CS 4780/6780: Fundamentals of Data Science © Spring 2019

Homework 2: Data Simulation

Instructor: Daniel L. Pimentel-Alarcón Due 02/19/2019

In this homework you will simulate and visualize data using Matlab (or your preferred coding language).

The ambient dimension of the data will be D = 4. You may choose the number of samples N; I suggest at least N = 100. The 4 rows in this matrix will contain information about height, weight, glucose level, and a label in $\{0, 1\}$ indicating healthy or diabetic.

(a) Generate a 4×100 data matrix $\mathbf{X}$, initially populated with zeros.

(b) Populate the first row with i.i.d. random variables $\mathcal{N}(165, 25)$ (simulating height), and the second row with i.i.d. random variables $\mathcal{N}(137, 100)$ (simulating weight).

(c) Does this data make sense? Are all values reasonable? How would you *clean/preprocess* this data?

(d) Visualize/plot your preprocessed data (in two dimensions).

(e) Compute the weight/height ratio of each sample, and show its histogram.

(f) Model/simulate the glucose level of each individual as a noisy version of its weight/height ratio (ratio + noise), and store this value in the third row of your data matrix $\mathbf{X}$. Let the noise be i.i.d. $\mathcal{N}(0, \sigma^2)$.

(g) We will model an individual as healthy (label = 0) if its glucose level (as defined above) is below a threshold τ, and diabetic otherwise (label = 1). Store these labels in the fourth row of your data matrix $\mathbf{X}$.

(h) Visualize/plot the clustered data (in two dimensions) for different values of σ and τ.

(i) How do σ and τ affect the data?