Michael DesRoches

CS 458

CS458 - Homework 1

13 September 2020

HW1-1. Data Preprocessing

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

Sampling is a technique that is used for data selection. Instead of using an entire data set, sampling uses a small part of that data set to save time and money. Feature selection also reduces the dimensionality of data just like sampling but the difference is the type of data retrieved. Sampling uses a subset contained in the data set that closely resembles the actual data set by use of that data's mean or in other words "Representative Data." Feature selection, from what I read in the book and PowerPoint, selects certain data based on criteria of the data sets using certain techniques which are beyond the scope of this question. The criteria, however, are certain attributes or descriptions of that data that are selected before the data mining algorithm is executed.

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

They both are a subset of the data set. The difference, however, is that dimensionality reduction captures the projection of the largest amount of variation. I believe I have already described Feature Selection to the best of my ability in the previous question.

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 for normalization: $Data\ value = v_i / 10^j$

$-100 / 10000 = -0.01$

$-9990 / 10000 = 0.999$

$x = 480/10000 = 0.048$

For studying purposes in future I have attached a few links where I acquired a better understanding of the used formula.

https://www.geeksforgeeks.org/data-normalization-in-data-mining/
https://t4tutorials.com/decimal-scaling-normalization-in-data-mining/

HW1-2. You are given a set of m objects that is divided into K groups, where the ith group is of size of 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.

This is a proportional sampling where a finite set of objects is divided into sub-objects.. Those sub-objects are then applied with random sampling techniques. The sample from every sub-object is proportional to the whole variety of objects.

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

This is the simple random sampling scheme where each member of a subset has equal opportunity to be picked.

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.

$$P(one\ group) = (1-(1-1/10)^{N})^{10} = where\ N = sample\ size.$$

Because each group is independent, we take the product of the probability for one group, 10 times.

(b) Plot the probability under different sample sizes.