Grant Davis
CS 458
Homework 1

HW1-1. Data Preprocessing

a. What is the main difference between sampling and Feature selection? What is the main similarity between them?

The main difference is feature selection will reduce the noise from the data set using only relevant features, while sampling does not reduce any features of the dataset. The main similarity is that both methods sample a given part of the dataset.

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

The main difference is that dimensionality reduction transforms features into a lower dimension, while feature selection simply excludes given features without transforming them. The main similarity is that both are methods for reducing the number of features in the dataset.

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?

Decimal scaling normalization: x = 480/10000 = 0.048
Min-Max normalization: x = (480 - (- 100))/ (9990 -(-100)) (1-(-1) + (-1) = -0.89

HW1-2. You are given a set of m objects that is divided into K groups, where the ith group is of size mi. 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 ∗ mi/m elements from each group.
(b) We randomly select n elements from the data set, without regard for the group to which an object
belongs.

(a) It is a proportional sampling and is proportionate. The sample from every cluster is proportional to its size relative to the whole variety of objects.
(b) It may be an easy random sampling theme. Proportional sampling generally has a benefit in the case once the objects in each group are undiversified.

HW1-3. 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

(a) The probability that there is at least one object from each group is 1-(1-1/10)^10 which approximately equals 0.65, which is a 65% chance that we will have one from each group.
(b) The probability increases as the number of groups and size of the groups increase. The probability decreases as the similarity between objects in different groups increases. The probability of one object from each group of 100 groups is 1-(1-1/100)^100 equals about 0.99 or 99% chance we have one from each group. With 10 groups each with 10 objects the probability that at least one object from each group is 1 -(1-1/10)^100 equals about 0.99 or 99% chance we have one from each group. Having 10 groups with 10 objects each that are very similar to each other, the probability that we will have at least one object from each group is 1-1(1-1/10)^10 which equals about 0.1 or 10% chance we have one from each group.