SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

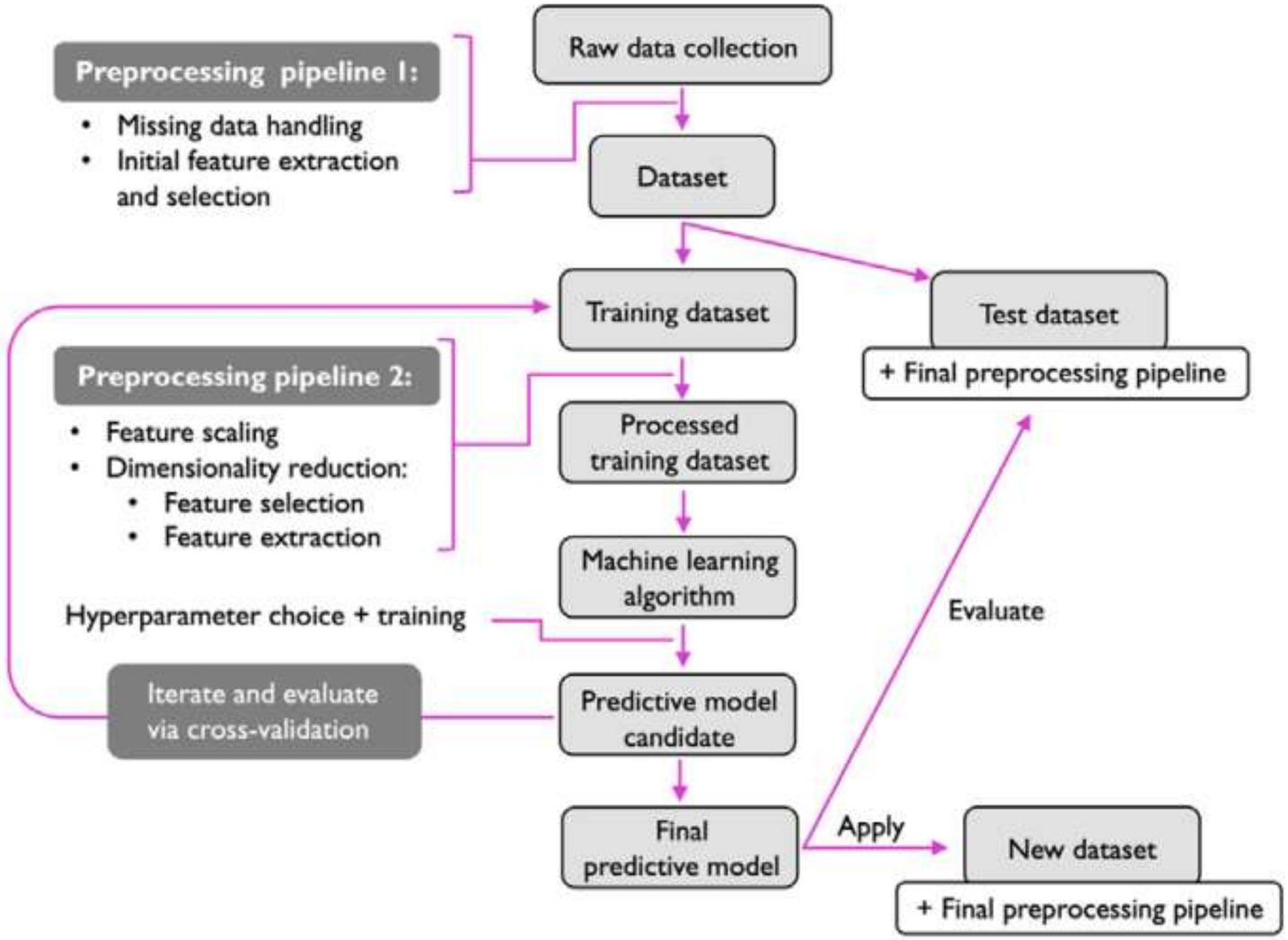
23AMB201 - MACHINE LEARNING

II YEAR IV SEM

UNIT I – INTRODUCTION

TOPIC 4– Turning data into Probabilities and Statistics for Machine Learning

Recall Training, Validation and Testing datasets

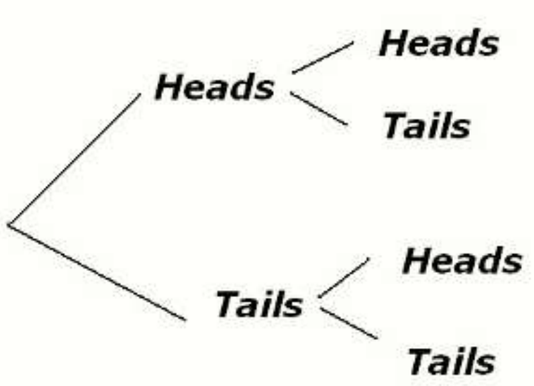


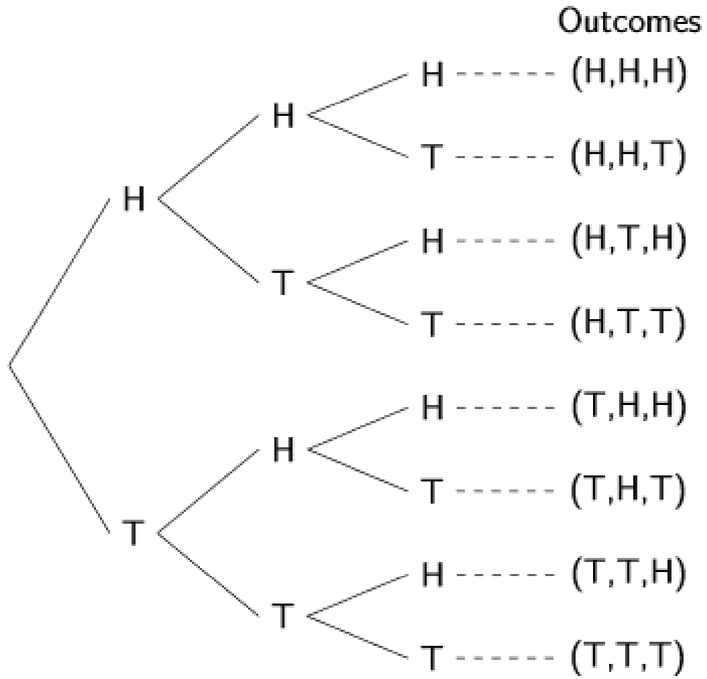
What is probability?

1. The ratio of the number of favourable outcomes to the total number of outcomes of an event is defined as probability.
2. The number of favourable outcomes for an experiment with 'n' outcomes is denoted by x.

Probability (Event) = Favourable Outcomes/Total Outcomes

= x/n





Types of Probabilities

1. Classical Probability
2. Empirical Probability
3. Subjective Probability
4. Axiomatic Probability

Classical Probability **Formula**: $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$

Outcomes in a sample space

Naive Bayes

Empirical Probability $P(A) = \frac{\text{Number of times event A occurs}}{\text{Total number of trials}}$

Observed data or historical frequencies

classification or regression tasks

Subjective Probability

Personal judgment, experience

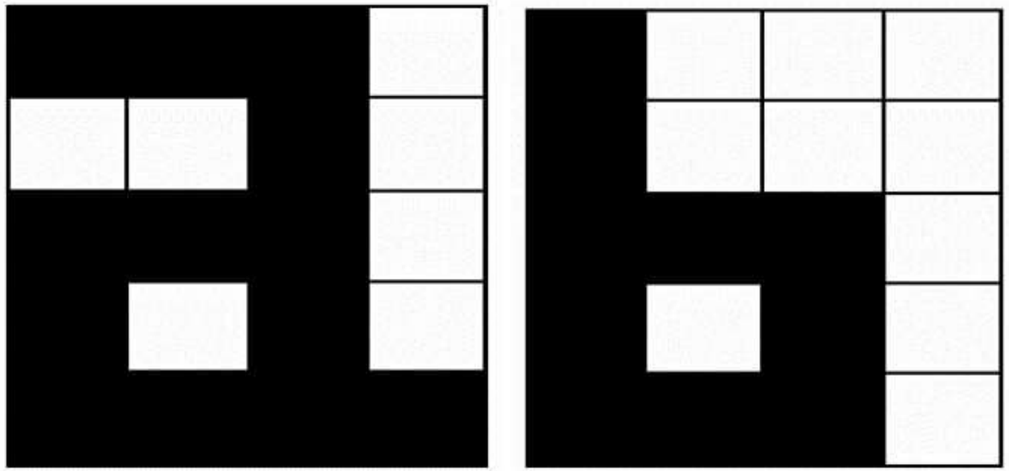
Reinforcement learning

Axiomatic Probability $P(A \cup B) = P(A) + P(B)$

set of axioms (rules)

Bayesian networks, hidden Markov models

Example



	A	B	C	D	E
1	Age	Income	Student	Creadit_Rating	Buys_Computer
2	<=30	high	no	fair	no
3	<=30	high	no	excellent	no
4	31-40	high	no	fair	yes
5	>40	medium	no	fair	yes
6	>40	low	yes	fair	yes
7	>40	low	yes	excellent	no
8	31-40	low	yes	excellent	yes
9	<=30	medium	no	fair	no
10	<=30	low	yes	fair	yes
11	>40	medium	yes	fair	yes
12	<=30	medium	yes	excellent	yes
13	31-40	medium	no	excellent	yes
14	31-40	high	yes	fair	yes
15	>40	medium	no	excellent	no

Statistics for Machine Learning

Core component of data analytics and machine learning. It helps you analyze and visualize data to find unseen patterns

Functions:

1. Collecting
2. Analyzing
3. Interpreting
4. Visualizing Data

Standard deviation measures how far apart numbers are in a data set. **Variance**, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean.

$$Mean = \frac{sum\ of\ all\ values}{total\ number\ of\ values}$$

$$Median = middle\ value\ (when\ the\ data\ are\ arranged\ in\ order)$$

$$Mode = most\ common\ value$$

Example: 5,9,4,7,8,6,3,5,5,6

Find Mean, Median, Mode

References

1. Aurélien Géron "Hands-On Machine Learning with Scikit-Learn and TensorFlow" Publisher(s): O'Reilly Media, Inc 2017.
2. https://medium.com/@ompramod9921/cross-validation-623620ff84c2
3. https://serokell.io/blog/machine-learning-testing