Due Date: See Webcampus
How to submit: Webcampus

HW1-1. Data Preprocessing
a. What is the main difference between sampling and Feature selection? What is the main similarity between them?

b. What is the main difference between feature selection and dimensionality reduction? What is the main similarity between them?

c. Given a number x = 480 in the range of [-100, 9990], we need to normalize and project the number into a new range [-1, 1]. What is the new value of x if we use decimal scaling for normalization? What is the new value of x if we use min-max normalization?

HW1-2. You are given a set of *m* objects that is divided into *K* groups, where the *i*th 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.)

(a) We randomly select $n * m_i/m$ elements from each group. stratified sampling

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

HW1-3. Curse of Dimensionality
Reproduce the figure in slide 37 in Chapter 2. Randomly generate 500 points under a given dimension, and then compute difference between max and min distance between any pair of points. Plot $\log_{10}\frac{\max-\min}{\min}$ under different number of dimensions.

HW1-4. Sampling

Given a set of data consisting of a small number of almost equal sized groups, find at least one representative point for each of the groups. Assume that the objects in each group are highly similar to each other, but not very similar to objects in different groups.

(a) Assume we have 10 independent groups, provide a formula to estimate the probability that there is at least one object from each of 10 groups.

(b) Plot the probability under different sample sizes.