缺漏資料的補值方法
Imputation Methods for Missing Data

J.-S. Roger Jang (張智星)
MIR Lab, CSIE Dept.
National Taiwan University
jang@mirlab.org, http://mirlab.org/jang

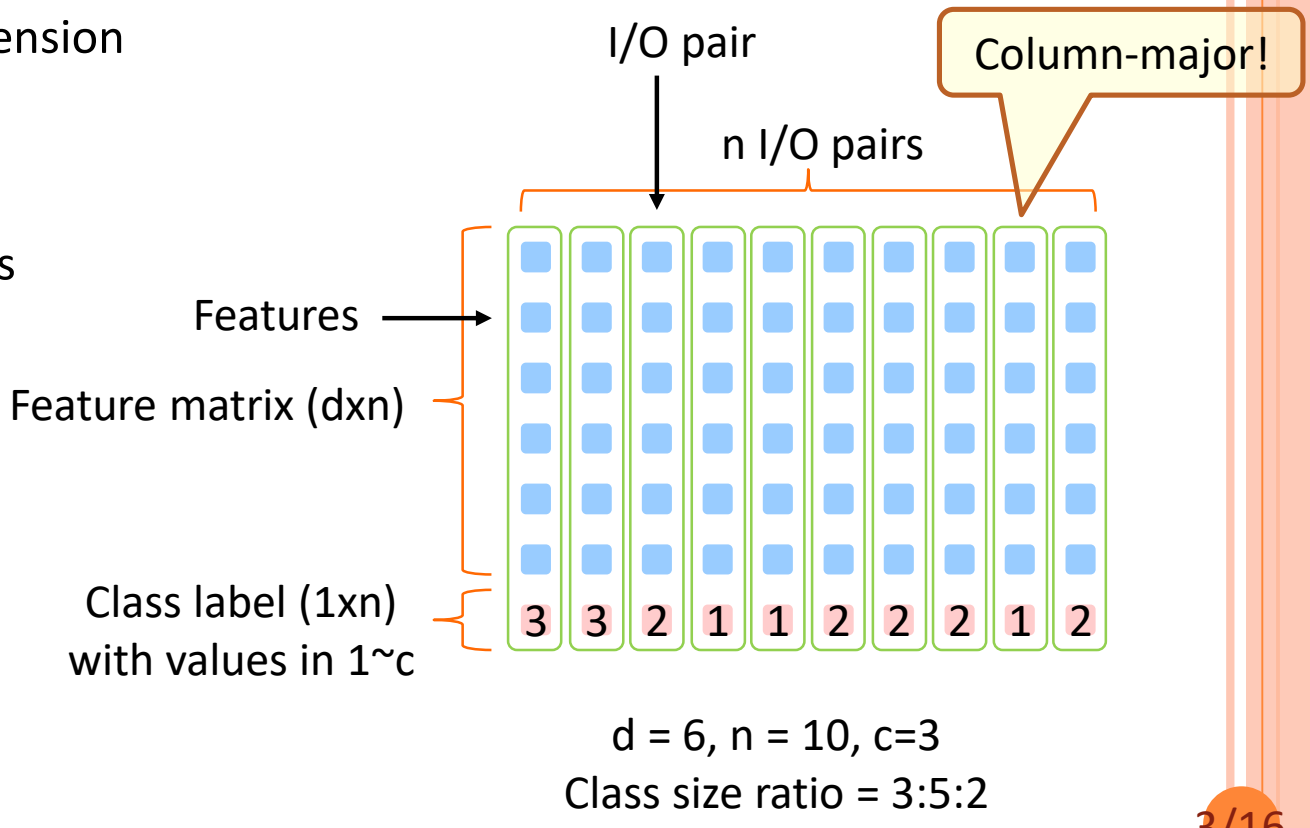
2023/11/29

Outline

- Introduction
- Imputation methods
  - Simple method
  - MICE method
- Summary

Format of Structured Dataset for Classification

- Format of structured dataset
  - Feature matrix: dxn
    - d: feature dimension
    - n: data count
  - Class label: 1xn
    - c: no. of classes



Anomaly in a Dataset

- Outlier detection
  - Via domain knowhow
    - Body temperature 364 degrees
  - Via box plot
  - Via multiple ML models for leave-one-out CV
- Other types of anomaly
  - Two different-class I/O pairs with the same feature vectors
  - A feature with the same value…
    - Across a dataset ➔ We can delete the feature since it has not discriminant power.
    - Across a class ➔ It will causes a error in naïve Bayes classifiers.

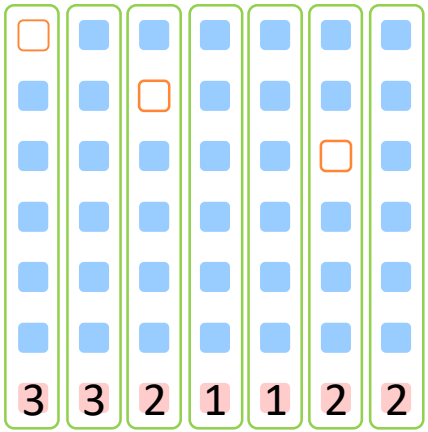
Why?

Simple Imputation for Missing Data

- Simple imputation
  - Categorical feature: Use “mode” of the feature.
  - Numerical feature: Use “mean” or “median” of the feature.
    - Mean: Easily affected by extreme values (“income” for example)
- Pros and cons of simple imputation

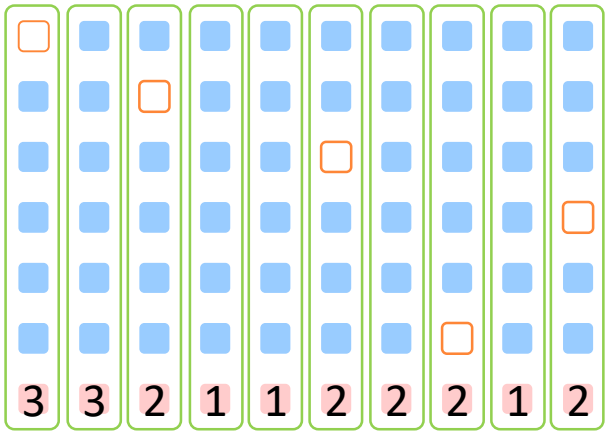
Quiz!

  - Pros: Quick and easy
  - Cons: Do not consider the correlation between features
    - Example: “income” of a 20-years-old man
- Extended simple imputation
  - Impute features based on “similar group”
    - Impute “income” based on “similar-age group”
    - Impute “height” based on “same-gender and similar-age group”



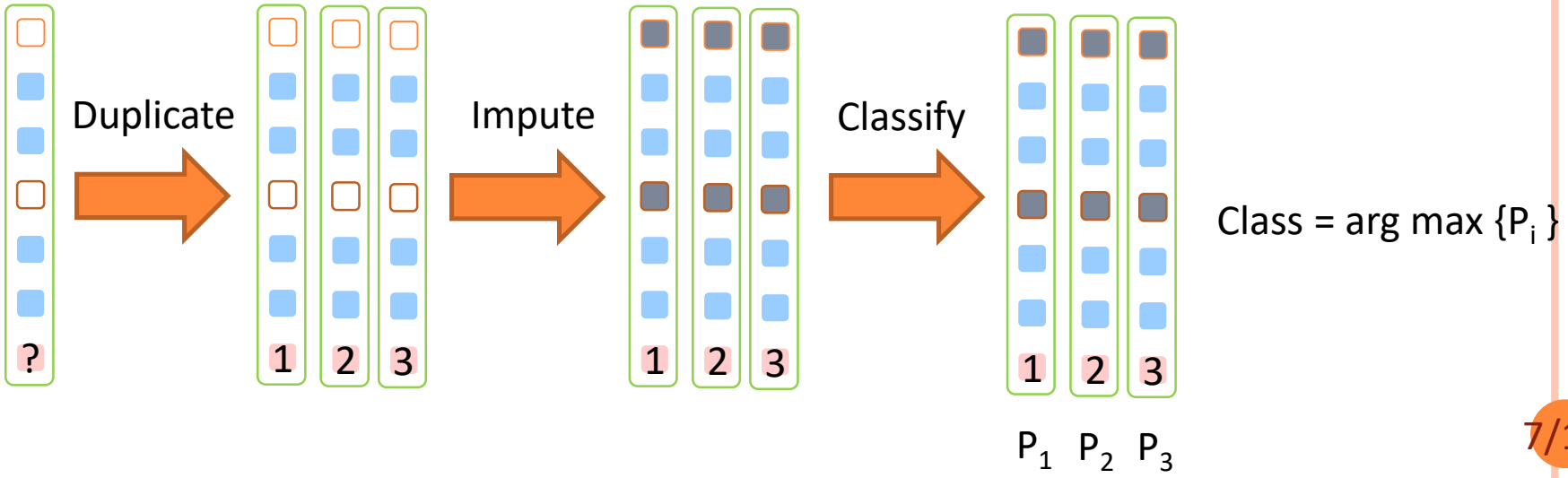
MICE Imputation for Missing Data

- MICE: Multiple imputation by chained equations
  - Use simple imputation at the beginning
  - For each of the missing value, perform imputation one by one
    - Categorical feature: Use “classification”
    - Numerical feature: Use “regression”
  - Repeat the above process until convergence.
- Characteristics
  - Iterative in nature ➔ Slow!
  - Wide selection of classifiers/regressors
  - For training set only, since we need to
    - Use the whole dataset
    - Use the labels as a feature

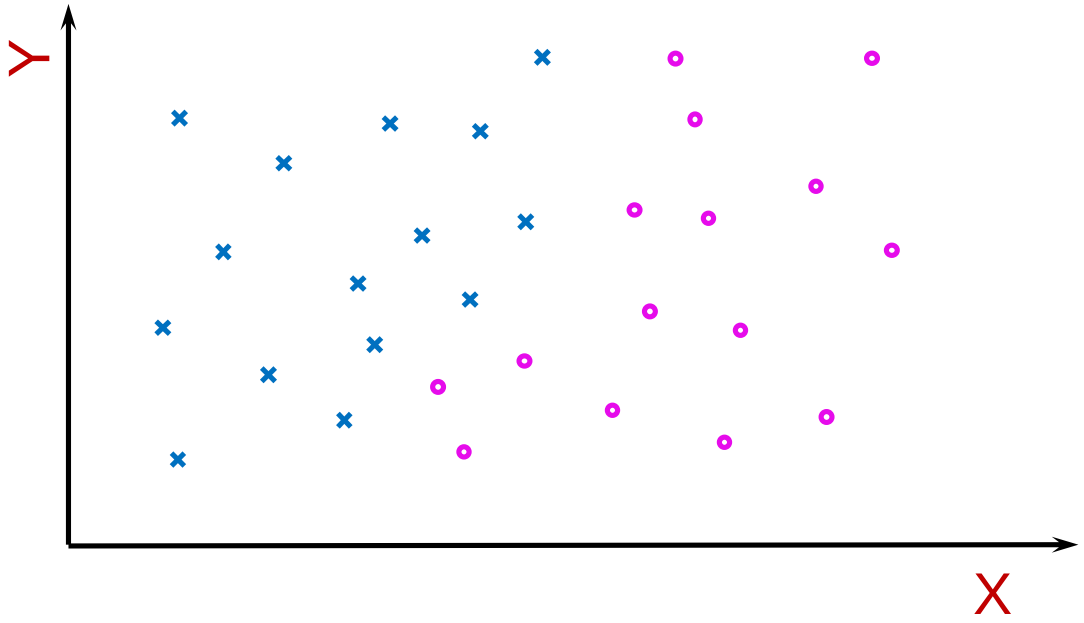


Missing Data Handling on Test/Validation Sets

- Imputation for test/validation sets
  - Duplicate the test input by adding each class labels ➔ c test inputs
  - Perform regular classification based on the imputed training set and the imputed c test inputs
  - Assign the class based on max. likelihood of these c inputs.

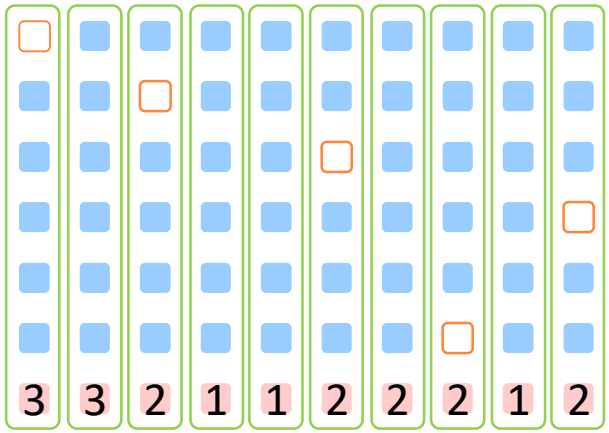


Visualization on Test Input Imputing



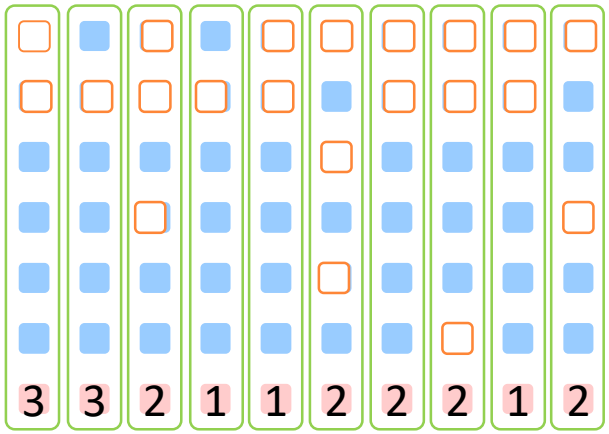
Caveat in Handling Missing Data during Cross Validation

- Approximate method: Quick
  1. Impute missing data
  2. Perform cross validation
- Exact method: Slow
  1. Hide a fold as the test set
  2. Impute the training set
  3. Use the hidden fold for evaluation (with imputation, if needed)
  4. Go back to step 1 until done.

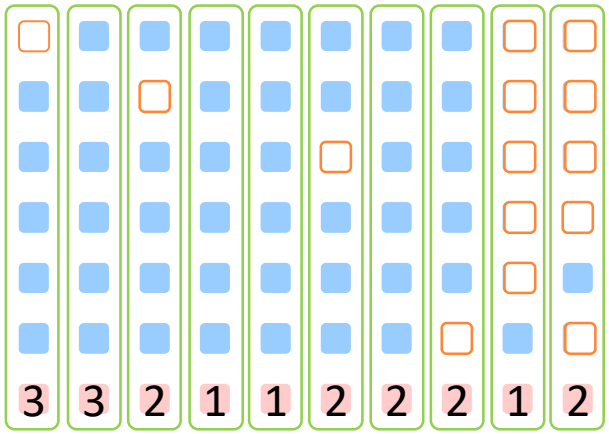


Delete Features or I/O Pairs

- If the missing data ratio is high and the dataset is large, we can simply delete features (rows) or data pairs (columns).
  - Try to plot missing data ratio vs. features or I/O pairs
  - For a given n, delete n features (or I/O pairs) to minimize overall missing data ratio ➔ Max coverage problem!



Missing data ratio = 21/60



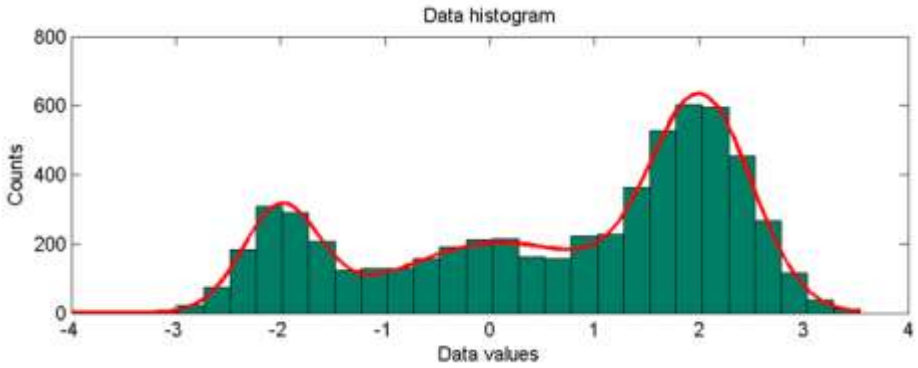
Missing data ratio = 14/60

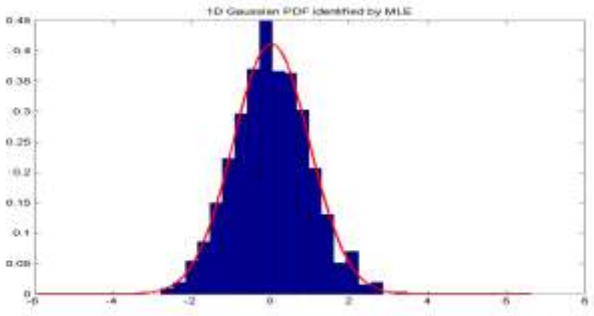
Summary

- Missing data imputation is an important first task in handling real-world data.
- We can use machine learning (classification or regression) for imputing missing data for the final machine learning task.
- Evaluation of test input with missing data is not very straightforward.

From Histogram to PDF

- To create PDF
  - Data ➔ Histogram ➔ PDF

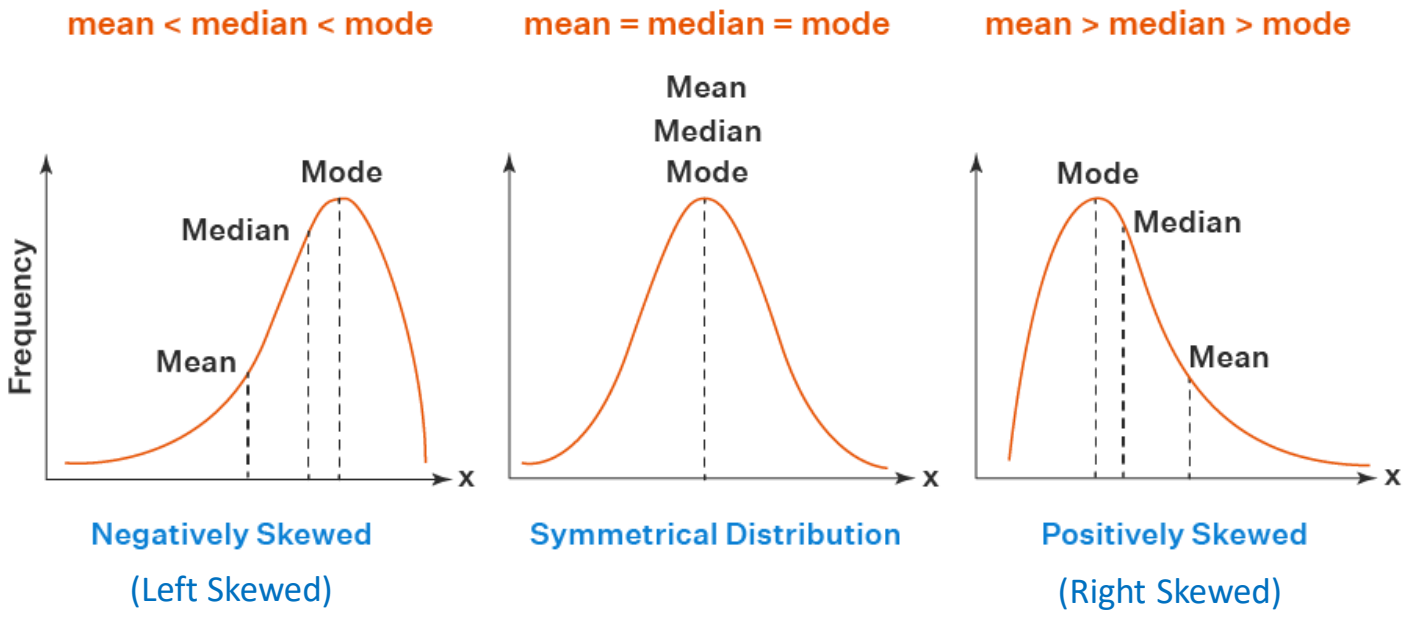




About Mean, Median, Mode

- Representative numbers of distributions
  - Mean: Average of samples
  - Median: Middle number of sample after sorting
  - Median: Most-appearing number of samples

Quiz!



https://www.cuemath.com/data/relation-between-mean-median-and-mode

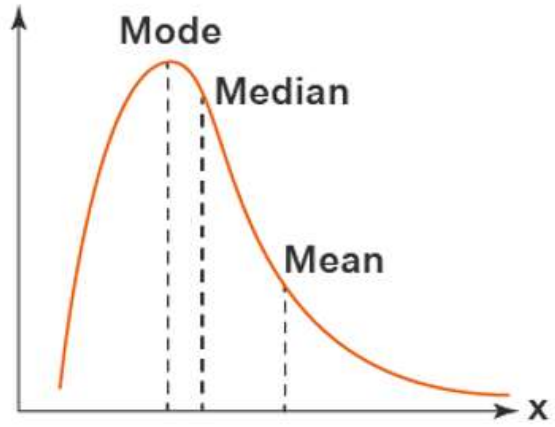
Definitions of Mean, Median, and Mode

- Definitions
  - Mode
  - Median
  - Mean

$$x_{mode} = \arg\max_x f(x)$$

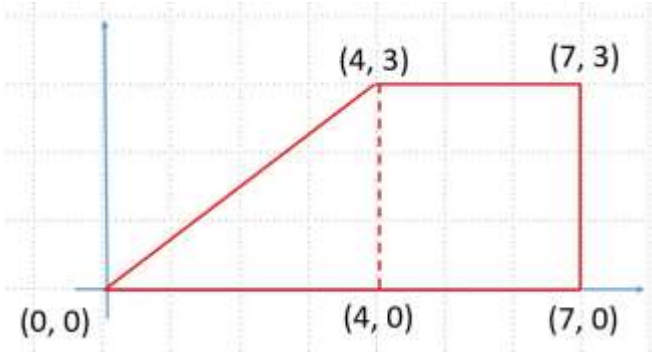
$$x_{median} = \arg\min_u \left| \int_{-\infty}^{u} f(x)dx - \int_{u}^{\infty} f(x)dx \right|$$

$$x_{mean} = \frac{\int_{-\infty}^{\infty} xf(x)dx}{\int_{-\infty}^{\infty} f(x)dx}$$



Exercise

- Compute the mean, median, and mode of the following PDF-like histogram.



Missing Value Imputation

- Steps
  - Determine neighborhood
  - Derive the value
- Neighborhood identification
  - 1D input
  - 2D input
  - K-nearest neighbor
  - All
- Value determination
  - Nearest
  - Interpolation
  - Model-based

Missing Value Imputation: Example

96. (20%) **Missing value imputation**: Given the following dataset with missing values denoted by $X_i, i = 1, 2$:

Feature name\Data index	1	2	3	4	5	6	7	8	9
Gender	♂	♀	♂	♂	♀	♀	♀	♂	♂
Age	15	30	40	65	45	11	55	20	50
Height (cm)	165	130	158	175	164	148	168	185	172
Weight (kg)	50	50	60	65	50	35	55	85	X_1
Monthly income (k)	2	80	60	120	X_2	2	100	30	90
BloodType	A	B	O	AB	A	B	O	AB	B
Class	1	3	1	2	3	1	3	2	1

Find those missing values based on the following guidelines:

a. Find X_1 based on same-gender average
b. Find X_1 based on same-gender/class average
c. Find X_1 based on same-gender height-dependent linear interpolation
d. Find X_2 based on age-dependent linear interpolation
e. Find X_2 based on same-gender age-dependent 2-nearest-neighbor average
f. Find X_2 based on same-gender age-dependent linear interpolation