Data Mining HW-1

Ehsan Mosadegh

Question#1

A. difference and similarity between sampling and feature selecting:

Similarity:
both techniques are used to pre-process the data and improve the data quality by selecting a sub set of the data set.

Difference:
sampling is used to select a subset of the data objects from the entire data set instead of processing the entire data set because collecting and/or processing the entire data might be too expensive or time consuming or might be basically impossible to collect like past climatology data. So, we take sample of the dataset to study the features. However, we should note that out data sample should be representative of the entire dataset meaning the collected sample has to have the same statistical properties of the entire dataset.
Feature subset selection is used to select a subset of features and it will help to reduce the dimensions of the data set by removing redundant or irrelevant features.

B. feature subset selection (FSS) and dimensionality reduction (DR):
Both techniques are used to reduce the dimensions of features or in other words to reduce feature space. However, in FSS we create new features by selecting a subset of features from old features, and in DR we create new features by combining the features from old features.

C. mapping x to a new range/domain by two normalization techniques:
X= 480
Range= [-100,9990]
Target range= [-1,+1]

1- decimal scaling
First, we find the maximum number in the range= 9990
Then, we divide x by 10^j, where j is the total digits in the max number.
So, we will have: 480/10^4= 0.048

2- min-max mapping
We will map x in the old range to a new range:
x - min_old/max_old - min_old = x - min_new/max_new - min_new
x_new = -0.8850

Question#2

a- for the first scheme, mi/m can be considered as the weight of each group in the sampling scheme. In other words, it represents how many elements we can sample from each group of data with respect to the total size of data objects m. If we collect data objects based on scheme A, our selected sample will have data objects from all K groups based on the weighting function of each group.
b- if we collect data objects based on scheme B, out collected data set might not include some data groups because we have selected n elements from the entire data set without regard to the group to which an object belongs.