Your name:

B565 HW2 (Fall 2022)

Submission instructions

Submit a PDF file to canvas. The PDF file includes your answers to the calculation problems, results of your exploratory analysis of the microbiome, and results from your simulation. In addition, submit your code/notebook for Q4 and Q5 to github.iu under HW2 folder in your B565 repository.

Questions

1. There are 150 students in our B565 class. Assume each student has a 2% of chance of carrying coronavirus. We also know that the Omicron variants dominate and account for most of the new cases (95%). What's the probability that the entire class is free of coronavirus and Omicron variants, respectively? Show your work. [10 pts]

2. You are given a set of m objects that is divided into G groups, where the i group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement is used). Show a couple of scenarios (or analysis goals), in which you will prefer one strategy over the other. [15 pts]

   (a) You randomly select $n \times m_i / m$ elements from each group.

   (b) You randomly select n elements from the data set, without regard for the group to which an object belongs.

3. Recall that given two vectors $\mathbf{x}$ and $\mathbf{y}$ of n dimensions, their cosine similarity, Euclidean distance and correlation can be computed as following. [15 pts]

   Cosine similarity, $\cos(\mathbf{x}, \mathbf{y}) = \dfrac{\mathbf{x}}{\|\mathbf{x}\|} \cdot \dfrac{\mathbf{y}}{\|\mathbf{y}\|}$

   Euclidean distance, $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})}$

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

   Sets can also be represented as vectors of zeros and ones, so for those vectors, Jaccard similarity (intersection over union) can be used.

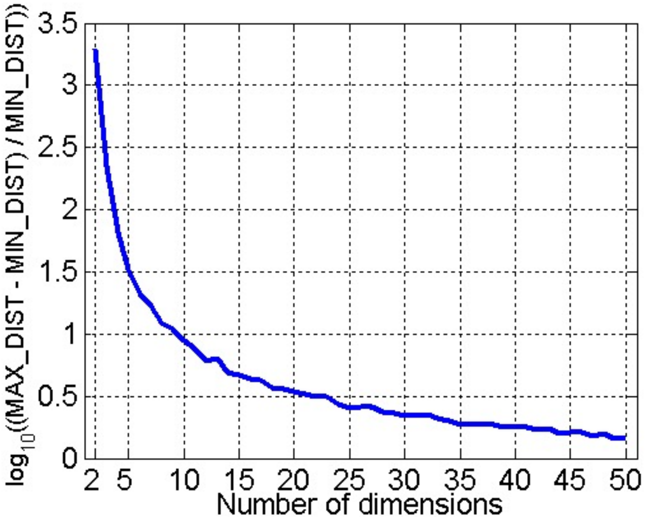
   For the following vectors $\mathbf{x}$ and $\mathbf{y}$, calculate the indicated similarity or distance measures. Show the steps.

   - $\mathbf{x} = (1, 0, 0, 1, 1, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Euclidean, Jaccard
   - $\mathbf{x} = (1, -2, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 2, -1, 0, 0, -1)$ cosine, correlation, Euclidean

4. Analyze a microbiome dataset. [30 pts]

   - The dataset is available here.
   - 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 class (with diabetes or not), and the other columns are for the relative abundances.
   - Perform PCA and t-SNE on the dataset and visualize the data in 2D space. In the plots, each data point is a user.
   - 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? Do you see clusters of people according to their disease status?
   - Does t-SNE result in a good dimensionality reduction of this dataset? Why or why not.

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. [30 pts]


$\log_{10}((\text{MAX\_DIST} - \text{MIN\_DIST}) / \text{MIN\_DIST})$
Number of dimensions