **=** $\frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} * \sqrt{\sum_{i=1}^{n}(y-\bar{y})^2}}$

Euclidean Distance: ***d(x,y)*** **=** $\sqrt{(\mathbf{x}-\mathbf{y}) * (\mathbf{x}-\mathbf{y})}$

Jaccard Similarity:

$$\frac{X_1 Y_1}{X_1 Y_1 + X_1 Y_0 + X_0 Y_1}$$

x = (1, 1, 1, 1), y = (2, 2, 2, 2)

$\bar{x}$ = (1+1+1+1)/4 = 4/4 = 1

$\bar{y}$ = (2+2+2+2)/4 = 8/4 = 2

Cosine Similarity:

$$= \frac{(1*2)+(1*2)+(1*2)+(1*2)}{\sqrt{1^2+1^2+1^2+1^2} * \sqrt{2^2+2^2+2^2+2^2}}$$

$$= \frac{8}{\sqrt{4} * \sqrt{16}}$$

$$= \frac{8}{2 * 4}$$

$$= 1$$

Correlation:

$$= \frac{(0)+(0)+(0)+(0)}{\sqrt{0+0+0+0} * \sqrt{0+0+0+0}}$$

$$= \frac{\mathbf{0}}{\mathbf{0}}$$ **(Undefined)**

Euclidean Distance:

$$= \sqrt{(1-2)^2 + (1-2)^2+(1-2)^2 + (1-2)^2}$$

$$= \sqrt{(-1)^2 + (-1)^2+(-1)^2 + (-1)^2}$$

$$= \sqrt{1+1+1+1}$$

$$= \mathbf{2}$$

x = (0, 1, 0, 1), y = (1, 0, 1, 0)

$\bar{x}$ = (0+1+0+1)/4 = 2/4 = 0.5 $\quad$ $\bar{y}$ = (1+0+1+0)/4 = 2/4 = 0.5

Cosine Similarity:

$$= \frac{(0*1)+(1*0)+(0*1)+(1*0)}{\sqrt{0^2+1^2+0^2+1^2} * \sqrt{1^2+0^2+1^2+0^2}}$$

$$= \frac{0}{\sqrt{2} * \sqrt{2}}$$

= 0

Correlation:

$$= \frac{((0-0.5)(1-0.5))+((1-0.5)(0-0.5))+((0-0.5)(1-0.5))+((1-0.5)(0-0.5))}{\sqrt{((0-0.5)(0-0.5))+((1-0.5)(1-0.5))+((0-0.5)(0-0.5))+((1-0.5)(1-0.5))} * \sqrt{((1-0.5)(1-0.5))+((0-0.5)(1-0.5))+((1-0.5)(1-0.5))+((0-0.5)(1-0.5))}}$$

$$= \frac{(-0.25)+(-0.25)+(-0.25)+(-0.25)}{\sqrt{(0.25)+(0.25)+(0.25)+(0.25)} * \sqrt{(0.25)+(0.25)+(0.25)+(0.25)}}$$

$$= \frac{\mathbf{-1}}{\sqrt{1} * \sqrt{1}}$$

= -1

Euclidean Distance: *d(x,y)* = $\sqrt{(\mathbf{x}-\mathbf{y}) * (\mathbf{x}-\mathbf{y})}$

$$= \sqrt{(0-1)^2 + (1-0)^2+(0-1)^2 + (1-0)^2}$$

$$= \sqrt{(-1)^2 + (1)^2+(-1)^2 + (1)^2}$$

$$= \sqrt{1+1+1+1}$$

= **2**

Jaccard:

	$x = 0$	$x = 1$
$y = 0$	0	2
$y = 1$	2	0

$= \frac{0}{0+2+2}$

$= \mathbf{0}$

x = (0, -1, 0, 1), y = (1, 0, -1, 0)

$\bar{x}$ = (0-1+0+1)/4 $\quad$ $\bar{y}$ = (1+0-1+0)/4
= 0/4 = 0 $\quad$ = 0/4 = 0

Cosine Similarity:

$$= \frac{(0*1)+(-1*0)+(0*(-1))+(1*0)}{\sqrt{0^2+1^2+0^2+1^2} * \sqrt{1^2+0^2+1^2+0^2}}$$

$$= \frac{0}{\sqrt{2} * \sqrt{2}}$$

= 0

Correlation:

$$= \frac{((0-0)(1-0))+((-1-0)(0-0))+((0-0)(-1-0))+((1-0)(0-0))}{\sqrt{((0-0)(0-0))+((-1-0)(-1-0))+((0-0)(0-0))+((1-0)(1-0))} * \sqrt{((1-0)(1-0))+((0-0)(1-0))+((-1-0)(-1-0))+((0-0)(1-0))}}$$

$$= \frac{(0)+(0)+(0)+(0)}{\sqrt{(0)+(1)+(0)+(1)} * \sqrt{(1)+(0)+(1)+(0)}}$$

$$= \frac{\mathbf{0}}{\sqrt{2} * \sqrt{2}}$$

= 0

Euclidean Distance:

$$= \sqrt{(0-1)^2 + (-1-0)^2+(0+1)^2 + (1-0)^2}$$

$$= \sqrt{(-1)^2 + (-1)^2+(1)^2 + (1)^2}$$

$$= \sqrt{1+1+1+1}$$

= **2**

x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1)

$\bar{x}$ = (1+1+0+1+0+1)/6 = 4/6 = 2/3 = 0.67

$\bar{y}$ = (1+1+1+0+0+1)/6 = 4/6 = 2/3 = 0.67

Cosine Similarity:

$$= \frac{(1*1)+(1*1)+(0*1)+(1*0)+(0*0)+(1*1)}{\sqrt{1^2+1^2+0^2+1^2+0^2+1^2} * \sqrt{1^2+1^2+1^2+0^2+0^2+1^2}}$$

$$= \frac{3}{\sqrt{4} * \sqrt{4}}$$

$$= \frac{3}{2 * 2} = \frac{3}{4} = 0.75$$

Correlation:

$$= \frac{((1-0.67)(1-0.67))+((1-0.67)(1-0.67))+((0-0.67)(1-0.67))+((1-0.67)(0-0.67))+((0-0.67)(0-0.67))+((1-0.67)(1-0.67))}{\sqrt{(1-0.67)^2+(1-0.67)^2+(0-0.67)^2+(1-0.67)^2+(0-0.67)^2+(1-0.67)^2} * \sqrt{(1-0.67)^2+(1-0.67)^2+(1-0.67)^2+(0-0.67)^2+(0-0.67)^2+(1-0.67)^2}}$$

$$= \frac{\mathbf{0.33}}{\mathbf{0.67}*\sqrt{3} * 0.67*\sqrt{3}}$$

= 0.25

Jaccard:

	$x = 0$	$x = 1$
$y = 0$	1	1
$y = 1$	1	3

$$= \frac{3}{3+1+1}$$

= 3/5 = 0

x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1)

$\bar{x}$ = (2-1+0+2+0-3)/6 = 0

$\bar{y}$ = (-1+1-1+0+0-1)/6 = -2/6 = -1/3

Cosine Similarity:

$$= \frac{(2*(-1))+((-1)*1)+(0*(-1))+(2*0)+(0*0)+((-3)*(-1))}{\sqrt{2^2+(-1)^2+0^2+2^2+0^2+-3)^2} * \sqrt{(-1)^2+1^2+(-1)^2+0^2+0^2+(-1)^2}}$$

$$= \frac{0}{\sqrt{4} * \sqrt{4}}$$

$$= \frac{0}{2 * 2} = 0$$

Correlation:

$$= \frac{((2-0)(-1+0.33))+((-1-0)(1+0.34))+((0-0)(-1+0.33))+((2-0)(0+033))+((0-0)(0+0.33))+((-3-0)(-1+0.33))}{\sqrt{(2-0)^2+(-1-0)^2+(0-0)^2+(2-0)^2+(0-0)^2+(-3-0)^2} * \sqrt{(-1+0.33)^2+(1+0.33)^2+(-1+0.33)^2+(0+0.33)^2+(0+0.33)^2+(-1+0.33)^2}}$$

$$= \frac{\mathbf{0}}{\boldsymbol{Denom}}$$

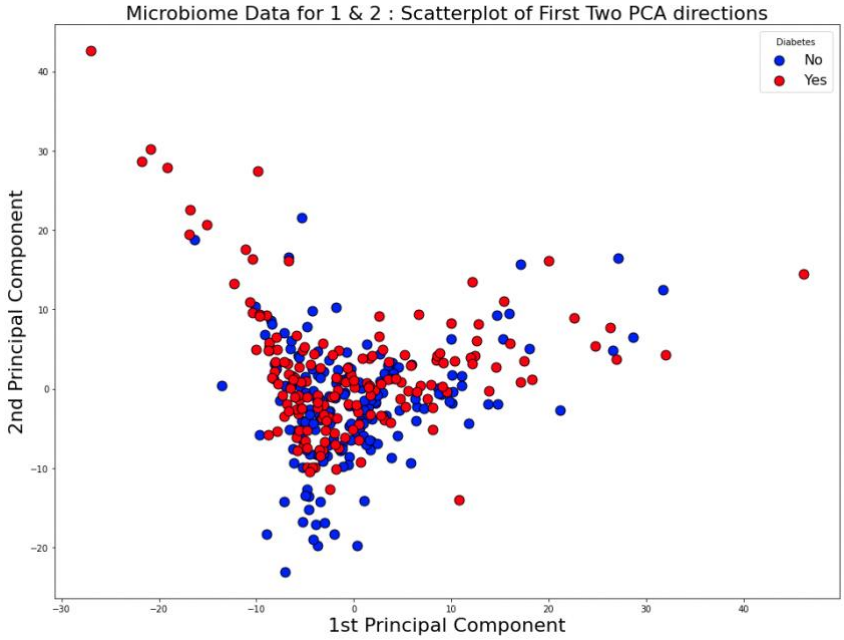
= 0

4) Analyze a microbiome dataset:

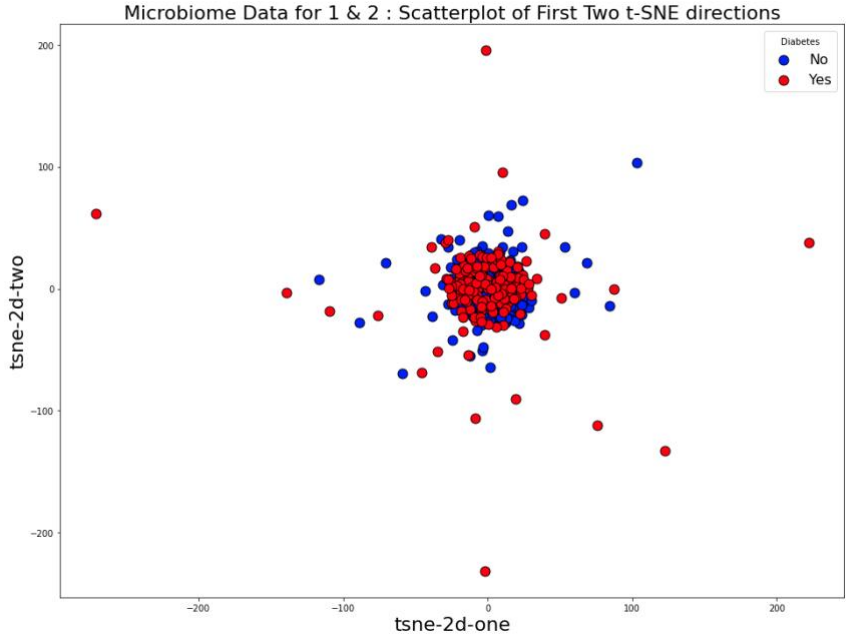
Perform PCA and t-SNE on the dataset and visualize the data in 2D space. In the plots, each data point is a user.

Answer:

2-D visualization of data after performing PCA:



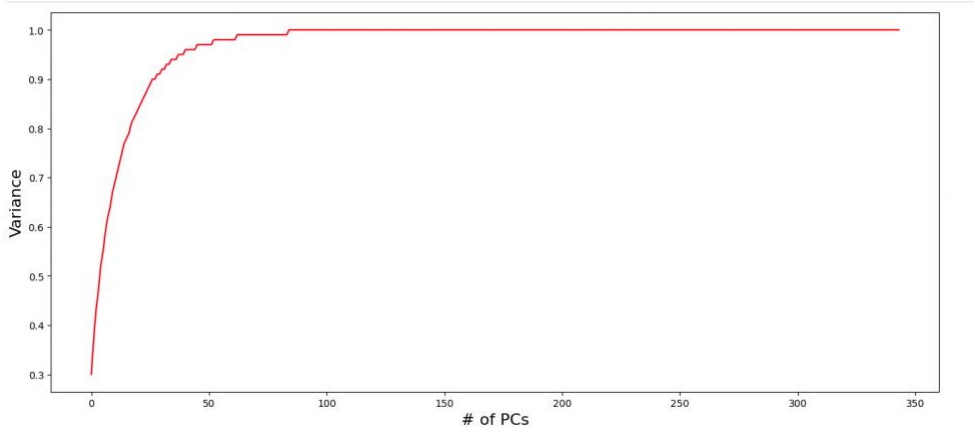
2-D visualization of data after performing t-SNE:



Report what you learn from the PCA analyses. Is PCA a good approach for dimensionality reduction for this dataset?
Answer:

Principal Components Analysis (PCA) *is an algorithm to transform the columns of a dataset into a new set of features called Principal Components. By doing this, a large chunk of the information across the full dataset is effectively compressed in fewer feature columns. This enables dimensionality reduction and ability to visualize the separation of classes or clusters if any. Source: https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/*

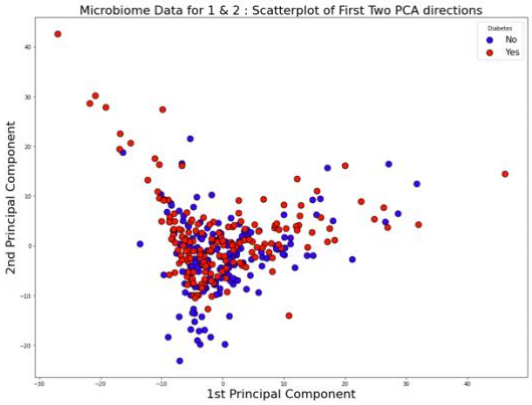
For the above dataset, we can consider PCA as a good dimensionality reduction approach as out of the 574 total features, the top 40 principal components are able to explain almost 95% variance in the dataset.



Hence reducing the dimensionality to 40 will give us the same result for the whole dataset. Reducing the dimensions and correlating the data would summarize whether a particular individual is diabetic or not.

How much variability of the data is captured by using only two dimensions? Do you see clusters of people according to their disease status?
Answer:

Using only two of the top principal components, we can see a decent amount of variability in the data. Yes, we are able to see clusters of Diabetic vs Non Diabetic people but their differentiation isn't clear. That could be the case due to lower dimensional visualization.



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

t-distributed stochastic neighbor embedding *(**t-SNE**) is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.*

t-SNE is a good dimensionality reduction technique for datasets with extremely high dimensions.

In our dataset, t-SNE first two principal components do a good job of clustering the data points.

5. 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.

Answer:

```
import numpy as np
import matplotlib.pyplot as plt
import os
import math

deltas = []
for N in range(2,50,1):
    P = [np.random.randint(-100, 100, N) for _ in range(500)]
    Q = np.random.randint(-100,100,N)
    diffs = [np.linalg.norm(p-Q) for p in P]
    mxd = max(diffs)
    mnd = min(diffs)
    delta = math.log10(mxd-mnd)/mnd
    deltas.append( delta )

plt.plot(range(2,50,1),deltas)
plt.show()
```