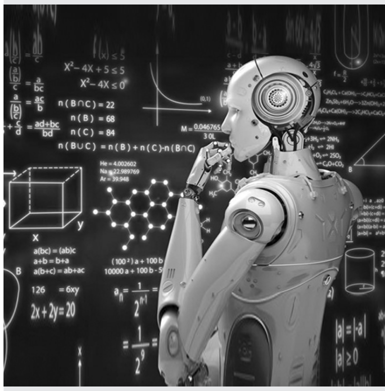
22AM204 MACHINE LEARNING



Source: https://www.forbes.com/sites/kalevleetaru/2019/01/15/why-machine-learning-needs-semantics-not-just-statistics/

Hours Per Week :

L	T	P	C
3	0	2	4

PREREQUISITE KNOWLEDGE: Probability & Linear Algebra, Python language.

COURSE DESCRIPTION AND OBJECTIVES:

This course provides a broad introduction to various machine learning concepts including Supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks) and Unsupervised learning (clustering, dimensionality reduction) methods. Students will get an understanding of various challenges of Machine Learning and will be able to decide on model complexity. Numerous case studies introduced in this course allow the students to apply machine-learning algorithms in computer vision, medical imaging, audio, and text domains. Laboratory experiments of this course will introduce students to advanced Machine Learning Python libraries such as Scikit-Learn, Matplotlib, and many other recent ML-related APIs. The course is designed such that the students get enough hands-on experience with a major focus onthe practical implementation of theoretical concepts.

MODULE-1

UNIT-1 **14L+0T+8P=22 Hours**

INTRODUCTION:

What is machine learning? Machine learning applications; Types of Learning: Supervised learning; Un-supervised learning; Reinforcement learning.

Model Training Essentials: Re-sampling methods: Bias–Variance Trade-off. Hypothesis Testing and Variable Selection, Sub sampling and Upsampling, SMOTE; Cross Validation (validation set, Leave-One-Cut (LOO), k-fold strategies) and bootstrap; Evaluation measures-Error functions, Confusion Matrix, Accuracy, Precision and Recall, F1 Score.

Regression Analysis: Linear Regression, Simple and Multiple Linear Regression, Polynomial Regression, Logistic Regression, Multi nominal Regression. Ordinary Least Squares Method, Model Shrinkage-Ridge, and LASSO regression.

UNIT-2 **10L+0T+8P=18 Hours**

FEATURE SELECTION:

Feature Selection Strategies: Problem statement and Uses, Filter methods, Wrapper methods, Embedded methods. Branch and bound algorithm, Sequential forward/backward selection algorithms.

Dimensionality Reduction: Singular value decomposition, matrix factorization, Linear discriminant analysis, Principal components analysis.

PRACTICES:

- Apply the following tasks to any given dataset:
  a. Load and visualize data
  b. Check out and replace missing values
  c. Encode the Categorical data
  d. Splitting the dataset into Training and Test set
  e. Splitting the dataset into k-folds
  f. Feature scaling
- House price prediction:
  a. Create a model that predicts a continuous value (price) from input features square footage, number of bedrooms and bathrooms.).
  b. Implement a univariate Model using Least Squares and plot best-fit line
  c. Implement a multivariate Model using Least Squares and plot best-fit line
  d. Retrieve model error and model coefficients.
  e. Observe Variance Inflation Factor(VIF)

SKILLS:

- ✓ *Statistical data analysis.*
- ✓ *Classify / Cluster data.*
- ✓ *Tool usage for developing ML applications.*

f. Implement Ridge regression model
g. Implement LASSO regression model
h. Report your observations on the above models for house prediction

- Heart disease prediction:
  a. Implement a logistic regression model to predict whether an individual is suffering fromheart disease or not
  b. Evaluate and compare model performance using the following validation approaches:
     i. Validation set approach
     ii. K-fold cross validation
     iii. Stratified K-fold cross validation
     iv. LOO strategy
  c. Plot Confusion matrix
  d. Report performance of the model in terms of the following metrics:
     i. Accuracy
     ii.Precision-Recall
     iii. F1 Score
  e. Report your observations and explain when to use what type of measures
- Implement the Polynomial Regression algorithm to fit data points. Select the appropriatedata set for your experiment and draw graphs.
- Working with imbalanced datasets:
  a. Load an imbalanced dataset and visualize imbalance in the data as a bar plot
  b. Implement KNN model for classification
  c. Balance the dataset using:
     i. Random Over sampling
     ii. Random Under sampling
     iii. SMOTE
  d. Implement KNN model for classifying data balanced in the above steps
  e. Report your observations on the performance of models trained using balanced and imbalanced data
- Perform effective feature selection in a given dataset using any one of the feature selection techniques.
- Dimension Reduction:
  a. Load a dataset and Implement Bayes classification model
  b. Apply dimension reduction using:
     i. Principal Component Analysis
     ii. Linear Discriminant Analysis
  c. Apply the model on data with reduced dimension
  d. Compare and contrast model performance in each case

MODULE-2

UNIT-1 **16L+0T+8P=24 Hours**

CLASSIFICATION:

Classification: Binary, Multi-class and Multi-label Classification; K-Nearest Neighbours, Support Vector Machines, Decision Trees, The Naïve Bayes' Classifier, Class Imbalance, Perceptron ANN model.

Ensemble Methods: Ensemble Learning Model Combination Schemes, Voting, Error-Correcting Output Codes, Bagging: Random Forest Trees, Boosting: Adaboost, Stacking.

UNIT-2 **8L+0T+8P=16 Hours**

CLUSTERING:

Clustering: Different distance functions and similarity measures, K-means clustering, Medoids, Hierarchical Clustering-Single linkage and Complete linkage clustering, Graph based Clustering -MST, DBSCAN, Spectral clustering.

PRACTICES:

- Implement and demonstratethe FIND-Salgorithm for finding the most specific hypothesis based on a given set of training data samples. Read the training data from a .CSV file.
- Implement the naïve Bayesian classifier for a sample training data set stored as a.csv file. Compute the accuracy of the classifier, considering few test data sets.
- Assuming a set of spam or not-spam mails that need to be classified, use the naïve Bayesian classifier model to perform this task. Calculate the accuracy, precision, and recall for your data set.
- Implement k-Nearest Neighbor algorithm to classify the iris data set. Print both correct and wrong predictions. Python ML library classes can be used for this problem.
- Demonstrate the working of the decision tree-based ID3 algorithm. Use an appropriate data set for building the decision tree and apply this knowledge to classify a new sample?
- Build a model using SVM with different kernels.
- Implement and build models using the following Ensemble techniques
  a. Bagging
  b. Boosting: Adaboost, Stacking
- Build a model to perform Clustering using K-means after applying PCA and determining the value of K using the Elbow method.
- Unsupervised Modeling:
  a. Cluster the data using the following models:
    i. Spectral Clustering
    ii. K-medoids
    iii. DBSCAN
    iv. Hierarchical Clustering
  b. Compare and contrast model performance in each case.

COURSE OUTCOMES:

Upon successful completion of this course, students will have the ability to:

CO No.	Course Outcomes	Blooms Level	Module No.	Mapping with POs
1	Apply a wide variety of learning algorithms such as Probabilistic, Discriminative and Generative algorithms for a given application.	Apply	1, 2	1
2	Design an end-to-end Machine-learning model to realize solutions for real-world problems.	Design	1	3
3	Implement various machine learning models using advanced ML tools.	Create	1, 2	5
4	Analyze and evaluate the performance of various machine learning models approaches on different kinds of data.	Analyze	2	2

TEXT BOOKS:

1. EthemAlpaydin, “Introduction to Machine Learning”, 3rd edition, The MIT Press, 2014
2. Flach, Peter. “Machine learning: the art and science of algorithms that make sense of data”. Cambridge University Press, 2012.

REFERENCE BOOKS:

1. Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
2. AurélienGéron, “Hands-on Machine Learning with Scikit Learn and Tensor Flow”, O’reilly, 2017.
3. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, “An Introduction to Statistical Learning with Applications in R”, Springer, 2013. (ISLR).