Midterm Review

What is Data Mining

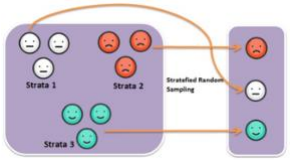
Data pre-processing: (80%: data processing; 20% modeling)

Data quality (missing values; duplicated data; etc)

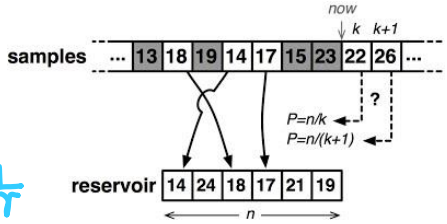
Sampling

1. simple random sampling
2. sampling without replacement
3. sampling with replacement
4. stratified sampling :

(split the data into several groups, then draw random samples from each group)



5. sampling for streaming data (reservoir sampling)



Similarity and dissimilarity measures

1. Euclidean distance
3. Minkowski distance (r=1; r=2; r=\infinity).
3. common properties of distances / similarities
4. similarity between binary vectors

simple matching and Jaccard coefficients

Association rule mining

Concepts: set, subset, itemset, support count, support, frequent itemset

Association rule:

An implication expression of the form X -> Y, where X and Y are itemsets.

Rule evaluation metrics:

support, confidence

Association rule mining task:

1. Given a set of transactions, the goal of association rule mining is to find all rules have support >= minsup threshold.
2. approaches:

Brute-force (very expensive)

Two-step approach

a. frequent itemset generation (very expensive)

frequent itemset generation strategies
a.1 Reduce the number of candidates
– Apriori principle / algorithm and its property
a.2 Reduce the number of transactions
a.3 Reduce the number of comparisons
– support counting using hash tree

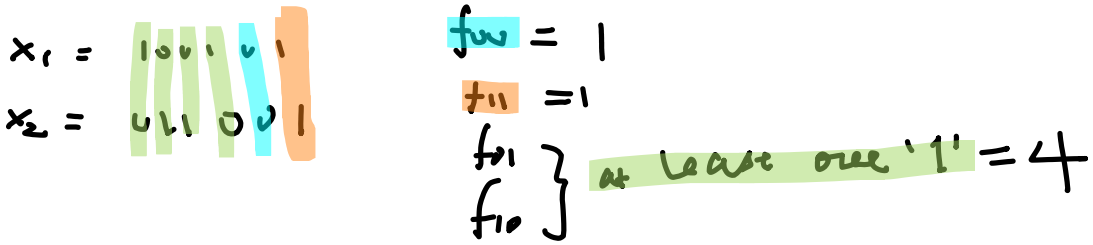
b. rule generation
(generate high confidence rules from each frequent itemset)

b.1 rule generation for Apriori algorithm
b.2 factors affecting complexity of Apriori

Compact representation of frequent itemsets
Find maximal frequent itemsets
closed

---

similarity between binary vectors
simple matching and Jaccard coefficients



==========-==========
Concepts: set, subset, itemset, support count, support, frequent itemset

{a, b, b} set ? × / order ×

S_1={a, b, c} find subset: $2^n = 2^3 = 8$

{} = ϕ
{a} {b} {c}
{a b} {a c} {b c}
{a b c}

K-itemset k = 3 is a set