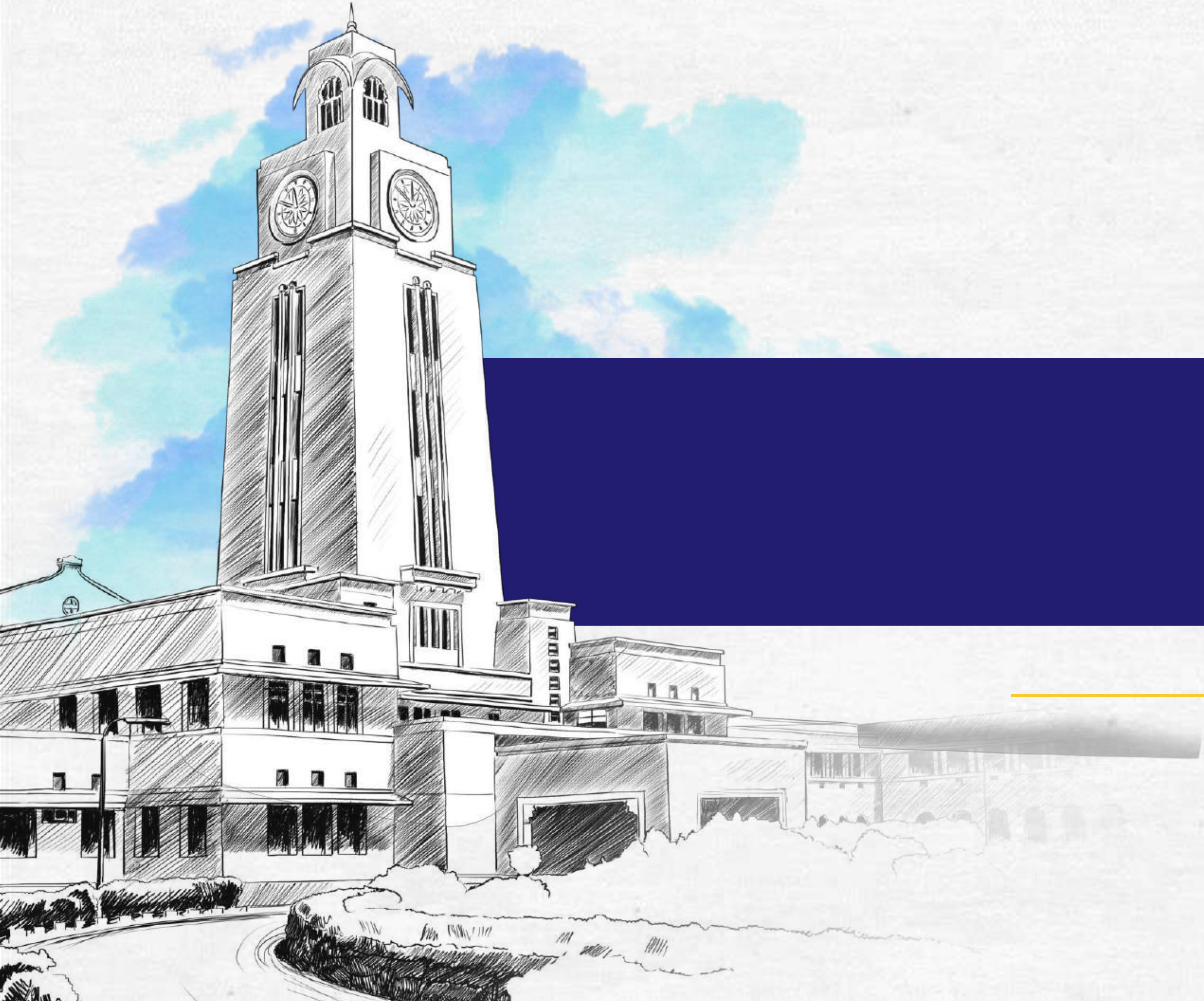


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Data classification

Prof. Aruna Malapati

Lecture Objectives

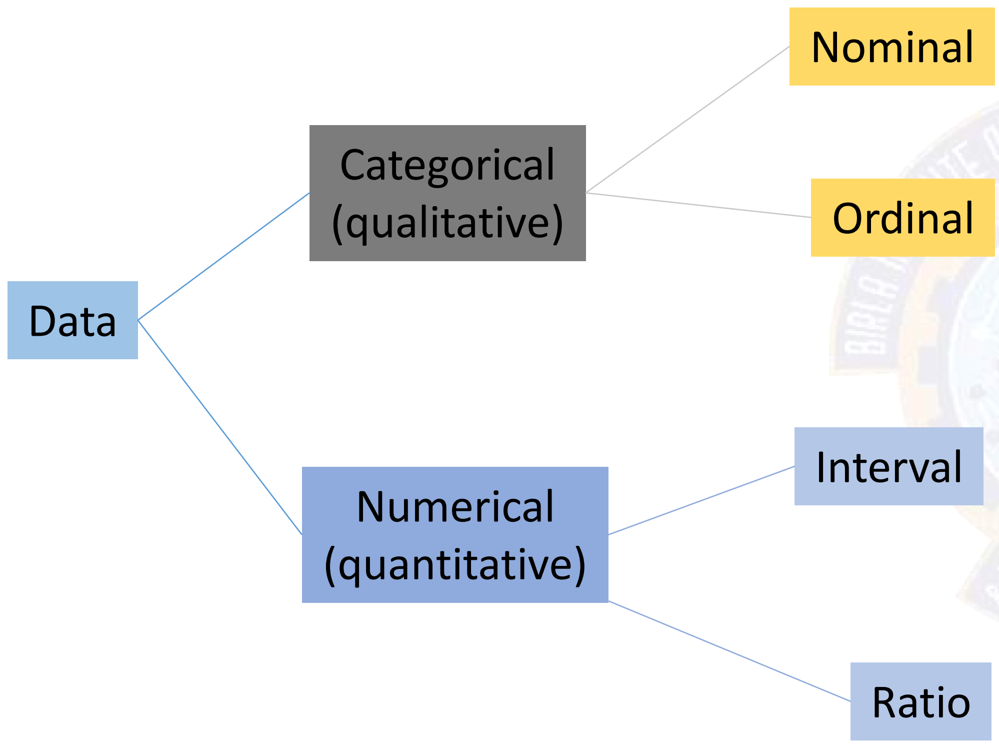
- Define data
- List different data types based on the values they take
- Define qualitative and quantitative data types

What is Data?

- Data is a collection of objects described using its attributes or features.

Building Area	Common Area	Type of Flooring	DistanceFrom BusDepot	Sale Price per square feet
11345	350	Marble	16503.22	6,715
2000	1334	Vitrified Tiles	16321.19	3,230
2544	924	Wood Vitrified Tiles	15619.92	6,588

Four Levels of data



- Used for names or label a series of values.

  Ex:Gender(dichotomous),Hair color, Country names

- Ordinal values represent discrete and ordered units.

  Ex: Shirt size, letter grades, food quality

- Interval values represent ordered units that have the same difference.

  Ex: temperature of a given place

- Ratio values represent ordered units that have the same difference. There's a value that means "nothing/none."

  Ex: Age, weight, test score

Flavors for Quantitative Attributes

Quantitative Attributes

- Discrete
- Continuous

Discrete

- Has a finite or countably infinite set of values
- Often integer values are used to denote the values
- Ex: No of employees, set of words in a document collection

Continuous

- Has real numbers as attribute values
- Often real numbers are used to denote the values
- Ex: height, Weight, Blood sugar levels

Attribute Values

Nominal	Ordinal	Interval	Ratio	Operations
	✓	✓	✓	Ordering
✓	✓	✓	✓	Counts
✓	✓	✓	✓	Mode
	✓	✓	✓	Median
		✓	✓	Mean
		✓	✓	Difference
		✓	✓	Add/Subtract
			✓	Multiply/ Divide
			✓	True Zero

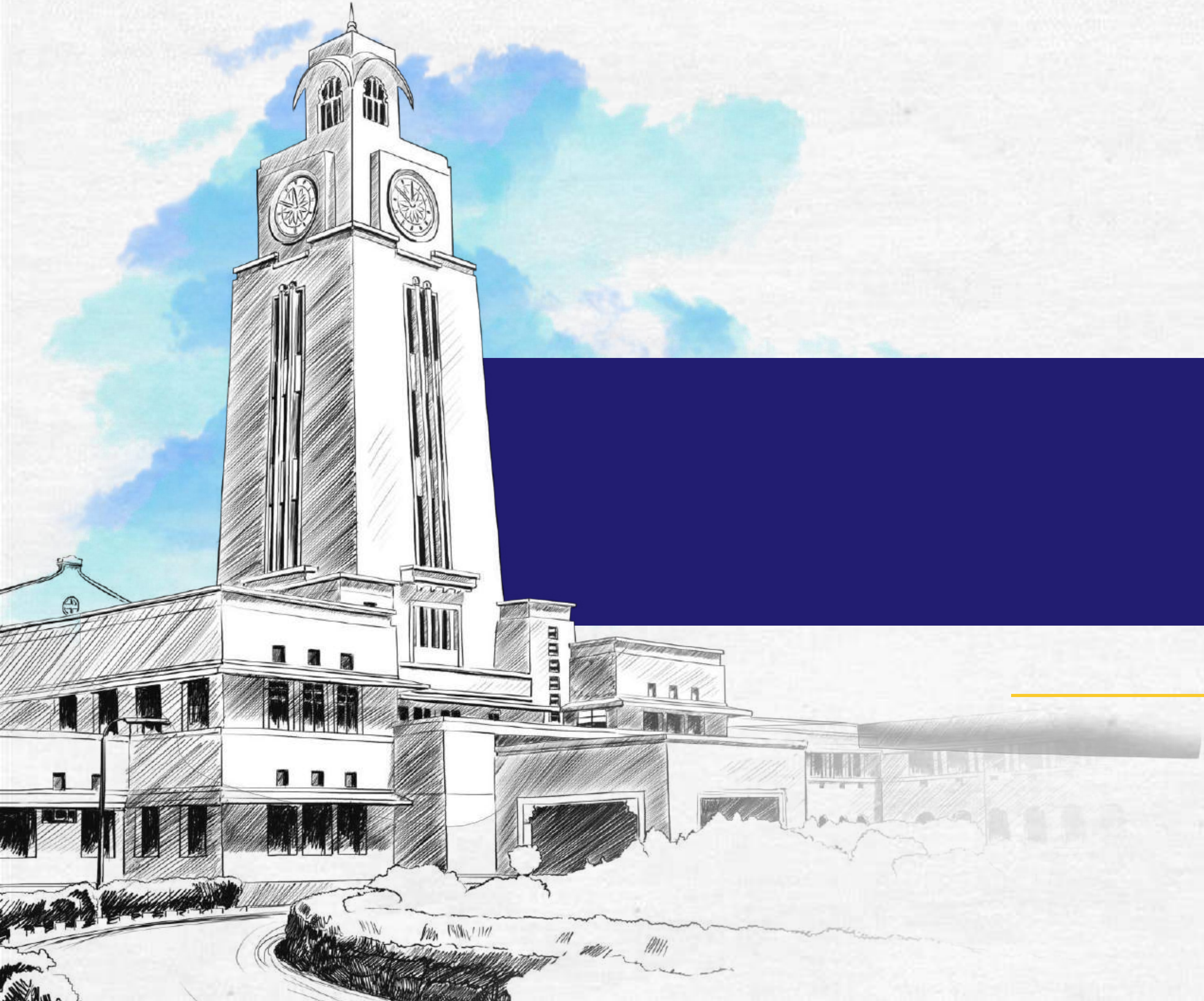
Important characteristics of data

- Dimensionality (number of attributes)
- Sparsity
- Resolution
- Size

Thank You!

In our next session: Types of data

BITS Pilani
Pilani | Dubai | Goa | Hyderabad



Types of Data

Prof. Aruna Malapati

Learning Objectives

- List different data types

Record Data/Data Matrix

Building Area	Common Area	Type of Flooring	DistanceFrom BusDepot	Sale Price per square feet
11345	350	Marble	16503.22	6,715
2000	1334	Vitrified Tiles	16321.19	3,230
2544	924	Wood Vitrified Tiles	15619.92	6,588

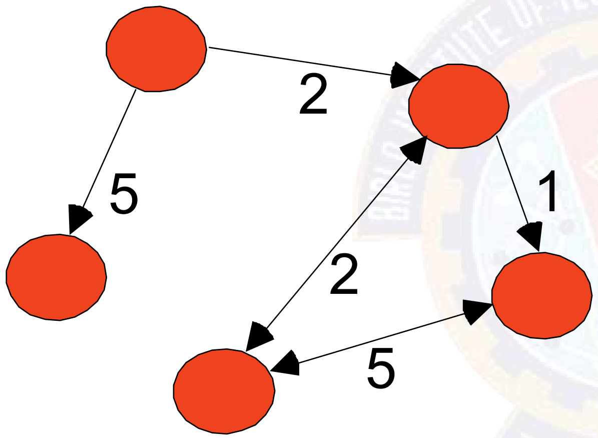
Document Data

	Term1	Term2	Termn
Document1	1	22	3
Document2	3	8	5
Document3	7	3	1
Documentn	8	5	0

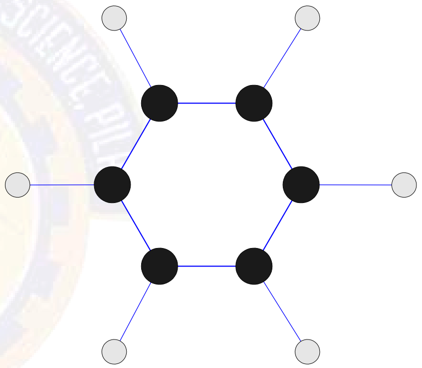
Transactional Data

TID	Items
1	Bread,Butter,Cheese
2	Shampoo, Coke, Milk
3	Toothpaste, hair oil
N	Lays, Kurkure, Pepsi

Graph Data



Web graph/Email structure



Benzene Molecule: $C6H6$

Sequence Data

- Sequences of transactions

Items/Events

(Lays,Pepsi)(bread)(Eggs,Milk)

(Pepsi,bread)(Eggs)(Milk)

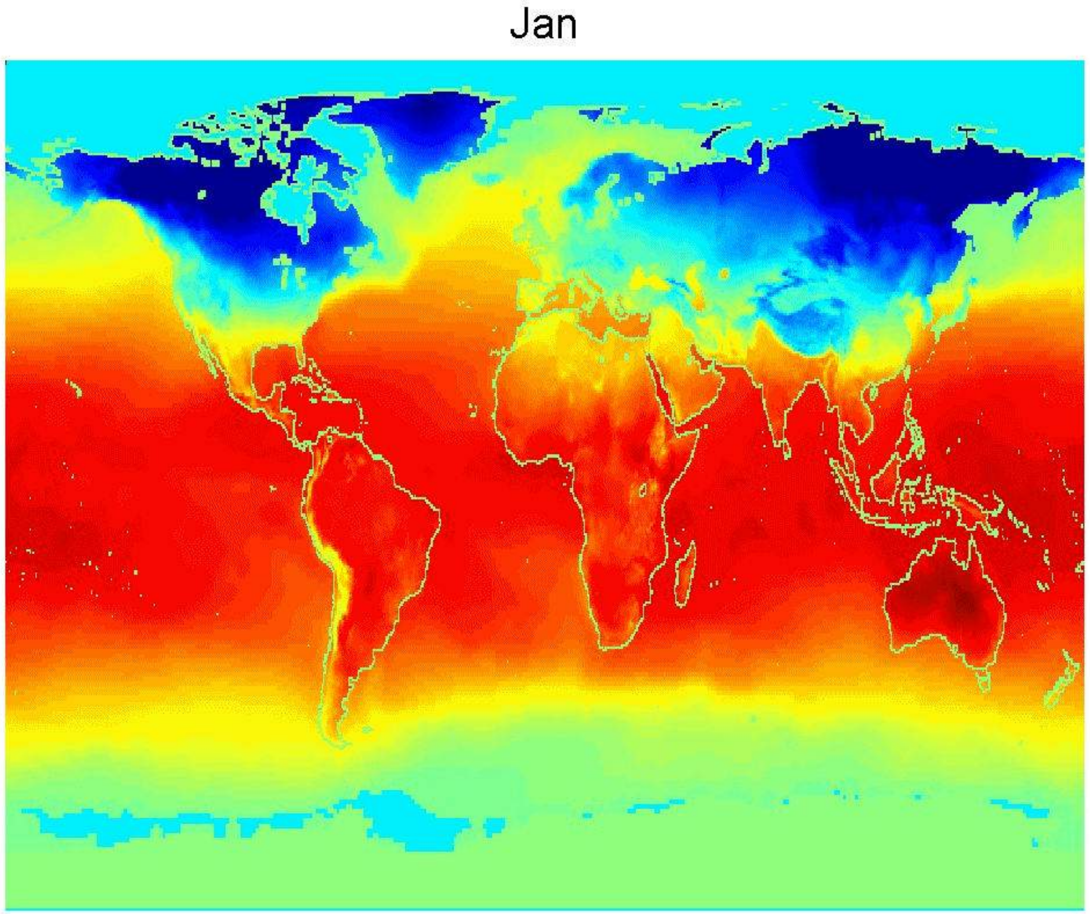
(Eggs, bread)(Pepsi)(Lays,Milk)

Ordered Data

aaaaaaaaaaaattaaaaaaaaaattaaataaaaaaaaa
aataaaaaaaaaaataaaaaaaaaaaaataaaataattata
aaaaattaaaaaaaaaaaaattaaaaaaaaaaattaaaaa
aaaaattaaataaaaaaaaaaataaaaaaaaaaataaaa
aaaaaaaataaaataattataaaaaattaaaaaaaaaaaat
taaaaaaaaaaaattaaaaaaaaaattaaataaaaaaaaa
aataaaaaaaaaaataaaaaaaaaaaaataaaataattata
aaaaattaaaaaaaaaaaaattaaaaaaaaaaattaaaaa
aaaaattaaataaaaaaaaaaataaaaaaaaaaataaaa
aaaaaaaataaaataattataaaaaattaaaaaaaaa

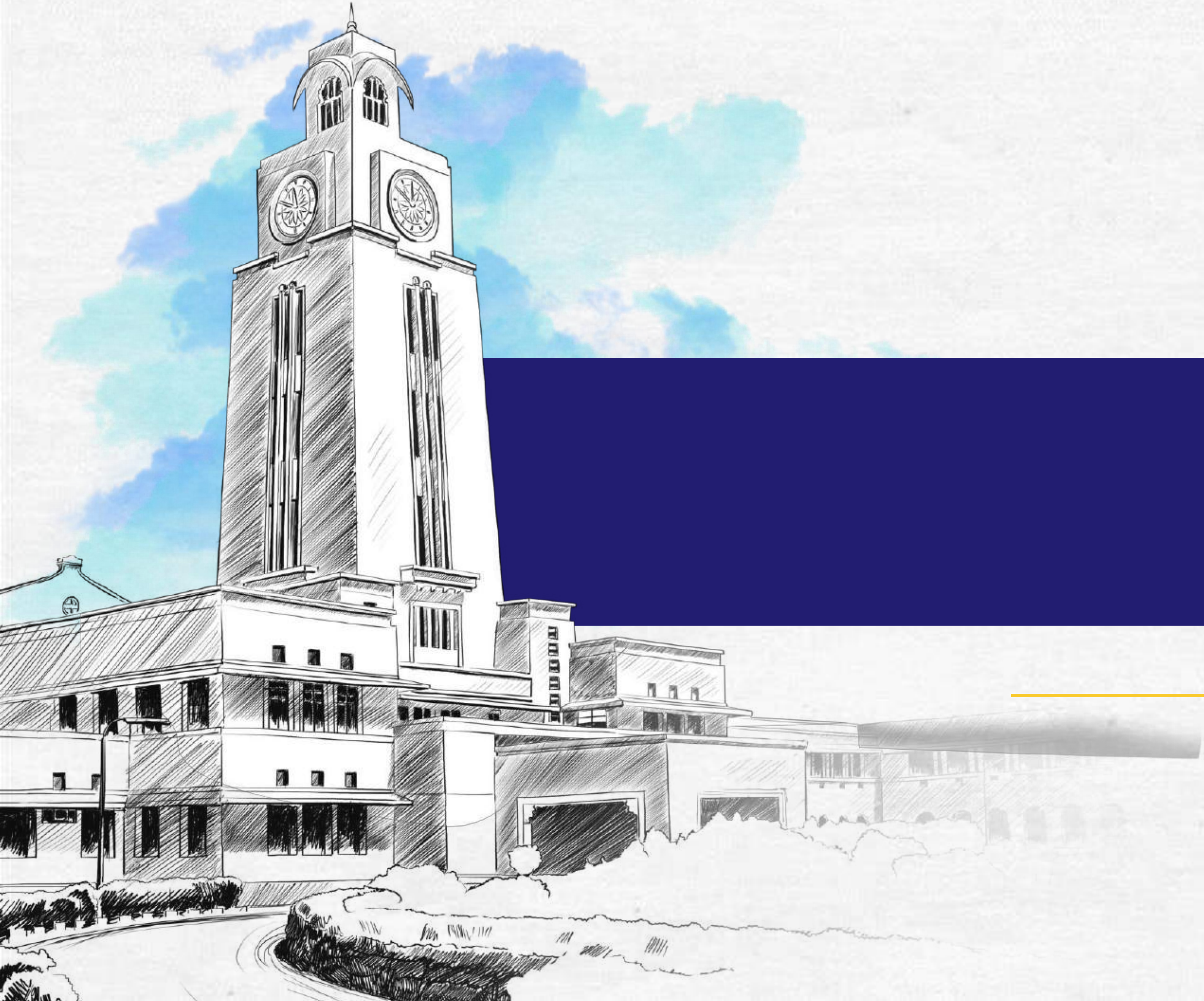
Spatiotemporal data

Average Monthly Temperature of land and ocean



Thank You!

In our next session:Data quality



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Types of Data

Prof. Aruna Malapati

Learning Objectives

- List different data types

Record Data/Data Matrix

Building Area	Common Area	Type of Flooring	DistanceFrom BusDepot	Sale Price per square feet
11345	350	Marble	16503.22	6,715
2000	1334	Vitrified Tiles	16321.19	3,230
2544	924	Wood Vitrified Tiles	15619.92	6,588

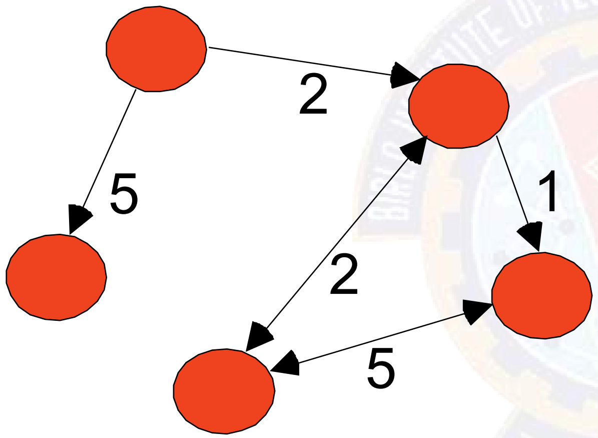
Document Data

	Term1	Term2	Termn
Document1	1	22	3
Document2	3	8	5
Document3	7	3	1
Documentn	8	5	0

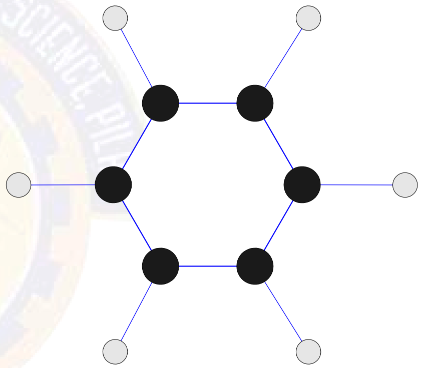
Transactional Data

TID	Items
1	Bread,Butter,Cheese
2	Shampoo, Coke, Milk
3	Toothpaste, hair oil
N	Lays, Kurkure, Pepsi

Graph Data



Web graph/Email structure



Benzene Molecule: C6H6

Sequence Data

- Sequences of transactions

Items/Events

(Lays,Pepsi)(bread)(Eggs,Milk)

(Pepsi,bread)(Eggs)(Milk)

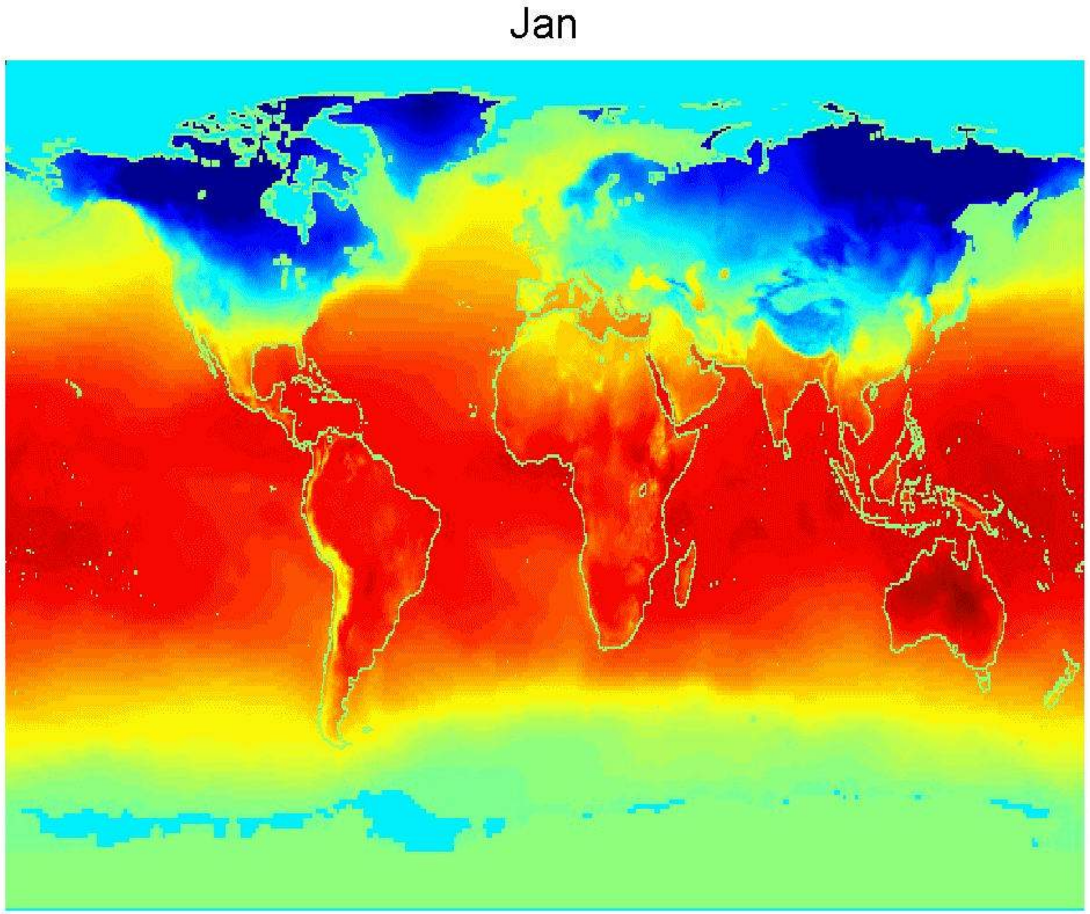
(Eggs, bread)(Pepsi)(Lays,Milk)

Ordered Data

aaaaaaaaaaaattaaaaaaaaaattaaataaaaaaaaa
aataaaaaaaaaaataaaaaaaaaaaataaaataattata
aaaaattaaaaaaaaaaattaaaaaaaaaaattaaaaa
aaaaattaaataaaaaaaaaataaaaaaaaaaataaaa
aaaaaaaataaaataattataaaaaattaaaaaaaaaaaat
taaaaaaaaaaaattaaaaaaaaaattaaataaaaaaaaa
aataaaaaaaaaaataaaaaaaaaaaataaaataattata
aaaaattaaaaaaaaaaattaaaaaaaaaaattaaaaa
aaaaattaaataaaaaaaaaataaaaaaaaaaataaaa
aaaaaaaataaaataattataaaaaattaaaaaaaaa

Spatiotemporal data

Average Monthly Temperature of land and ocean



Thank You!

In our next session:Data quality

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

DATA PREPROCESSING

Prof.Aruna Malapati

Learning objectives

- Explain the importance of data pre-processing
- List the objectives of data pre-processing
- Identify the data quality issues

Data Preprocessing

GOAL: Engineer the data suitable for building faster and simple models

CLEAN

FORMAT
- **Select appropriate features**
- **Data Transformation**

Data Cleaning

Data Quality issues

- Missing values
- Duplicate data
- Inconsistent / Invalid data
- Noise
- Outliers

Missing Values

Customer Name	Age	Income
Rahul	35	12,00,000
Shravya	--	8,00,000
Mehul	40	22,00,000
Vaishali	25	--
Shiva	65	--

Duplicate Data

- **Duplicate data occurs when your data set has redundant data objects**

Customer Name	Address
Shravya	Flat no 450, Street no – 2 Celebrity homes,Mumbai
Shravya	Flat no 304, Street no – 215, Lave view homes, Bangalore
Mehul	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli
Shiva	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli

Inconsistent / Invalid data

- Impossible value for a feature
  - ✓ Ex: age -10
  - ✓ 7 letter Income -10000
  - ✓ zip code in India
- Primarily occur due to data entry error

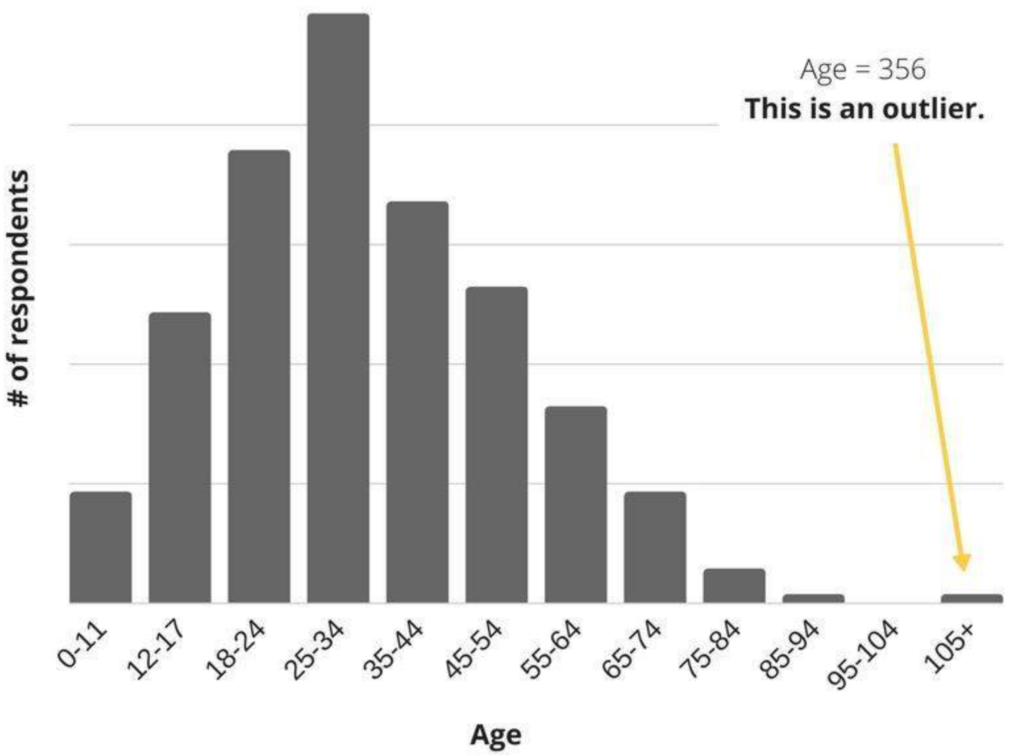
Noise

- **It is meaningless or distorted data.**

Customer Name	Address
Shravya	Flat no 450, Street no – 2 Celebrity homes,Mumbai
Shravya	Flat no ϱђЋѭҖ, Street no – 215, Lave view homes, Bangalore
Mehul	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli
Shivă	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli

Outliers

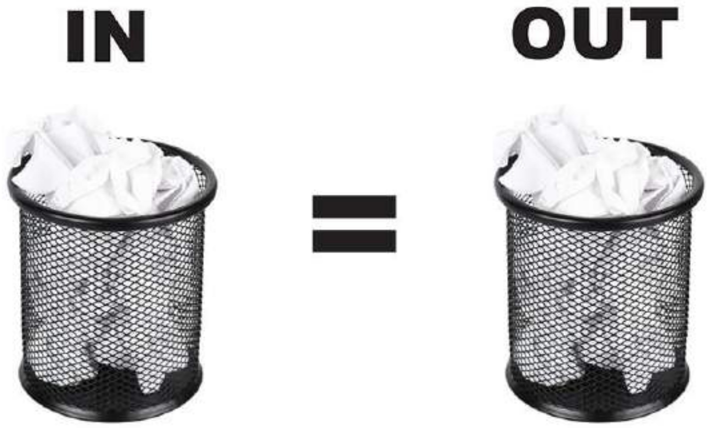
- A data object that is considerable different from others general behavior of the data.



Data Preprocessing techniques

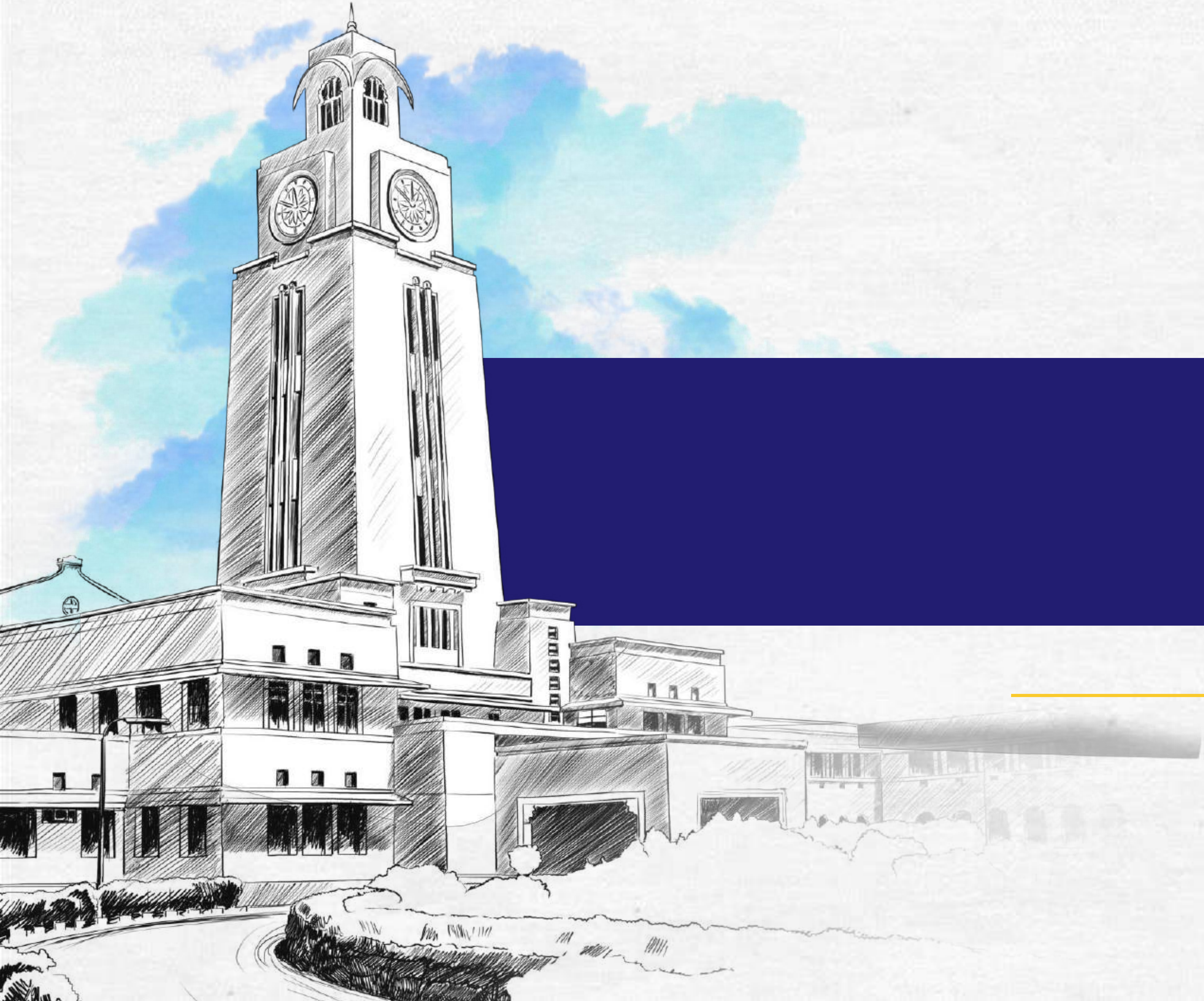
- ➢Feature selection
  - ✓Adding or removing features
- ➢Feature Transformation
  - ✓Scaling
  - ✓Dimension reduction

Data Preprocessing is a very important and tedious process



Thank You!

In our next session: Data Quality



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Data Quality Issues

Prof.Aruna Malapati

Learning objectives

- Identify and impute missing values
- List the reasons for missing values

Major issue with real word data sets

- **Real world data is often dirty**

Missing data

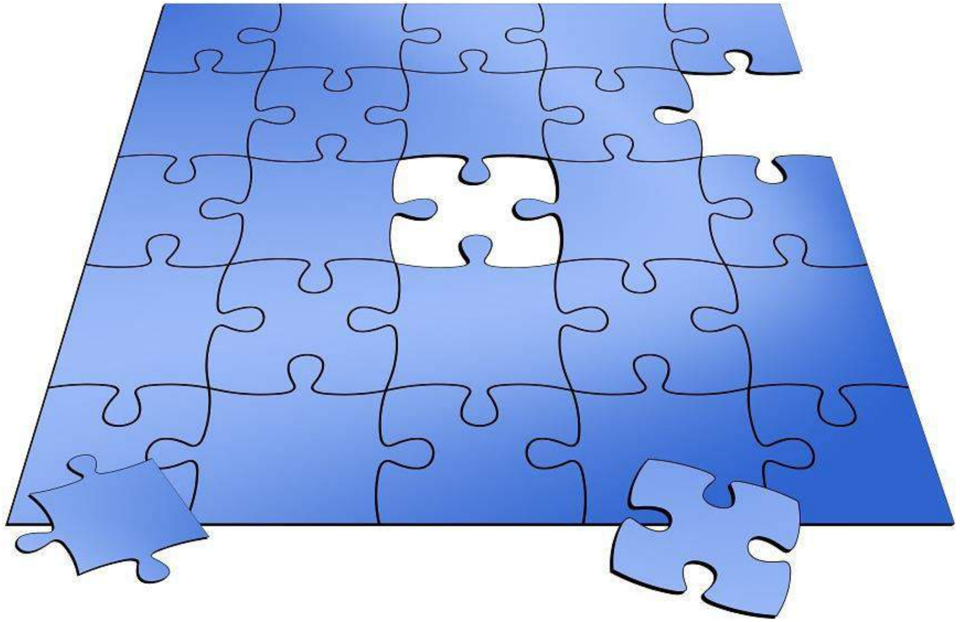
Customer Name	Age	Income
Rahul	35	12,00,000
Shravya	--	8,00,000
Mehul	40	22,00,000
Vaishali	25	--
Shiva	65	--

Impact of Missing data

- Incompatible in Scikitlearn
- Missing data imputation may distort variable distribution
- Affect the performance of Machine Learning Models

Missing Data : Mechanism

- **Understanding the mechanism of missing data will help us choose appropriate imputation method.**
  - ✓Missing completely at random (MCAR)
  - ✓Missing at random (MAR)
  - ✓Not missing at random (NMAR):



Missing completely at random (MCAR)

- The probability of missing is same for all the observations.
- There is no relationship between the missing values and any other values in the dataset.
- Removing such missing values will not effect the inferences made.

Missing Data at Random(MAR)

- The **probability of a missing values depends on available information** i.e it depends on other variables in the dataset.

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

33% males

50% Females

Missing Data Not at Random(MNAR)

- **The missing values exist as an indication of a certain class.**

No of clinical visits	No of sports classes attended	Depression
1	NA	Yes
NA	NA	Yes
NA	0	Yes
4	2	Yes
NA	1	Yes
3	NA	Yes
0	0	No
NA	5	No
1	2	No
1	1	No
2	1	No
NA	2	No

Imputation Techniques for numeric values

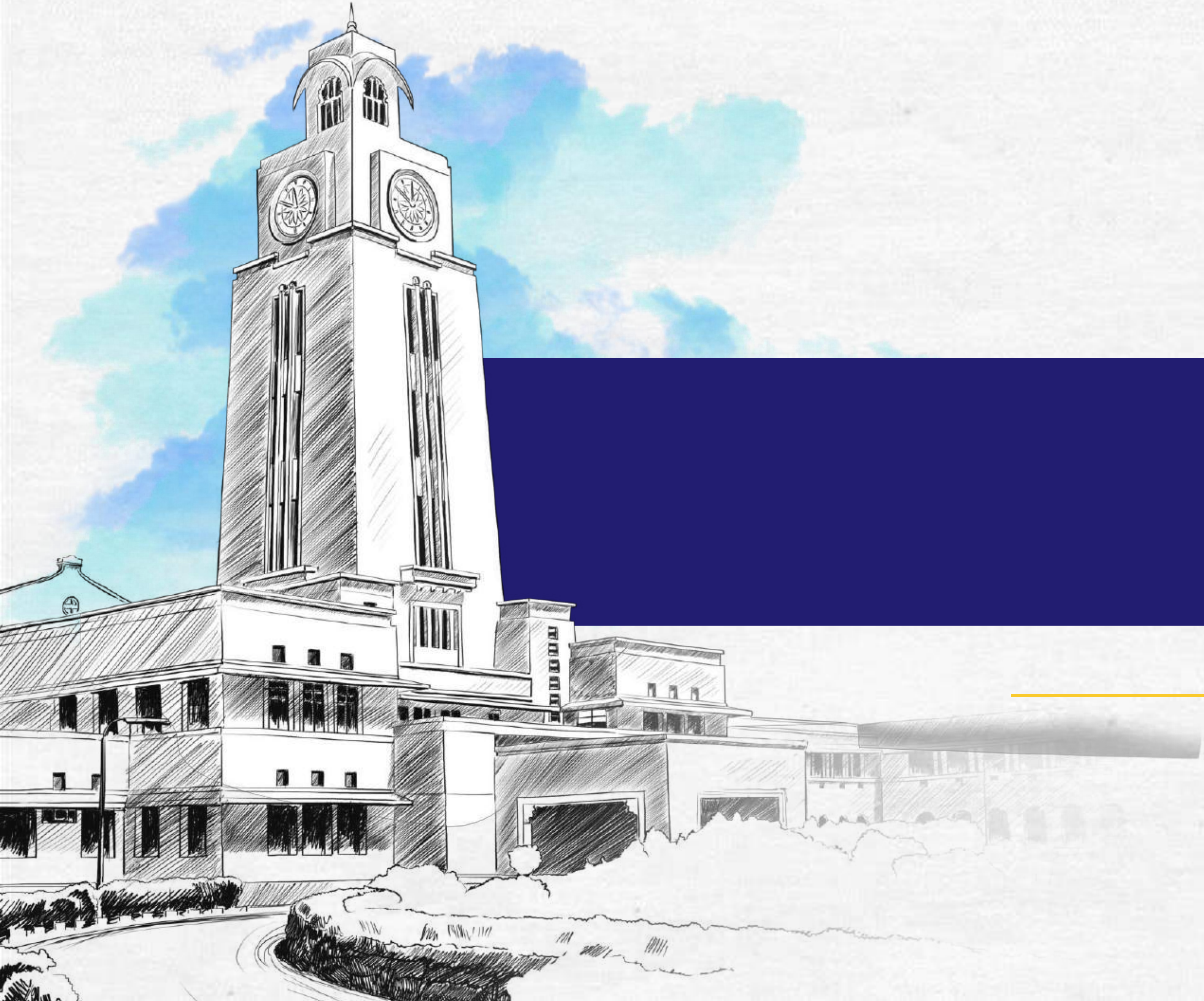
- **Mean / Median Imputation**
- **Random Sampling Imputation**
- **Adding a new variable to indicate missingness**
- **Imputation of NA by values at the end of distribution**
- **Imputation of NA by arbitrary values**

Imputation Techniques for categorical values

- Imputation by most frequent category
- In categorical variables treating NA as an additional category

Thank You!

In our next session: Data Quality Issues



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Data Quality Issues

Prof.Aruna Malapati

Learning objectives

- Identify different type of data quality issues in real world datasets
- Choose appropriate techniques to overcome data quality issues

Duplicate Data

- **Delete old data**
- **Merge duplicate records**

Customer Name	Address
Shravya	Flat no 450, Street no – 2 Celebrity homes,Mumbai
Shravya	Flat no 304, Street no – 215, Lave view homes, Bangalore
Mehul	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli
Shiva	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli

Customer Name	Address
Shravya	Flat no 304, Street no – 215, Lave view homes, Bangalore
Shiva	Plot no 80, APARNA SAROVAR ZENITH, Nallagandla, Gachibowli

Invalid Data

- **Use external knowledge bases to get the right values**
- **Apply reasoning and domain knowledge to come with a reasonable estimate**

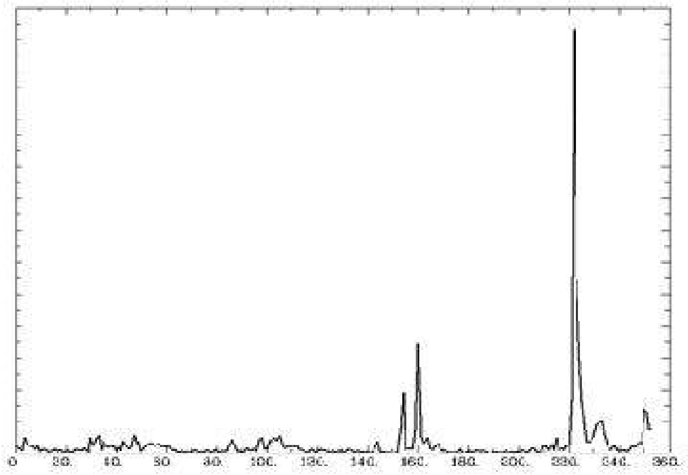
Location	Pincode	State	District
Aliabad	**500015**	Telangana	Hyderabad
Ambernagar	**500044**	Telangana	Hyderabad
Amberpet	**500013**	Telangana	Hyderabad
Anandnagar	**500004**	Telangana	Hyderabad
Anantagiri	**5012015**	Telangana	Hyderabad

Noise

- Filter out the noise component
- This may result in partial loss of data if not done carefully.

Outliers

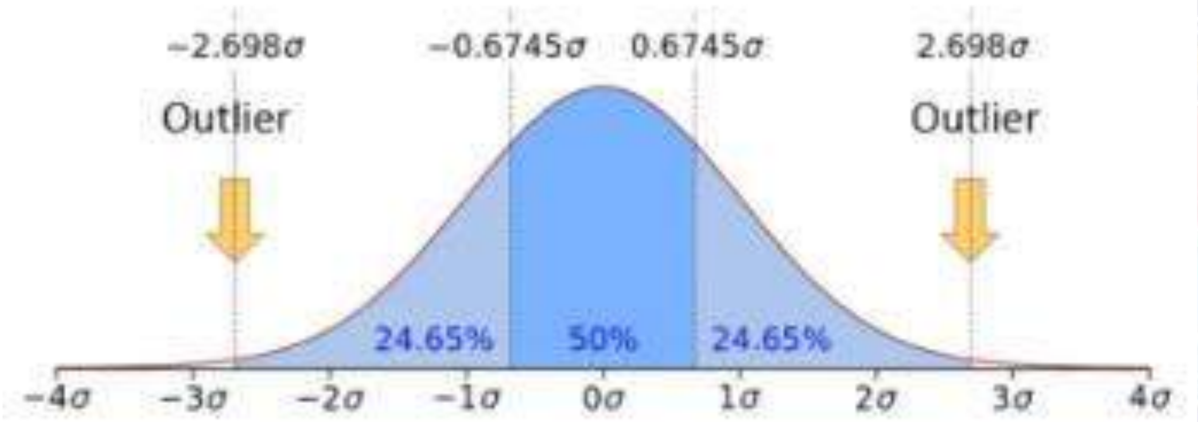
- Algorithms like Linear Regression, K-Nearest Neighbor, Adaboost are sensitive to noise.
- Outlier can significantly skew the distribution of your data.
- Outliers can be identified using summary statistics and plots of the data.



Credit Card transactions of a customer

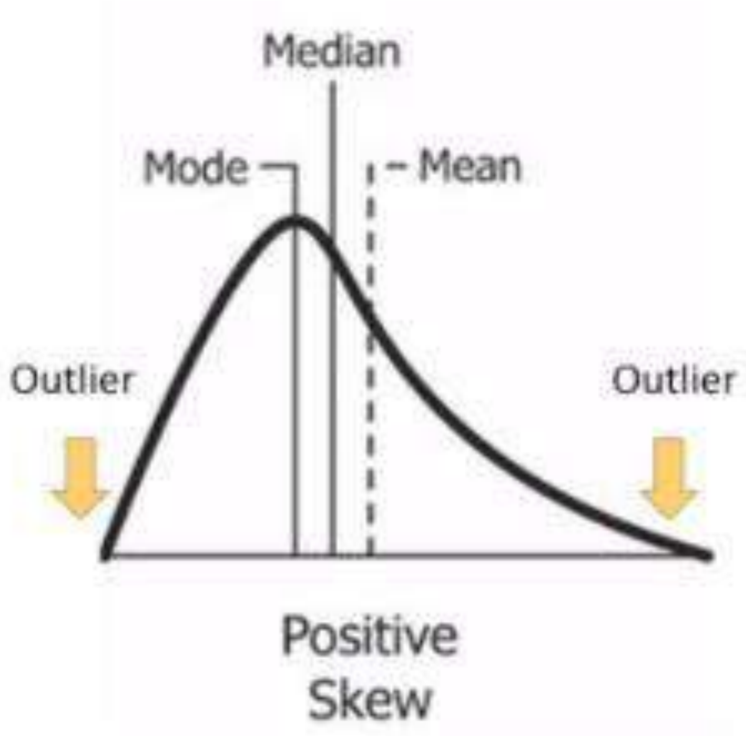
Outliers (cond..)

➢ **Detecting outliers using Normal distribution**



➢ **99% of the observations of a variable following a normal distribution lie within mean +/- 3 X standard deviation**

Outliers (cond..)



- Calculate the quantiles and the Inter-quantile range(IQR)

IQR = 75th Quantile – 25th Quantile

Upperlimit = 75th Quantile + IQR x 1.5

Lowerlimit = 25th Quantile + IQR x 1.5

Thank You!

In our next session: Imputation of Missing values

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Missing Value Imputation

Prof.Aruna Malapati

Learning objectives

- List different type Imputation techniques
- Choose appropriate techniques to impute missing values

Imputation Techniques for Numeric values

Mean / Median Imputation

- **Used when MCAR / MAR**
- **Assumes that the feature follows normal distribution**

- **Mean age = 33.14**
- **Median age = 32**

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

Pros and cons of mean/median imputation

- ➢ **Advantages**

✓ **Easy to implement**

✓ **Faster way of obtaining complete dataset**

- ➢ **Disadvantages**

✓ **Mean imputation reduces the variance of the imputed variables.**

✓ **Mean imputation does not preserve relationships between variables such as correlations.**

Random Sampling Imputation

- Used when MCAR / MAR
- Aim to preserve the statistical parameters of the feature
- Number of random samples are at least as many as missing values

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

Adding a new variable to indicate missingness

- **Used when MCAR / MAR**

Gender	Age	Age Missing?
Male	42	0
Male	NA	1
Male	24	0
Male	NA	1
Male	36	0
Male	57	0
Female	32	0
Female	NA	1
Female	NA	1
Female	18	1
Female	NA	0
Female	23	1

Imputation of NA by values at the end of distribution

- **Used when NMAR**

- **Imputed value can be 18 or 57 depending on other feature observation**

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

Imputation of NA by arbitrary values

- Used when NMAR
- Use any value except mean/median value

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

Imputation by most frequent category for categorical values

- **Used when NMAR**

Gender	Age
Male	42
Male	43
Male	24
NA	63
Male	36
Male	57
Female	32
NA	33
Female	45
Female	18
NA	55
Female	23

  - **Mode = Male**

In categorical variables treating NA as an additional category

- **Encode as unique category as unknown or missing**
- **Use mode to fill missing value**

Gender	Gender_new	Gender_new_value
Male	Male	Male
Male	Male	Male
Male	Male	Male
NA	Missing	Male
Male	Male	Male
Male	Male	Male
Female	Female	Female
NA	Missing	Male
Female	Female	Female
Female	Female	Female
NA	Missing	Male
Female	Female	Female

Thank You!

In our next session: Aggregation and Sampling

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Aggregation and Sampling

Prof.Aruna Malapati

Learning objectives

- Explain the effect of aggregation on variance in the data set.
- Define sampling and list various methods for sampling.

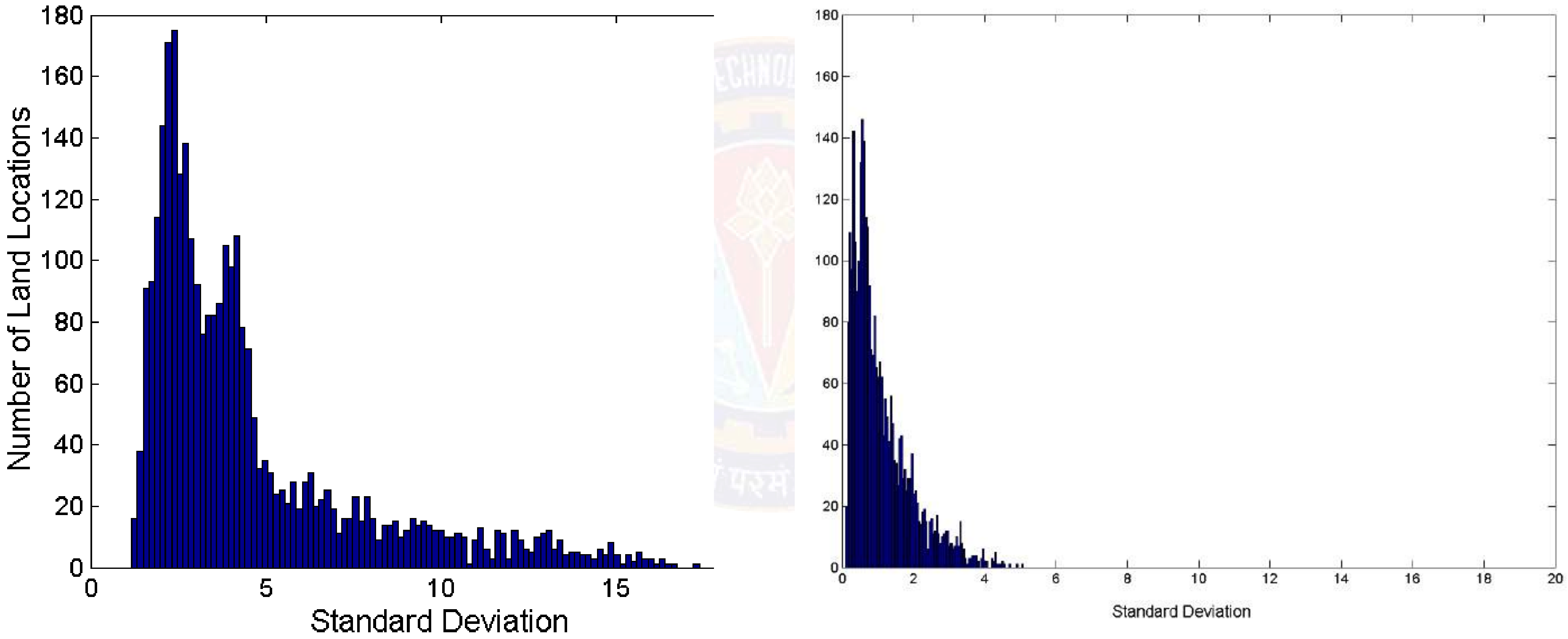
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)

- Purpose
  - ✓Data reduction
  - ✓Change of scale
  - ✓More stable data

Aggregation – Change in Variability

Less variability at “higher-level” view

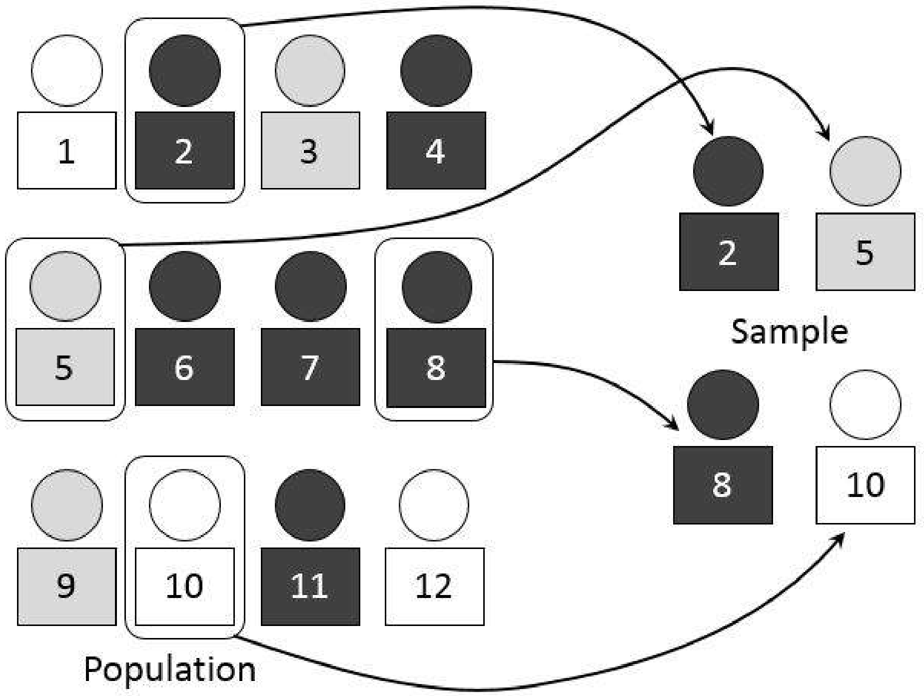


Standard Deviation of Average Monthly Precipitation

Standard Deviation of Average Yearly Precipitation

Sampling

- Processing the entire dataset may be too expensive and time consuming.
- Using a sample will work almost as well as using the entire data set, if the sample is representative.
- A sample is representative if it has approximately the same properties (of interest) as the original set of data.



Types of Sampling

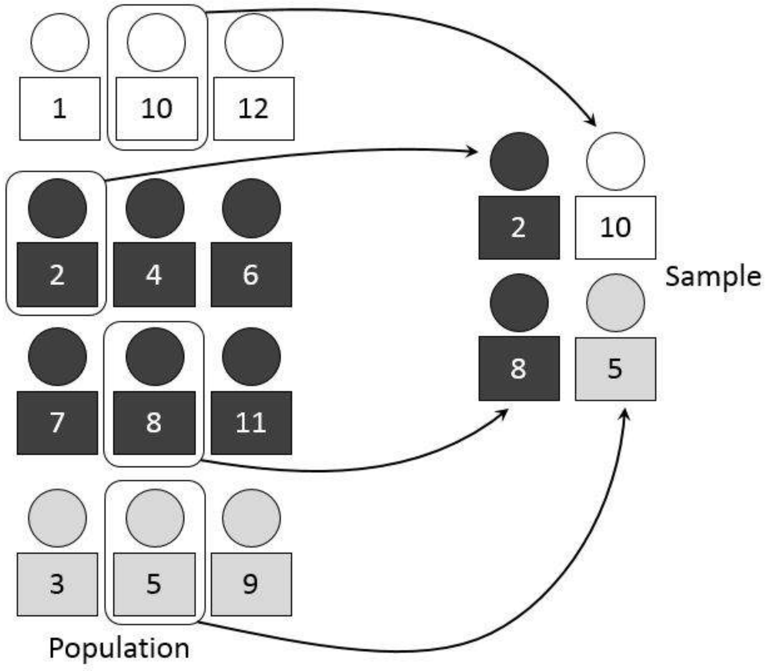
- Simple Random Sampling
- Stratified sampling

Simple Random Sampling

- **Sampling without replacement:** As each item is selected, it is removed from the population
- **Sampling with replacement:** Objects are not removed from the population as they are selected for the sample.
  - ✓ In sampling with replacement, the same object can be picked up more than once.

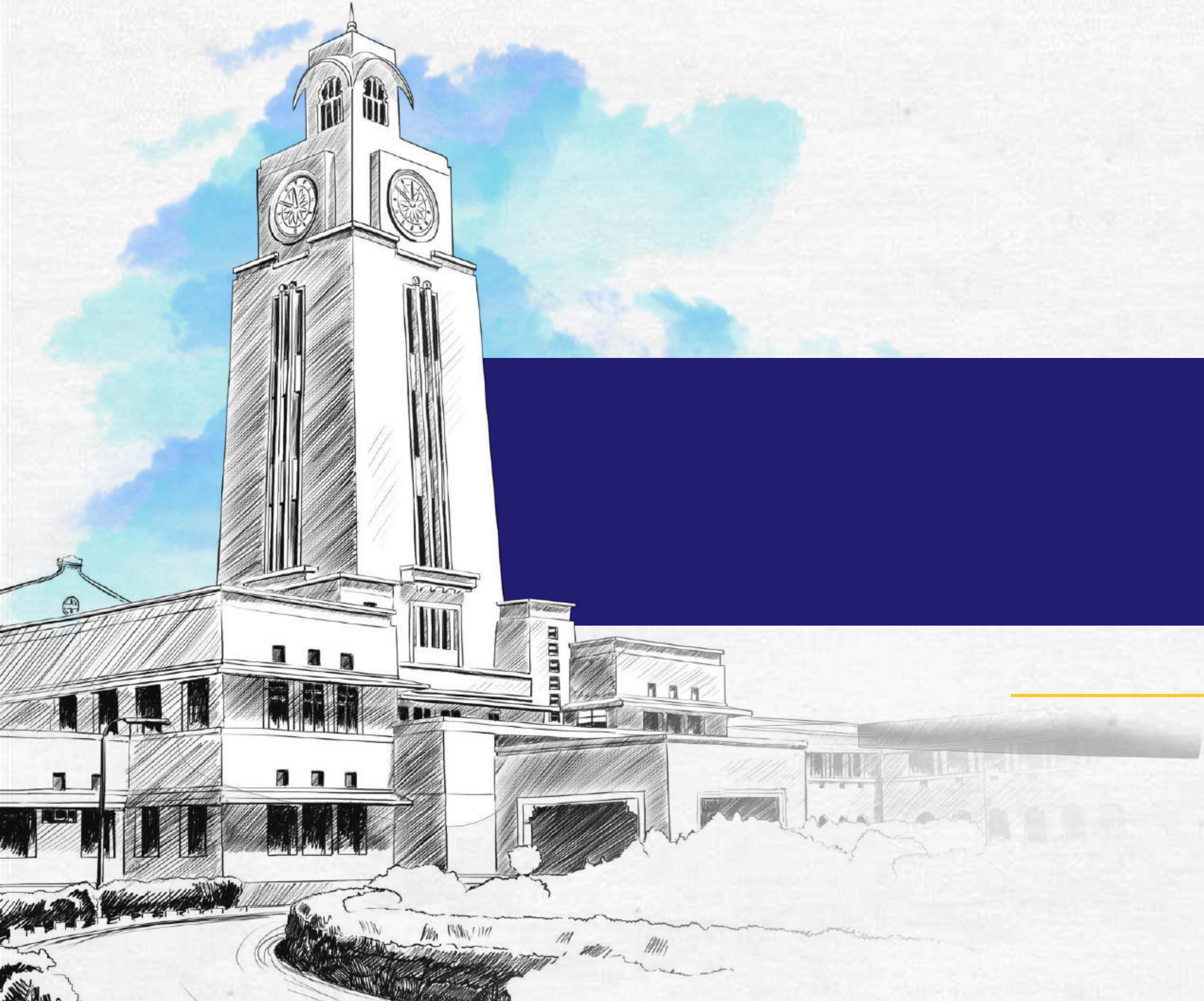
Stratified sampling

- Split the data into several partitions, then draw random samples from each partition



Thank You!

In our next session: Feature Creation



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Feature Creation

Prof.Aruna Malapti

Learning Objectives

- Define Feature creation.
- List the commonly used methodologies for Feature Creation.
- Identify appropriate techniques to be used given a data set.

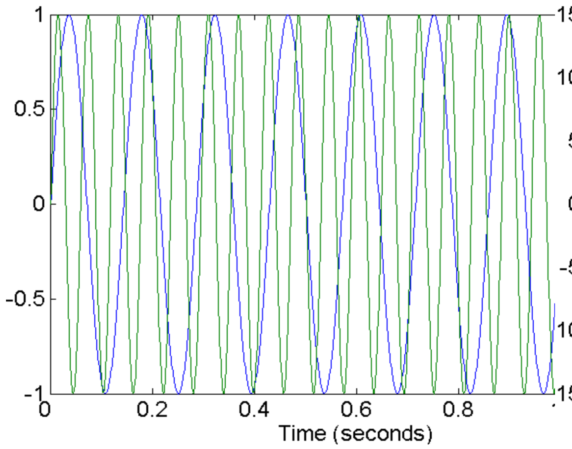
Feature Creation

- Create new attributes that can capture important information in a data set much more efficiently than the original attributes.

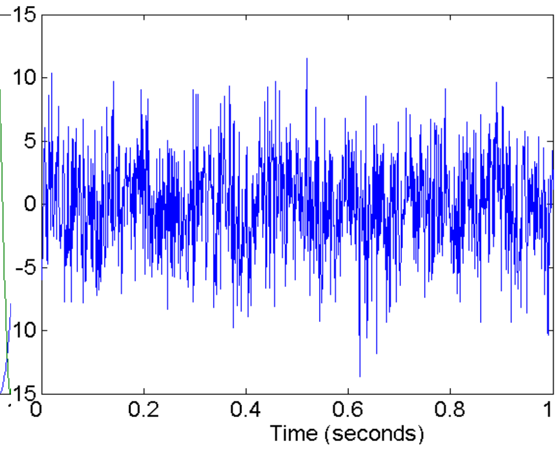
- Three general methodologies:
  - Feature Extraction
  - Mapping Data to New Space
  - Feature Construction

Mapping Data to a New Space

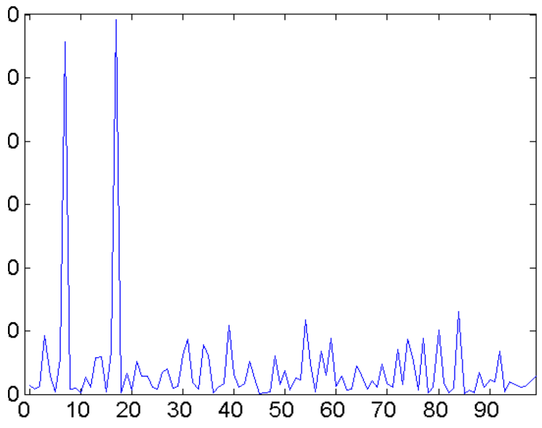
- Fourier transform
- Wavelet transform
- Scale-Invariant Feature Transform (SIFT)



Two Sine Waves



Two Sine Waves + Noise



Frequency

Feature construction

- Create dummy features: Often used to **convert categorical variable into numerical variables.**

Customer_ID	Gender	Paymet_Method	Online Banking	Credit Card	Debit Card
C001	FEMALE	Online Banking	1	0	0
C002	MALE	Online Banking	1	0	0
C003	FEMALE	Credit Card	0	1	0
C004	MALE	Debit Card	0	0	1

Feature construction

- Create derived features

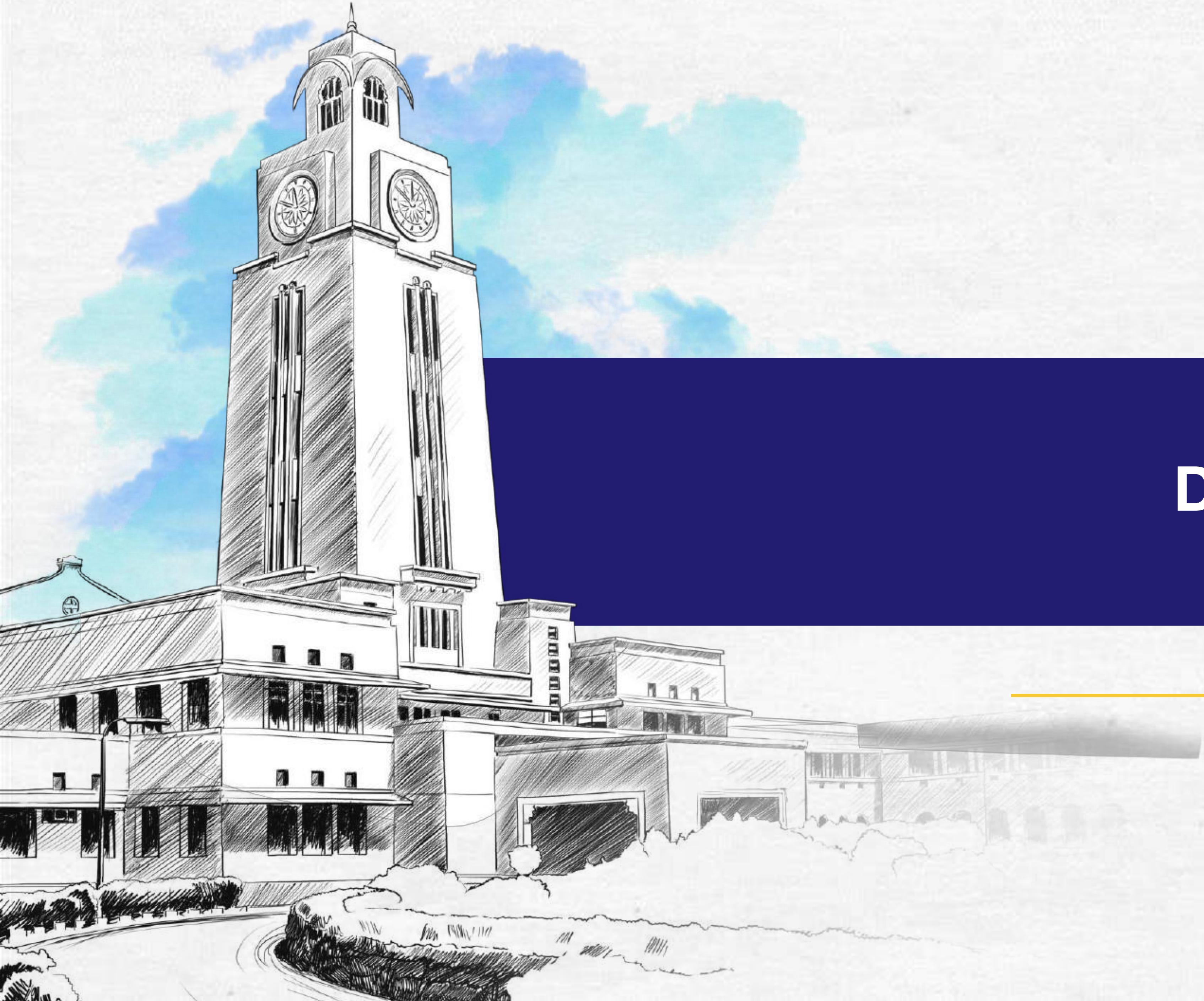
Customer_ID	Gender	Session_Begin	Session_End	Session_Duration
C001	FEMALE	15-06-2019 10.30	15-06-2019 11.15	45
C002	MALE	13-06-2019 8.00	13-06-2019 8.03	3
C003	FEMALE	2-06-2019 16.25	2-06-2019 18.35	126
C004	MALE	1-06-2019 11.20	1-06-2019 1.00	100

Derived features examples

- **Commonly used tricks**
  - ✓Handling date, time and addresses
  - ✓Handling sales and Marketing data
  - ✓Handling large range of values using standard transformations
  - ✓Encoding special objects as influencing person

Thank You!

In our next session: Discretization



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Data Transformation

Prof.Aruna Malapati

Learning Objectives

- List and define various data transformation methods.
- Articulate the need for feature scaling.
- Make the calculations that are necessary to get meaningful transformations.

Data Transformation

- **Data transformation tasks:**
  - ✓Normalization
  - ✓Attribute construction
  - ✓Aggregation
  - ✓Attribute Subset Selection
  - ✓Discretization
  - ✓Generalization

Linear Models

- $Y = W_0 + W_1 X$
- W indicates the change in Y per unit change of X
- If X changes scale, W will change its value
- Regression coefficients depend on the magnitude of the variable
- Features with bigger magnitude dominate over the features with smaller magnitudes
- Euclidian distances are sensitive to feature magnitude
- Hence it is a good practice to have all variables within a similar scale.

Algorithms that are sensitive to feature magnitude

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-Means Clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

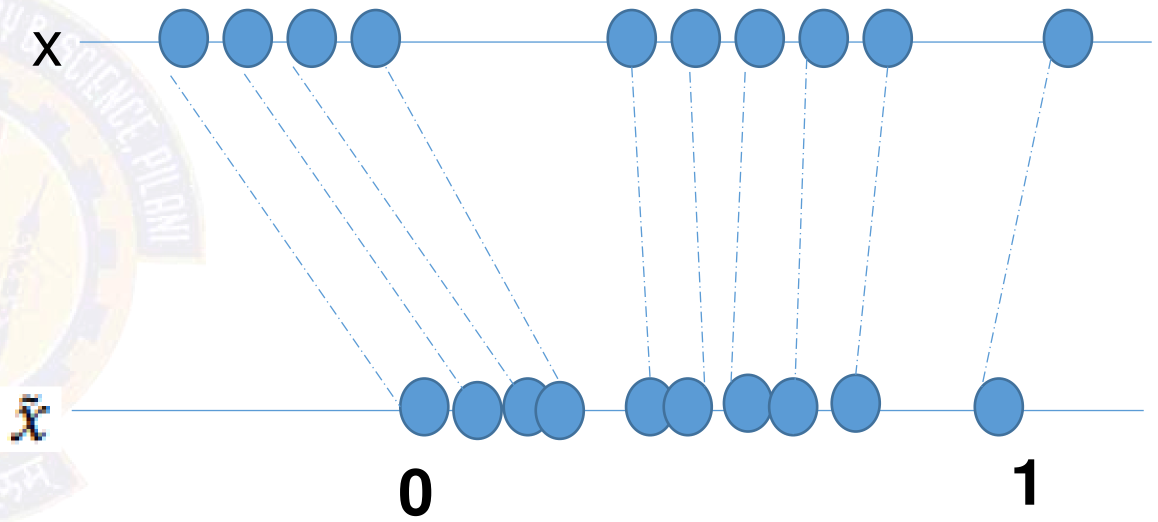
Normalization

- Types of common scaling operations or Normalization methods
  - ✓Min-max normalization
  - ✓z-score normalization
  - ✓Normalization by decimal scaling

Min-Max Scaling

➢ Min-max scaling squeezes (or stretches) all feature values to be within the range of [0, 1].

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$



x - feature value

min(x) - minimum value of feature x

max(x) - maximum value of feature x

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}(new\_max - new\_min) + new\_min$$

Example: Min-max Normalization

Let income range \$12,000 to \$98,000 be normalized to [0.0, 1.0].

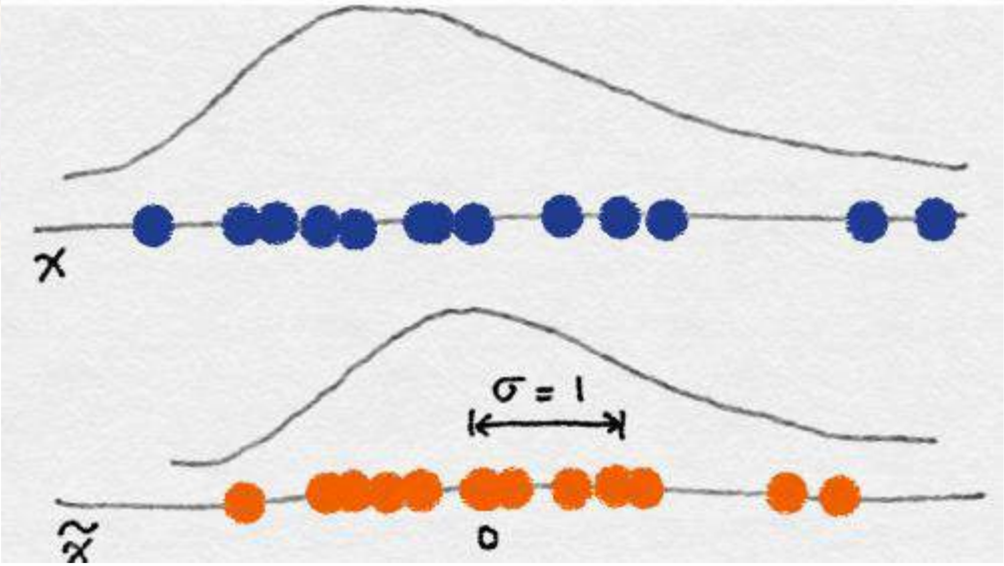
Then \$73,000 is mapped to ?

$$\frac{73{,}600-12{,}000}{98{,}000-12{,}000}(1.0-0)+0=0.716$$

z-score normalization

- In z-score normalization (or zero-mean normalization)
- The values for an attribute, x, are normalized based on the mean (x) and standard deviation (σ_x) of x.

$$\tilde{x} = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))}$$



- The resulting scaled feature has a mean of 0 and a variance of 1.

Example: z-score Normalization

Let μ_x= 54,000, σ_x = 16,000, for the attribute income

With z-score normalization, a value of $73,600 for income is transformed to:

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Z=score normalization can change the original data quite a bit.

Decimal Scaling

- Normalizes by moving the decimal point of values of attribute A.
- The number of decimal points moved depends on the maximum absolute value of A.
- A value, v, of A is normalized to v′ by computing

$$V' = \frac{V}{10^J}$$

where j is the smallest integer such that Max(|v′|) < 1.

Examples of Decimal Scaling

Example-1

CGPA	Formula	Normalized CGPA
2	2/10	0.2
3	3/10	0.3

Example-2

Bonus	Formula	Normalized Bonus
400	4/1000	0.4
310	3/1000	0.31

Example-3

Salary	Formula	Normalized Salary
40000	4/100000	0.4
31000	3/100000	0.31

Thank You!

In our next session: Feature subset selection

Discretization

Prof.Aruna Malapati

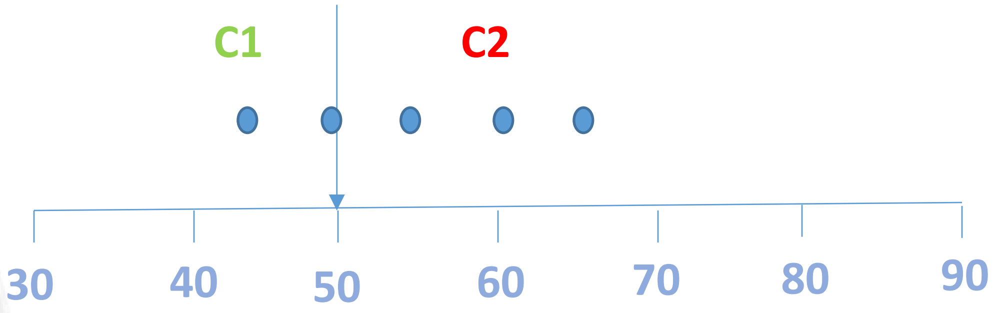
Learning Objective

- Define and list the discretization methods
- List and apply unsupervised discretization methods to any data set

Discretization

➢ **Convert continuous attribute into a discrete one**

Age	Alzimer
60	Yes
65	Yes
45	No
55	Yes
50	No



Main issues

- ✓ How to choose the number of intervals K ?
- ✓ How to define the cut points ?
- ✓ … which are relevant according to the studied problem….

Discretization

- Unsupervised discretization
  - ✓ Equal-interval binning
  - ✓ Equal-frequency binning

• Class labels are ignored
• The best number of bins k is determined experimentally

- Supervised discretization
  - ✓ Entropy-based discretization
  - ✓ It tries to maximize the "purity" of the intervals (i.e. to contain as less as possible mixture of class labels)

Unsupervised Discretization

- User specifies the number of intervals and/or how many data points should be included in any given interval.
- The following heuristic is often used to choose intervals:
  - ✓ The number of intervals for each attribute should not be smaller than the number of classes (if known).
  - ✓ The other popular heuristic is to choose the number of intervals, n_{Fi}, for each attribute, F_i (i=1,…,n,) where n is the number of attributes), as follows:

    $n_{Fi} = M/3 * C$ where M is the number of training examples and C is the number of known classes.

Methods for Binning Numeric Predictor Variables

- Equal width binning
- Equal frequency binning
- Binning by clustering

Original data:	53 56 57 63 66 67 67 67 68 69 70 70 70 70 72 73 75 75 76 76 78 79 80 81			
Method		**Bin1**	**Bin2**	**Bin3**
Equi Width	81-53=28 28/3=9.33	**[53,62)= {53,56,57}**	**[62,72)= {63,66,67,67,67,68, 69,70,70,70}**	**[72,81]={72,73,75,75,76, 76,78,79,80,81}**
Equi Frequency	24/3=8	**{53,56,57,63,66,67, 67,67}**	**{68,69,70,70,70, 72,73,75}**	**{75,76,76,78,79,80,81}**
Find natural gaps in the data	some variation	**{53,56,57,63,66,67, 67,67,68,69}**	**{70,70,70, 72,73,75,75}**	**{76,76,78,79,80,81}**

Thank You!

In our next session: Supervised Discretization

Supervised Discretization

Prof.Aruna Malapati

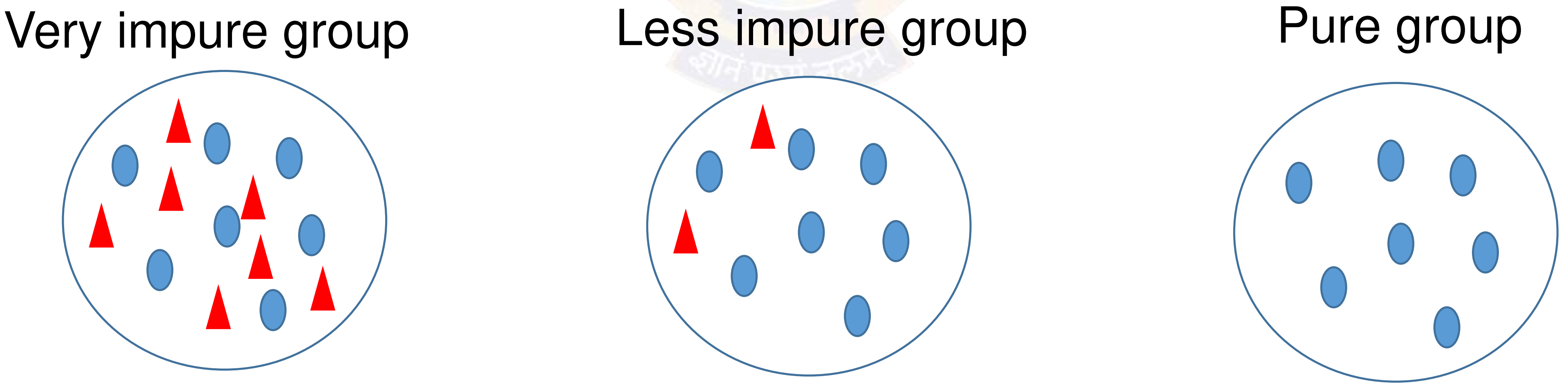
Learning Objective

- Apply supervised discretization methods to any data set
- Define Entropy and Information Gain

Entropy

- **Entropy E(S) - measure of the impurity/uncertainty in a group of examples**
  - ✓S - training set with $C_1,\ldots,C_C$ classes
  - ✓p_i - proportion of C_c in S

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$



Information Gain

- Information gain (IG) measures how much “information” a feature gives us about the class.
  - ✓ Features that perfectly partition should give maximal information.
  - ✓ Unrelated features should give no information.
- It measures the reduction in entropy.

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

Where S is the number of samples in the training set, with S_v instances belonging to class i, where i = 1, … , c.

Supervised Discretization - Entropy Based

- **Entropy Based Discretization**

1. Sort examples in increasing order
2. Each value forms an interval (m intervals)
3. Calculate the entropy measure of this discretization $E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$
4. Calculate "Entropy" for the target given a bin. $E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$
5. Calculate "Information Gain" given a bin. E(S) - E(S,A)
6. Apply the process recursively until some stopping criterion is met.

E(S) - E(S,A) > δ

Supervised Discretization Example

Runs	53	56	57	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	80	81
Matches Won	Y	Y	Y	N	N	N	N	N	N	N	N	Y	Y	Y	N	N	N	Y	N	N	N	N	N	N

In this example we are discretizing the feature Runs using 2 bins <=60 and >60.

Step 3: Calculate "Entropy" for the target.

Runs	
Y	N
7	17

E (Runs) = E(7, 17) = E(0.29, .71)
$= -0.29 \times \log_2(0.29) - 0.71 \times \log_2(0.71)$
= 0.871

Step 4: Calculate "Entropy" for the target given a bin.

		Matches Won	
		Y	N
Runs	<= 60	3	0
	> 60	4	17

E (Matches Won,Runs) = P(<=60) x E(3,0) + P(>60) x E(4,17)
= 3/24 x 0 + 21/24 x 0.7= 0.615

Step 5: Calculate "Information Gain" given a bin.

Information Gain (Matches Won,Runs) = 0.256

Supervised Discretization Example (Contd..)

Runs	53	56	57	63	66	67	67	67	68	69	70	70	70	70	72	73	75	75	76	76	78	79	80	81
Matches Won	Y	Y	Y	N	N	N	N	N	N	N	N	Y	Y	Y	N	N	N	Y	N	N	N	N	N	N

		Matches Won	
		Y	N
Runs	<= 60	3	0
	> 60	4	17

Information Gain = 0.256

		Matches Won	
		Y	N
Runs	<= 70	6	8
	> 70	1	9

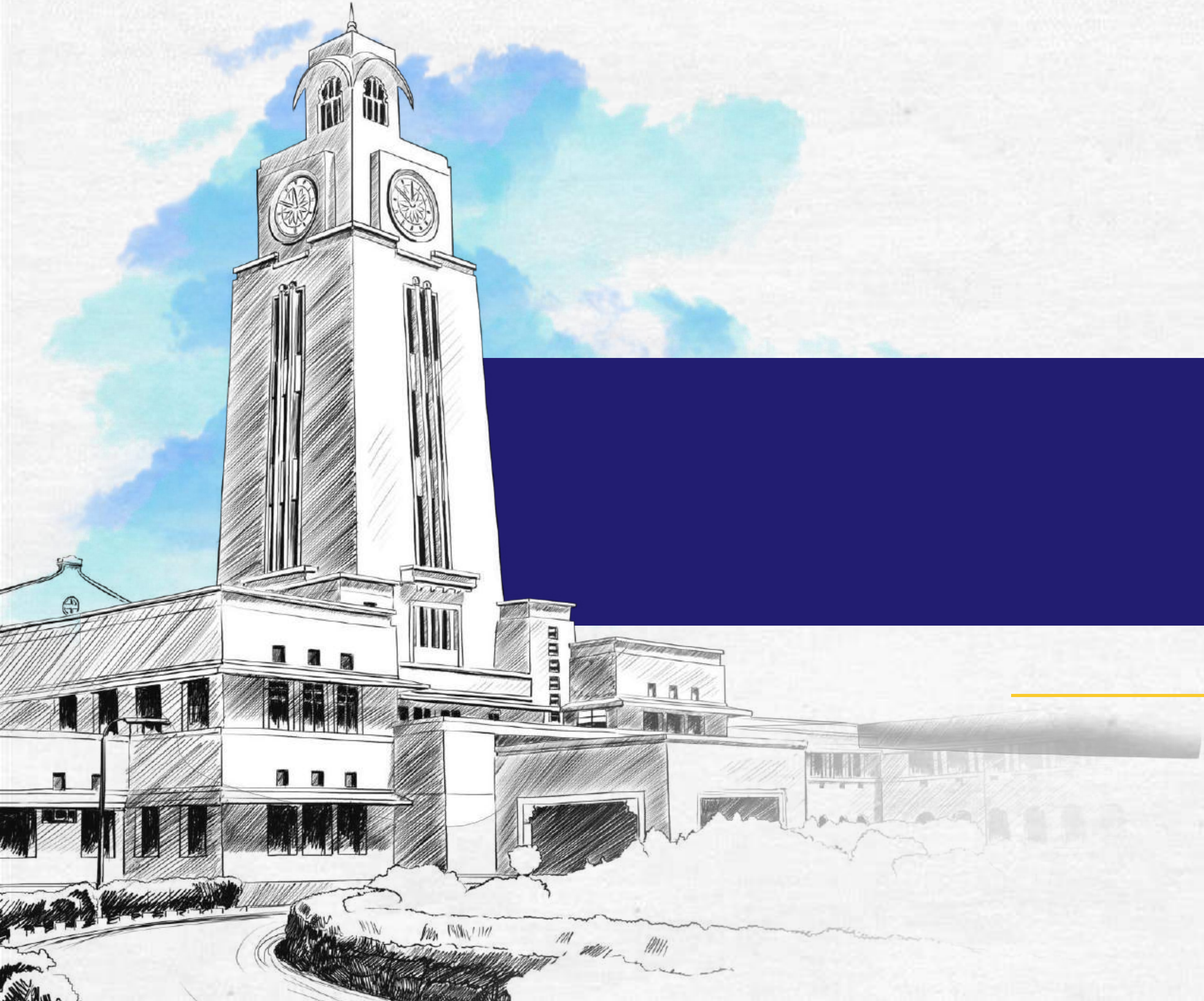
Information Gain = 0.101

		Matches Won	
		Y	N
Runs	<= 75	7	11
	> 75	0	6

Information Gain = 0.148

Thank You!

In our next session: Binarization



Binarization

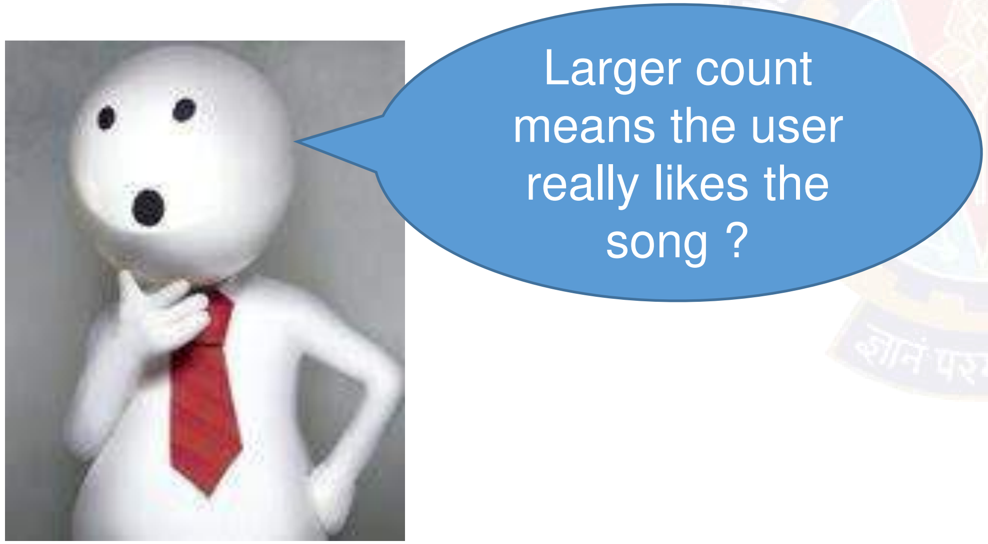
Prof.Aruna Malapati

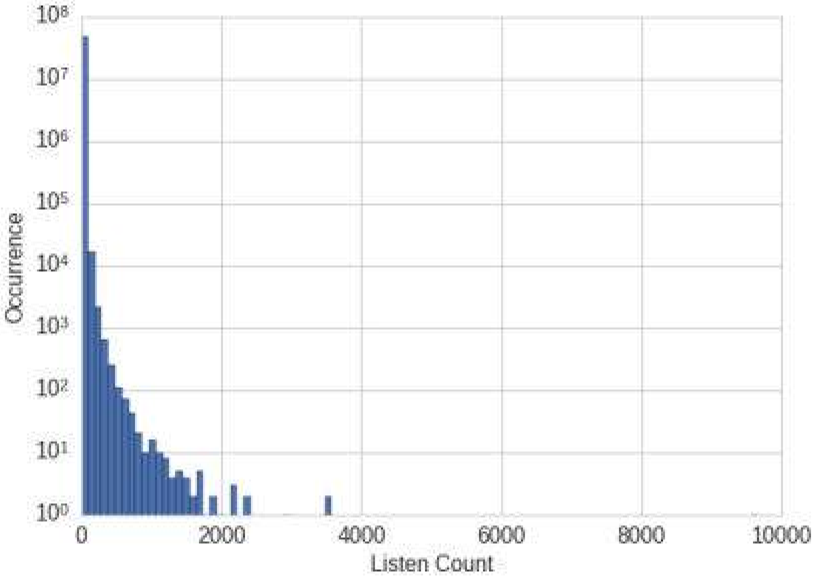
Learning Objective

- Apply Binarization
- List the issues encountered during binarization

Motivation for Binarization

- **Echo Nest Taste Profile Dataset:**
  - **There are more than 48 million triplets of user ID, song ID, and listen count.**
  - **The full dataset contains 1,019,318 unique users and 384,546 unique songs.**
- Build a recommender system to recommend songs to users.





99% of the listen counts are 24 or lower

Raw listen count is not a robust measure of user taste.

Binarization

- ➢Binarization maps a continuous or categorical attribute into one or more binary variables.
- ➢Must maintain ordinal relationship
- ✓Assume an ordinal attribute for representing service of a restaurant: ({Awful,Poor,OK,Good,Great})

Service quality	Integer Value	X1	X2	X3
Awful	0	0	0	0
Poor	1	0	0	1
Ok	2	0	1	0
Good	3	0	1	1
Great	4	1	0	0

Unintended relationships: X2 and X3 are now correlated because “good” is encoded using both attributes

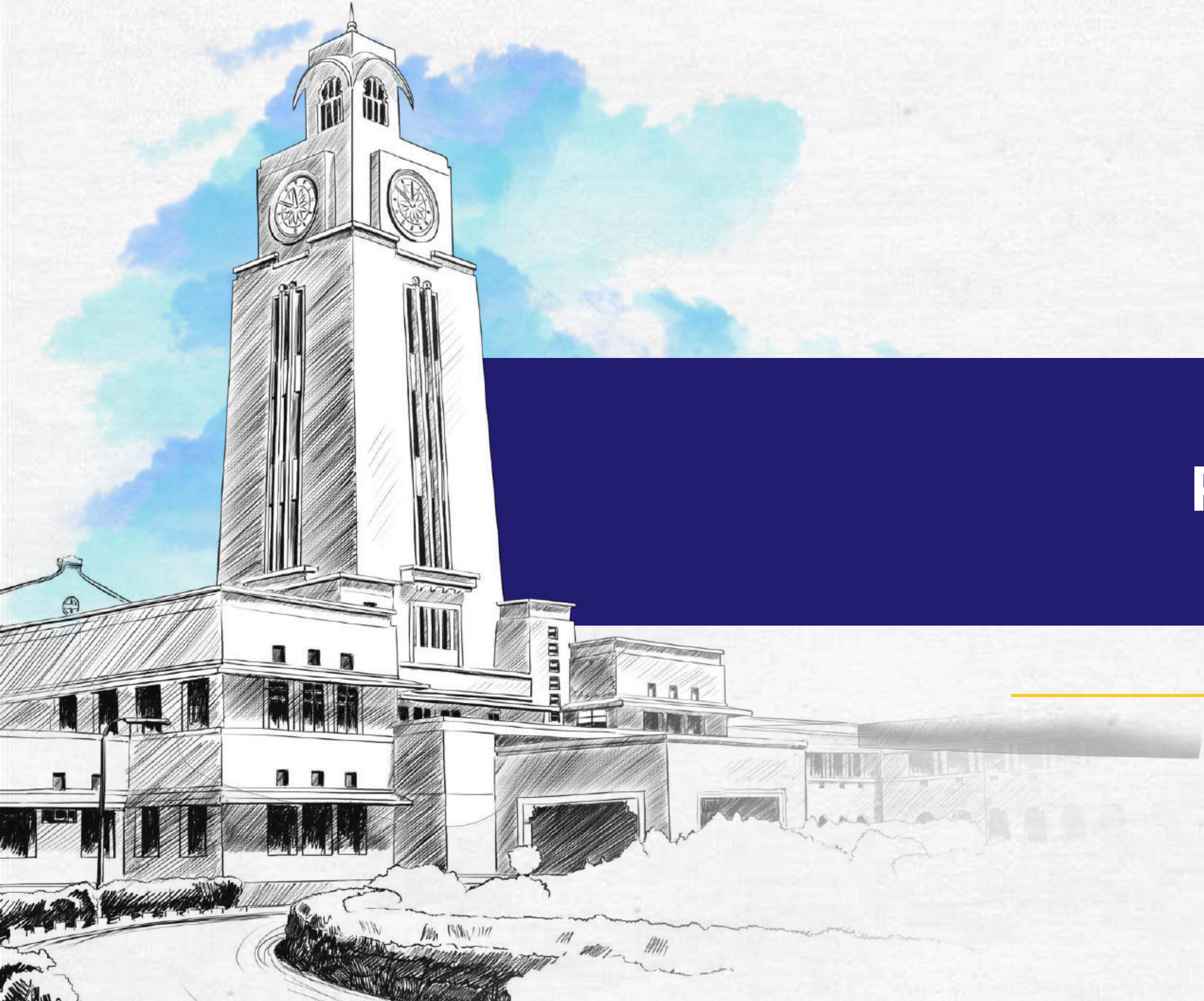
Binarization

Service quality	Integer Value	X1	X2	X3	X4	X5
Awful	0	1	0	0	0	0
Poor	1	0	1	0	0	0
Ok	2	0	0	1	0	0
Good	3	0	0	0	1	0
Great	4	0	0	0	0	1

- ✓Binary attributes, where only the presence of 1 is important
- ✓One binary attribute for each categorical value
- ✓Be Careful: Number of resulting attributes may become too large

Thank You!

In our next session: Proximity measures for binary attributes



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Proximity Measures

Prof.Aruna Malapati

Learning Objective

- Define proximity/similarity
- Dissimilarity/Similarity for Binary Attributes and its variants

Proximity

- For many problems we need to quantify how close two objects are.
- Examples:
  - ✓ For an item bought by a customer, find other similar items
  - ✓ Group together the customers of site so that similar customers are shown the same ad.
  - ✓ Group together web documents so that you can separate the ones that talk about politics and the ones that talk about sports.
  - ✓ Find all the near-duplicate mirrored web documents.
  - ✓ Find credit card transactions that are very different from previous transactions.
- To solve these problems we need a definition of similarity, or distance.

Proximity (contd..)

- Similarity
  - ✓Numerical measure of how **alike** two data objects are.
  - ✓Is higher when objects are more alike.
  - ✓Often falls in the range [0,1]
  - ✓Examples: Cosine,Jaccard,Tanimoto
- Dissimilarity
  - ✓Numerical measure of how **different** two data objects are
  - ✓Lower when objects are more alike
  - ✓Minimum dissimilarity is often 0
  - ✓Upper limit varies

Proximity Measures for Single Nominal attribute

- ✓ Suppose a binary attribute Gender = {Male, female} where Male is equivalent to binary 1 and female is quivalent to binary 0.
- ✓ The similarity value(p) is 1 if the two objects contains the same attribute value, 0 otherwise.

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

- ✓ $p(Ram, sita) = 0$
- ✓ $p(Ram, Laxman) = 1$

- ✓ Note : In this case, if q denotes the dissimilarity between two objects $i \ and \ j$ with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

Proximity Measures for Two or more Nominal attribute

✓We define the contingency table summarizing the different matches and mismatches between any two objects x and y, which are as follows.

Contingency table with binary attributes

Object x \ Object y	1	0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

Here, f_{11}= the number of attributes where x=1 and y=1.

f_{10}= the number of attributes where x=1 and y=0.

f_{01}= the number of attributes where x=0 and y=1.

f_{00}= the number of attributes where x=0 and y=0

✓Note : $f_{00} + f_{01} + f_{10} + f_{11}$ will total number of binary attributes.

✓Two cases of binary attributes may arise: symmetric and asymmetric binary attributes.

Similarity Measure for Symmetric Binary attribute

- Symmetric binary coefficient($\mathcal{S}$) is used to measure the similarity between two objects and is defined as

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

- The dissimilarity measure($\mathcal{D}$) is defined as

$$\mathcal{D} = \frac{Number\ of\ mismatched\ attribute\ values}{Total\ number\ of\ attributes}$$

or

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Similarity Measure with Symmetric Binary

Consider the following two dataset, where objects are defined with symmetric binary attributes.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I}, Hobby = {T, C}, Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{S}(\text{Hari, Ram}) = \frac{2+1}{2+2+1+1} = 0.5$$

	1	0
1	1	1
0	2	2

Proximity Measure with Asymmetric Binary

- **Jaccard Coefficient** is used to measure the similarity between two objects is symbolized by $\mathcal{J}$ and is defined as follows

$$\mathcal{J} = \frac{Number\ of\ matching\ presence}{Number\ of\ attributes\ not\ involved\ in\ 00\ matching}$$

or

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity Measure with Asymmetric Binary

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I}, Hobby = {T, C}, Job = {Y, N}

Compute the Jaccard coefficient between Ram and Hari assuming that all binary attributes are asymmetric and for each pair values for an attribute, first one is more important than the second.

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

	1	0
1	1	1
0	2	2

Note: $\mathcal{J}$(Hari, Ram) = $\mathcal{J}$(Ram, Hari)

Thank You!

In our next session: Proximity measures for categorical attributes

Proximity Measures for Categorical Attributes

Prof.Aruna Malapati

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Learning Objectives

- Define and compute proximity measures between objects when the attributes are categorical

Proximity Measures for Categorical Attribute

- Attributes with three or more states (e.g. color = {Red, Green, Blue}) are called nominal.
- If $s(x, y)$ denotes the similarity between two objects $x \; and \; y$, then

$$s(x, y) = \frac{Number \; of \; matches}{Total \; number \; of \; attributes}$$

- and the dissimilarity $d(x, y)$ is

$$d(x, y) = \frac{Number \; of \; mismatches}{Total \; number \; of \; attributes}$$

- If m = number of matches and a = number of the categorical attribute for object x and y then s and D are defined as

$$s(x, y) = \frac{m}{a} \quad \text{and} \quad d(x, y) = \frac{a-m}{a}$$

Proximity Measures for Categorical Attribute

Object	Color	Position	Distance
1	R	L	L
2	B	C	M
3	G	R	M
4	R	L	H

Proximity Measure for Ordinal Attribute

- Ordinal attribute is a special kind of categorical attribute, where the values of attribute follows a sequence (ordering) e.g. Grade = {Ex, A, B, C} where Ex > A >B >C.
- Suppose, *A* is an attribute of type ordinal and the set of values of $A = \{a_1, a_2, \ldots\ldots, a_n\}$. Let n values of A are ordered in ascending order as $a_1 < a_2 < .. < a_n$. Let *i-th* attribute value a_i be ranked as *i, i=1,2,..n.*
- The normalized value of a_i can be expressed as

$$\hat{a}_i = \frac{i-1}{n-1}$$

- Thus, normalized values lie in the range $[0..1]$.
- As a_i is a numerical value, the similarity measure, then can be calculated using any similarity measurement method for numerical attribute.
- For example, the similarity measure between two objects $x \; and \; y$ with attribute values a_i and a_j, then can be expressed as

$$s(x, y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

where $\hat{a}_i$ and $\hat{a}_i$ are the normalized values of $\hat{a}_i$ and $\hat{a}_i$, respectively.

Proximity Measure for Ordinal Attribute

Consider the following set of records, where each record is defined by two ordinal attributes size={S, M, L} and Quality = {Ex, A, B, C} such that S<M<L and Ex>A>B>C.

Object	Size	Quality
A	S (0.0)	A (0.66)
B	L (1.0)	Ex (1.0)
C	L (1.0)	C (0.0)
D	M (0.5)	B (0.33)

Find the dissimilarity matrix, when each object is defined by only one ordinal attribute say size (or quality).

Thank You!

In our next session: Proximity Measures for continuous attributes

Proximity Measures for continuous attributes

Prof.Aruna Malapati

Learning Objectives

- Define and compute proximity measures between objects when the attributes are of continuous type

Properties of distance measures

- Distance $d\,(p, q)$ between two points p and q is a dissimilarity measure if it satisfies:
  1. **Positive definiteness:**

     $d\,(p, q) \geq 0$ for all p and q and

     $d\,(p, q) = 0$ only if $p = q$.
  2. **Symmetry:** $d\,(p, q) = d\,(q, p)$ for all p and q.
  3. **Triangle Inequality:**

     $d\,(p, r) \leq d\,(p, q) + d\,(q, r)$ for all points p, q, and r.

Proximity Measure with Interval Scale

- The generic formula to express distance d between two objects $x \; and \; y$ in *n*-dimensional space.

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{\frac{1}{r}}$$

Here, r is any integer value, $x_i \; and \; y_i$ denote the values of *i*th attribute of the objects $x \; and \; y$ respectively

- This distance metric most popularly known as Minkowski metric.

Proximity Measure with Interval Scale

Manhattan distance (L_1 Norm: r = 1)

The Manhattan distance is expressed as

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

where $|\ldots|$ denotes the absolute value.

This metric is also alternatively termed as **Taxicabs metric, city-block metric**.

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance is $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

- As a special instance of Manhattan distance, when attribute values $\in [0, 1]$ is called Hamming distance.
- Alternatively, Hamming distance is the number of bits that are different between two objects that have only binary values (i.e. between two binary vectors).

Proximity Measure with Interval Scale

Euclidean Distance (L_2 Norm: r = 2)

This metric is same as Euclidean distance between any two points $x \ and \ y \ in \ \mathcal{R}^n$.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Euclidean distance between $x \ and \ y$ is

$$d(x, y) = \sqrt{(7-3)^2 + (3-2)^2 + (5-6)^2} = \sqrt{18} \approx 2.426$$

Proximity Measure with Interval Scale

Chebychev Distance ($L_\propto$ Norm: $r \in \mathcal{R}$)

This metric is defined as

$$d(x, y) = \max_{\forall i}\{|x_i - y_i|\}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance = $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

The chebychev distance = $\text{Max}\ \{|7 - 3|, |3 - 2|, |5 - 6|\} = 4.$

Proximity Measure for Ratio scale

The proximity between the objects with ratio-scaled variable can be carried with the following steps:

1. Apply appropriate transformation to the data to bring it into a linear scale. (e.g. logarithmic transformation to data of the form $X = Ae^{B}$.
2. The transformed values can be treated as interval-scaled values. Any distance measure discussed for interval-scaled variable can be applied to measure the similarity.

Proximity Measure for Ratio scale

Normalization:

- A major problem when using the similarity (or dissimilarity) measures (such as Euclidean distance) is that the large values frequently swamp the small ones.
- For example, consider the following data.

Make	Cost 1	Cost 2	Cost 3
X	2,00,000	70	10
Y	2,50,000	100	5

- Here, the contribution of Cost 2 and Cost 3 is insignificant compared to Cost 1 so far the Euclidean distance is concerned.
- This problem can be avoided if we consider the normalized values of all numerical attributes.

Proximity Measure for Mixed Attributes

- The previous metrics on similarity measures assume that all the attributes were of the same type. Thus, a general approach is needed when the attributes are of different types.
- One straightforward approach is to compute the similarity between each attribute separately and then combine these attribute using a method that results in a similarity between 0 and 1.
- Typically, the overall similarity is defined as the average of all the individual attribute similarities.

Proximity Measure with Vector Objects

Suppose, the objects are defined with $A_1, A_2, \ldots\ldots, A_n$ attributes.

1. For the *k-th* attribute (*k* = 1, 2, . . . , *n*), compute similarity $s_k(x, y)$ in the range [0, 1].
2. Compute the overall similarity between two objects using the following formula

$$similarity\ (x, y) = \frac{\sum_{i=1}^{n} s_i(x, y)}{n}$$

3. The above formula can be modified by weighting the contribution of each attribute. If the weight w_k is for the *k-th* attribute, then

$$w\_similarity\ (x, y) = \frac{\sum_{i=1}^{n} w_i s_i(x,y)}{n}$$ such that $\sum_{i=1}^{n} w_i = 1.$

4. The definition of the Minkowski distance can also be modified as follows:

$$d(x, y) = \left(\sum_{i=1}^{n} w_i |x_i - y_i|^r \right)^{\frac{1}{r}}$$

Proximity Measure with Mixed Attributes

Consider the following set of objects.

Object	A (Binary)	B (Categorical)	C (Ordinal)	D (Numeric)	E (Numeric)
1	Y	R	X	475	10^8
2	N	R	A	10	10^{-2}
3	N	B	C	1000	10^5
4	Y	G	B	500	10^3
5	Y	B	A	80	1

Non-Metric Similarity

- In many applications (such as information retrieval) objects are complex and contains a large number of symbolic entities (such as keywords, phrases, etc.).
- To measure the distance between complex objects, it is often desirable to introduce a non-metric similarity function.

Cosine similarity

Suppose, *x* and *y* denote two vectors representing two complex objects. The cosine similarity denoted as $\cos(x, y)$ and defined as

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

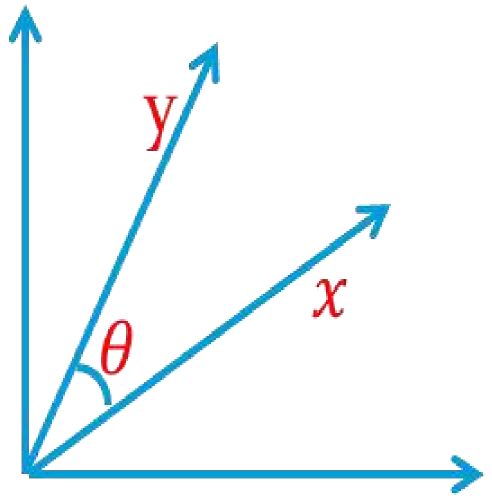
- where $x \cdot y$ denotes the vector dot product, namely $x \cdot y = \sum_{i=1}^{n} x_i \cdot y_i$ such that $x = [x_1, x_2, .., x_n]$ and $y = [y_1, y_2, .., y_n]$.
- $\|x\|$ and $\|y\|$ denote the Euclidean norms of vector *x* and *y*, respectively (essentially the length of vectors *x* and *y*), that is
    - $\|x\| = \sqrt{{x_1}^2 + {x_2}^2 + .. + {x_n}^2}$ and $\|y\| = \sqrt{{y_1}^2 + {y_2}^2 + .. + {y_n}^2}$

Cosine Similarity

- In fact, cosine similarity essentially is a measure of the (cosine of the) angle between *x* and *y*.
- Thus if the cosine similarity is 1, then the angle between *x* and *y* is 0° and in this case, *x* and *y* are the same except for magnitude.
- On the other hand, if cosine similarity is 0, then the angle between *x* and *y* is 90° and they do not share any terms.
- Considering, this cosine similarity can be written equivalently

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} = \hat{x} \cdot \hat{y}$$

where $\hat{x} = \frac{x}{\|x\|}$ and $\hat{y} = \frac{y}{\|y\|}$. This means that cosine similarity does not take the magnitude of the two vectors into account, when computing similarity.



- It is thus, one way normalized measurement.

Non-Metric Similarity

Cosine Similarity

Suppose, we are given two documents with count of 10 words in each are shown in the form of vectors *x* and *y* as below.

$$x = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0] \text{ and } y = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

Thus, $x \cdot y$ = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$$\|x\| = \sqrt{3^2 + 2^2 + 0 + 5^2 + 0 + 0 + 0 + 2^2 + 0 + 0} = 6.48$$

$$\|y\| = \sqrt{1^2 + 0 + 0 + 0 + 0 + 0 + 0 + 1^2 + 0 + 2^2} = 2.24$$

$$\therefore \quad \cos(x, y) = 0.31$$

Thank You!

In our next session: Curse of Dimensionality

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Curse of Dimensionality

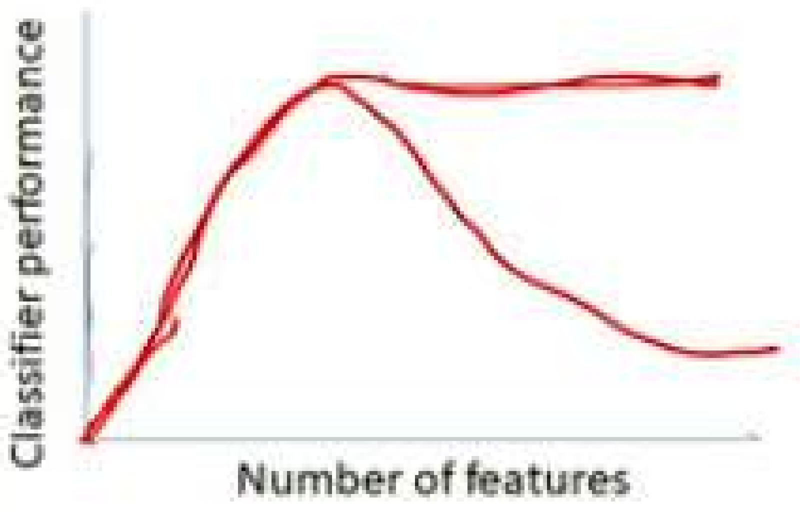
Prof.Aruna Malapati

Learning Objective

- Articulate the effects of curse of dimensionality

Curse of Dimensionality

- As dimensionality increases the number of data points required for a classification model also increase exponentially.
- Hughes Phenomenon: For a fixed number of training samples(N) in the data set the performance of the models decreases as dimensionality increase.



Reasons for this phenomenon:

- ✓ Redundant Features – Carry same data in some other form
- ✓ Correlation between features – the presence of one feature influence the other.
- ✓ Irrelevant Features - those that are simply unnecessary

Curse of Dimensionality

✓ The intuitions of distances in 3D are invalid in higher dimensions.

✓ For example consider a data point x_i from N samples in 1D

$dist_{min}(x_i) = \min_{xi \neq xj}\{dist(x_i, x_j)\}$ The minimum of distance between x_i and x_j such that $x_i \neq x_j$

$dist_{max}(x_i) = \max_{xi \neq xj}\{dist(x_i, x_j)\}$ The maximum of distance between x_i and x_j such that $x_i \neq x_j$

$$\lim_{d \to \infty} \frac{\text{dist}_{\max} - \text{dist}_{\min}}{\text{dist}_{\min}} = 0$$

✓ $dist_{max}(x_i) \approx dist_{min}(x_i)$ that means every pair of points are approximately at the same distance from each other.

➢ **Distance measures become meaningless in higher dimensions.**

Euclidean distance VS Cosine similarity

- Euclidean distance in high dimensionality does not make the sense solution for this is using cosine similarity for high dimensional spaces.
- Impact of dimensionality on cosine similarity is lower as compared to the Euclidean distance.
- If the data is dense then it's impact will be high and if it is sparse then impact will be lower that means in sparse most of values are 0 so data is non uniformly spread.

Feature Subset Selection
Dimension Reduction

Thank You!

In our next session:Feature Subset Selection

Feature Subset Selection

Prof.Aruna Malapati

Importance of feature subset selection

The objective of feature selection is three-fold:

- ✓Improving the prediction performance of the models
- ✓Reduction in the training time required to build model
- ✓Providing a better understanding of the underlying process that generated the data

What is Feature Selection for classification?

- Given: A set of predictors (“features”) $F=\{f_1,f_2,f_3 \dots f_D\}$ and target class label T.
- Find: Minimum subset $F'=\{f_1',f_2',f_3' \dots f_M'\}$ that achieves maximum classification performance where $F' \subseteq F$.

Feature subset selection

- ✓ Given D initial set of features
- ✓ There are 2^D possible subsets.
- ✓ Need a criteria to decide which subset is the best:
  - ✓ Classifier based on these m features has the lowest probability of error of all such classifiers.
- ✓ Evaluating 2^D possible subsets is time consuming and expensive.
- ✓ Use heuristics to reduce the search space.

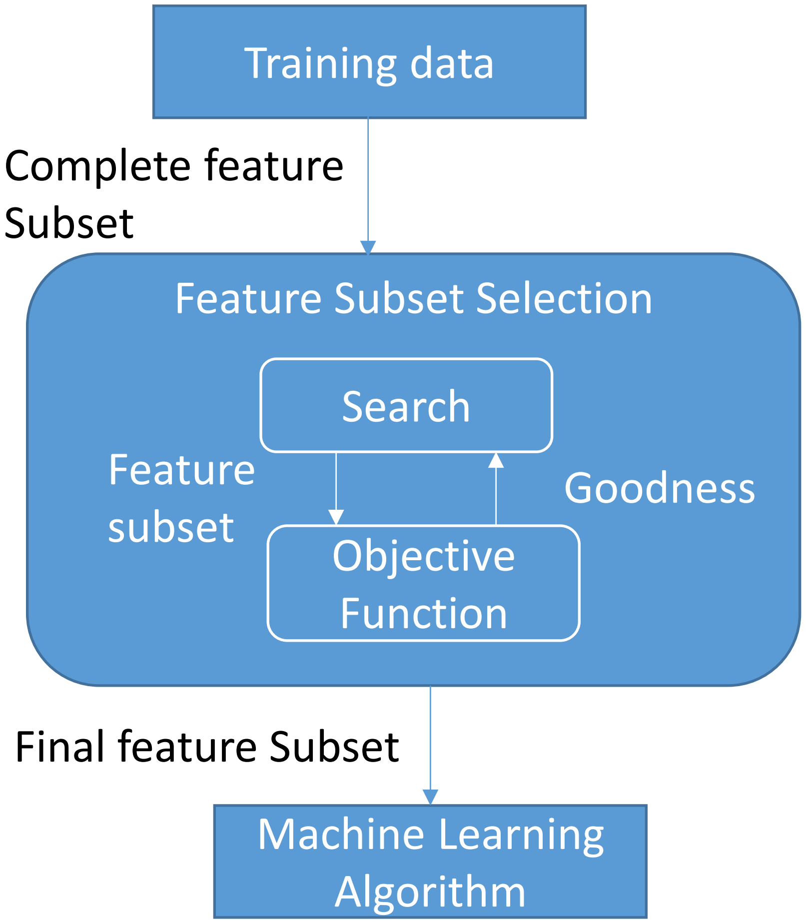
Feature Selection approaches

Three approaches to evaluate 2^D possible subsets

- Unsupervised (Filter Methods)
  - ✓ Use only features/predictor variables
  - ✓ Select the features that have the most information
- Supervised: Wrapper Methods
  - ✓ Train using the selected subset
  - ✓ Estimate error on the validation set
- Embedded Methods
  - ✓ Feature selection is done while training the model

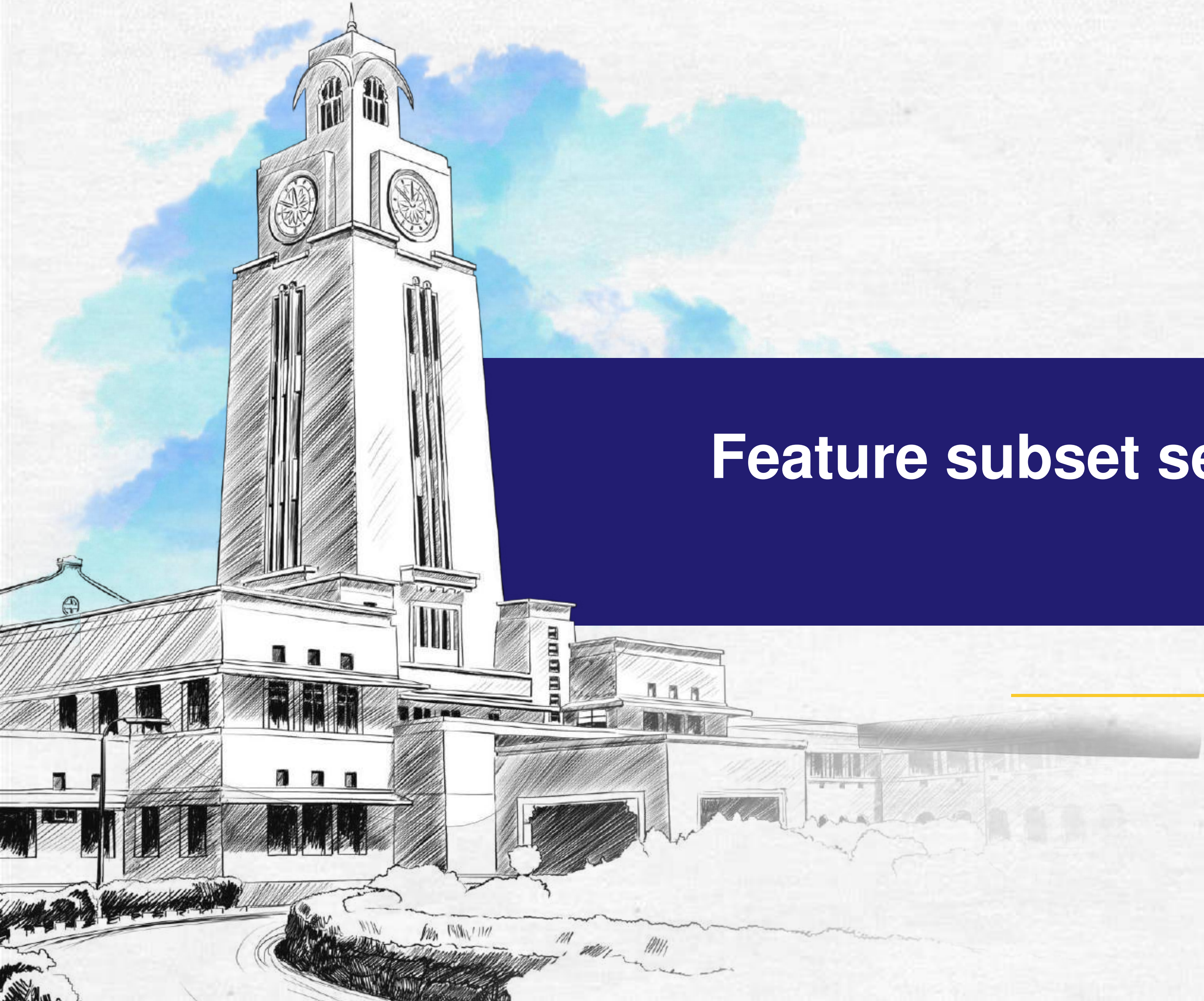
Steps in Feature Selection

- Feature selection is an optimization problem having the following steps:
- Step1: Search the space of all possible features
- Step2: Pick the optimal subset using an objective function



Thank You!

In our next session:Feature selection using Filter Methods



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Feature subset selection using Filter methods

Prof.Aruna Malapati

Learning Objectives

- Formulate the problem of filter methods for feature subset selection
- List various filters
- Define the Pearson correlation filter for regression
- Explain no free lunch theorem

Filter Methods

- The Predictive power of individual feature is evaluated.
- Rank each feature according to some univariate metric and select the highest ranking features.
- The score should reflect the discriminative power of each feature.

Input: large feature set Ω

1 Identify candidate subset $S \subseteq \Omega$

2 While !stop criterion()

Evaluate utility function J using S.

Adapt S

3 Return S.

Pros: fast, provides generically useful feature set

Cons: cause higher error than wrappers

Types of Filters

- **Univariate filters** evaluate each feature independently with respect to the target variable.
  - ✓ Correlation
  - ✓ Fisher Score
  - ✓ Mutual Information (Information Gain)
  - ✓ Gini index
  - ✓ Gain Ratio
  - ✓ Chi-Squared test
- **Multivariate filters** evaluate features in context of others.

Types of filters

- Correlation-based
  - ✓ Pearson product-moment correlation
  - ✓ Spearman rank correlation
  - ✓ Kendall concordance
- Statistical/probabilistic independence metrics
  - ✓ Chi-square statistic
  - ✓ F-statistic
  - ✓ Welch's statistic
- Information-theoretic metrics
  - ✓ Mutual Information (Information Gain)
  - ✓ Gain Ratio
- Others
  - ✓ Fisher score
  - ✓ Gini index
  - ✓ Cramer's V

How “useful” is a single feature? : Univariate filters

Trying to predict someone’s ML exam grade from various possible indicators (a.k.a. features):

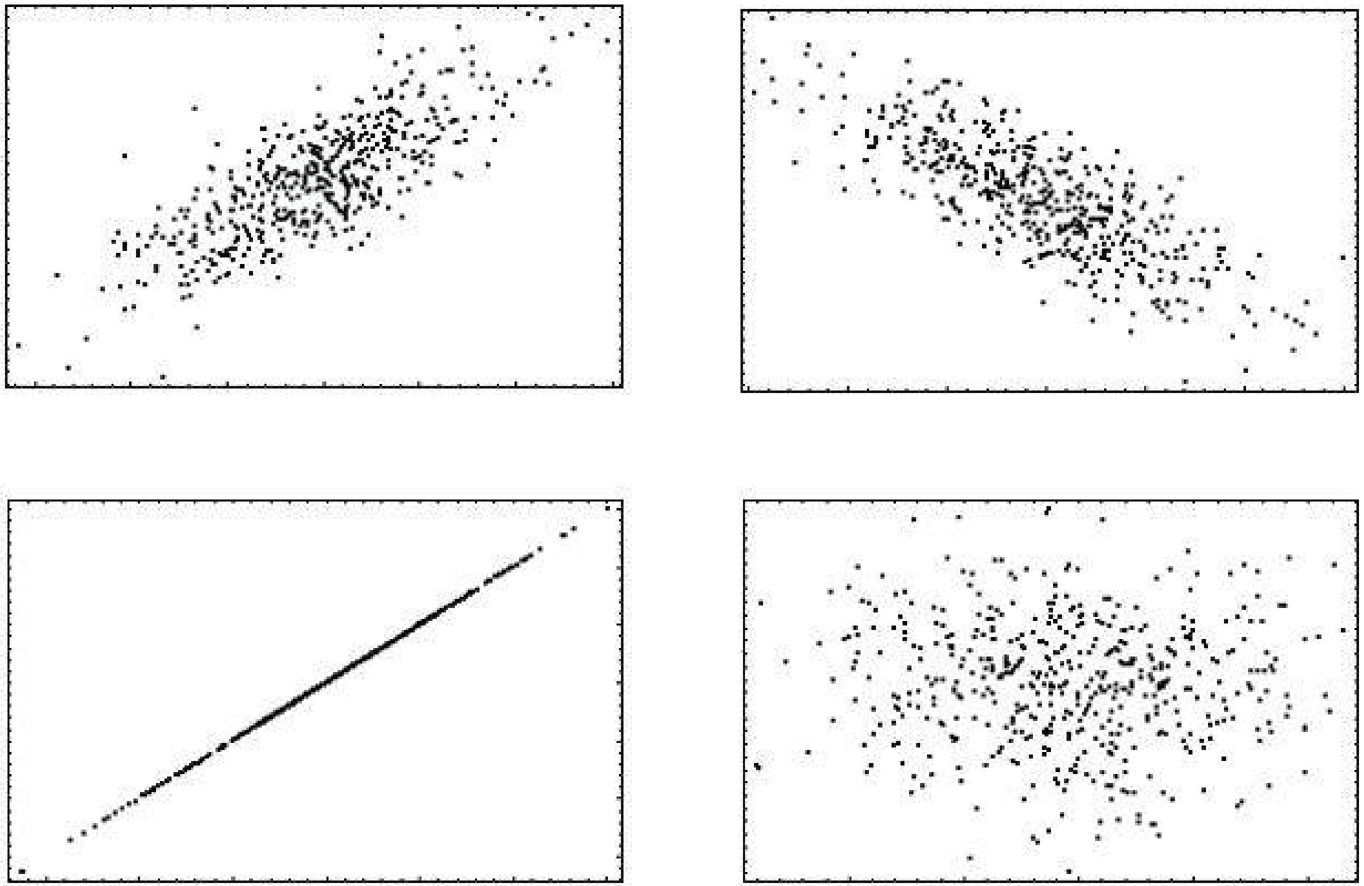
1) Statistics grade,

2) Biology grade,

3) Linear Algebra grade, or

4) Height ...

Which one would you pick?



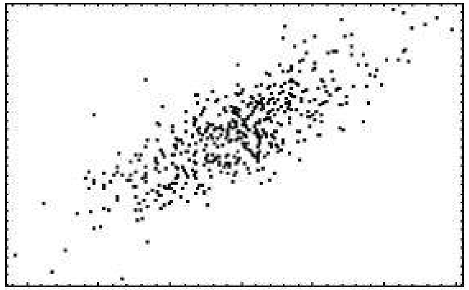
Pearson's Correlation Coefficient

- Used to measure the strength of association between two continuous random variables.

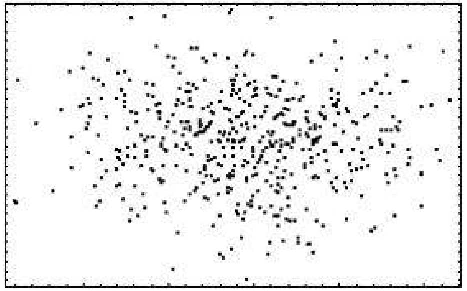
Feature : $\mathbf{x}_k = \{x_k^{(1)}, ..., x_k^{(N)}\}^T$

Target : $\mathbf{y} = \{y^{(1)}, ..., y^{(N)}\}^T$

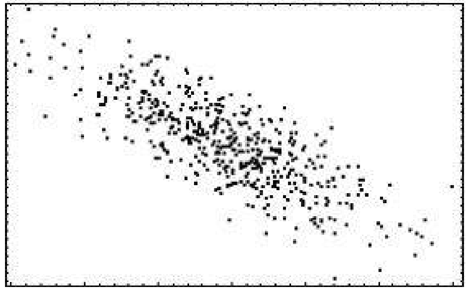
$$r(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x^{(i)} - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y^{(i)} - \bar{y})^2}}$$



$r = +0.5$



$r = 0.0$



$r = -0.5$

Both positive and negative correlation is useful!

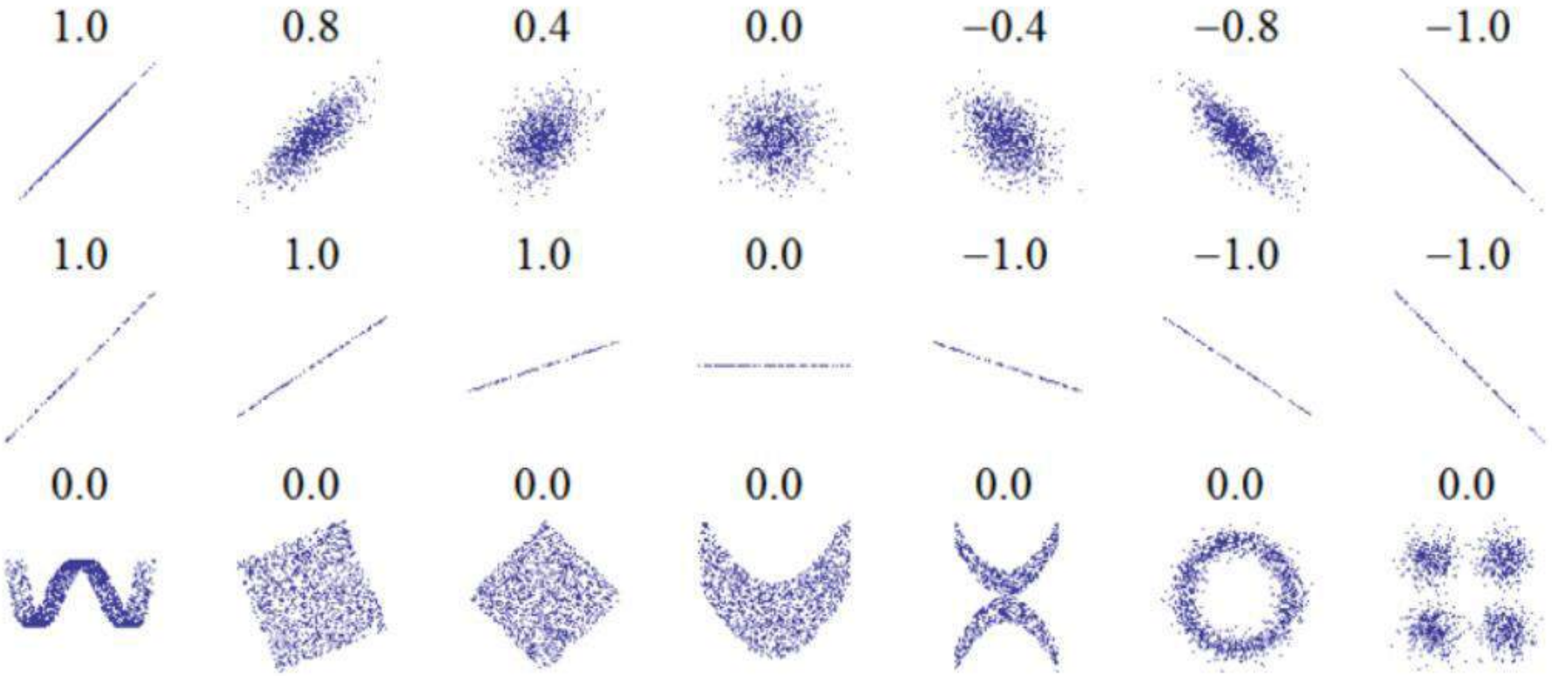
Ranking with Filter Criteria

- ➢Rank features X_i , $\forall i$ by their values of $J(X_k)$.
- ➢Retain the highest ranked features, discard the lowest ranked.

Cut-off point decided by user, e.g. $|S| = 5$,
$S = \{35, 42, 10, 654, 22\}$.

k	$J(X_k)$
35	0.846
42	0.811
10	0.810
654	0.611
22	0.443
59	0.388
...	...
212	0.09
39	0.05

Limitation: Pearson assumes all features are INDEPENDENT ! and... only identifies LINEAR correlations.



There are LOTS of ranking criteria...

Pearson, Fisher, Mutual Info, Jeffreys-Matsusita, Gini Index, AUC, F-measure, Kolmogorov distance, Chi-squared, CFS, Alpha-divergence, Symmetrical Uncertainty,.... etc, etc

- How do I pick the right filter ? Unfortunately, quite complex.... depends on:
  - ✓type of variables/targets (continuous, discrete, categorical).
  - ✓class distribution
  - ✓degree of nonlinearity/feature interaction

- The “**No Free Lunch**” theorem states that there is no universal model that works best for every problem.

Thank You!

In our next session:Feature selection using Chi Squared Test

Feature selection using Chi-Squared Test of Independence

Prof.Aruna Malapati

Learning Objectives

- Explain and formulate Chi-Squared test of independent between two variables
- Apply Chi-Squared test of independence for categorical variables

Hypothesis Testing

- Hypothesis is a premise or claim that we want to investigate.
- Test whether the two random variables (categorical) are independent or not.

- Test Statistic
  - ✓Chi-Squared Test
  - ✓T-Test
  - ✓ANNOVA-Test

Test Statistic

Example

A group of customers were classified in terms of personality (introvert, extrovert or normal) and in terms of color preference (red, yellow or green) with the purpose of seeing whether there is an association (relationship) between personality and color preference.

Data was collected from 400 customers and presented in the 3 (rows) x 3 (cols) contingency table below:

(Observed counts)	Colors			
Personality	**Red**	**Yellow**	**Green**	**Totals**
Introvert personality	11	5	1	17
Extrovert personality	8	6	8	22
Normal	3	10	12	25
Total	22	21	21	64

Five-step approach for Chi-Squared test of independence

- **Step 1.** Set up hypotheses and determine level of significance.
  - ✓Null hypothesis(H0): Color preference is independent of personality.
  - ✓Alternative hypothesis(H_A): Color preference is dependent on personality
  - ✓α=0.05

Five-step approach for Chi-Squared test of independence (contd..)

Step 2. Compute the expected frequency (under the null hypothesis) in each cell using E = (Row Total * Column Total)/N

(Expected counts)	Colors			
Personality	**Red**	**Yellow**	**Green**	**Totals**
Introvert personality	**5.8**	**5.6**	**5.6**	**17**
Extrovert personality	**7.6**	**7.2**	**7.2**	**22**
Normal	**8.6**	**8.2**	**8.2**	**25**
Total	**22**	**21**	**21**	**64**

Step 3:Select the test statistic

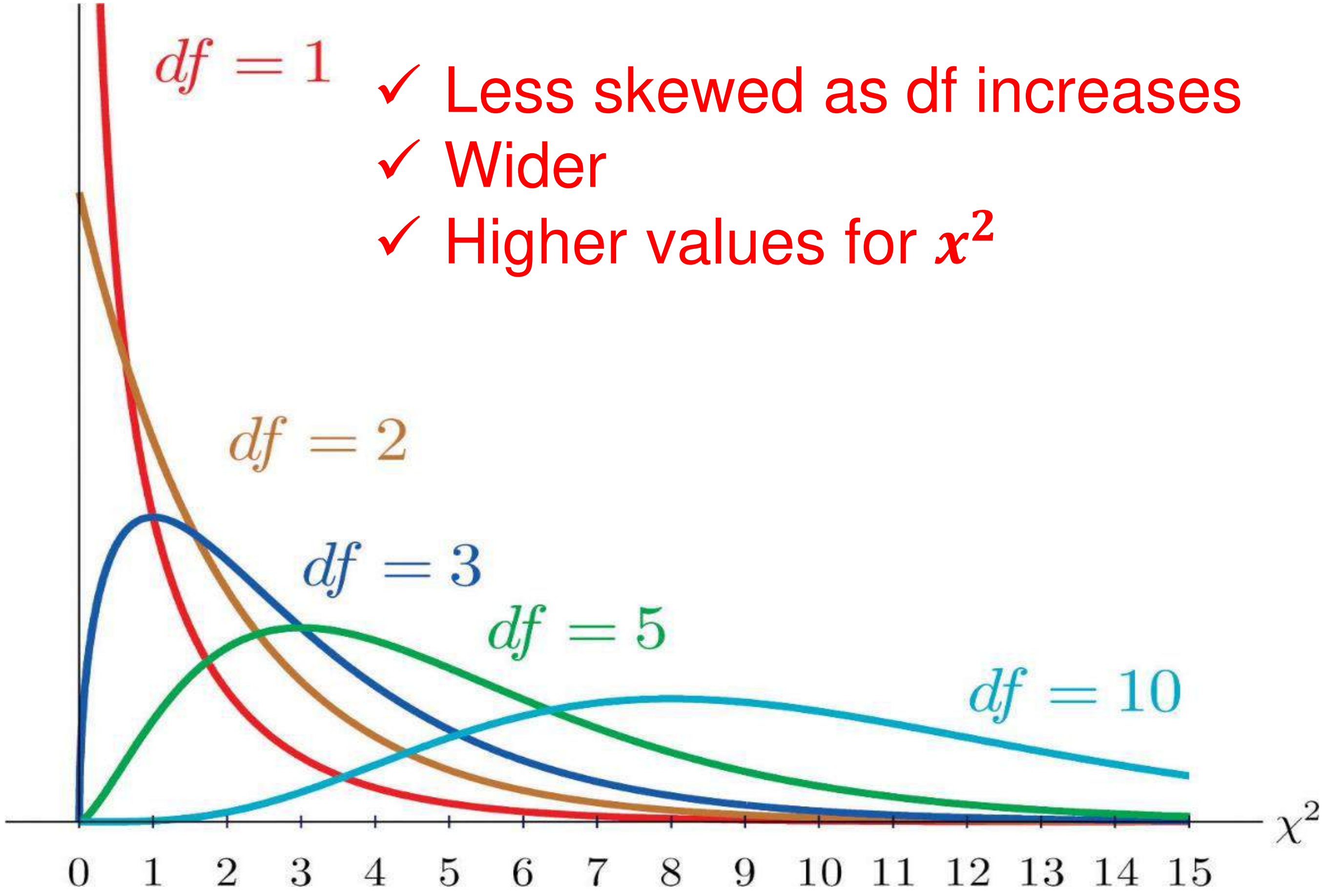
$$x^2 = \Sigma \frac{(0 - E)^2}{E}$$

$$x^2 = \frac{(11-5.8)^2}{5.8} + \frac{(5-5.6)^2}{5.6} + \frac{(1-5.6)^2}{5.6} + \ldots\ldots + \frac{(12-8.2)^2}{8.2} = 14.5$$

Chi-Squared distribution

The probability density function for the x^2 distribution with r degrees of freedom(df) is given by

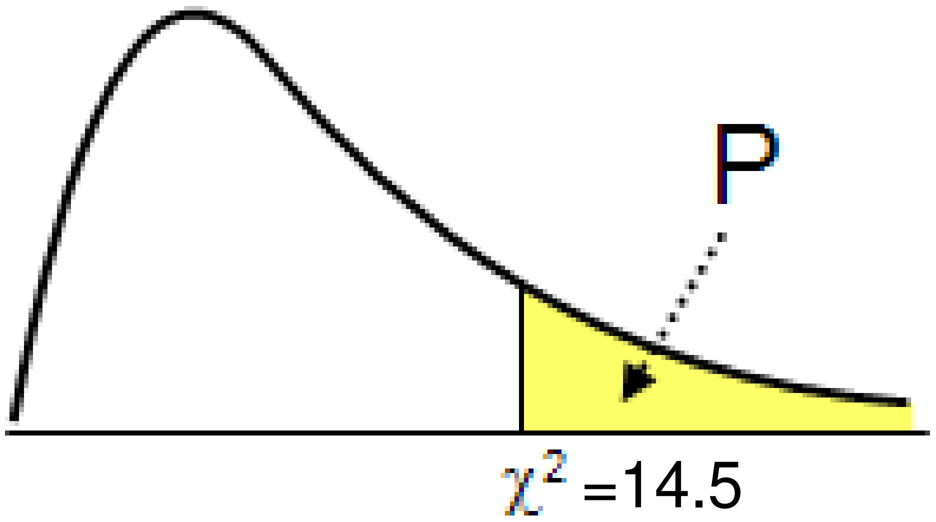
$$P_r(x) = \frac{x^{r/2-1}\, e^{-x/2}}{\Gamma\left(\frac{1}{2}r\right) 2^{r/2}}$$



- ✓ Less skewed as df increases
- ✓ Wider
- ✓ Higher values for x^2

Always positive

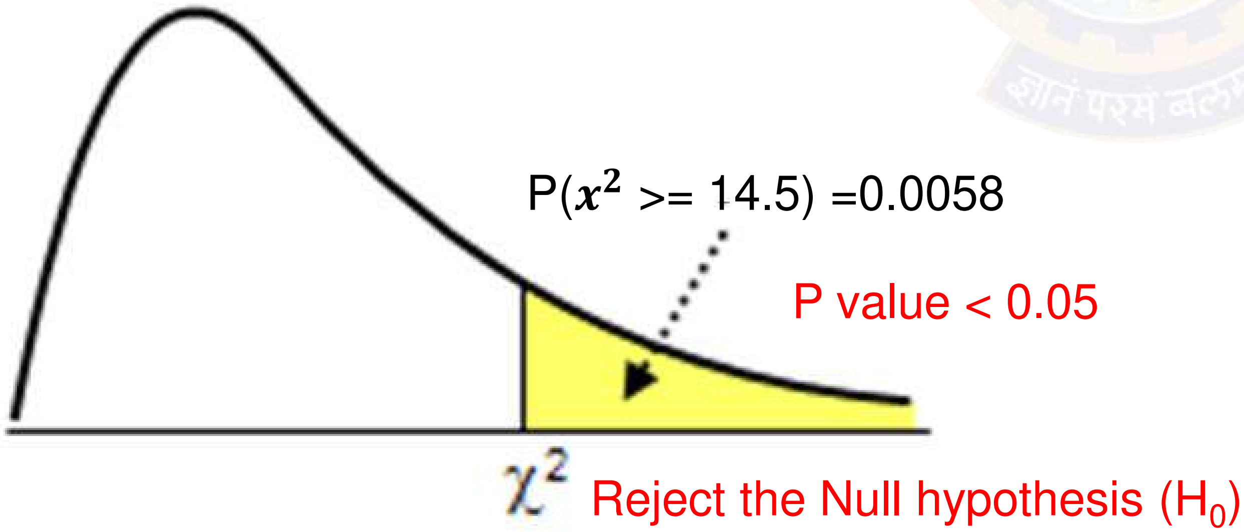
Significance of P value



Five-step approach for Chi-Squared test of independence (Contd..)

- Step 4: Use a probability table to find P-Value associated with x^2 value for with degrees of freedom df = (r − 1) (c − 1), r is the number of categories in one variable and c is the number of categories in the other.

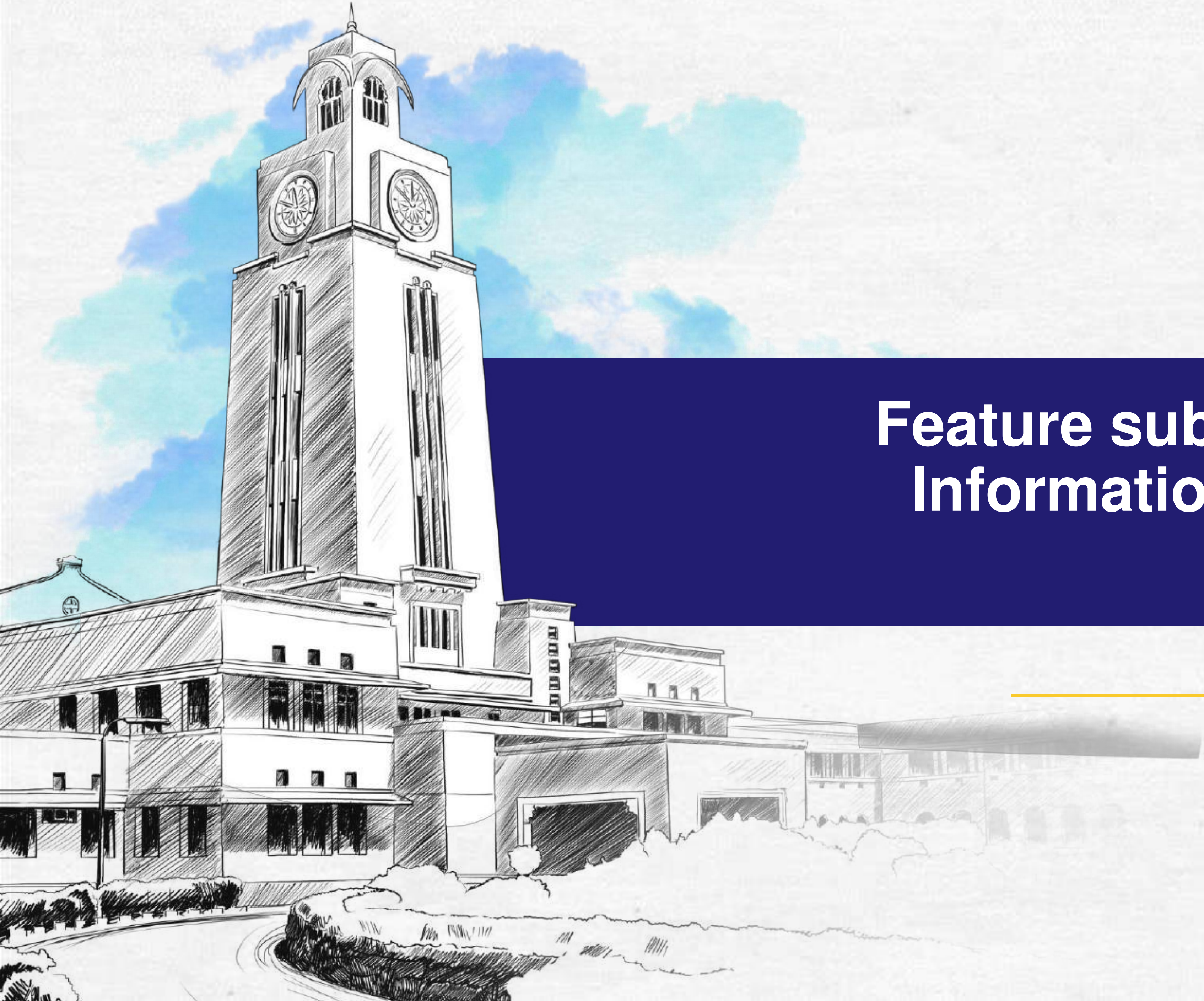
- Step 5: Make a conclusion using P-value



	Significance Level				
df	0.10	0.05	0.025	0.01	0.005
1	2.7055	3.8415	5.0239	6.6349	7.8794
2	4.6052	5.9915	7.3778	9.2104	10.5965
3	6.2514	7.8147	9.3484	11.3449	12.8381
4	7.7794	9.4877	11.1433	13.2767	14.8602
5	9.2363	11.0705	12.8325	15.0863	16.7496
6	10.6446	12.5916	14.4494	16.8119	18.5475
7	12.017	14.0671	16.0128	18.4753	20.2777

Thank You!

In our next session:Feature selection using Information Theoretic Measures



Feature subset selection using Information Theoretic Metrics

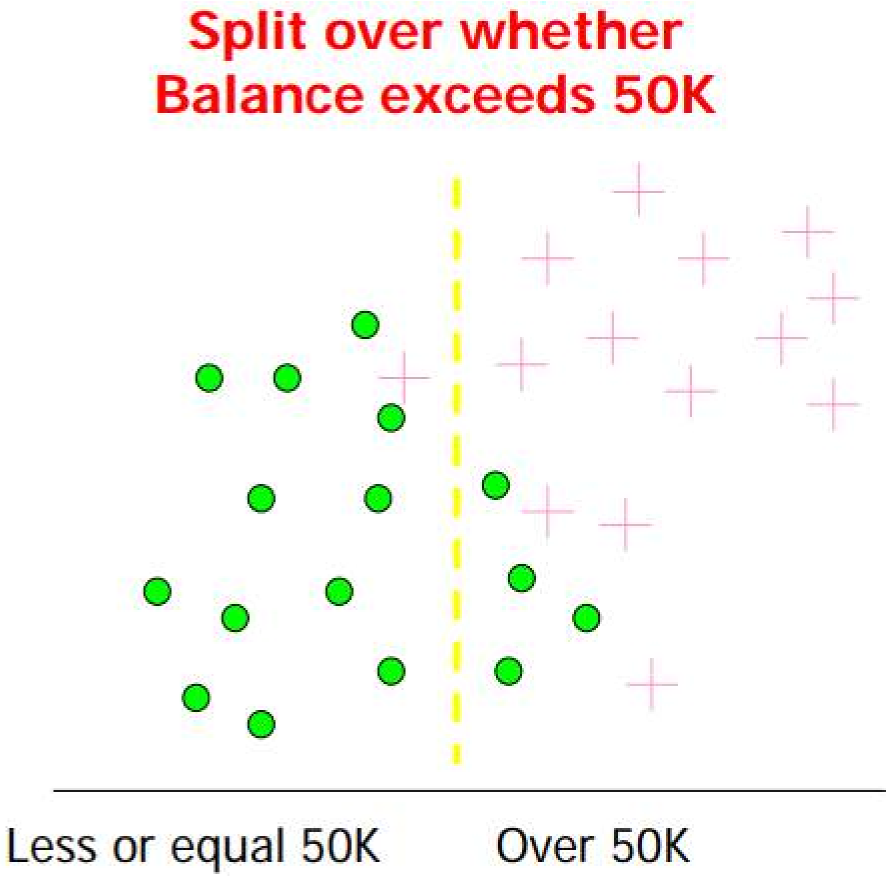
Prof.Aruna Malapati

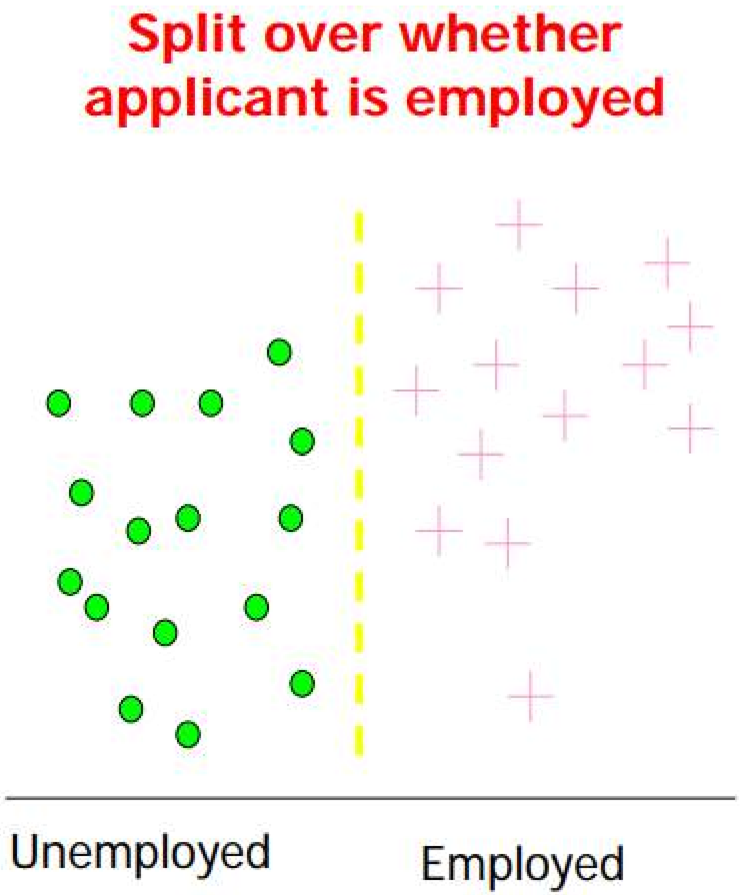
Learning Objectives

- Define Entropy and Mutual Information

Information Theoretic approaches for Feature selection

Which test is more informative?





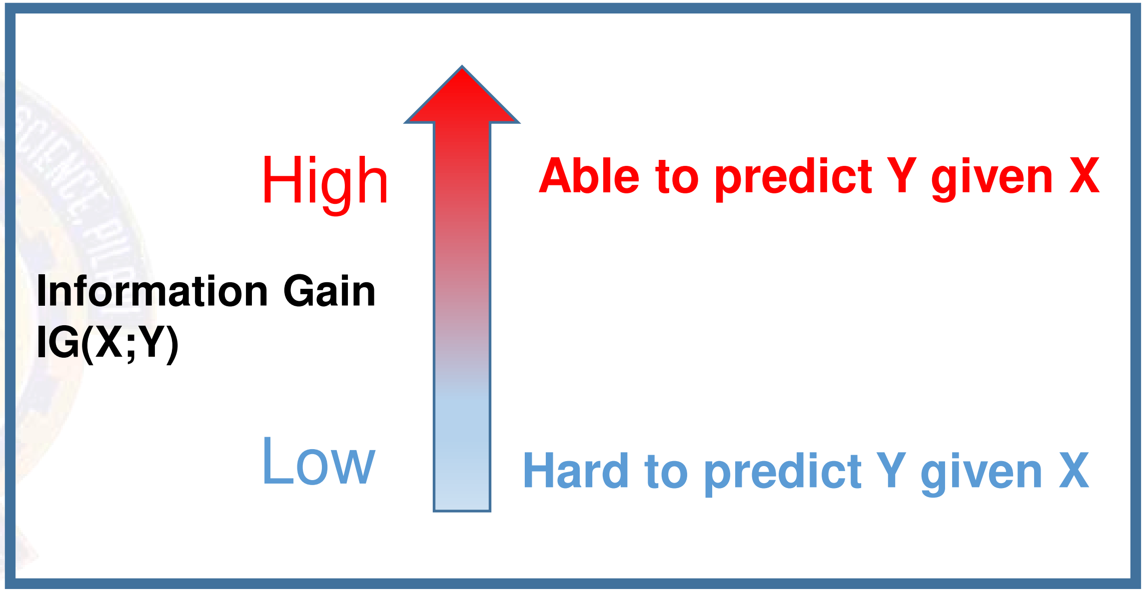
- Information-theoretic concepts can only be applied to discrete variables.

Entropy and Conditional entropy

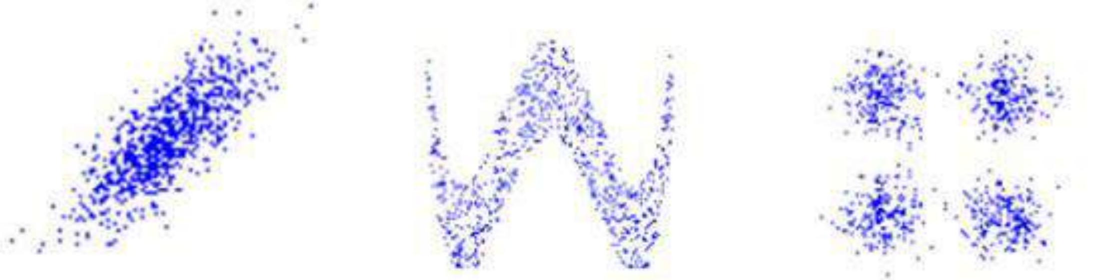
- **Entropy: A common way to measure impurity or uncertainty**

Information Gain

Information Gain IG(X;Y) is a measure of the mutual independence between two random variables X and Y.



- ✓Symmetric I(X,Y) = I(Y,X)
- ✓Measures non-linear dependencies
- ✓I(X;Y)=0 if X and Y are independent
- ✓Biased towards the features having large number of values



Pearson r = 0.8
IG = 0.5

Pearson r = 0.0
IG = 0.7

Pearson r = 0.0
IG = 0.1

Gain Ratio

- **The gain ratio "normalizes" the information gain**

$$Gain\ Ratio(Attribute) = \frac{IG(Attribute)}{H(Attribute)}$$

✓reduces the bias toward attributes with many values.

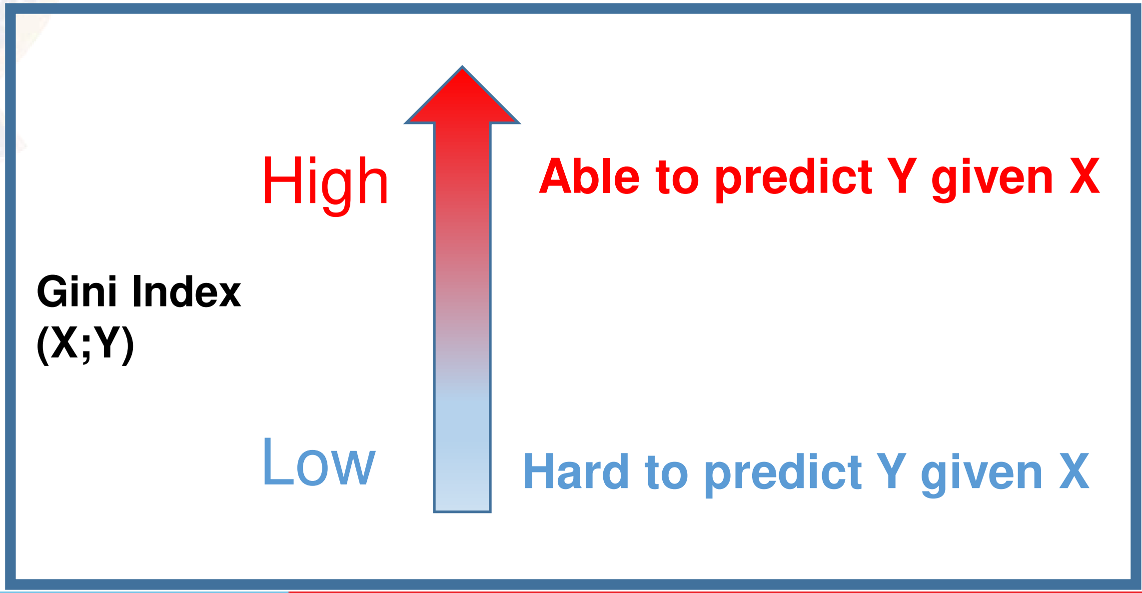
✓The feature with the maximum gain ratio is selected as the best feature.

Gini Index

➢Gini index minimize the probability of misclassification

$$\text{Gini} = 1 - \sum_{i=1}^{K} {p_k}^2$$

where p_k denotes the proportion of instances belonging to class k K = 1, … , k.



Example

ATTRIBUTES				CLASS LABEL
Gender	**CAR Ownership**	**Travel Cost(Rs/Km)**	**Income Level**	**Transport Mode**
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

Transport Mode		
Bus	**Train**	**Car**
4	3	3

H(Transport Mode) = H(4,3,3)

$= -(4/10 \log_2 4/10) - (3/10 \log_2 3/10)$
$\quad - (3/10 \log_2 3/10)$

$= 1.571$

Example (Contd..)

	Class Label			
Attribute values		Bus	Train	Car
	Cheap	4	1	0
	Expensive	0	0	3
	Standard	0	2	0

H(Transport Mode)= 1.571

IG(Transport Mode, Travel Cost) = H(Transport Mode)-H(Transport Mode| Travel Cost)
= 1.571 - H(4,3,2)
= 1.571 - (-5/10 (4/5 log 4/5 +1/5log1/5) - (3/10 (3/3 log 3/3 + 0 log 0)) - (3/10 (2/2 log 2/2 + 0 log 0)))
= 1.571 – 0.36 = 1.211

Gain Ratio (Travel Cost) = $\frac{IG(Attribute)}{H(Attribute)} = \frac{1.211}{H(Attribute)}$

H(Attribute) = -(5/10 log 5/10)-(3/10 log 3/10)- (2/10 log 2/10) = 1.48

$= \frac{1.211}{1.48} = 0.818$

Gini Index(Transport Mode,|Travel Cost=cheap) = $1-(0.8^2+0.2^2) = 0.32$
Gini Index(Transport Mode,|Travel Cost=Expensive) = $1-(1^2+0^2)=0$
Gini Index(Transport Mode,|Travel Cost=Standard) = $1-(1^2+0^2)=0$
Gini Index(Transport Mode,|Travel Cost=cheap) = 5/10*0.32+3/10*0+2/10*0 = 0.16

Example (Contd..)

	Information Gain	Gain Ratio	Gini
Gender	0.147	0.147	0.6
Car Ownership	0.544	0.368	0.453
Travel Cost	1.21	0.818	0.16
Income Level	0.696	0.458	0.366

Thank You!

In our next session: Feature subset selection using Fisher Score

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Feature subset selection using Fisher Score

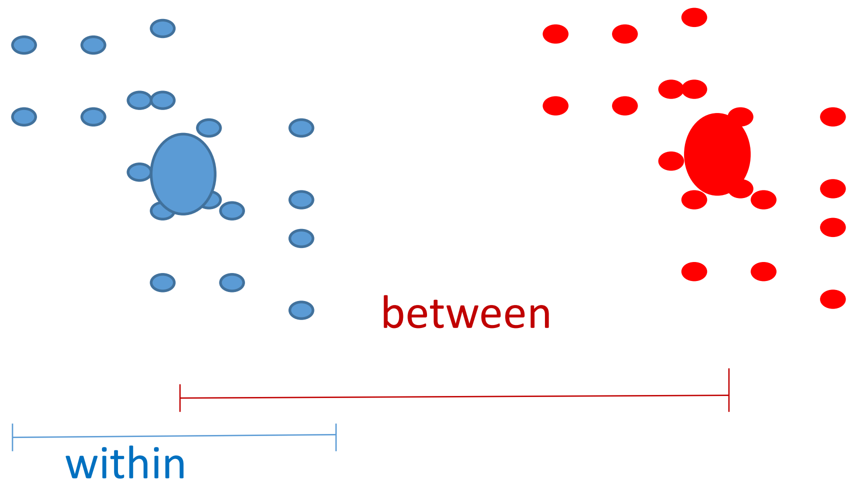
Prof.Aruna Malapati

Learning Objectives

- Define the inter and intraclass distances
- Formulate the Fischer score filter

How class information is useful?

- Applicable for classification problems with numeric features.
- Between-class distance – Distance between the centroids of different classes
- Within-class distance – Accumulated distance of an instance to the centroid of its class



Fisher Score

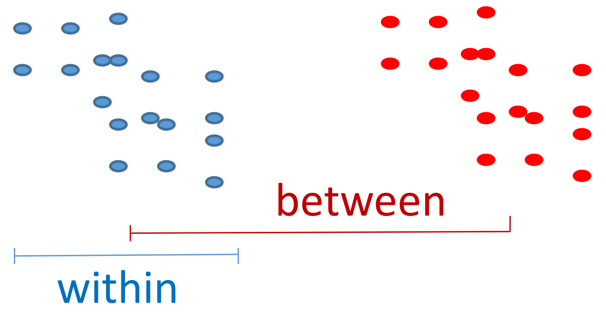
- ➢Fisher score is the measure the ratio of the average interclass separation to the average intraclass separation.

- ➢The larger the Fisher score, the greater the discriminatory power of the attribute.

$$F = \frac{\sum_{j=1}^{k} p_j(\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2}$$

μ_j - mean of the data points belonging to class j for a particular feature,
σ_j - standard deviation of data points belonging to class j for a particular feature,
p_j - the fraction of data points belonging to class j.
μ - the global mean of the data on the feature

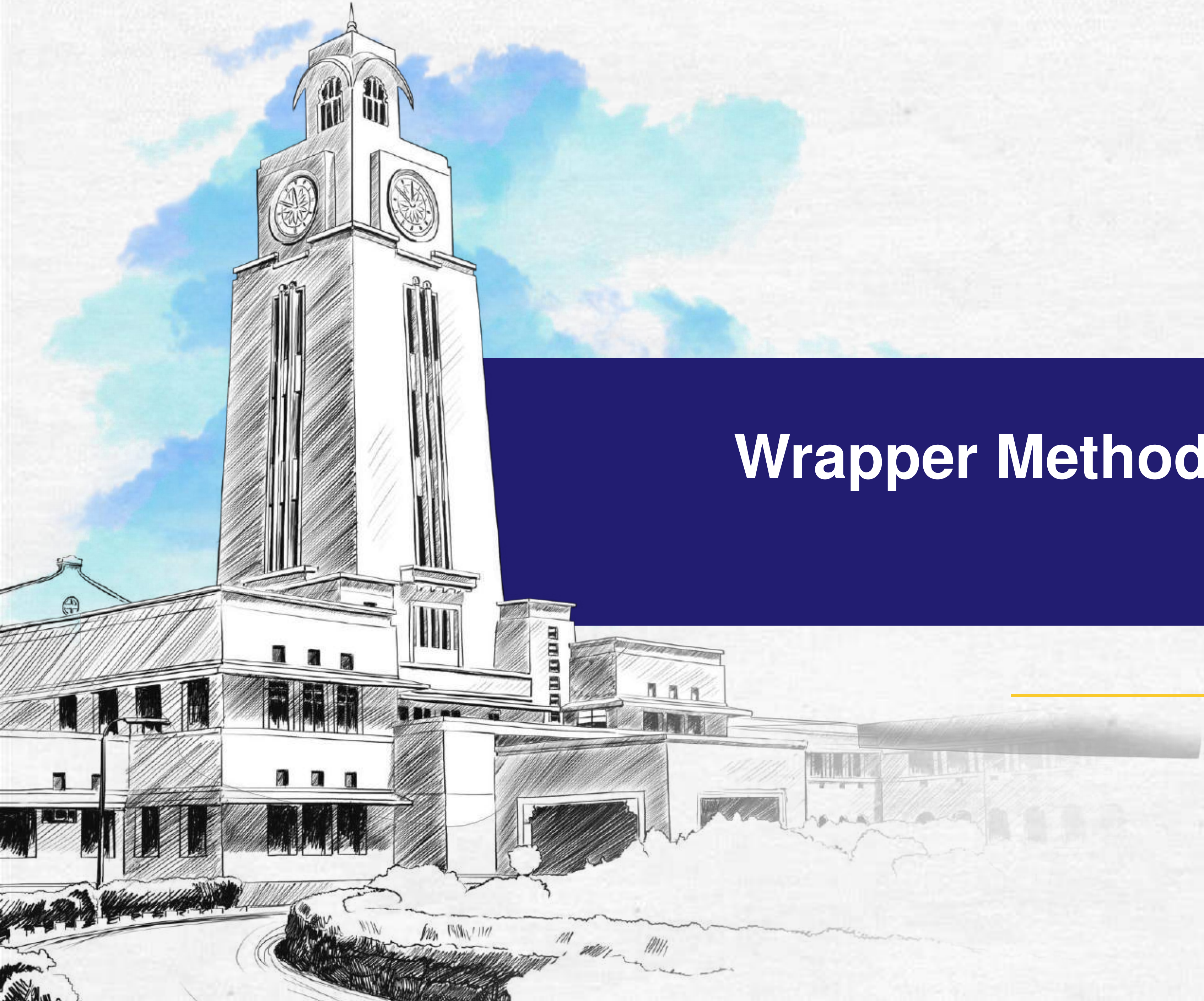
Good class separation



- ➢This score is often referred as signal to noise ratio

Thank You!

In our next session:Feature selection using wrapper methods



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Wrapper Method for Feature Subset Selection

Prof.Aruna Malapati

Learning Objective

- Formulate the problem of wrapper based subset selection
- List and apply wrapper based subset selection

Wrapper Based Methods

- Greedy Based algorithms
- Agnostic to the machine learning models chosen.
- Sequential feature selection algorithm add or remove one feature at a time based on the classifier performance until a desired criterion is met.

Algorithm for Wrapper based method

Input: large feature set Ω

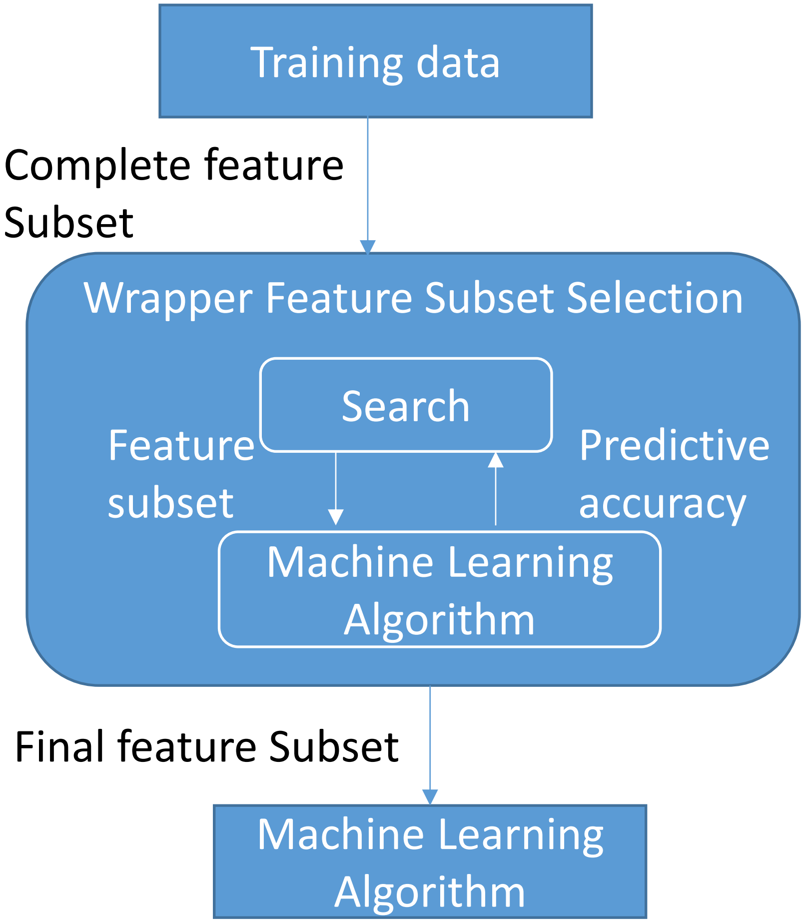
Identify candidate subset $S \subseteq \Omega$

While !stop_criterion()

Evaluate error of a classifier using S.

Adapt subset S.

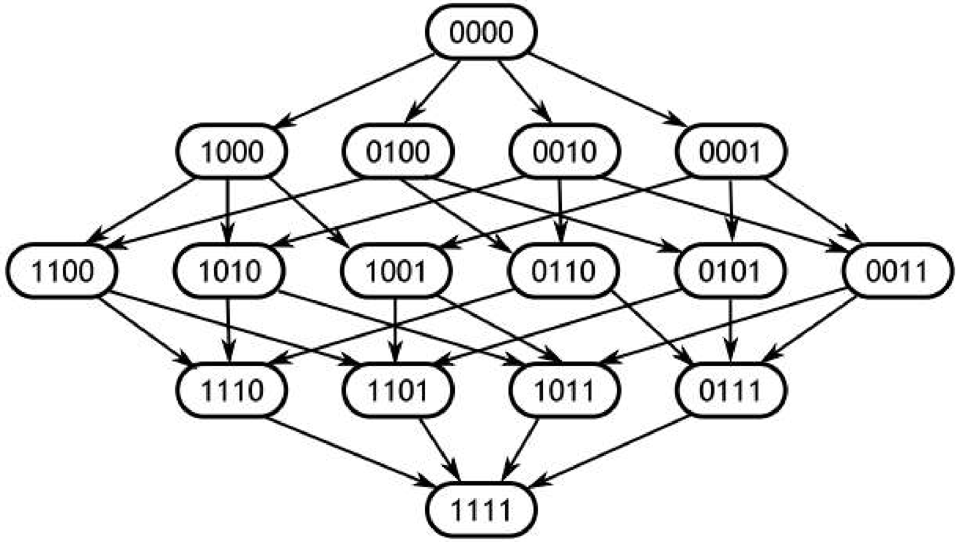
Return S.



- Commonly used Stop criterions
  - ✓Increase / Decrease in Predictive accuracy
  - ✓Predefined number of features is reached

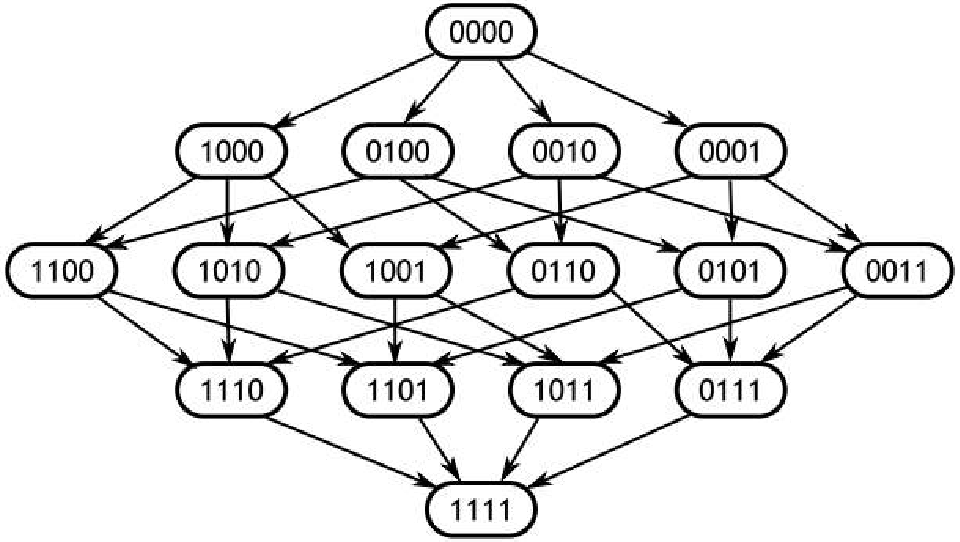
Sequential forward selection(SFS)

1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \underset{x \notin Y_k}{\arg\max}\, J(Y_k + x)$
3. Update $Y_{k+1} = Y_k + x^+$; $k = k + 1$
4. Go to 2



Sequential Backward selection(SBS)

1. Start with the full set $Y_0 = \{X\}$
2. Remove the worst feature $x^- = \arg\max_{x \in Y_k} J(Y_k - x)$
3. Update $Y_{k+1} = Y_k - x^-$; $k = k + 1$
4. Go to 2



- Backwards selection is frequently used with random forest models

Pros and Cons of Greedy Sequential Algorithms

- Pros
  - ✓ Highest performance
- Cons
  - ✓ Computationally expensive
  - ✓ Memory intensive

Thank You!

In our next session:Implementing Feature selection using Python