Books

1. Introduction to ML with python : A guide for Data Scientist — O'Reilly
   Andreas Muller — 2016
2. Hands on ML with Scikit-learn, Keras & Tensorflow — O'Reilly
   Aurelien Geron

Unit-1 Introduction to ML
Unit-2 Clustering in ML
Unit-3 Classification in ML
" -4 Ensemble learning & random forest
" 5 Dimensionality reduction.

Students should able to learn

⇒ Describes theories, methods & algo. in ML
⇒ find & analyze the optimal hyper parameters of ML algo.
⇒ Examine the nature of a problem at hand & determine whether ML can solve it efficiently
⇒ Solve & Implement real world problems using ML.

CO1- describe in-depth about theories, methods & algo in ML
CO2- Clustering
CO3- Classification
CO4- Understand Random forest as an ensemble method specifically using DT
CO5- Students will grasp why reducing no of features in a dataset is important

learning CD-602 (Summary Unit-1)

roduction to ML —

⇒ What is Machine learning
⇒ ML life cycle
⇒ Types of ML (supervised & unsupervised)
⇒ Batch & online learning
⇒ Instance based & model based learning
⇒ Scope & limitations of ML
⇒ challanges of ML
⇒ Data Visulization, hypothesis function & testing
⇒ Data preprocessing
⇒ Data augmentation
⇒ Normalizing data sets.
⇒ Bias-Variance trade off
⇒ Relationship b/w AL, ML, DL, DS

⇒ What is Machine Learning

* Science of getting Computer to learn automatically without being Explicitly programmed.
* ML works on Simple Concept, Understand with Experience.

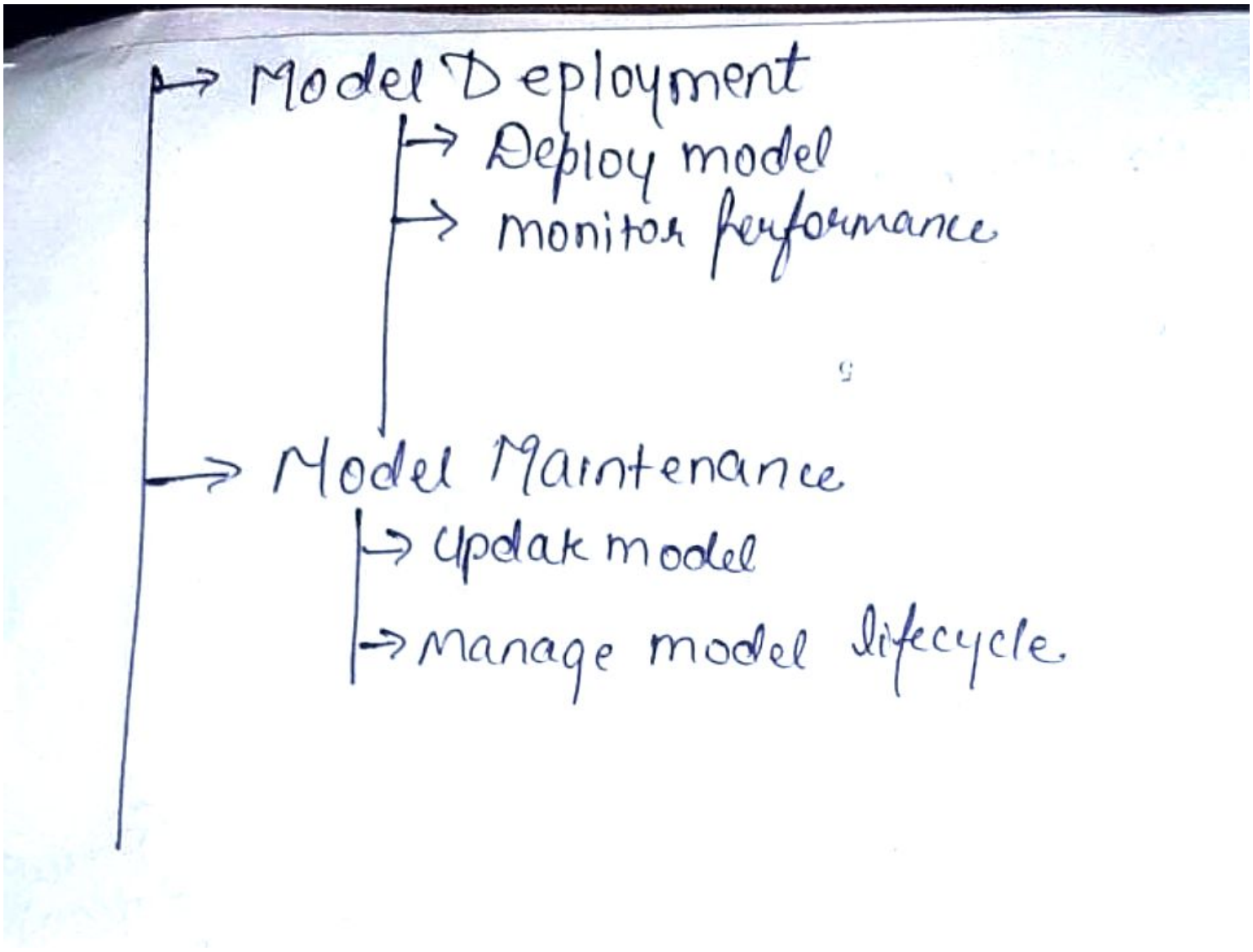
Eg:- Ads Recomendation
Tag Friends on facebook
Searc Engine
Education

Need of ML

- Handling Big data
- Automation
- Speech recognition
- Prediction & Pattern Recognition
- Adaptability to learn new data.
- Solve complex problems
- Recommendation System in e-commerce (Personalization)
- Innovation (Developing Intelligent Systems)

ML life Cycle

- Problem Definition
  - Identify the problem
  - Define the success Criteria
- Data Collection
  - Gather Data
  - Understanding Data
- Data pre-processing (Data Preparation)
  - (Handling missing values)
  - Data cleaning
  - Data Transformation – Normalize Sale, encode
  - Feature Engineering – create new features
  - Data Splitting – split dataset onto train & test
- Model Development
  - Select Algorithms
  - model training
  - Hyperparameter tuning – optimize hyperparameters to improve model performance
- Model Evaluation
  - Evaluate models
  - Compare models


Model Deployment
Deploy model
monitor performance
Model Maintenance
Updak model
Manage model lifecycle.

Traditional Programming v/s Machine learning.

Traditional

Data → Compute → output
Program →

Machine learning

Data → Computer → program
output →

Working of Machine learning

- Input training data
- Analyze the data
- Finding Patterns
- Predictions
- Results

- Re-train the model (in case accuracy is not good)
- This enables continually learn on its own. and improve accuracy.

Types of Machine learning

- Supervised ML (classification / Regression)
- Unsupervised ML (Data driven → clustering) Association
- Reinforcement ML (learn by mistakes)
- Batch & online learning
  - Batch (offline learning)
    Involves training a model using entire dataset at once., used in traditional ML
  - Online learning (Incremental learning)
    Involves training a model incrementally, processing one training examples or small batch of examples at a time, Used where data evolves over time
- Instance based and model based learning
  - Instance based
    (Also known as memory based, do not construct a general internal model during training)

Model based learning

↳ Involves creating a predictive model from the training data, which is then used to make predictions on new data.

Scope & Limitations of ML

Scope

- ML in Education
- ML in Search Engine
- ML in Digital Market
- ML in Health Care
- ML in Spam Detection
- ML in Traffic alert
- ML in Google translate

Limitations

- accuracy depends on training & learning which is not always avl.
- Require accurate details on many past data
- Requires large amount data
- M/c needs different dataset to learn Meaningful insights.

Challanges in ML

- Data Quality and Quantity
  - Data Collection
  - Data Quality
- Feature Engineering
  - feature Selection
  - feature Extraction
- Model Selection
  - Choosing the right Algorithm
  - Hyperparameter Tuning
- Overfitting & underfitting
- Computational Resources
  - Processing Power
  - Scalability
- Interpretability
  - Model Transparency
  - Explainability

Deployment & Maintenance
- Model Deployment
- model maintenance

Ethical and Bias Issues
- Bias in Data
- Ethical considerations

Security & privacy
- Data privacy
- Model Security

Data Visualization

→ Involves representing data in graphical or pictorial formats to facilitate understanding and insight extraction

→ Primary goal

↳ To communicate information clearly & effectively through visual means, making it easier for individuals to comprehend complex data patterns

→ Data visulization tools

Charts, Graphs, maps, Dashboards
Power BI, Qlika, Tableau, looker, Qlik Sense

→ Types of Data Visulization

→ Numerical Data Visulaizatia
→ Categorical Data Visulization

→ Types of Data Visulization Techniques

→ Bar chart
→ Line chart
→ Pie chart
→ Scatter Plot
→ Histogram
→ Heatmap
→ Box plot
→ Bubble chart
→ Area chart
→ Tree maps
→ Word Clouds
→ 3D Surface plot
→ N/w Graphs

→ Advantages

Enhanced comparison, Improved Methodology, Efficient Data Sharing, Sales Analysis

→ Disadvantages

Can be time-consuming, Can be misleading
Can be difficult to interpret

Hypothesis function and Testing

Hypothesis in ML

→ Hypothesis is an assumption made by scientists, whereas a model is a mathematical representation that is used to test the hypothesis

Hypothesis fn

→ In ML hypothesis fn, denoted as $h(x)$ represents the model or fn that we use to make predictions based on input features x.

→ Goal of ML algo is to find best hypothesis fn that appropriate the true r/s b/w i/p features & target variables.

⇒ Hypothesis is defined as the supposition or proposed explanation based on insufficient evidence or assumptions.

Hypothesis Testing

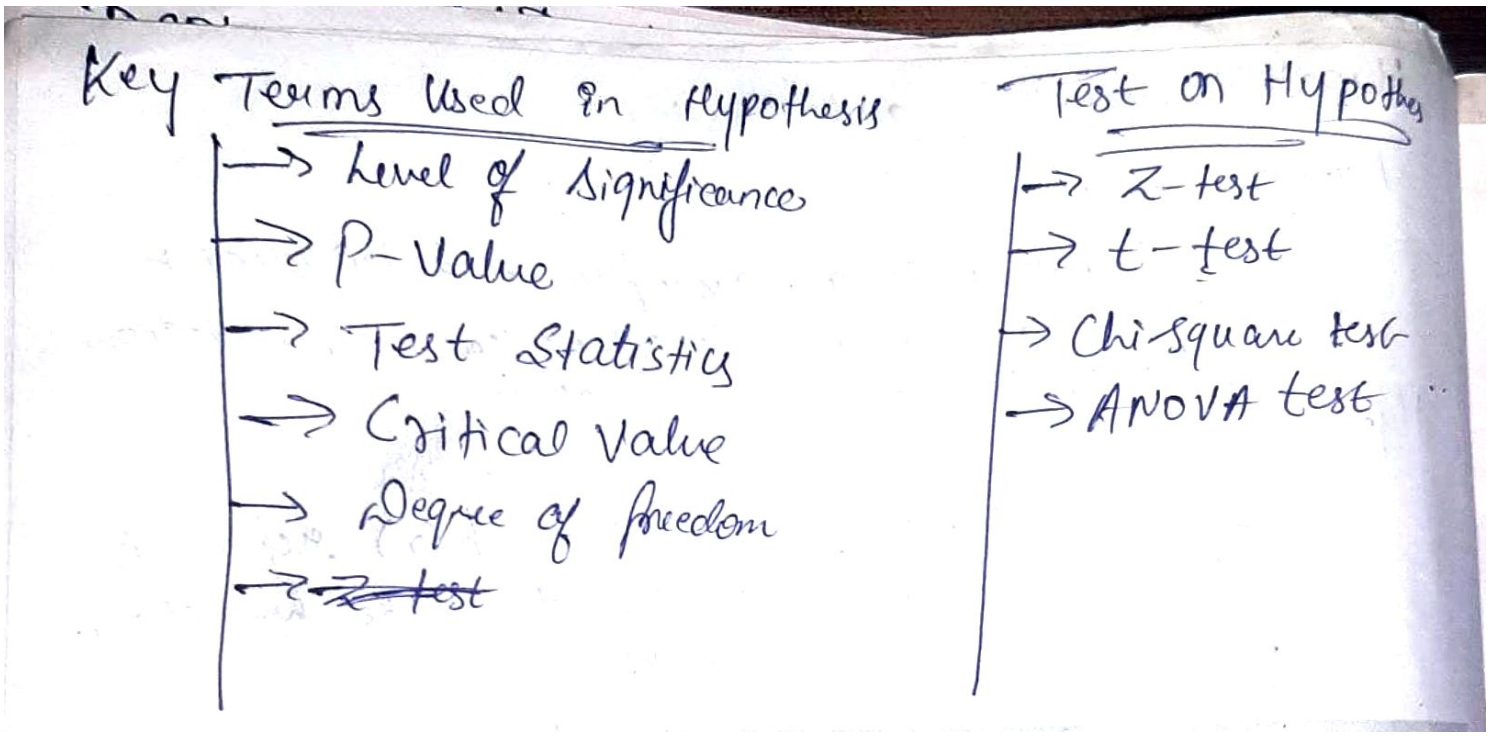
→ It is a statistical method that is used to make a statistical decision using experimental data. Hypothesis testing is basically an assumption that we make about population parameter.

Types of Hypothesis

→ Null Hypothesis (H_0) General statement that there is no r/s b/w two measures cases or groups.

→ Alternative Hypothesis (H_1)

The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to null hypothesis.


Key Terms Used in Hypothesis
→ Level of Significance
→ P-Value
→ Test Statistics
→ Critical Value
→ Degree of freedom
→ Z-test
Test on Hypothes
→ Z-test
→ t-test
→ Chi-square test
→ ANOVA test

Z-Test

↳ Statistical method used to determine whether two population means are different when the variance are known & the sample size is large.

Steps involved in z-test

- formulate Hypothesis
  - Null Hypothesis
  - Alternative Hypothesis
- Calculate z-score
- Determine Critical value
- Compare z-score & Critical Value

T-test

↳ Statistical tool used to compare the means of two groups & determine if they are statistically different from each other.

Types of t-test

- One Sample t-test
- Independent two-sample t-test
- Paired Sample t-test

Steps involved in t-test

- formulate hypothesis
  - Null hypothesis
  - alternative hypothesis
- Choose the significance level (α)
- Cal. the Test Statistic (t-score)
  - Apply one sample test
  - Independent two sample
  - Paired sample

→ Determine degree of freedom (df)
→ Compare the t-score with the Critical value.

Chi-Square Test

→ Non-parametric Statistical test
→ Used to determine whether there is a significance association b/w Categorical Variables or not.

Types of chi-square Test

→ Chi-square Test for Independence
→ Chi-Square Goodness-of-Fit Test

Steps involved in Chi-Square test

→ formulate hypothesis
  → Null hypothesis
  → Alternative hypothesis
→ Construct a Contingency Table

→ Calculate Expected frequencies
→ Calculate the Chi-square Statistic
→ Determine degree of freedom
→ Compare Chi-square Statistic with Critical value

ANOVA

Analysis of Variance
↳ Statistical test used to compare the means of three or more groups to determine if there is a statistical significance difference among them.

Types of ANOVA

→ One way Anova
→ Two way Anova

Roll No. ..... Scholar No. .....

Bias - Variance Tradeoff

→ describes the balance b/w two sources of error that affect a model's performance : bias & variance

→ Bias

Error occured b/w model's predicted value and actual value.

→ Bias is a systematic error that occurs due to wrong assumptions in ML process

↳ High Bias - makes strong assumptions about data

Variance

→ Error introduced by the model's sensitivity to fluctuations in the training data

High Variance - leads to Overfitting

Bias - Variance Tradeoff

→ There is a trade off b/w bias and variance, as you decrease bias, variance usually increases, and vice-versa.

→ High Bias, low Variance

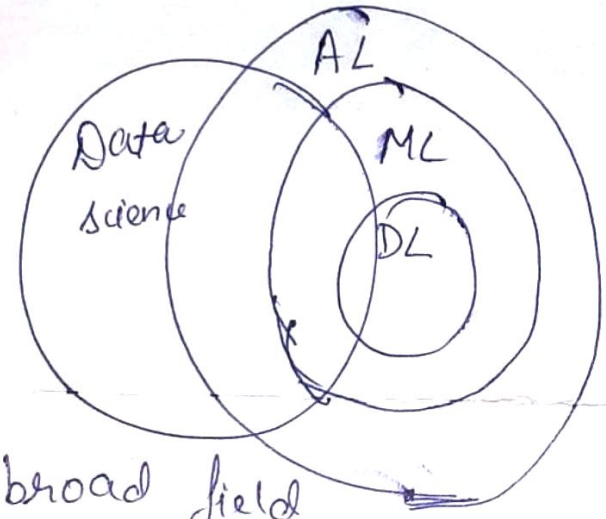
→ Low Bias, High Variance

→ Optimal Point

Mitigating the Bias-Variance Tradeoff

- → Cross Validation
- → Regularization
- → Ensemble Methods

Difference b/w AI, ML, DL, DS



AI – broad field Can perform task that required human intelligence

ML – Subset of AI focused on algorithms that learn from data

DL – A further subset of ML involving deep neural n/ws

DS – field of Extracting insights from data, often using ML & DL techniques