MCA Semester-3 Machine Learning

Introduction

Will learn ...

- Unit-1 Introduction
  - 1.1 Applications and Future Scope of Machine Learning
  - 1.2 Types of Learning
    - Supervised and Unsupervised
  - 1.3 Training versus Testing
  - 1.4 Data Processing
    - Missing Data, Categorical Data, Feature Scalling

Will learn ...

- Unit-2 Regression
  - 2.1 Simple Linear Regression
  - 2.2 Multiple Linear Regression
  - 2.3 Polynomial Regression
  - 2.4 Decision Tree Regression
  - 2.5 Random Forest Regression

Prof.(Dr.) Dharmendra Bhatti 3

Will learn ...

- Unit-3 Classification
  - 3.1 Logistic Regression
  - 3.2 K-Nearest Neighbors (K-NN)
  - 3.3 Support Vector Machine (SVM)
  - 3.4 Naive Bayes

Prof.(Dr.) Dharmendra Bhatti 4

Will learn ...

- Unit-4 Clustering
  - 4.1 K-Means Clustering
  - 4.2 Hierarchical Clustering
  - 4.3 Apriori

Prof.(Dr.) Dharmendra Bhatti 5

Will learn ...

- Unit-5 Reinforcement and Deep Learning
  - 5.1 Upper Confidence Bound (UCB)
  - 5.2 Thompson Sampling
  - 5.3 Artificial Neural Networks
  - 5.4 Convolutional Neural Networks

Prof.(Dr.) Dharmendra Bhatti 6

Will learn ...

- Unit-6 Dimensionality Reduction
  - 6.1 Principal Component Analysis (PCA)
  - 6.2 Linear Discriminant Analysis (LDA)
  - 6.3 Kernel PCA

Prof.(Dr.) Dharmendra Bhatti 7

What is Machine Learning?

Human	Machine
Learn from experience	Learn from past data

Prof.(Dr.) Dharmendra Bhatti 8

What is Machine Learning?

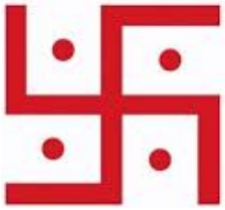
- What is the name of this symbol???

Prof.(Dr.) Dharmendra Bhatti 9

What is Machine Learning?

Prof.(Dr.) Dharmendra Bhatti 10

What is Machine Learning?



- Is this swastik??? => YES...

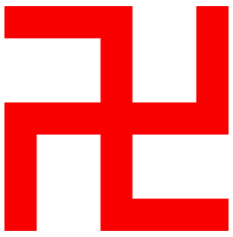
Prof.(Dr.) Dharmendra Bhatti 11

What is Machine Learning?

- Is this swastik??? => YES...

Prof.(Dr.) Dharmendra Bhatti 12

What is Machine Learning?



- Is this swastik??? => NO...

Prof.(Dr.) Dharmendra Bhatti 13

What is Machine Learning?

- Conclusion: Human learn from experience.
- Can machine learn from experience???

14

What is Machine Learning?

- According to Wikipedia,
  - Machine learning (ML) is the study of computer algorithms that improve automatically through experience.
  - https://en.wikipedia.org/wiki/Machine_learning

Prof.(Dr.) Dharmendra Bhatti 15

What is Machine Learning?

- Arthur Lee Samuel (1959)
  - "Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"
  - https://en.wikipedia.org/wiki/Arthur_Samuel

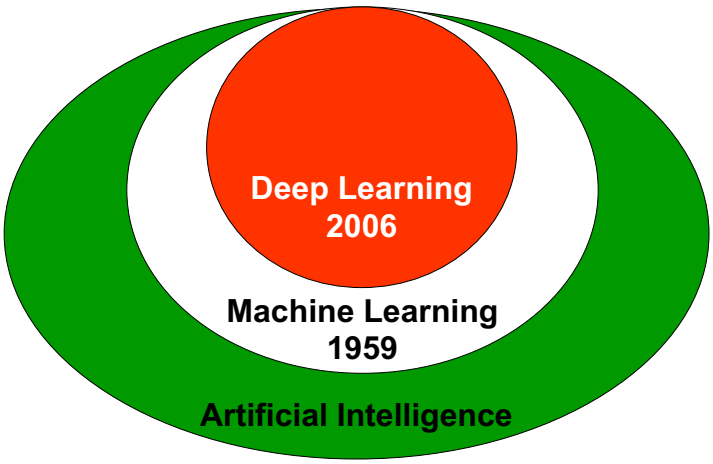
Prof.(Dr.) Dharmendra Bhatti 16

What is Machine Learning?

- Tom M. Mitchell (1998)
  - "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

Prof.(Dr.) Dharmendra Bhatti 17

What is Machine Learning?



Prof.(Dr.) Dharmendra Bhatti 18

What is Machine Learning?

- Artificial Intelligence
  - Wikipedia – “In computer science, artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals.”
- Deep Learning
  - Deep learning is part of a broader family of machine learning methods based on artificial neural networks with representation learning.

Prof.(Dr.) Dharmendra Bhatti 19

Why Machine Learning?

- Volume and variety of data
- Cheaper computational processing
- Affordable data storage
- Classical programming limitation
- Solving complex problem which is even difficult for human

Prof.(Dr.) Dharmendra Bhatti 20

Applications of Machine Learning

- Image Recognition
  - **automatic** - image tags, keywording of photos, and categorization on the basis of visual topics
  - Stock photography and video websites
  - Visual product search
  - Image and face recognition on social media
  - Security devices
  - Vehicle number plate reading

Prof.(Dr.) Dharmendra Bhatti 21

Applications of Machine Learning

- Speech Recognition
  - Google Assistant, Apple Siri, Amazon Alexa, Microsoft Cortona
  - Telecommunication: customer care, voice calling
  - Office: Voice navigation, voice dictation
  - Manufacturing: data entry
  - Medical digital dictation system
  - Voice commands – Game, Automobile, Fighter aircraft (Mirage, Rafale)

Prof.(Dr.) Dharmendra Bhatti 22

Applications of Machine Learning

- Traffic prediction
  - Highway/Metro cities: traffic police, transportation companies, people
- User interest and product recommendations
  - E-commerce/shopping websites
- Self-driving cars
- Email Spam and Malware Filtering
- Cyber fraud detection
  - Online payment/order authorization

Prof.(Dr.) Dharmendra Bhatti 23

Applications of Machine Learning

- Virtual Personal Assistant
  - software agent that can perform tasks or services for an individual based on commands or questions
    - Managing social media, calendars, appointments, and emails, Preparing documents
    - Booking hotels and restaurants
- Stock Market Trading
- Medical Diagnosis
- Automatic Language Translation

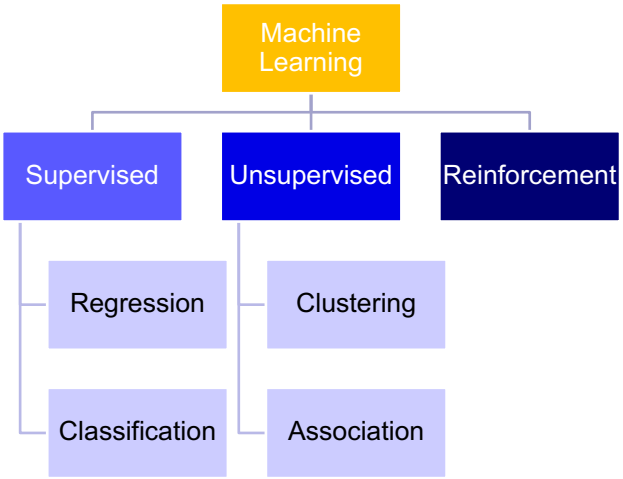
Prof.(Dr.) Dharmendra Bhatti 24

Life Cycle of Machine Learning

1. Collect the Data
2. Data Preprocessing
3. Build Model
4. Train Model
5. Test Model
6. Deploy Model

Prof.(Dr.) Dharmendra Bhatti 25

Types of Machine Learning



Prof.(Dr.) Dharmendra Bhatti 26

Supervised Learning

- **Supervised learning** is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- It infers a function from labeled training data consisting of a set of training examples.

Prof.(Dr.) Dharmendra Bhatti 27

Supervised Learning

- It requires input variables (X) and an output variable (Y) and use an algorithm to learn the mapping function from the input to the output.
- Y = f(X)
- Once mapping function f( ) is ready, it can predict Y for unseen X value.

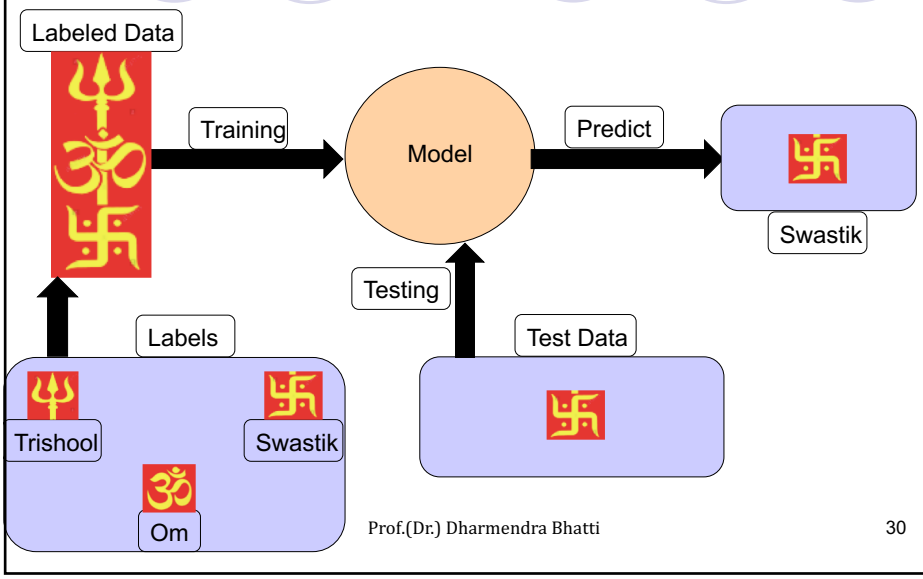
Prof.(Dr.) Dharmendra Bhatti 28

Supervised Learning

- Models are trained using labelled dataset, where the model learns about each type of data
- Once the training process is completed, the model is tested with test data (a subset of the training set)
- Finally, trained and tested model predicts the output

Prof.(Dr.) Dharmendra Bhatti 29

Supervised Learning



Prof.(Dr.) Dharmendra Bhatti 30

Supervised Learning – Step-by-step

- Determine the type of training examples.
  - Decide what kind of data is to be used as a training set.
  - i.e. In the case of handwriting analysis, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.

Prof.(Dr.) Dharmendra Bhatti 31

Supervised Learning – Step-by-step

- Gather a training set.
  - The training set needs to be representative of the real-world use of the function.
  - Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.
- Split the training dataset into training dataset, test dataset, and validation dataset.

Prof.(Dr.) Dharmendra Bhatti 32

Supervised Learning – Step-by-step

- Determine the input feature representation of the learned function.
  - The accuracy of the learned function depends strongly on how the input object is represented.
  - Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object.

Supervised Learning – Step-by-step

- Determine the input feature representation of the learned function.
  - The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

Supervised Learning – Step-by-step

- Determine the structure of the learned function and corresponding learning algorithm.
  - i.e. use support vector machines or decision trees

Supervised Learning – Step-by-step

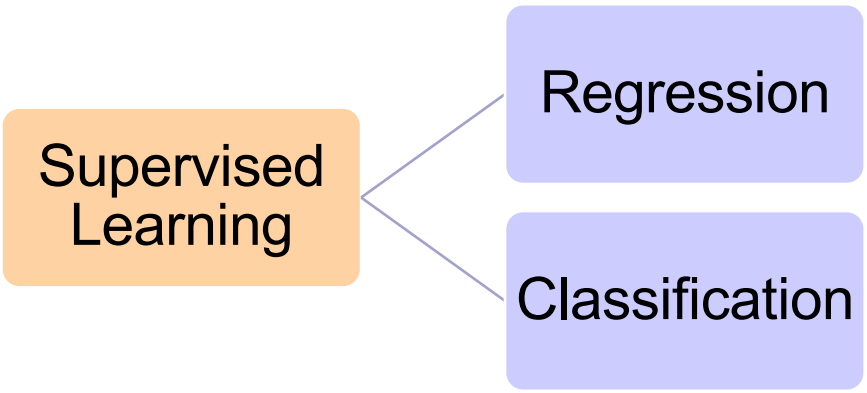
- Execute the algorithm on the training dataset.
  - Some supervised learning algorithms require the user to determine certain control parameters.
  - These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.

Supervised Learning – Step-by-step

- Evaluate the accuracy of the learned function.
  - After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

Prof.(Dr.) Dharmendra Bhatti 37

Types of Supervised Learning



Prof.(Dr.) Dharmendra Bhatti 38

Types of Supervised Learning

- Regression
  - Output variable in regression is numerical or continues
  - Regression algorithms attempt to estimate the mapping function (f) from the input variables (x) to numerical or continuous output variables (y).

Types of Supervised Learning

- Regression algorithms
  - Simple Linear Regression
  - Multiple Linear Regression
  - Polynomial Regression
  - Decision Tree Regression
  - Random Forest Regression
  - Non-Linear Regression
  - Bayesian Linear Regression

Types of Supervised Learning

- Classification
  - Output variable in classification is categorical or discrete
  - Classification algorithms attempt to estimate the mapping function (f) from the input variables (x) to discrete or categorical output variables (y).

Types of Supervised Learning

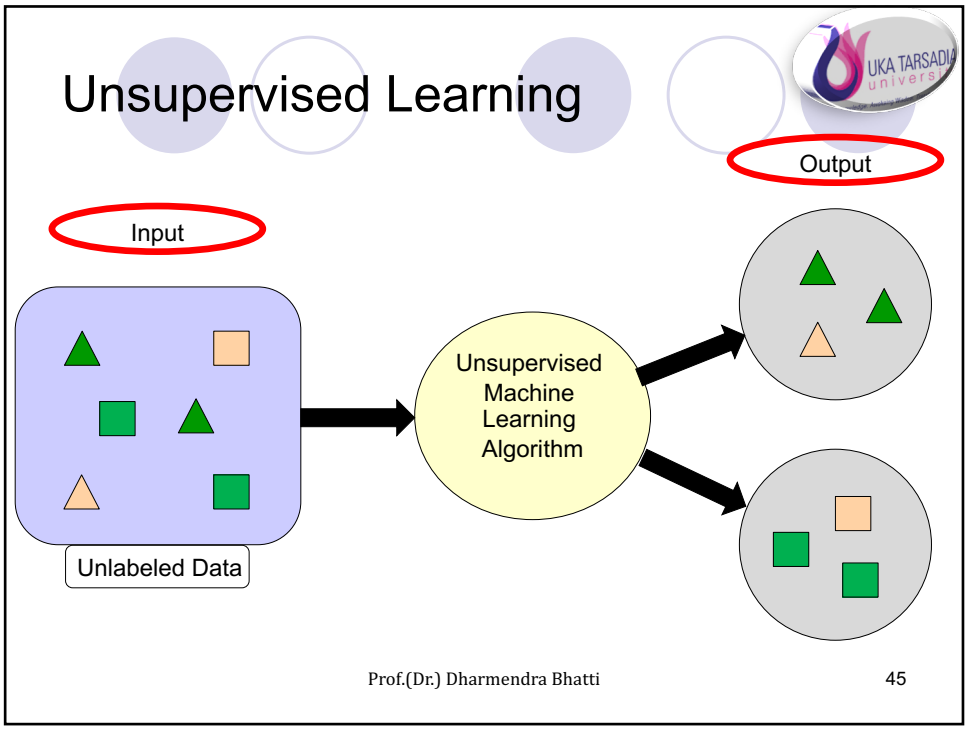
- Classification algorithms
  - Logistic Regression
  - K-Nearest Neighbors (K-NN)
  - Support Vector Machine (SVM)
  - Naive Bayes
  - Decision Trees

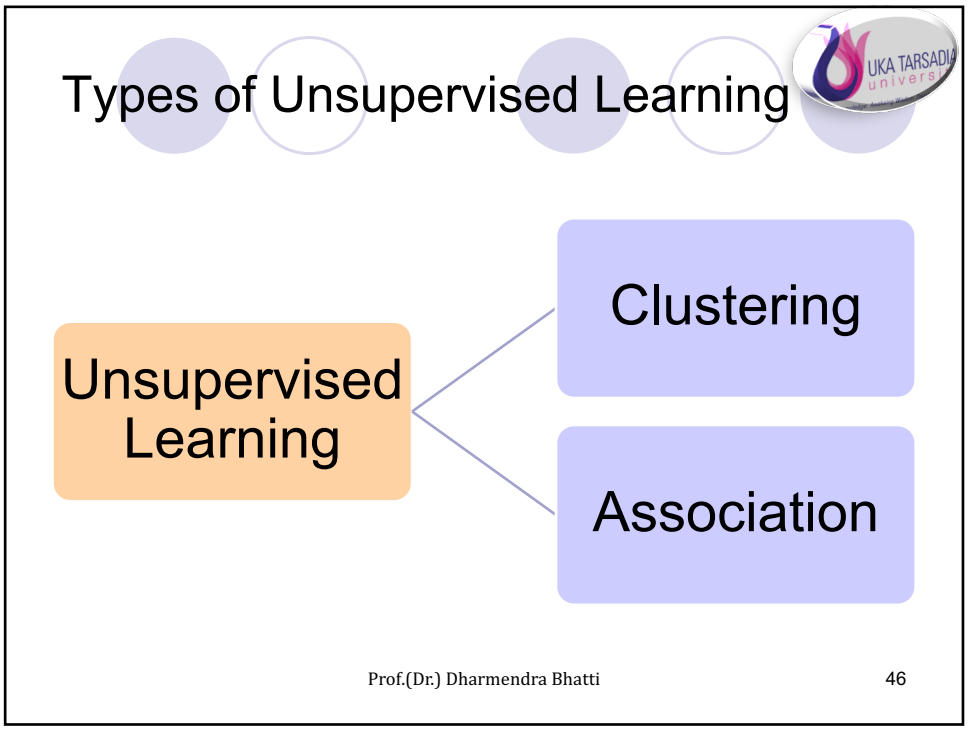
Unsupervised Learning

- Absence of supervisor/labelled training data
- Unsupervised learning is where you only have input data (X) and no corresponding output variables.

Unsupervised Learning

- Machine learning algorithms need to discover the pattern in data for learning
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.


Unsupervised Learning
Output
Input
Unsupervised Machine Learning Algorithm
Unlabeled Data
Prof.(Dr.) Dharmendra Bhatti
45


Types of Unsupervised Learning
Unsupervised Learning
Clustering
Association
Prof.(Dr.) Dharmendra Bhatti
46

Types of Unsupervised Learning

- Clustering
  - Find similarity among data
  - No accuracy measurement but comparative performance

Prof.(Dr.) Dharmendra Bhatti 47

Types of Unsupervised Learning

- Clustering Algorithms
  - K-Means Clustering
  - Hierarchical Clustering
  - Mean-Shift Clustering
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

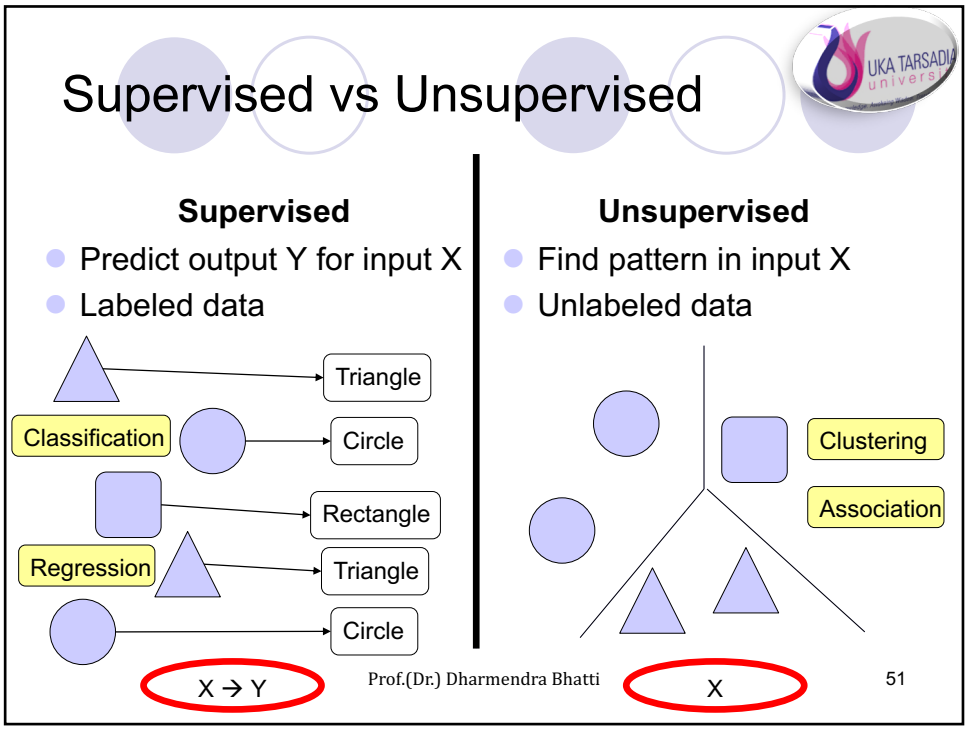
Prof.(Dr.) Dharmendra Bhatti 48

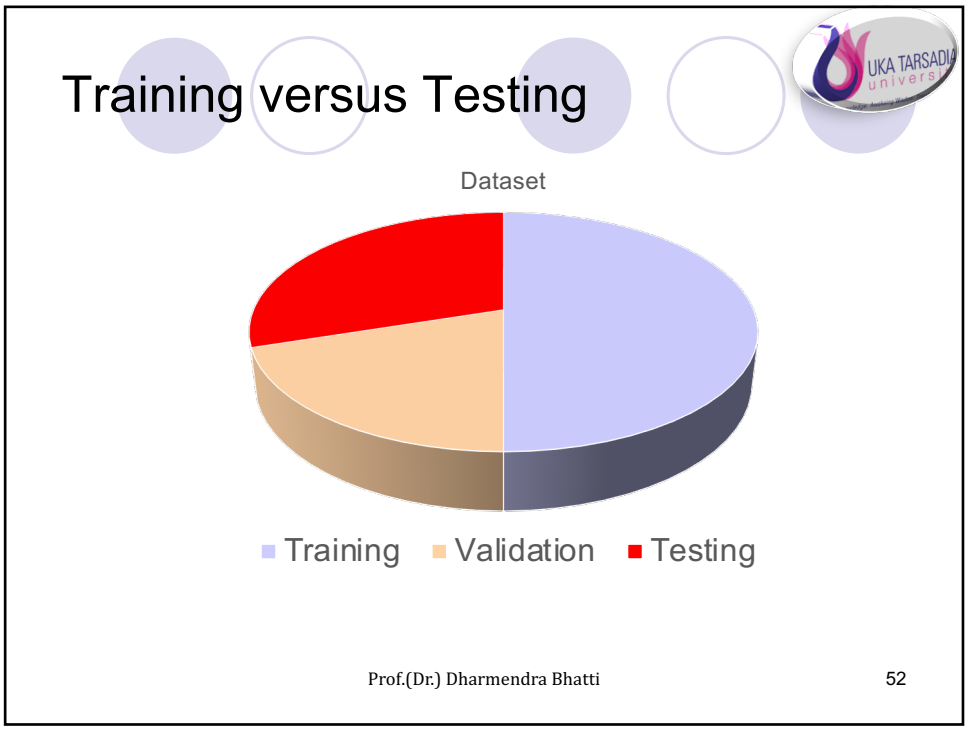
Types of Unsupervised Learning

- Association
  - Find associations and relations among data items
  - i.e. which items frequently purchased together

Types of Unsupervised Learning

- Association Algorithms
  - Apriori
  - Eclat
  - FP-growth
  - ASSOC
  - OPUS


Supervised vs Unsupervised
Supervised
Predict output Y for input X
Labeled data
Unsupervised
Find pattern in input X
Unlabeled data
Triangle
Classification
Circle
Rectangle
Regression
Triangle
Circle
Clustering
Association
X → Y
X
Prof.(Dr.) Dharmendra Bhatti
51


Training versus Testing
Dataset
Training
Validation
Testing
Prof.(Dr.) Dharmendra Bhatti
52

Training dataset

- A training dataset is a dataset of examples used during the learning process and is used to fit the parameters.
- **Objective**: train the model

Validation dataset

- A validation dataset is a dataset of examples used to tune the hyperparameters.
- **Objective**: provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters

Testing dataset

- A test dataset is a dataset that is **independent** of the training dataset, but that follows the same probability distribution as the training dataset.
- **Objective**: evaluate the performance of the model using some performance metric

Testing dataset

- If training data consists of testing data
- Then ???

Testing dataset

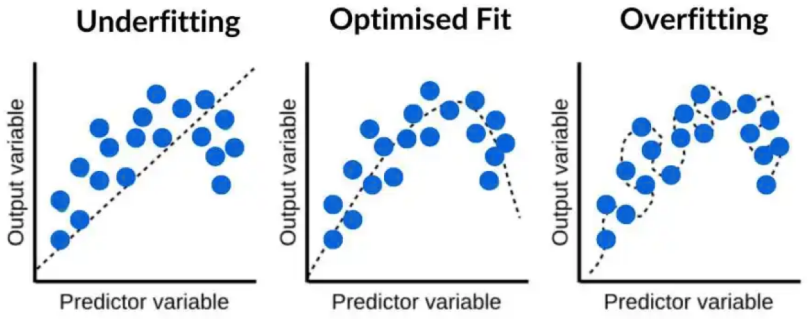
- If training data consists of testing data
- Then excellent result due to memorize/over-fitting.

Testing dataset

- Memorize (over-fitting)
  - Accurate with training set, but will fail to predict the value of the response variable for new examples.
- **Regularization** may be applied to many models to reduce over-fitting.

Testing dataset

- Over-fitting vs under-fitting (losing important properties)



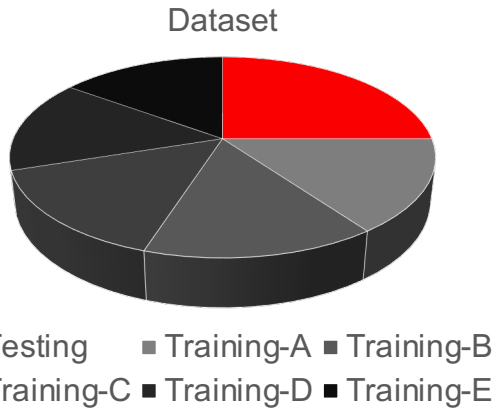
Prof.(Dr.) Dharmendra Bhatti 59

Cross-validation

- A dataset can be repeatedly split into a training dataset and a validation dataset: this is known as cross-validation.

Prof.(Dr.) Dharmendra Bhatti 60

Cross-validation



Prof.(Dr.) Dharmendra Bhatti 61

Training versus Testing

- Performance measures used to evaluate machine learning algorithms
  - Accuracy, Precision and Recall

		Predicted class	
		P	N
Actual class	P	TP	FN
	N	FP	TN

where: P = positive; N = Negative; TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative.

Prof.(Dr.) Dharmendra Bhatti 62

Training versus Testing

- Accuracy
  - **ACC = (TP + TN)/(TP + TN + FP + FN)**

Training versus Testing

- Precision or Positive Predictive Value
  - **PREC = TP/(TP + FP)**

Training versus Testing

- Recall or Sensitivity or True Positive Rate
  - **R = TP/(TP + FN)**

Data Preprocessing

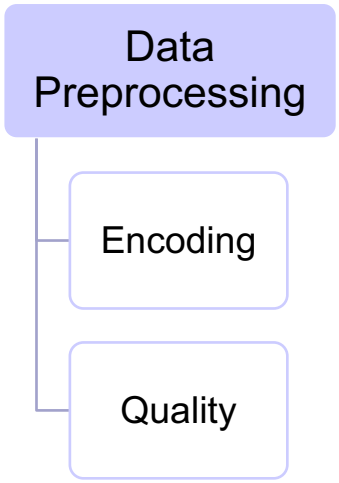
- In any Machine Learning process, Data Preprocessing is transforming or encoding data so that the machine can easily parse it.
- In other words, the *features* of the data can be easily interpreted by the algorithm.

Data Preprocessing

- Data preprocessing is the technique to convert raw data in to clean data before passing it to machine learning algorithm.
- “Garbage in, garbage out.”

Prof.(Dr.) Dharmendra Bhatti 67

Data Preprocessing



Prof.(Dr.) Dharmendra Bhatti 68

Data Preprocessing

- Encoding
  - Machine don't understand
    - Free text
    - Tables
    - Images
    - Audio files
    - Videos
  - Machine understands 1s and 0s
  - Transform or encode data so that machine can understand it.

Prof.(Dr.) Dharmendra Bhatti 69

Data Preprocessing

- A dataset can be viewed as collection of
  - Objects
  - Records
  - Vectors
  - Events
  - Cases
  - Samples
  - Entities
  - Observations

Described by features

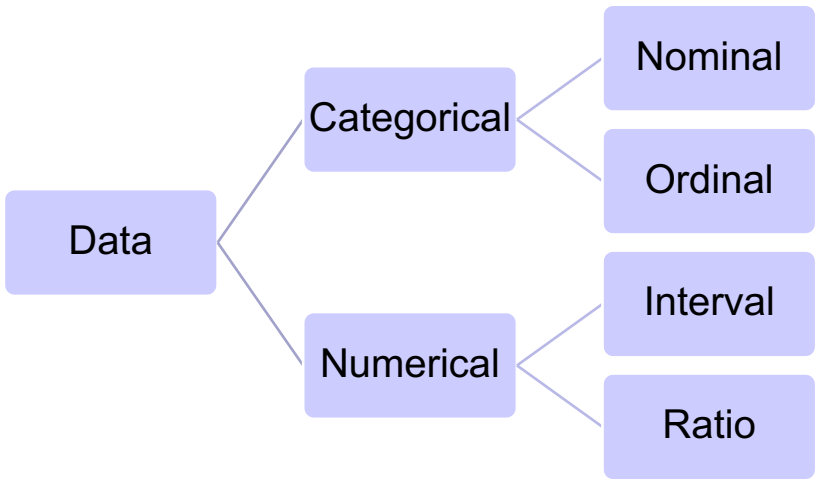
Prof.(Dr.) Dharmendra Bhatti 70

Data Preprocessing

- Features
  - Captures basic characteristics of data object
  - Also called variables, characteristics, attributes, fields, dimensions
  - A feature is an individual measurable property or characteristic of a phenomenon being observed

Prof.(Dr.) Dharmendra Bhatti 71

Data Preprocessing



Prof.(Dr.) Dharmendra Bhatti 72

Data Preprocessing

- Categorical
  - Features whose values are taken from a defined set of values.
  - Months: {Jan, Feb, ..., Dec}
  - {True, False}

Data Preprocessing

- Categorical
  - Nominal
    - No implied order
    - Mobile colour: Red, Blue, Black
  - Ordinal
    - Natural implied order without scale of difference
    - Size: Extra Small < Small < Medium < Large < Extra Large but (Large – Medium) do not equal to (Medium – Small)

Data Preprocessing

- Numerical
  - Features whose values are numeric or integer.
  - Number of students in a class
  - Vehicle speed

Data Preprocessing

- Numerical
  - Interval
    - Numerical variables with a defined unit of measurement and differences between values are meaningful
    - Dates, Celsius
  - Ratio
    - Numerical variables with a defined unit of measurement and differences and ratios are meaningful. Absolute zero.
    - Age, Length

Data Preprocessing

- Quality
  - Missing values
    - NULL
    - NaN
    - Blank field

Data Preprocessing

- Quality
  - Out-of-range values
    - Income: -10,000
    - Age: 412
    - Exam Result: 101%

Data Preprocessing

- Quality
  - Impossible data combinations
    - Sex: Male, Pregnant: Yes

Data Preprocessing

- Quality
  - Inconsistent data
    - Address: 9825339693

Data Preprocessing

- Quality
  - Duplicate values
    - Remove bias

Data Preprocessing

Independent Variables

Dependent Variable

City	Age	Salary	Purchased
Bardoli	45	151000	Yes
Surat	51	164000	Yes
Valsad	42	134000	No
Surat	24	89000	Yes
Valsad	36		No
Bardoli	25	95000	Yes
Surat		156000	No
Bardoli	43	128000	No
Valsad	30	103000	Yes
Bardoli	27	91000	Yes

Missing Values

Data Preprocessing

- **Missing Data**
- Categorical Data
- Feature Scaling

Data Preprocessing

- Missing Data
  - Values not available while collecting data
  - How to handle missing data?
    - Delete the row
    - Generate the missing data
      - mean
      - median
      - most_frequent (mode)

Data Preprocessing

- The **mean** is the **average** of a data set.
- The **median** is the middle of the set of numbers.
- The **mode** is the most common number in a data set.

Data Preprocessing

- **Mean:** The “average” number; found by adding all data points and dividing by the number of data points.
- Example: The mean of 4, 2, and 3 is (4 + 2 + 3) / 3 = 9 / 3 = 3.

Data Preprocessing

- What is the arithmetic mean of the following numbers?
- 4, 9, 7, 3, 2

Data Preprocessing

- **Median:** The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).
- Example: The median of 4, 2, and 3 is 3 because when the numbers are put in order (2, 3, 4), the number 3 is in the middle.

Data Preprocessing

- What is the arithmetic median of the following numbers?
- 4, 9, 7, 3, 2

Data Preprocessing

- What is the arithmetic median of the following numbers?
- 4, 6, 9, 7, 3, 2

Data Preprocessing

- **Mode:** The most frequent number - that is, the number that occurs the highest number of times.
- Example: The mode of {3, 2, 4, 3, 2, 2}, is 2 because it occurs three times, which is more than any other number.

Data Preprocessing

- What is the mode of the following numbers?
- 4, 3, 3, 2, 3, 2

Data Preprocessing

- Missing Data
- **Categorical Data**
- Feature Scaling

Data Preprocessing

- Which is categorical data?

City	Age	Salary	Purchased
Bardoli	45	151000	Yes
Surat	51	164000	Yes
Valsad	42	134000	No
Surat	24	89000	Yes
Valsad	36		No
Bardoli	25	95000	Yes
Surat		156000	No
Bardoli	43	128000	No
Valsad	30	103000	Yes
Bardoli	27	91000	Yes

Data Preprocessing

- Which is categorical data?

City	Age	Salary	Purchased
Bardoli	45	151000	Yes
Surat	51	164000	Yes
Valsad	42	134000	No
Surat	24	89000	Yes
Valsad	36		No
Bardoli	25	95000	Yes
Surat		156000	No
Bardoli	43	128000	No
Valsad	30	103000	Yes
Bardoli	27	91000	Yes

Prof.(Dr.) Dharmendra Bhatti 95

Data Preprocessing

- Categorical Data
  - Qualitative in nature
  - Machine don't understand text and thus requires to encode the categorical data

Prof.(Dr.) Dharmendra Bhatti 96

Data Preprocessing

- Categorical Data
  - LabelEncoder class will convert Bardoli, Surat, Valsad to 0, 1, 2.

Data Preprocessing

- Categorical Data
  - LabelEncoder class will convert Bardoli, Surat, Valsad to 0, 1, 2.
  - What is the problem with this encoding?

Data Preprocessing

- Categorical Data
  - LabelEncoder class will convert Bardoli, Surat, Valsad to 0, 1, 2.
  - What is the problem with this encoding?
  - Valsad > Surat as 2 > 1 which doesn't make any sense.

Prof.(Dr.) Dharmendra Bhatti 99

Data Preprocessing

- Categorical Data
  - **Solution**: instead of having one column with n number of categories, we will use n number of columns with only 1s and 0s to represent whether the category occurs or not.
  - Use library called OneHotEncoder

Prof.(Dr.) Dharmendra Bhatti 100

Data Preprocessing

- Categorical data

Bardoli	Surat	Valsad	Age	Salary	Purchased
1	0	0	45	151000	Yes
0	1	0	51	164000	Yes
0	0	1	42	134000	No
0	1	0	24	89000	Yes
0	0	1	36		No
1	0	0	25	95000	Yes
0	1	0		156000	No
1	0	0	43	128000	No
0	0	1	30	103000	Yes
1	0	0	27	91000	Yes

Prof.(Dr.) Dharmendra Bhatti 101

Data Preprocessing

- Missing Data
- Categorical Data
- **Feature Scaling**

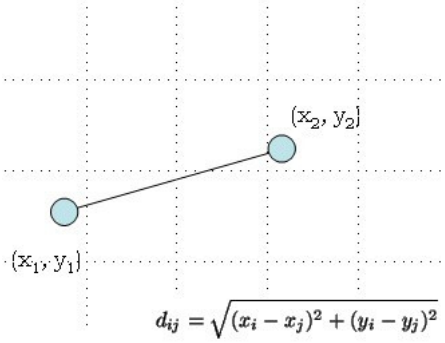
Prof.(Dr.) Dharmendra Bhatti 102

Data Preprocessing

- Feature Scaling
  - standardize the range of independent variables or features of data.
  - Use the class StandardScaler

Data Preprocessing

- Feature Scaling
  - If the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared.



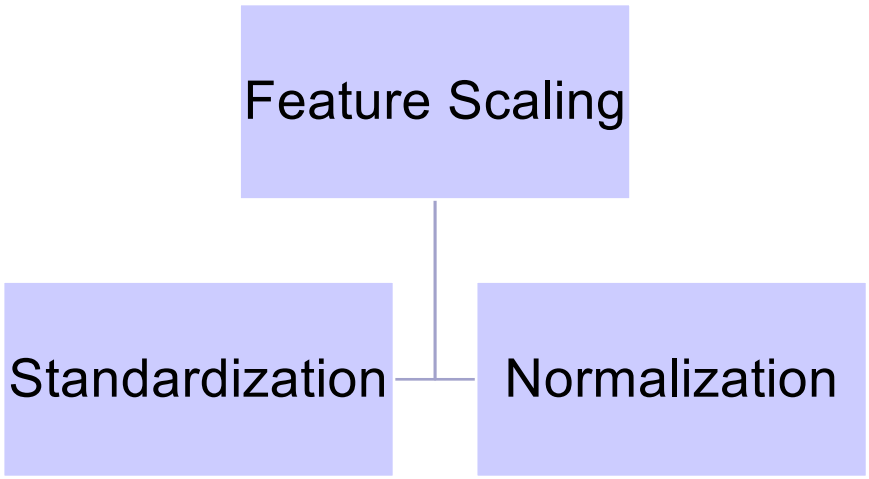
Data Preprocessing

- Salary variable dominates Age variable

Bardoli	Surat	Valsad	Age	Salary	Purchased
1	0	0	45	151000	Yes
0	1	0	51	164000	Yes
0	0	1	42	134000	No
0	1	0	24	89000	Yes
0	0	1	36		No
1	0	0	25	95000	Yes
0	1	0		156000	No
1	0	0	43	128000	No
0	0	1	30	103000	Yes
1	0	0	27	91000	Yes

Prof.(Dr.) Dharmendra Bhatti 105

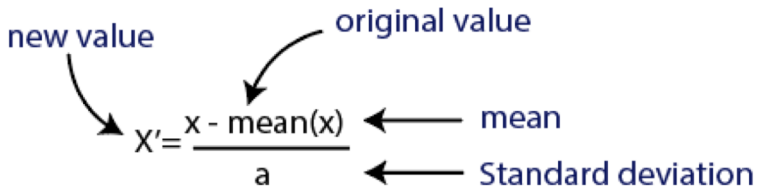
Data Preprocessing



Prof.(Dr.) Dharmendra Bhatti 106

Data Preprocessing

- Feature Scaling – Standardization



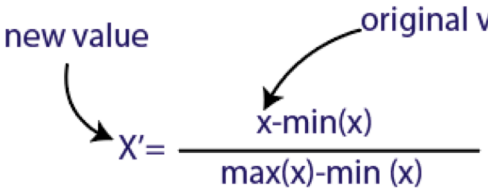
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation
N = the size of the population
x_i = each value from the population
μ = the population mean

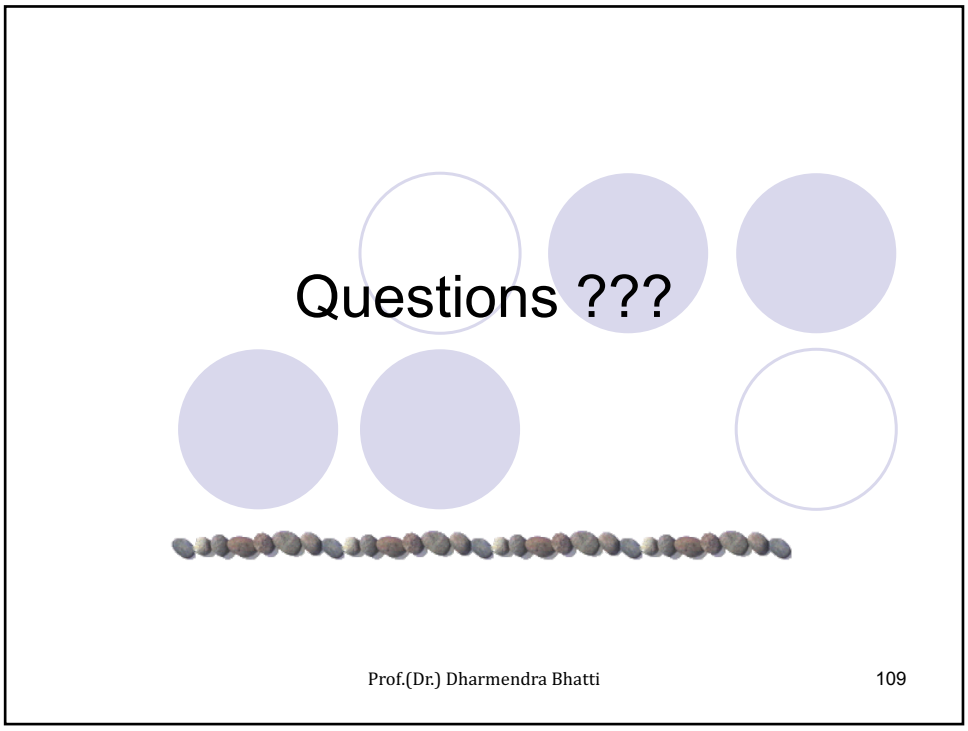
Prof.(Dr.) Dharmendra Bhatti

Data Preprocessing

- Feature Scaling – Normalization



Prof.(Dr.) Dharmendra Bhatti 108


Questions ???
Prof.(Dr.) Dharmendra Bhatti
109