Machine Learning

Introduction to Machine Learning

Satishkumar L. Varma

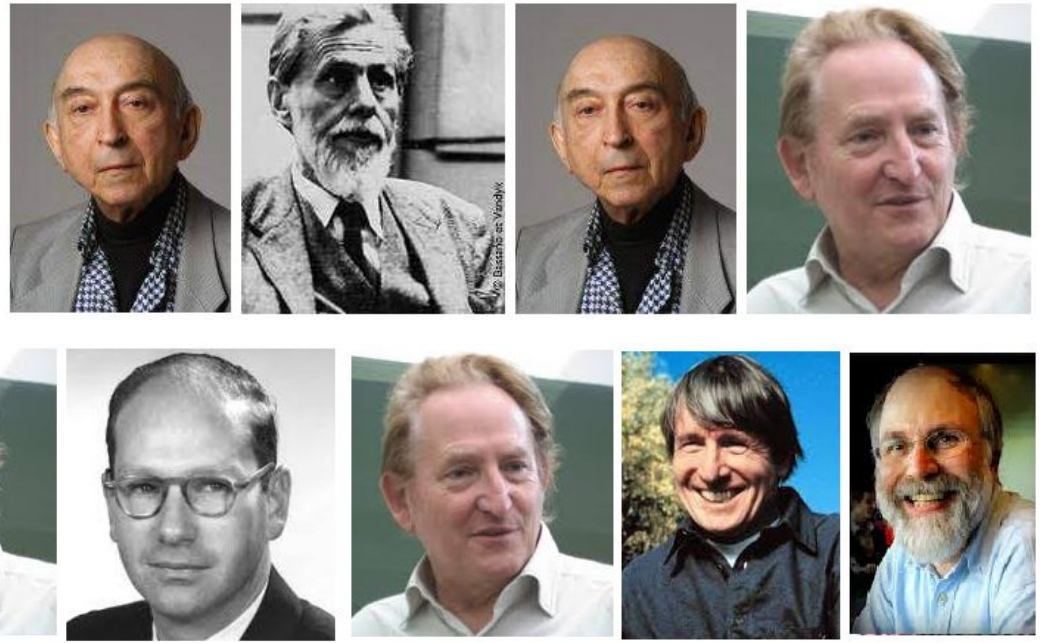
Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Introduction to Machine Learning
  - Types of Machine Learning
  - Steps involved in developing a Machine Learning Application
  - Evaluating a Learning Algorithm
  - Evaluating Hypothesis
  - Model Selection and Train/ Validation/ Test Sets
  - Bias Vs variance: Regularization and Bias/ Variance, Learning Curve, Error Analysis
  - Handling Skewed Data: Error Matrices for Skewed Classes
  - Trade-off between Precision and recall
- Issues in Machine Learning
- Application of Machine Learning

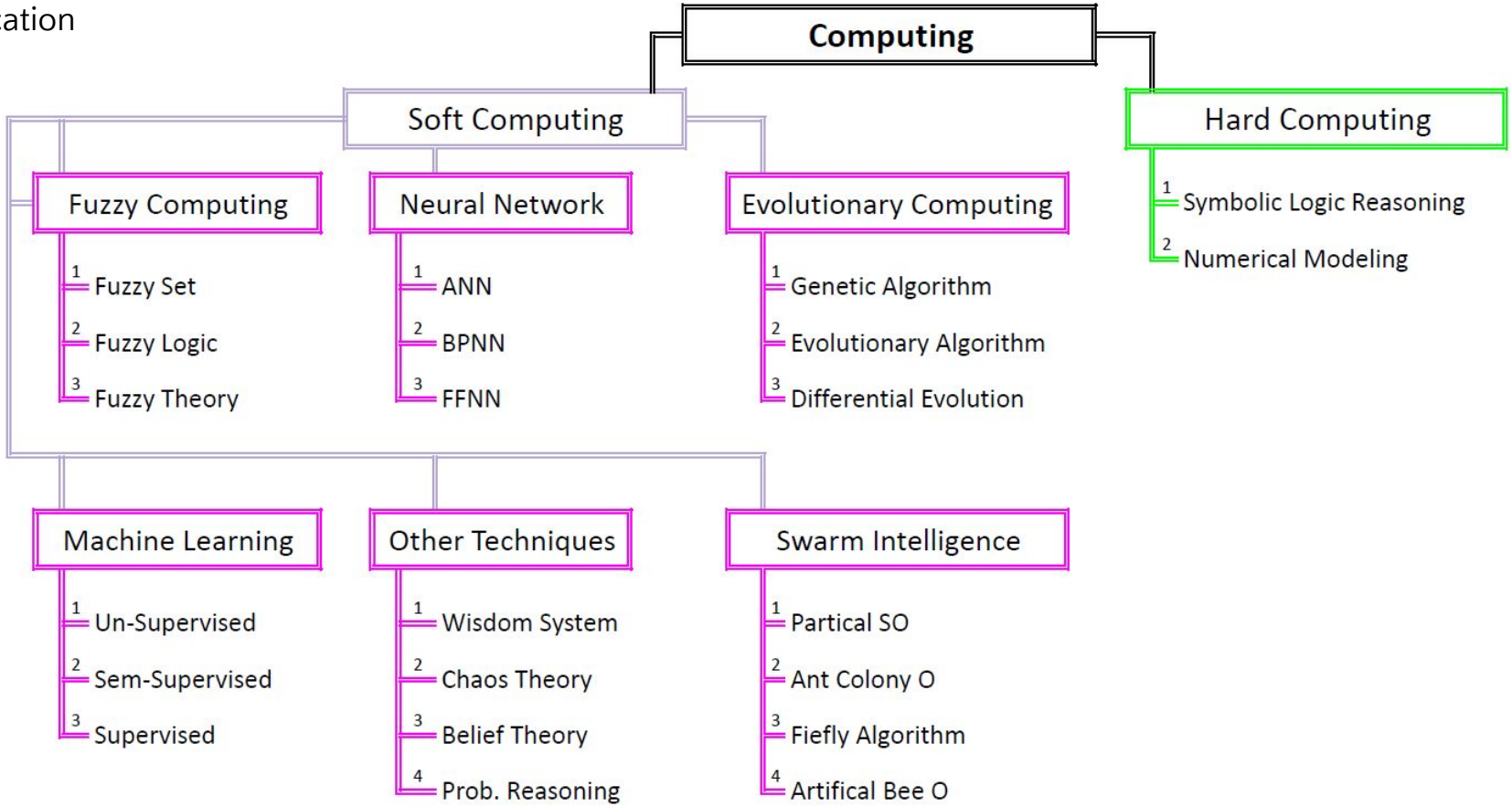
Introduction to Machine Learning

- Lotfi A. Zadeh, 1997, defined SC into one multidisciplinary system as the fusion of
  - **Neural Network**: for learning and adaptation
  - Fuzzy Logic: for knowledge representation via fuzzy If – Then rules.
  - Genetic Computing: for evolutionary computation
- History and Classification
- SC(Zadeh, 1981) =
  - [McCulloch(NN, 1943) +
  - Zadeh(FL, 1965) +
  - Rechenberg(EC, 1960)]
- EC(Rechenberg, 1960) =
  - [Fogel(EP, 1962) +
  - Rechenberg(ES, 1965) +
  - Holland(GA, 1970) +
  - Koza(GP, 1992)]



Introduction to Machine Learning

- Classification



Introduction to Machine Learning

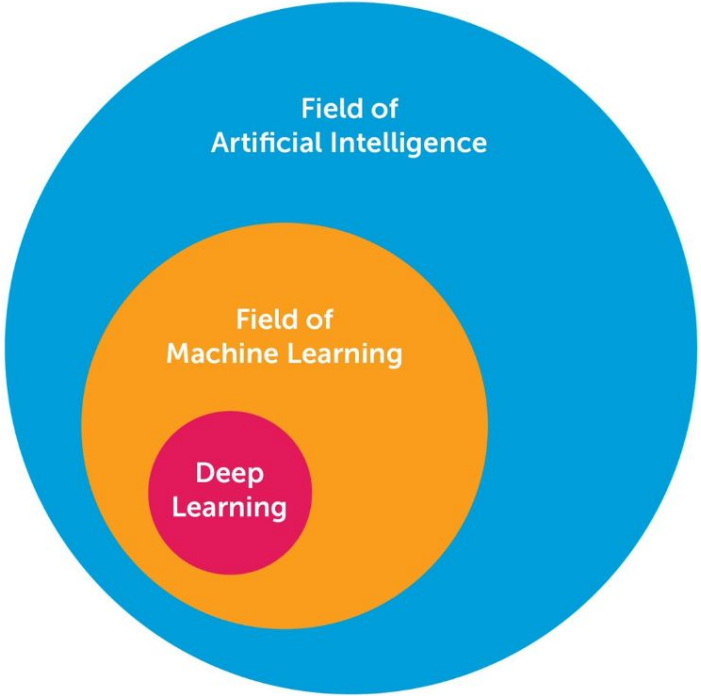
- ML Classification

Learning Algorithms

Supervised Learning		Unsupervised Learning		Reinforcement Learning	
Classificaiton		**Clustering**		1	Robot Navigation
1	Customer Retention	1	Recommender Systems	2	Skill Acquistion
2	Fraud Detection	2	Customer/Market segmentation	3	Learning Tasks
3	Object Classification	3	Social Network Nnalysis	4	Real-time Decisions
4	Diagnostics	4	Image Segmentation	5	Game AI
5	Sentiment Analysis	5	Anomaly Detection		
Regression		**Dimensionality Reduction**			
1	Market / Weather Forecasting	1	Compression		
2	Advertizing Popularity Prediction	2	Learning Obervations/ Behavior		
3	Polution Growth Prediction	3	Structure Discovery		
4	Estimating Life Expectancy	4	Big Data Visualization		

Introduction to Machine Learning

- Data Science: Interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data
  - Artificial intelligence
  - Machine learning
  - Deep learning
- Data Science ?



Introduction to Machine Learning

Parameters	Supervised Machine Learning	Unsupervised Machine Learning
Input Data	Labeled data	Unlabeled data
Output	Desired output is given	Desired output is not given
Training data	Use training data to infer model	No training data is used
Model	We can test our model	We can not test our model
Accuracy	Highly accurate	Less accurate
No. of classes	Known	Not known
Data Analysis	Uses offline analysis	Uses real-time analysis of data
Called as	Also called as classification	Also called as clustering
Comp. Complexity	Simpler method	Computationally complex
Complex model	Can't learn larger/more complex models	Can learn larger/more complex models
Algorithms used	LR, NB, RF, SVM, NN, K-NN, DT, etc.	K-Means, H-clustering, Apriori, etc.
Examples of problems	Classification and regression	Clustering and generative modeling
Examples of algorithms	Logistic regression and random forest	K-means and **Generative Adversarial Networks**
Applications	OCR, Classification, Regression, Object Detection, Captioning, etc.	Find face in image, Clustering, Dim. Reduction, Density Estn., Feature Learning.

Introduction to Machine Learning

- ML/DL Classification

ML/DL Learning Algorithms

Supervised Learning Algorithms	
1	Multilayer Perceptrons (MLPs)
2	Convolutional Neural Network (CNN)
3	Long Short-Term Memory (LSTM)
4	Recurrent Neural Network (RNN)
5	Radial Basis Function Network (RBFN)

Unsupervised Learning Algorithms	
1	Self Organizing Maps (SOMs)
2	Radial Basis Function Network (RBFN)
3	Restricted Boltzmann Machines (RBM)
4	Deep Belief Network (DBN)
5	Variational AutoEncoder (VAE)
6	Generative Adversarial Network (GAN)

Introduction to Machine Learning

- Definition: Machine Learning
  - ML is a subset of AI
  - Automatically learn from data, improve performance from past experiences, and make predictions.
  - It contains a set of algorithms that work on a huge amount of data.
  - Data is fed to these algorithms to train them,
  - An on the basis of training, they build the model & perform a specific task.
- ML algorithms help: Regression, Classification, Forecasting, Clustering, and Associations, etc.
- Types of Machine Learning
  - Supervised Machine Learning (House Price Prediction, Medical Image Processing)
  - Unsupervised Machine Learning (Customer Segmentation, Market Basket Analysis)
  - Semi-Supervised Machine Learning (Text Classification, Lane Fining on GPS Data)
  - Reinforcement Learning (Driverless Cars, Optimization Problems)

Supervised Machine Learning

- **Categories** of Supervised Machine Learning
- Classification Algorithms
  - Naive Bayes
  - K-Nearest Neighbors (KNN)
  - Random Forest; Gradient Boosting
  - Decision Tree
  - Logistic Regression
  - Support Vector Machine
  - Problems: Fraud Detection, Spam Detection, Email filtering, etc.
- Regression: Definition: Predict continuous output variables to solve linearly related input and output variables.
- Regression algorithms:
  - Simple Linear Regression; Polynomial Regression
  - Multivariate Regression
  - Decision Tree; Random Forest
  - Lasso Regression; Ridge Regression
- Applications of Supervised Learning
- Image Segmentation; Medical Diagnosis; Fraud Detection; Spam detection; Speech Recognition

Supervised Machine Learning

- Advantages:
  - Supervised learning allows collecting data and produces data output from previous experiences.
  - Helps to optimize performance criteria with the help of experience.
  - Supervised machine learning helps to solve various types of real-world computation problems.
  - It performs classification and regression tasks.
  - It allows estimating or mapping the result to a new sample.
  - We have complete control over choosing the number of classes we want in the training data.
- Disadvantages:
  - Classifying big data can be challenging.
  - Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
  - Supervised learning cannot handle all complex tasks in Machine Learning.
  - Computation time is vast for supervised learning.
  - It requires a labelled data set.
  - It requires a training process.

Unsupervised Machine Learning

- **Categories** of Unsupervised Machine Learning
  - Clustering
  - Association
- Clustering Algorithms
  - K-Means Clustering
  - Mean-shift
  - DBSCAN
  - Principal Component Analysis
  - Independent Component Analysis
- Association: Finds interesting relations among variables within a large dataset.
  - Association Problems: Market Basket analysis, Web usage mining, continuous production, etc.
  - Association Algorithms: Apriori Algorithm, Eclat, FP-growth algorithm.
- Applications of Unsupervised Learning
  - Clustering; Anomaly detection; Dimensionality reduction; Network Analysis
  - Recommendation Systems; Image and video compression; Genomic data analysis; Image segmentation
  - Anomaly Detection; Market basket analysis; Community detection in social networks;
  - Singular Value Decomposition; Content recommendation;

Unsupervised Machine Learning

- Advantages of unsupervised learning:
  - It does not require training data to be labeled.
  - Dimensionality reduction can be easily accomplished using unsupervised learning.
  - Capable of finding previously unknown patterns in data.
  - Flexibility: USL is flexible and can be applied to for clustering, anomaly detection, and asso.rule mining.
  - Exploration: USL allows for the exploration of data and the discovery of novel and useful patterns.
  - Low cost: USL is often less expensive than supervised learning
- Disadvantages of unsupervised learning
  - Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
  - The results often have lesser accuracy.
  - The user needs to spend time interpreting and label the classes which follow that classification.
  - Lack of guidance: Lacks the guidance and feedback provided by labeled data.
  - Sensitivity to data quality: Sensitive to data quality, including missing values, outliers, and noisy data.
  - Scalability: Unsupervised learning can be computationally expensive, particularly for large datasets or complex algorithms, which can limit its scalability.

Semi-Supervised Learning: SL + USL

- It uses the combination of labelled and unlabeled datasets during the training period.
- It mostly consists of unlabeled data as labels are costly.
- Applications of Semi-Supervised Learning
  - Image Classification and Object Recognition; Natural Language Processing (NLP)
  - Speech Recognition; Recommendation Systems; Healthcare and Medical Imaging
- **Example:**
  - SL: Student is under the supervision of an faculty at home and college.
  - USL: Student is self-analysing the same concept without any help from the faculty .
  - SSL: Student has to revise after analyzing the same concept under faculty guidance in college.
- Advantages of Semi-supervised Learning
  - It is simple and easy to understand the algorithm.
  - It is highly efficient.
  - It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.
- Disadvantages of Semi-supervised Learning
  - Iterations results may not be stable.
  - We cannot apply these algorithms to network-level data.
  - Accuracy is low.

Reinforcement Learning

- Works on a feedback-based process using an AI agent (A software component);
- Agent gets rewarded for each good action and get punished for each bad action;
- Hence the goal of reinforcement learning agent is to maximize the rewards.
- No labelled data like supervised learning, and agents learn from their experiences only.
  - Example: A child learns various things by experiences in day-to-day life just like a game playing.
  - **Game** - environment; **States** - moves of an agent at each step; and **Goal** - is to get a high score.
  - Agent receives feedback in terms of punishment and rewards.
- Reinforcement Learning Algorithms
  - Q-learning
  - SARSA (State-Action-Reward-State-Action)
  - Deep Q-learning

Reinforcement Learning

- Types of Reinforcement Machine Learning
- Positive Reinforcement
  - Rewards the agent for taking a desired action.
  - Encourages the agent to repeat the behavior.
  - Examples: Giving a treat to a dog for sitting, providing a point in a game for a correct answer.
- Negative Reinforcement
  - Removes an undesirable stimulus to encourage a desired behavior.
  - Discourages the agent from repeating the behavior.
  - Examples: Turning off a loud buzzer when a lever is pressed, avoiding a penalty by completing a task.
- Applications of Reinforcement Learning
  - Game AI; Video Games; Resource Management; Text Mining
  - Industrial Control; Autonomous Vehicles; Robotics; Education; Agriculture
  - Recommendation Systems; Natural Language Processing (NLP)
  - Healthcare; Finance and Trading; Supply Chain and Inventory Management
  - Energy Management; Adaptive Personal Assistants;
  - Virtual Reality (VR) and Augmented Reality (AR);

Steps involved in developing a Machine Learning Application

- Steps to Build a Machine Learning Model
  - Step 1: Data Collection for Machine Learning
  - Step 2: Data Preprocessing and Cleaning
  - Step 3: Selecting the Right Machine Learning Model
  - Step 4: Training Your Machine Learning Model
  - Step 5: Evaluating Model Performance
  - Step 6: Tuning and Optimizing Your Model
  - Step 7: Deploying the Model and Making Predictions
- Implement Machine Learning Steps in Python

Training Dataset vs Testing Dataset vs Validation Dataset

- Splitting the dataset is to assess how effective will the trained model be in generalizing to new data.
  - This split can be achieved by using train_test_split function of scikit-learn.
- Training Set
  - This is the actual dataset from which a model trains .i.e.
  - the model sees and learns from this data to predict the outcome or to make the right decisions.
  - Generally, this data is more than 60% of the total data available for the project.
- Testing Set
  - This dataset is independent of the training set but
  - It has a somewhat similar type of probability distribution of classes and
  - It is used as a benchmark to evaluate the model,
  - It is used only after the training of the model is complete.
- Validation Set
  - The validation set is used to fine-tune the hyperparameters of the model and
  - It is considered a part of the training of the model.
  - The model only sees this data for evaluation but
  - The model does not learn from this data, providing an objective unbiased evaluation of the model.

Common issues in Machine Learning

- Inadequate Training Data: Noisy Data; Incorrect data; Generalizing of output data
- Poor quality of data
- Irrelevant features
- Data Bias
- Non-representative training data
- Overfitting and Underfitting
- Monitoring and maintenance
- Getting bad recommendations
- Lack of skilled resources
- Lack of Explainability
- Customer Segmentation
- Slow implementations and results
- Process Complexity of Machine Learning

Applications

- Large amount of data, Ex. recommendation system
- Image recognition: Face recognition, Handwritten digit recognition
- Image classification: Dataset constructed from the National Institute of Standards and Technology (NIST), called MNIST (M stands for modified, which means data is pre-processed for the ease of machine learning processes
- Image-based search engines: Image classification
- Computer vision / Machine vision: Self-driving cars as an example, which interprets 360° camera views to make decisions in real time
- Color restoration from black and white photos
- Image generation: Handwriting, cat images, and even video game images. For example, interesting playground, or to create handwritings of the title of a book in three different styles

Applications

- Application of Deep Learning: Natural language processing (NLP)
  - DL with RNN are appropriate for sequences of inputs, such as natural language and text
- Machine translation
  - Example: the sentence-based Google Neural Machine Translation system (GNMT)
  - GNMT utilizes deep RNNs to improve accuracy and fluency
- Sentiment analysis, information retrieval, theme detection and many other common NLP applications
  - where DL models have achieved state-of-the-art performance thanks to word embedding techniques
- Text generation:
  - RNNs learn the intricate relationship between words (including punctuation) in sentences and
  - to write text, to become an author or a virtual Shakespeare
- Image captioning generation: Known as image to text, couples recent breakthroughs in computer vision and NLP
  - It leverages CNNs to detect and classify objects in images, and assigns labels to those objects.
  - It then applies RNNs to describe those labels in a comprehensible sentence.

Applications

- Sound and speech: A field of sequential learning, where ML to predict time series or label sequence data.
- Speech recognition: Apple's Siri, Amazon's Alexa, Google Home, Skype Translator.
- Music composer: Besides an author writing text, train RNNs to produce music.
- Accurate motion detection: Real-time behavior analysis in surveillance videos
  - Scientists from Google, DeepMind, and Oxford even built a computer lip reader called LipNet
- Besides supervised and unsupervised learning cases, deep learning is heavily used in reinforcement learning.
  - Robots who can handle objects, climb stairs, operate in kitchens are not new to us.
  - Recently, Google's AlphaGo beating the world's elite Go players received widespread media coverage.
  - Seeing self-driving cars being out in the market in just one or two years.
  - These have all benefited from the advance of deep learning in reinforcement learning.
- Bioinformatics, drug discovery, recommendation systems, finance (stock market), insurance and IoT.

Applications

- Applications of supervised learning and the different types of NN used

Input (X)	Output (y)	Application	Type of NN Used
Home Features	Price	Real Estate	Standard Neural Network
Ad, user info	Click prediction (0/1)	Online Advertising	Standard Neural Network
Image	Image Class	Photo Tagging	CNN
Audio	Text Transcript	Speech Recognition	RNN
English	Chinese	Machine Translation	RNN
Satellite Image	Position of car	Autonomous Driving	Custom / Hybrid NN

Evaluating a Learning Algorithm

- Refer slide
  - vsat2k_ML_Ch1a Evaluation of Learning Algorithms [PDF]

References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.