6/1/2017

Lab1

1. remove redundant data

2. remove irrelevant features

3. for each feature, using select and group to find out the number of distinct values.

/count duplicates/

select Count(E.dno) from employee: 8

/count unique value only/

select Count(distinct E.dno) from employee: 3

/count do NOT count NULL/

Select Count(E.superSSN) from employee

Select E.Dno, count(E.dno) from XXXXX group by E.Dno: NULL 0

Select E.Dno, count(\*) from XXXXX group by E.Dno: NULL 1

Because count do NOT count NULL

Discretization is a must for continuous numeric value or two many discrete values. (done by histogram: equal interval width / equal frequency)

Outlier: 1.5 of IQR, IQR = Q3 – Q1.

Histogram is ordered ranked interval.

BikeBuyer column is an outcome/class column, it has nothing to do with the attribute value.

Need to have an ID to identify each row, it is not part of feature. Include a set of features for classifier. It should also include class column to let classifier to learn the outcomes.

Lab 1 is to create training data set

Age, commute distance, salary,

1-5, 6-15, 16-20, 21-50, 51-80, 81-110

Normalization applied on numeric data (interval or ratio data)

EnglishOccupation: binarization for nominal (categorical data)

Cosine similarity is a better option for ordinal data than extended jaccard coeffient.

Create disiimilarity matrix, decide what to use.