22/01/21 Lec-11

Numerosity reduction → parametric
→ non-parametric

→ parametric (Regression and log-linear)

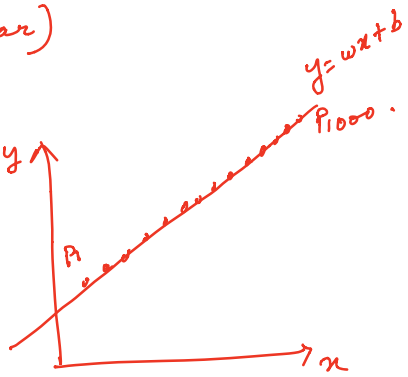
$y = wx + b$

predictor var.

response variable

regression coefficient

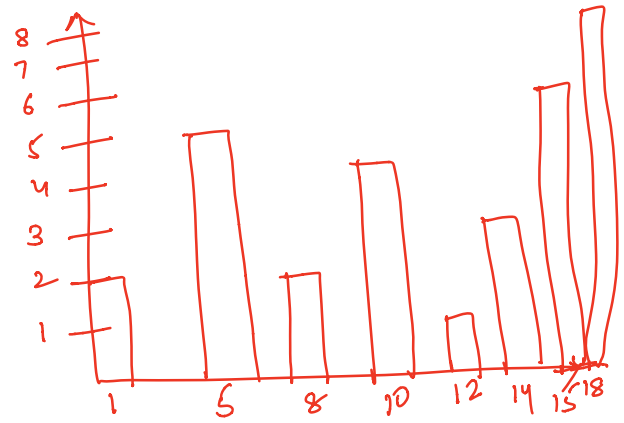
$\langle x_1, \ldots, x_{10} \rangle$
$w \quad b$



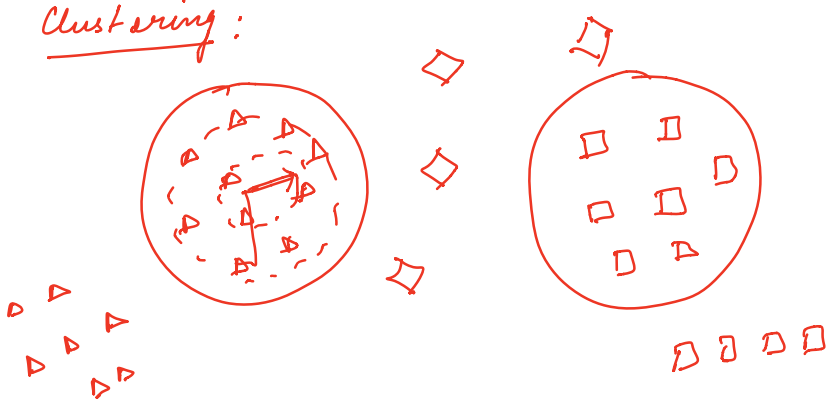
Non-parametric approaches:

1) Histogram:

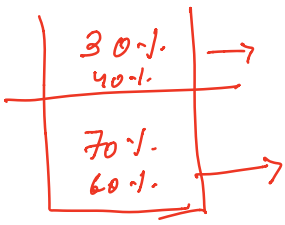
f(1) → 2
f(5) → 5
f(8) → 2
f(10) → 4
f(12) → 1
f(14) → 3
f(15) → 06
f(18) → 08



2) Clustering:

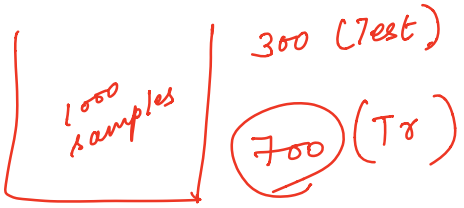


3) Sampling : a) simple random sample without replacement (SRS WOR) :



→ Test data (To check the efficiency of model)

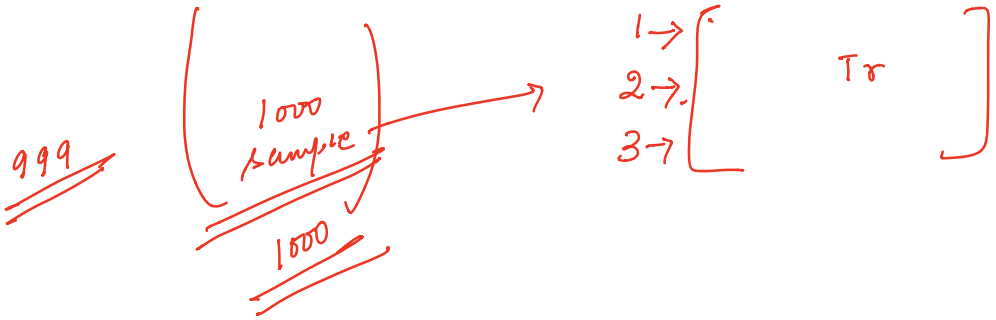
→ Training data (to construct M2 model)



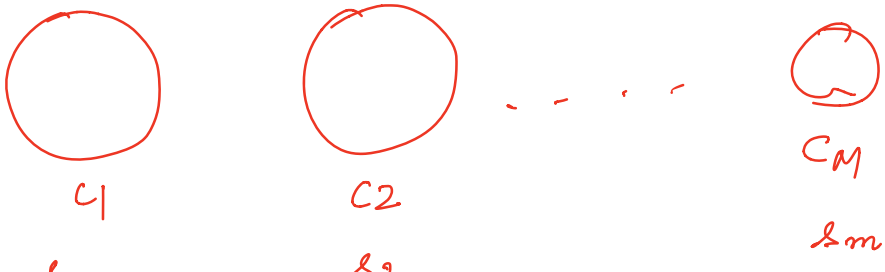
prob. of a tuple getting selected $= \frac{1}{N}$

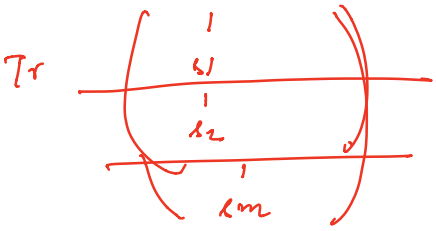
$= \frac{1}{1000}$

b) SRS WR : (simple random sample with replacement)

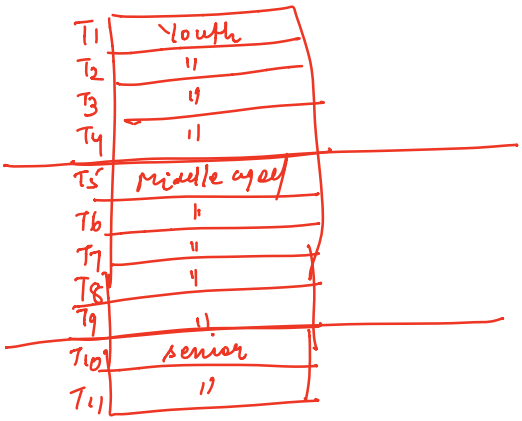


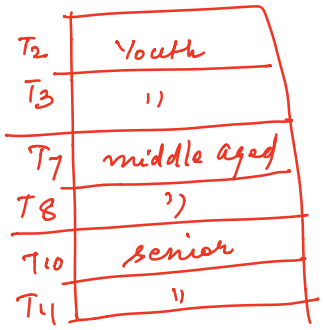
c) Cluster sample :



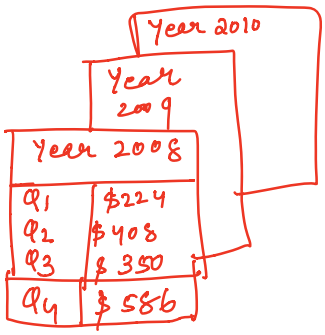


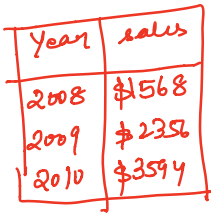
d) Stratified sample:





4) data cube aggregation:





Year	sales
2008	$1568
2009	$2356
2010	$3594

Normalization:

The attribute data are scaled so as to fall within a small specified range such as -1 to +1 or 0 to 1.

v. imp

3 ways for data normalization:

1) Min-Max normalization:

(performs a linear transformation on the original data)

$$v' = \frac{v - min_A}{max_A - min_A} (new\ max_A - new\ min_A) + new\text{-}min_A$$

Eg: min - income: $12000

max - income: $98000

Map to [0, 1] a income value of $73,600.

Soln:
$$v' = \frac{73,600 - 12000}{98000 - 12000}(1 - 0) + 0$$

$$= 0.716$$

2) Z-score normalization: (or Zero mean normalization)

The values for an attribute A are normalized based on the mean and std. deviation of A.

$$v' = \frac{v - \bar{A}}{\sigma_A}$$ → mean

σ_A → std. devn.

Eg: Let mean and std. devn. of attribute 'income' are $54,000 and $16,000 resp.

Using Z-score normalization, find the normalized value for $73,600.

Soln:
$$v' = \frac{73,600 - 54000}{16,000} = 1.225$$

3) Decimal scaling:

This normalizes by moving the decimal points of values of attribute A. The no. of decimal points moved depends on the maximum absolute value of A.

$$v' = \frac{v}{10^j}$$

where j is the smallest integer s.t $\max(|v'|) < 1$

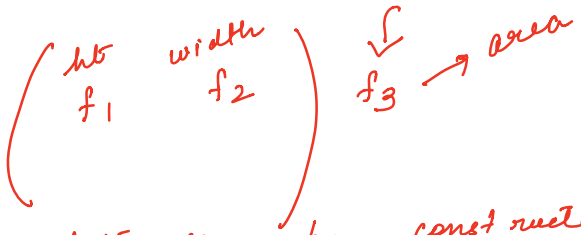
Eg: Suppose the recorded values of A range from -986 to 917.

The max. abs. value of A is 986.

To normalize by decimal scaling, we divide the value by 10^3 (ie 1000).

∴ Normalized value: -0.986 and 0.917.

✳ Attribute construction:



New attributes can be constructed from the given attributes and added in order to improve the accuracy and understanding of high dimensional data.