Data Mining & Machine learning

40% assignments

20% midsemester exam

40% final exam

Moodle

www.cmi.ac.in/~madhavan

→ Teaching

Debajit Paria, Kapil Pawse

Data Mining

└ Fetching data

Mining from data

└ information - data cleaning

Machine Learning

Data in → Prediction out

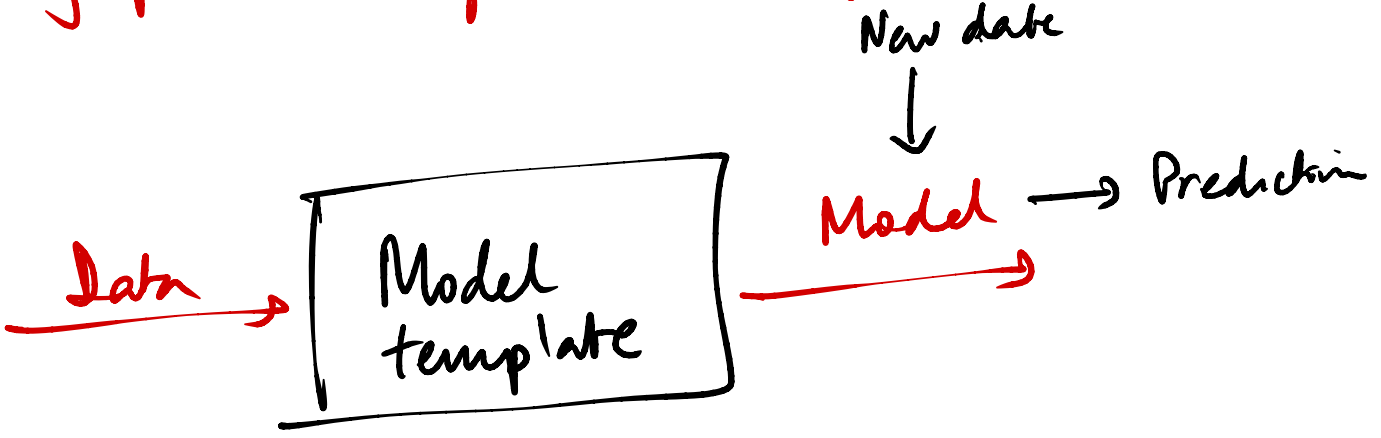
Model

Linear function

$z = ax + by + c$

Machine learning

Estimating parameters of a model from large sets of data



Supervised learning

Known / labelled past data

Unsupervised learning

Clustering

Market - Basket Analysis

People who buy X, also tend to buy Y

Students who fail X also fail Y

Items $I = \{i_1, i_2, \ldots, i_N\}$

Baskets = Transactions $T = \{t_1, \ldots, t_M\}$

Each $t_i \subseteq I$ – set of items

$I = \{i_1, \ldots, i_N\}$ $T = \{t_1, \ldots, t_M\}$ $t_i \subseteq I$

People who buy X also buy Y

Assume $X, Y \in I$
Single items, $X \neq Y$

$$\frac{\#\text{ transactions with } X, Y}{\#\text{ transactions with } X} \geq \text{threshold}$$

How often does this happen?

Infrequent correlations are not useful

Which items are frequent? Say > 1% of T

More generally - frequent sets of items (itemsets)

Given I, T and a frequency threshold f, $0 \le f \le 1$

Which $X \subseteq I$ appear in at least $f \cdot M$ transactions?

How to do this?

Long sequence of numbers 12, 37, 22, 12, 37, ------N

Say $N = 10^6$

Which numbers appear at least $\frac{10000}{K}$ times?

Maintain a dictionary

count [n] = # of times n appears

More simplistic version.

Array or list

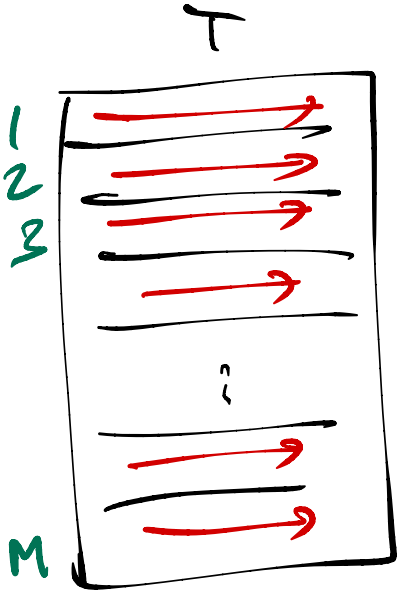
Position i records frequency of i

Suppose our original problem is restricted to individual items

for i in range(M):

    for each x in t_i:

        increment count for x



N items - dictionary size $\leq N$

$\leq \sum_{i=1}^{M} \text{size}(t_i)$

For subsets - count 2^N quantities

Space is a bottleneck

$I = \{i_1 \ldots, i_N\}$ $N = 10^6$

$T = \{t_1 \ldots, t_M\}$ $M = 10^9$ Assume each t_i has size ≤ 10

$f = 0.01$ (i.e. 1%) – only frequent items

If i_k is frequent, it appears in $0.01 \times 10^9 = 10^7$ transactions

M transactions — $10 \times M$ items overall $= 10^{10}$

$\frac{10^{10}}{10^7} \Rightarrow 10^3$ – max number of frequent items

Simple observation

A-PRIORI

If $\{i_k, i_\ell\}$ is a frequent set,

so are $\{i_k\}$ & $\{i_\ell\}$

If either $\{i_k\}$ or $\{i_\ell\}$ is infrequent, $\{i_k, i_\ell\}$ cannot be frequent

1000 frequent items $\rightarrow$ $\frac{1000 \cdot 999}{2}$ frequent pairs