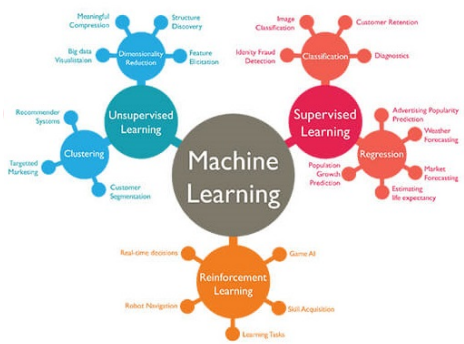
Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis



Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Supervised Machine Learning Workshop: Basic Level

Dr. Reza Belaghi

SLU,
Ultuna, Sweden

November 19, 2025

Outline

1. Introduction to Machine Learning
2. Logistic Regression
3. Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
4. Discrimination Analysis
5. K-Nearest Neighborhood (KNN)
6. Decision Trees
7. Random Forests
8. Boosting Methods
9. Support Vector Machine (SVM)
10. Artificial Neural Networks
11. Missing Data
12. Imbalanced Data
13. Variable Selection
14. Boruta Algorithm
15. Comparison of ML algorithms

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

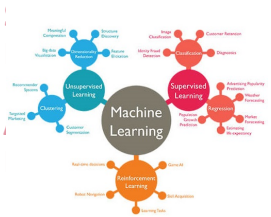
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Lecture 1

International evaluation of an AI system for breast cancer screening

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

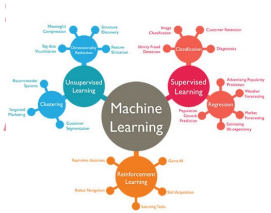
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Missing Data

Article

International evaluation of an AI system for breast cancer screening

https://doi.org/10.1038/s41586-019-1799-6

Received: 27 July 2019

Accepted: 5 November 2019

Published online: 1 January 2020

Scott Mayer McKinney[1,14]*, Marcin Sieniek[1,14], Varun Godbole[1,14], Jonathan Godwin[2,14], Natasha Antropova[2], Hutan Ashrafian[3,4], Trevor Back[2], Mary Chesus[2], Greg S. Corrado[1], Ara Darzi[3,4,5], Mozziyar Etemadi[6], Florencia Garcia-Vicente[6], Fiona J. Gilbert[7], Mark Halling-Brown[8], Demis Hassabis[2], Sunny Jansen[9], Alan Karthikesalingam[10], Christopher J. Kelly[10], Dominic King[10], Joseph R. Ledsam[2], David Melnick[6], Hormuz Mostofi[1], Lily Peng[1], Joshua Jay Reicher[11], Bernardino Romera-Paredes[2], Richard Sidebottom[12,13], Mustafa Suleyman[2], Daniel Tse[1]*, Kenneth C. Young[8], Jeffrey De Fauw[2,15] & Shravya Shetty[1,15]*

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the clinical setting, we curated a large representative dataset from the UK and a large enriched dataset from the USA. We show an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. We provide evidence of the ability of the system to generalize from the UK to the USA. In an independent study of six radiologists, the AI system outperformed all of the human readers: the area under the receiver operating characteristic curve (AUC-ROC) for the AI system was greater than the AUC-ROC for the average radiologist by an absolute margin of 11.5%. We ran a simulation in which the AI system participated in the double-reading process that is used in the UK, and found that the AI system maintained non-inferior performance and reduced the workload of the second reader by 88%. This robust assessment of the AI system paves the way for clinical trials to improve the accuracy and efficiency of breast cancer screening.

Introduction to Machine Learning

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

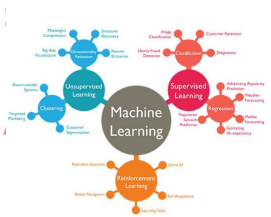
Discrimination Analysis

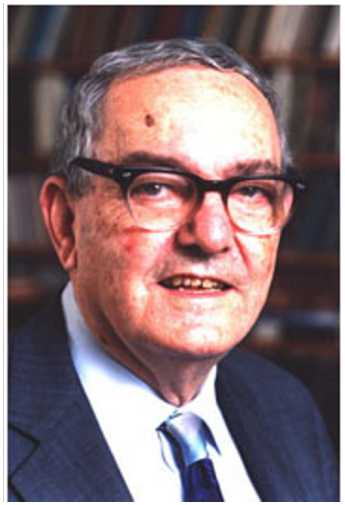
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods





Herbert Alexander Simon (1978 Nobel Prize for Economics):

- Learning is any process by which a system improves performance from experience.
- Machine Learning is concerned with computer programs that automatically improve their performance through experience.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

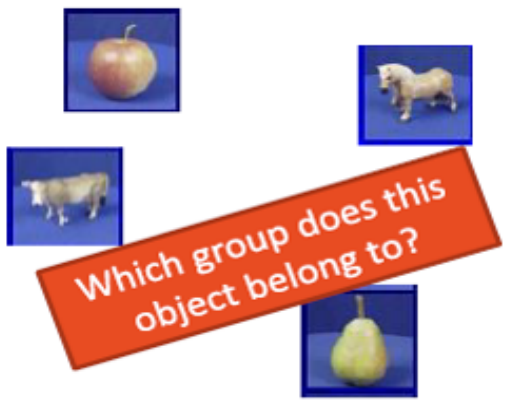
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

The Learning



Advantages of Artificial Intelligence

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

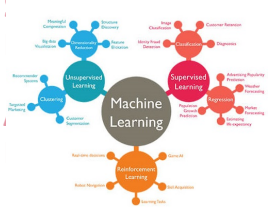
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

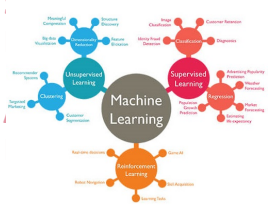
Boosting Methods



1. **Automation of Repetitive Tasks:** AI systems can automate repetitive tasks, allowing humans to focus on more complex and creative aspects of their work. This increases efficiency and reduces the risk of errors.

Advantages of Artificial Intelligence

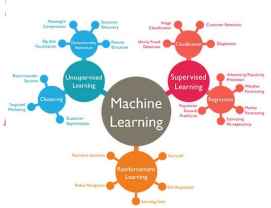
Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



1. **Automation of Repetitive Tasks:** AI systems can automate repetitive tasks, allowing humans to focus on more complex and creative aspects of their work. This increases efficiency and reduces the risk of errors.
2. **Efficient Decision Making:** AI algorithms can analyze vast amounts of data at incredible speeds, leading to more informed and data-driven decision-making processes. This is particularly beneficial in sectors such as finance, healthcare, and business.

Advantages of Artificial Intelligence

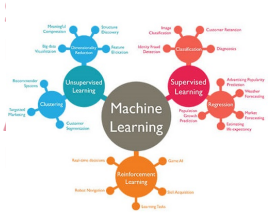
Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



1. **Automation of Repetitive Tasks:** AI systems can automate repetitive tasks, allowing humans to focus on more complex and creative aspects of their work. This increases efficiency and reduces the risk of errors.
2. **Efficient Decision Making:** AI algorithms can analyze vast amounts of data at incredible speeds, leading to more informed and data-driven decision-making processes. This is particularly beneficial in sectors such as finance, healthcare, and business.
3. **24/7 Availability:** AI systems can operate continuously without the need for breaks, providing round-the-clock availability for tasks and services.

Advantages of Artificial Intelligence

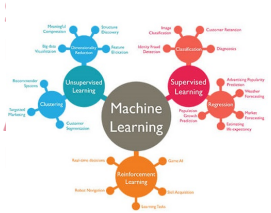
Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



1. **Automation of Repetitive Tasks:** AI systems can automate repetitive tasks, allowing humans to focus on more complex and creative aspects of their work. This increases efficiency and reduces the risk of errors.
2. **Efficient Decision Making:** AI algorithms can analyze vast amounts of data at incredible speeds, leading to more informed and data-driven decision-making processes. This is particularly beneficial in sectors such as finance, healthcare, and business.
3. **24/7 Availability:** AI systems can operate continuously without the need for breaks, providing round-the-clock availability for tasks and services.
4. **Handling Complex Tasks:** AI can handle complex tasks that may be challenging or impossible for humans, such as processing large datasets, recognizing patterns, and understanding natural language.

Advantages of Artificial Intelligence

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



1. **Automation of Repetitive Tasks:** AI systems can automate repetitive tasks, allowing humans to focus on more complex and creative aspects of their work. This increases efficiency and reduces the risk of errors.
2. **Efficient Decision Making:** AI algorithms can analyze vast amounts of data at incredible speeds, leading to more informed and data-driven decision-making processes. This is particularly beneficial in sectors such as finance, healthcare, and business.
3. **24/7 Availability:** AI systems can operate continuously without the need for breaks, providing round-the-clock availability for tasks and services.
4. **Handling Complex Tasks:** AI can handle complex tasks that may be challenging or impossible for humans, such as processing large datasets, recognizing patterns, and understanding natural language.

Advantages of Artificial Intelligence (Cont'd)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

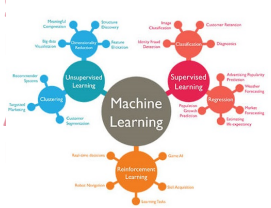
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

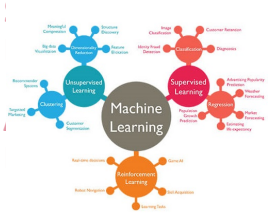
Boosting Methods



5 **Medical Advances:** AI is making significant contributions to healthcare by assisting in diagnostics, drug discovery, and personalized treatment plans. AI systems can analyze medical data more quickly and accurately than traditional methods.

Advantages of Artificial Intelligence (Cont'd)

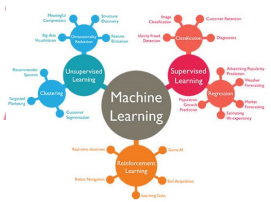
Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



5. **Medical Advances:** AI is making significant contributions to healthcare by assisting in diagnostics, drug discovery, and personalized treatment plans. AI systems can analyze medical data more quickly and accurately than traditional methods.
6. **Benefits to Natural Sciences:** AI aids in ecological modeling, biodiversity assessment, and environmental monitoring, contributing to our understanding and preservation of the natural world.

Advantages of Artificial Intelligence (Cont'd)

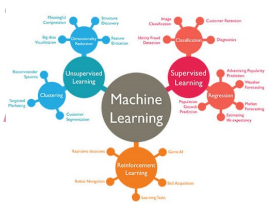
Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



5. **Medical Advances:** AI is making significant contributions to healthcare by assisting in diagnostics, drug discovery, and personalized treatment plans. AI systems can analyze medical data more quickly and accurately than traditional methods.
6. **Benefits to Natural Sciences:** AI aids in ecological modeling, biodiversity assessment, and environmental monitoring, contributing to our understanding and preservation of the natural world.
7. **Benefits to Health Sciences:** AI accelerates drug discovery, enables personalized medicine, and enhances diagnostic accuracy, leading to improved healthcare outcomes.

Advantages of Artificial Intelligence (Cont'd)

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



5. **Medical Advances:** AI is making significant contributions to healthcare by assisting in diagnostics, drug discovery, and personalized treatment plans. AI systems can analyze medical data more quickly and accurately than traditional methods.
6. **Benefits to Natural Sciences:** AI aids in ecological modeling, biodiversity assessment, and environmental monitoring, contributing to our understanding and preservation of the natural world.
7. **Benefits to Health Sciences:** AI accelerates drug discovery, enables personalized medicine, and enhances diagnostic accuracy, leading to improved healthcare outcomes.
8. **Benefits to Forest Sciences:** AI assists in forest management, monitoring deforestation, and assessing environmental impact, contributing to sustainable forestry practices and conservation efforts.

How exactly do we teach machines?

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

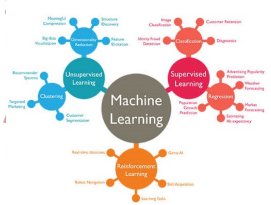
Discrimination Analysis

K-Nearest Neighborhood (KNN)

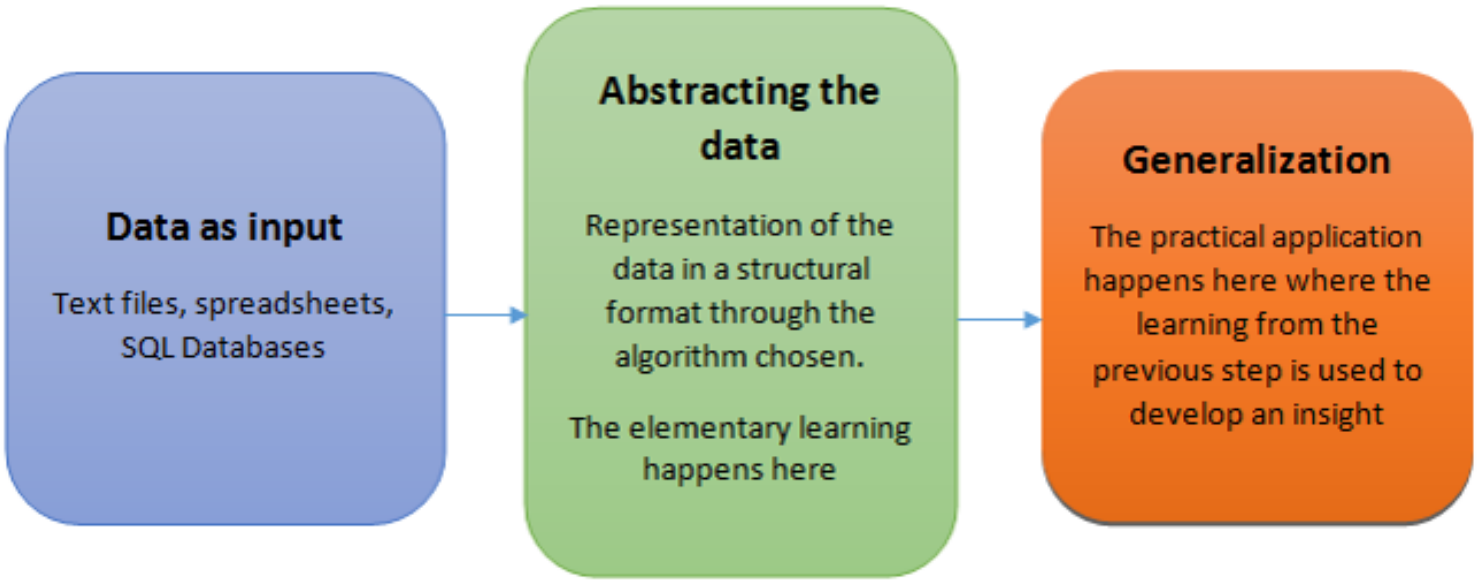
Decision Trees

Random Forests

Boosting Methods



Teaching the machines involve a structural process where every stage builds a better version of the machine. For simplification purpose, the process of teaching machines can broken down into 3 parts:



Steps to Develop a Machine Learning Model

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

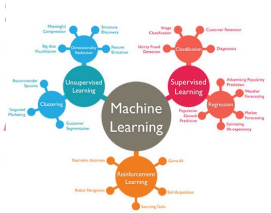
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



There are 5 basic steps used to perform a machine learning task

Steps to Develop a Machine Learning Model

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

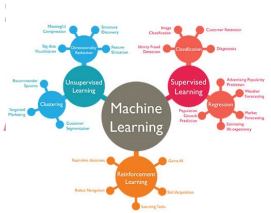
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



There are 5 basic steps used to perform a machine learning task

0. **Goal (research question)**: Variables, imputes, outcomes,....

Steps to Develop a Machine Learning Model

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

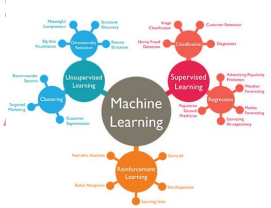
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



There are 5 basic steps used to perform a machine learning task

0. **Goal (research question)**: Variables, imputes, outcomes,....
1. **Collecting data**: Raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning. The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

2. **Training a model:** This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods
Support Vector Machine (SVM)
Artificial Neural Networks
Missing Data

2. **Training a model:** This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.
3. **Evaluating the model:** To test the accuracy, the second part of the data (holdout / test data) is used. This step determines the precision in the choice of the algorithm based on the outcome. A better test to check accuracy of model is to see its performance on data which was not used at all during model build.

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods
Support Vector Machine (SVM)
Artificial Neural Networks
Missing Data

2. **Training a model:** This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model. The second part (test data), is used as a reference.
3. **Evaluating the model:** To test the accuracy, the second part of the data (holdout / test data) is used. This step determines the precision in the choice of the algorithm based on the outcome. A better test to check accuracy of model is to see its performance on data which was not used at all during model build.
4. **Improving the performance:** This step might involve choosing a different model altogether or introducing more variables to augment the efficiency. That's why significant amount of time needs to be spent in data collection and preparation.

Summary of the Previous Steps

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

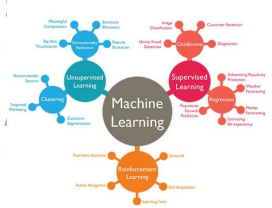
Discrimination Analysis

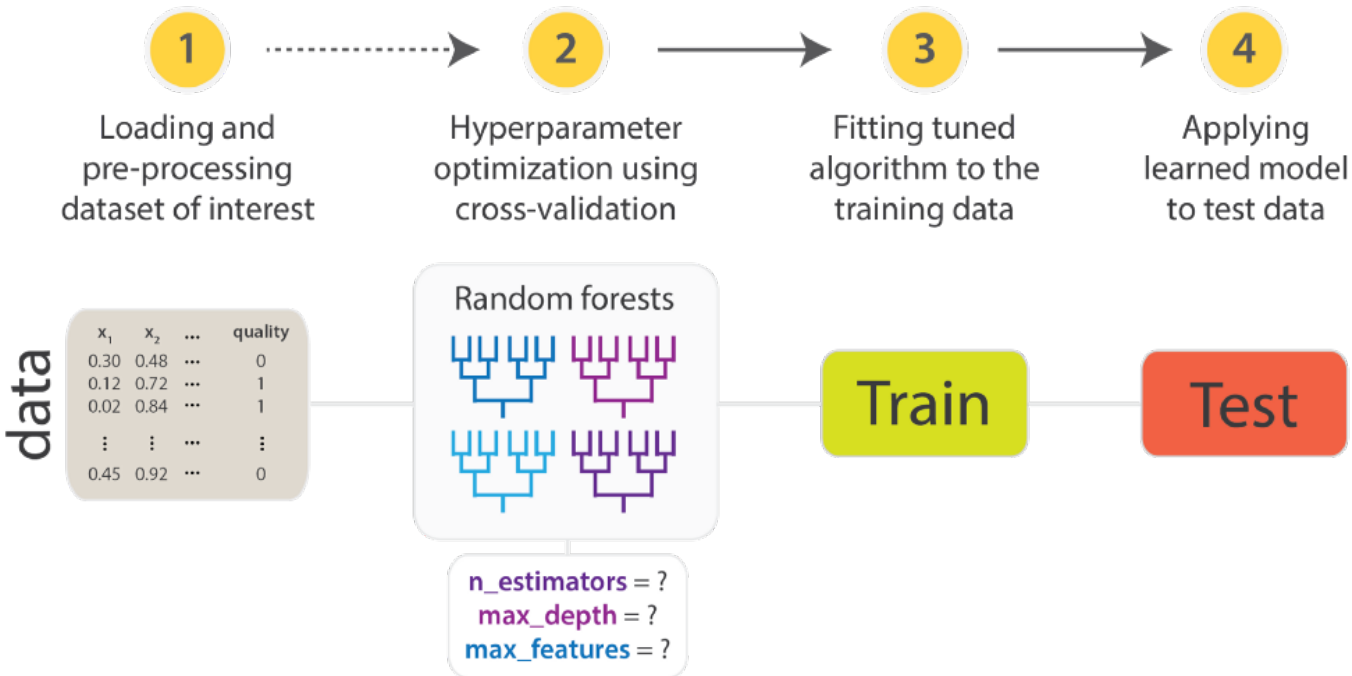
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods





Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

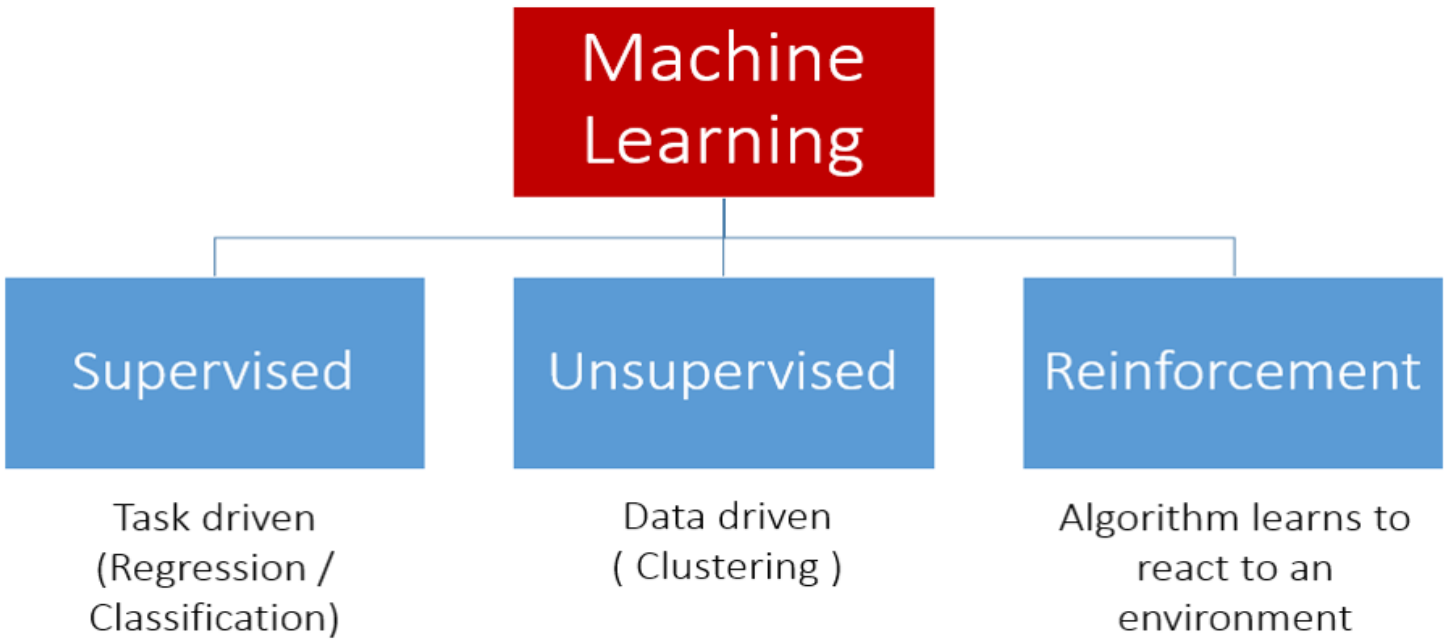
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

The Types of Machine Learning Algorithms



Supervised Learning / Predictive models:

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

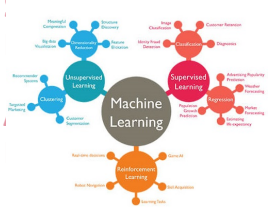
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



In a supervised learning (Predictive model), an algorithm learns from **labeled training data**, where each example in the training set consists of **input features and their corresponding output labels**. The goal is for the algorithm to generalize its learning to make accurate predictions or classifications on new, unseen data.

Supervised Learning / Predictive models:

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

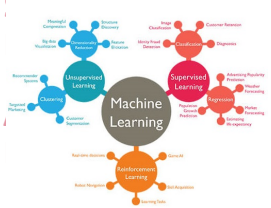
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



In a supervised learning (Predictive model), an algorithm learns from **labeled training data**, where each example in the training set consists of **input features and their corresponding output labels**. The goal is for the algorithm to generalize its learning to make accurate predictions or classifications on new, unseen data.

Supervised Learning Algorithm consist of

- Input: Labeled training data, where each example has known input features and the correct output.

Supervised Learning / Predictive models:

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

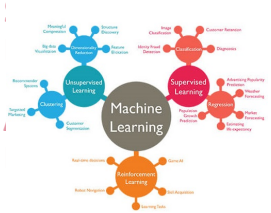
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



In a supervised learning (Predictive model), an algorithm learns from **labeled training data**, where each example in the training set consists of **input features and their corresponding output labels**. The goal is for the algorithm to generalize its learning to make accurate predictions or classifications on new, unseen data.
Supervised Learning Algorithm consist of

- Input: Labeled training data, where each example has known input features and the correct output.
- Process: The algorithm learns to map input features to output labels based on the provided training data.

Supervised Learning / Predictive models:

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

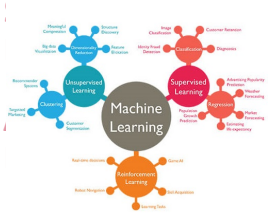
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



In a supervised learning (Predictive model), an algorithm learns from **labeled training data**, where each example in the training set consists of **input features and their corresponding output labels**. The goal is for the algorithm to generalize its learning to make accurate predictions or classifications on new, unseen data.

Supervised Learning Algorithm consist of

- Input: Labeled training data, where each example has known input features and the correct output.
- Process: The algorithm learns to map input features to output labels based on the provided training data.
- Output: The trained model can then make predictions or classifications on new, unseen data.

Unsupervised learning / Descriptive models

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

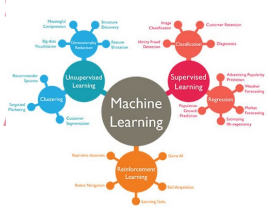
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Unsupervised learning / Descriptive models:

- Unsupervised learning is a machine learning paradigm where the algorithm is given data without explicit instructions on what to do with it.

Unsupervised learning / Descriptive models

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

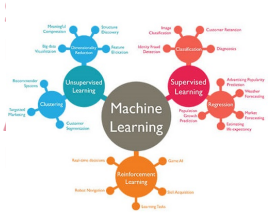
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Unsupervised learning / Descriptive models:

- Unsupervised learning is a machine learning paradigm where the algorithm is given data without explicit instructions on what to do with it.
- The system tries to find patterns, relationships, or structures in the data without labeled outputs.

Unsupervised learning / Descriptive models

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

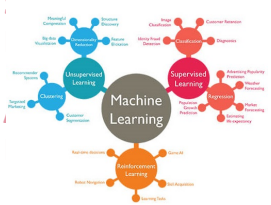
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Unsupervised learning / Descriptive models:

- Unsupervised learning is a machine learning paradigm where the algorithm is given data without explicit instructions on what to do with it.
- The system tries to find patterns, relationships, or structures in the data without labeled outputs.
- In other words, it explores the data on its own to identify hidden patterns or groupings, making it a form of self-discovery within the data set.

Unsupervised learning / Descriptive models

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

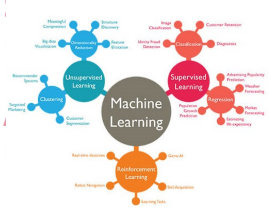
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Unsupervised learning / Descriptive models:

- Unsupervised learning is a machine learning paradigm where the algorithm is given data without explicit instructions on what to do with it.
- The system tries to find patterns, relationships, or structures in the data without labeled outputs.
- In other words, it explores the data on its own to identify hidden patterns or groupings, making it a form of self-discovery within the data set.
- Clustering and dimensionality reduction are common tasks in unsupervised learning.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Reinforcement learning (RL): Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or punishments based on the actions it takes. The goal of reinforcement learning is for the agent to learn the optimal sequence of actions that maximizes cumulative rewards over time. It involves a continuous learning process through trial and error, with the agent adjusting its behavior based on the consequences of its actions in the environment.

Model Selection for a Supervised Learning Algorithem

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

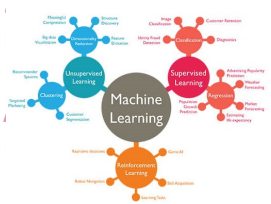
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



It is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set. This way, we can evaluate how well the model will generalize to a new set of data that were not seen during the training phase.

However, if we were just to perform a train/test split, the results can depend on a particular random choice for the pair of train/test sets.

One solution to this problem is a procedure called cross-validation.

Cross Validation

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

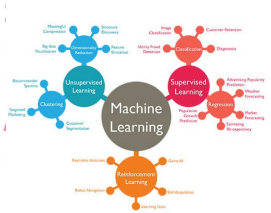
Discrimination Analysis

K-Nearest Neighborhood (KNN)

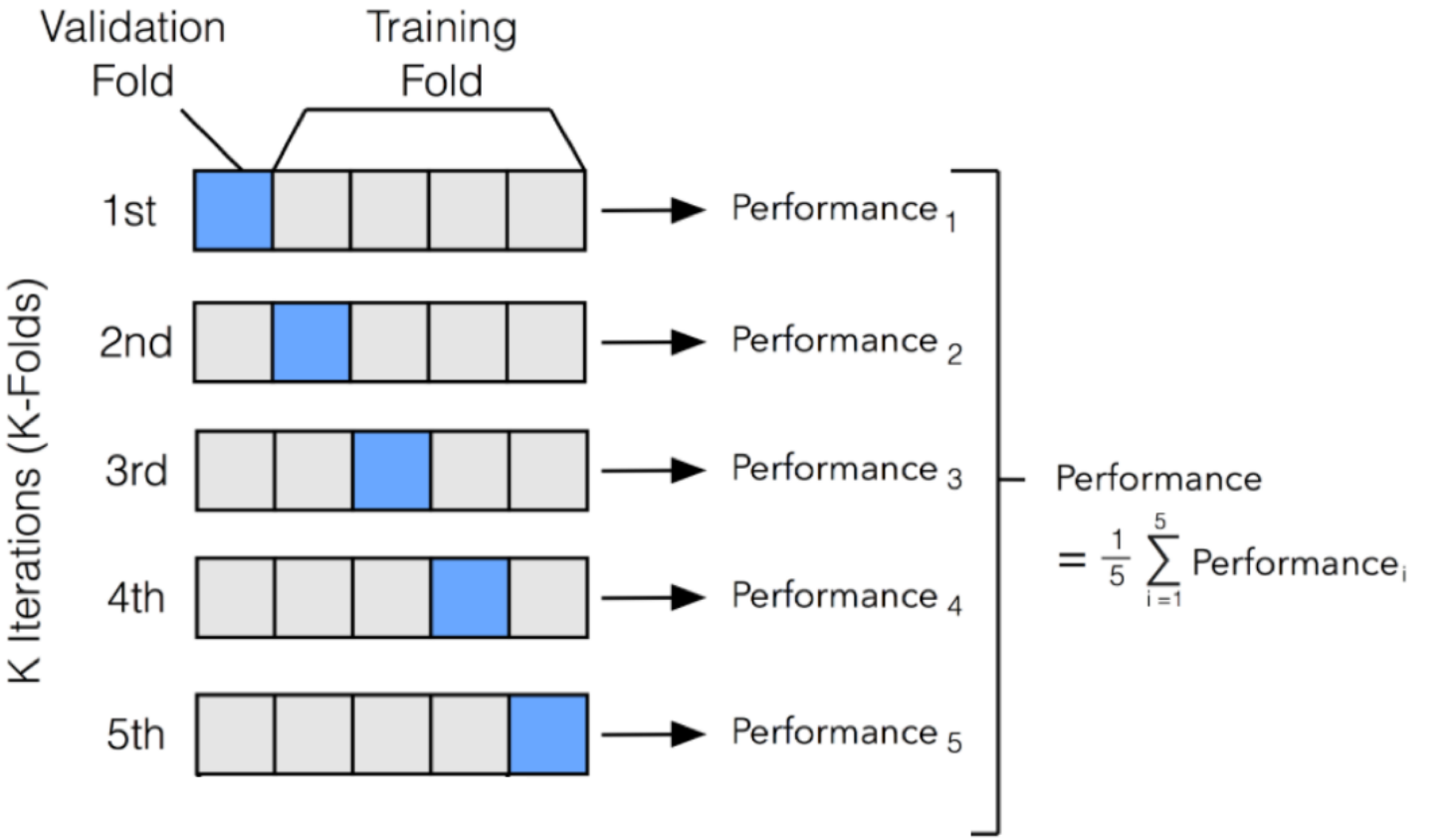
Decision Trees

Random Forests

Boosting Methods



- K-Fold Cross Validation



Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

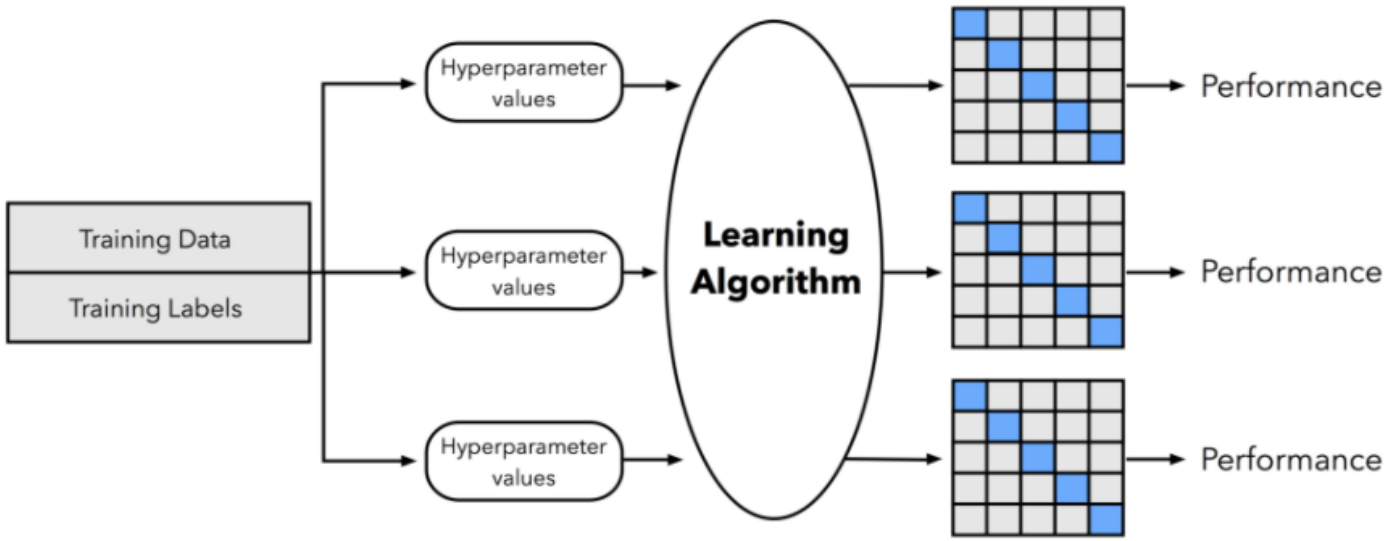
Random Forests

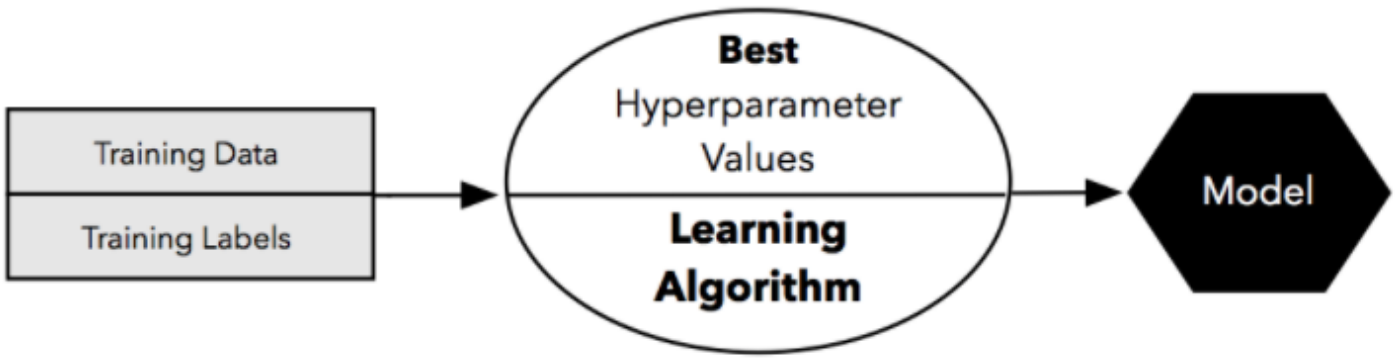
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data





Model Evaluation for a Binary Classification

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

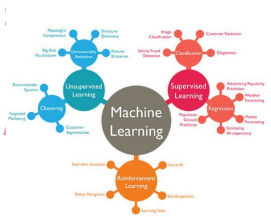
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



1. AUC (area under the curve) and ROC (receiver operating characteristics)
2. Classification Accuracy (Multiple Categories)
3. Sensitivity And Specificity
4. Positive Predictive Value
5. Negative Predictive Value

Predicted \ Actual		
	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precission = TP/(TP+FP)

Recall = TP/(TP+FN)

True Positive Rate = TP/(TP+FN)

False Positive Rate = FP/(FP+TN)

Supervised Learning Algorithms

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

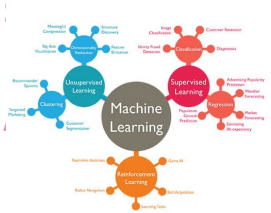
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Some of the Most Common Supervised Learning Algorithms

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Logistic Regression

Introduction to Logistic Regression

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

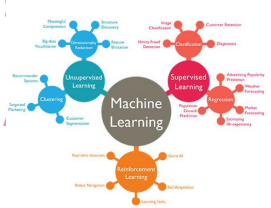
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Definition:** Logistic Regression is a statistical method used for binary classification.
- **Key Features:**
  - Outcome variable: Binary (0/1 or Yes/No)
  - Predictors: Continuous or categorical variables
  - Model: Uses the logistic function to map linear combinations of predictors to a probability between 0 and 1.
- **Formula:** Logistic function:

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n)}}$$

Examples and Applications

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

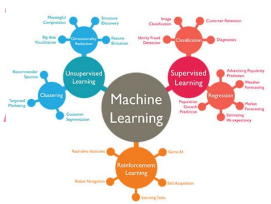
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Example: Predicting Email Spam**
  - Binary outcome: Spam (1) or Not Spam (0)
  - Predictors: Email content features, sender, etc.
- **Model:** Logistic Regression estimates the probability of an email being spam based on features.
- **Interpretation:** If predicted probability $>$ threshold (e.g., 0.5), classify as spam; otherwise, classify as not spam.
- **Performance Evaluation:** Common metrics include accuracy, Sensitivity, Specificity, precision, recall, calibration, and area under the ROC curve.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

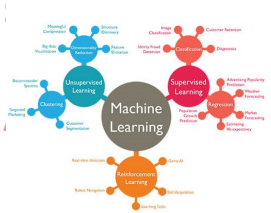
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- Logistic regression provides a good method for classification by modeling the probability of membership of a class with transforms of linear combinations of explanatory variables.
- Classical logistic regression does not work for micro-arrays because there are far more variables than observations.
- Particular problems are multi-collinearity and over fitting
- **A solution**: use penalized logistic regression.

Ridge

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

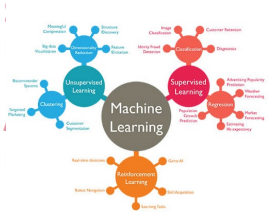
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Regularization in order to avoid over-fitting

One way to ensure shrinkage is to add the penalty term, $\lambda \sum \beta_j^2$, to the likelihood function. This penalty term is also known as the L2 norm or L2 penalty.

This term will help shrink the coefficients in the regression towards zero. The new loss function is as follows, where j is the number of parameters/coefficients in the model and L_{log} is the log likelihood function.

$$L_{log} + \lambda \sum \beta_j^2$$

This penalized likelihood function is called "ridge regression" (Hoerl and Kennard 1970).

Lasso

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

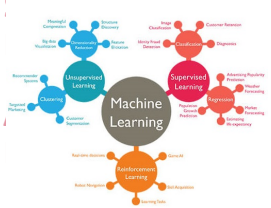
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



A small modification to the penalty is to use the absolute values of β_j instead of squared values. This penalty is called the "L1 norm" or "L1 penalty". The regression method that uses the L1 penalty is known as "Lasso regression" (Tibshirani 1996).

$$L_{log} + \lambda \sum |\beta_j|$$

Elastic Net

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

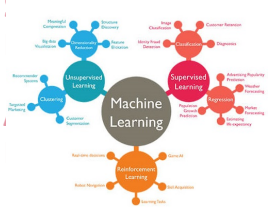
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



More recently, a method called "elastic net" was proposed to include the best of both worlds (Zou and Hastie 2005). This method uses both L1 and L2 penalties.
The equation below shows the modified likelihood function by this penalty.
As you can see the λ parameter still controls the weight that is given to the penalty.
This time the additional parameter α controls the weight given to L1 or L2 penalty and it is a value between 0 and 1.

$$L_{log} + \lambda \sum (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|)$$

Schematically representation of Lasso and Ridge

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

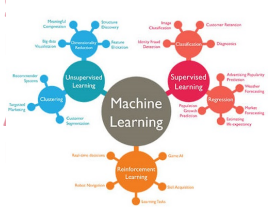
Discrimination Analysis

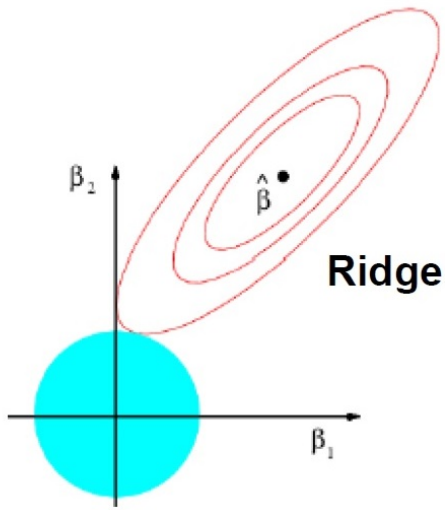
K-Nearest Neighborhood (KNN)

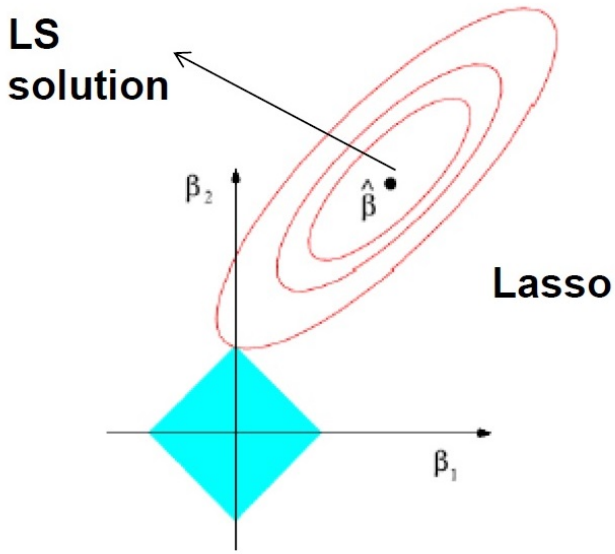
Decision Trees

Random Forests

Boosting Methods







Lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region

Comparison between Ridge and Lasso

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

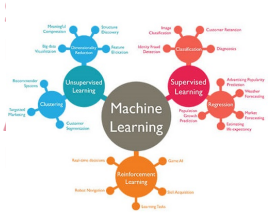
Discrimination Analysis

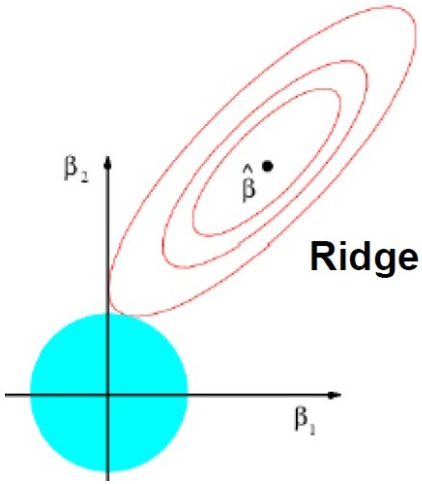
K-Nearest Neighborhood (KNN)

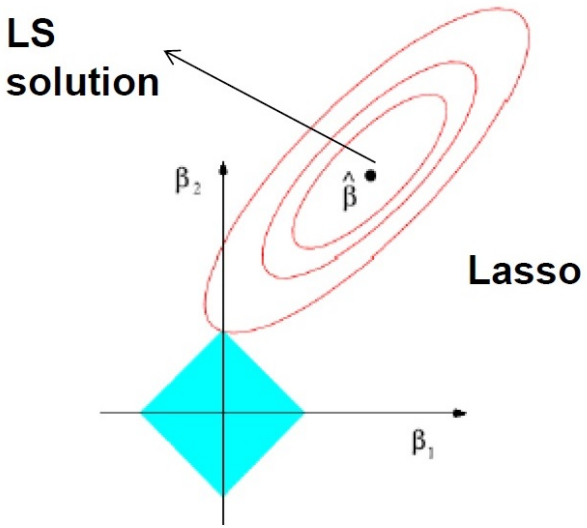
Decision Trees

Random Forests

Boosting Methods







Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero.

the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero

Comparison between Ridge and Lasso

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

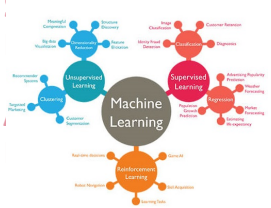
Discrimination Analysis

K-Nearest Neighborhood (KNN)

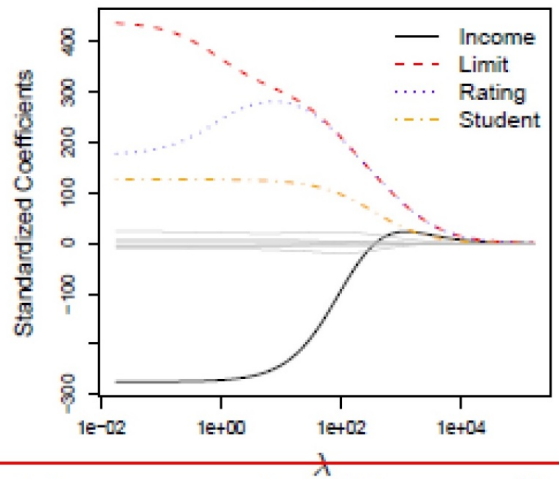
Decision Trees

Random Forests

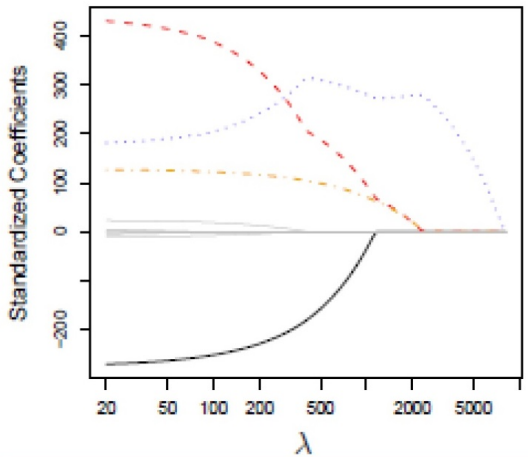
Boosting Methods



Ridge



Lasso



As λ increases, the ridge coefficient estimates shrink towards zero.
When λ is extremely large, then all of the ridge coefficient estimates are basically zero;

the ℓ1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Lecture 2

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

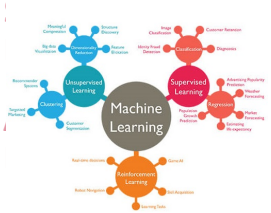
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

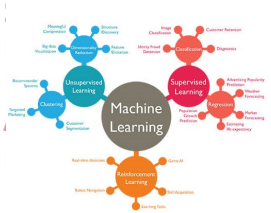
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.
- **Multiple Variables:** Involves multiple independent variables (features) to predict the group to which an observation belongs.

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

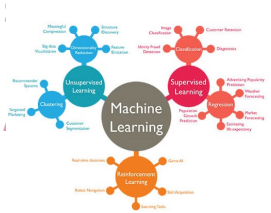
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.
- **Multiple Variables:** Involves multiple independent variables (features) to predict the group to which an observation belongs.
- **Assumptions:**
  - Assumes data follows a multivariate normal distribution.

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

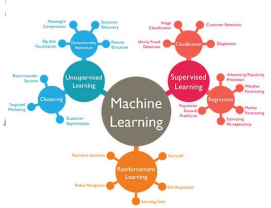
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.
- **Multiple Variables:** Involves multiple independent variables (features) to predict the group to which an observation belongs.
- **Assumptions:**
  - Assumes data follows a multivariate normal distribution.
  - Assumes variance-covariance matrices are equal across groups.

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

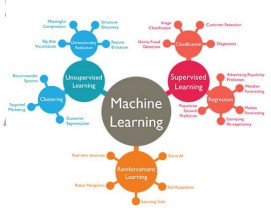
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.
- **Multiple Variables:** Involves multiple independent variables (features) to predict the group to which an observation belongs.
- **Assumptions:**
  - Assumes data follows a multivariate normal distribution.
  - Assumes variance-covariance matrices are equal across groups.
- **Output:** Includes discriminant functions, linear combinations of input variables used for classification.

Discriminant Analysis

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

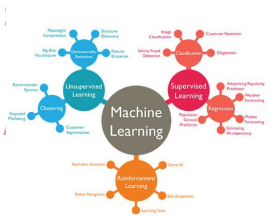
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Objective:** Discriminant Analysis aims to determine the variables that best discriminate between different groups.
- **Multiple Variables:** Involves multiple independent variables (features) to predict the group to which an observation belongs.
- **Assumptions:**
  - Assumes data follows a multivariate normal distribution.
  - Assumes variance-covariance matrices are equal across groups.
- **Output:** Includes discriminant functions, linear combinations of input variables used for classification.
- **Types:** Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA).

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

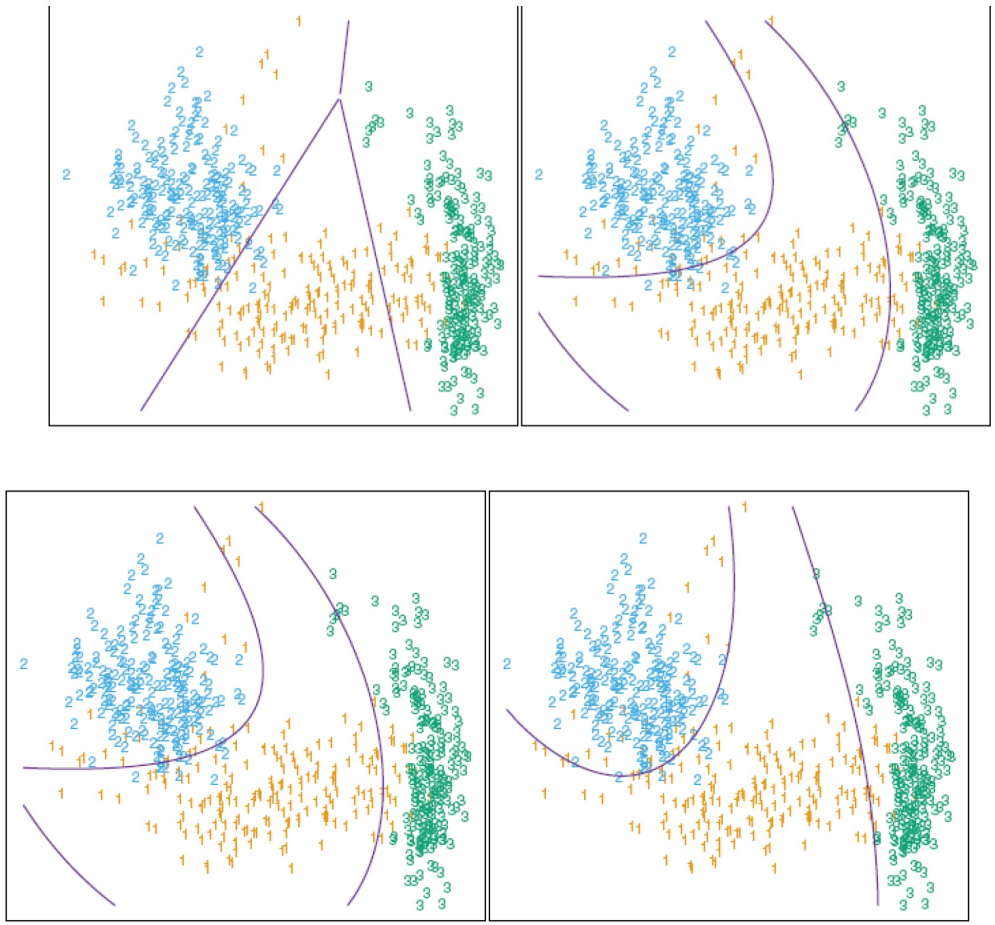
Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data



K-Nearest Neighborhood (KNN)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

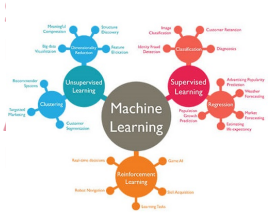
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **K-Nearest Neighbors (KNN):**
  - **What is it?** K-Nearest Neighbors is a simple but effective algorithm used to classify or predict things based on what similar things are like.

K-Nearest Neighborhood (KNN)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

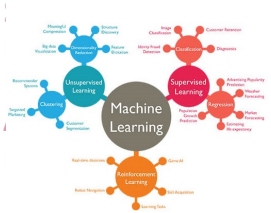
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **K-Nearest Neighbors (KNN):**
  - **What is it?** K-Nearest Neighbors is a simple but effective algorithm used to classify or predict things based on what similar things are like.
  - It works like asking your neighbors for advice – if most of your neighbors have similar preferences, chances are you'll like what they like.

K-Nearest Neighborhood (KNN)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

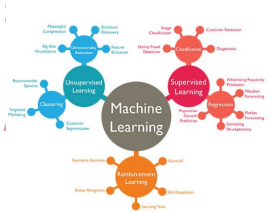
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **K-Nearest Neighbors (KNN):**
  - **What is it?** K-Nearest Neighbors is a simple but effective algorithm used to classify or predict things based on what similar things are like.
  - It works like asking your neighbors for advice – if most of your neighbors have similar preferences, chances are you'll like what they like.
  - **How does it work?** It decides based on similarities to nearby examples.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

* Advantages The algorithm is simple and easy to implement. There's no need to build a model, tune several parameters, or make additional assumptions. The algorithm is **versatile**. It can be used for classification, regression.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

* Advantages The algorithm is simple and easy to implement. There's no need to build a model, tune several parameters, or make additional assumptions. The algorithm is **versatile**. It can be used for classification, regression.
* Disadvantages The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

KNN Algorithm Example

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

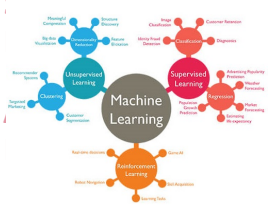
Discrimination Analysis

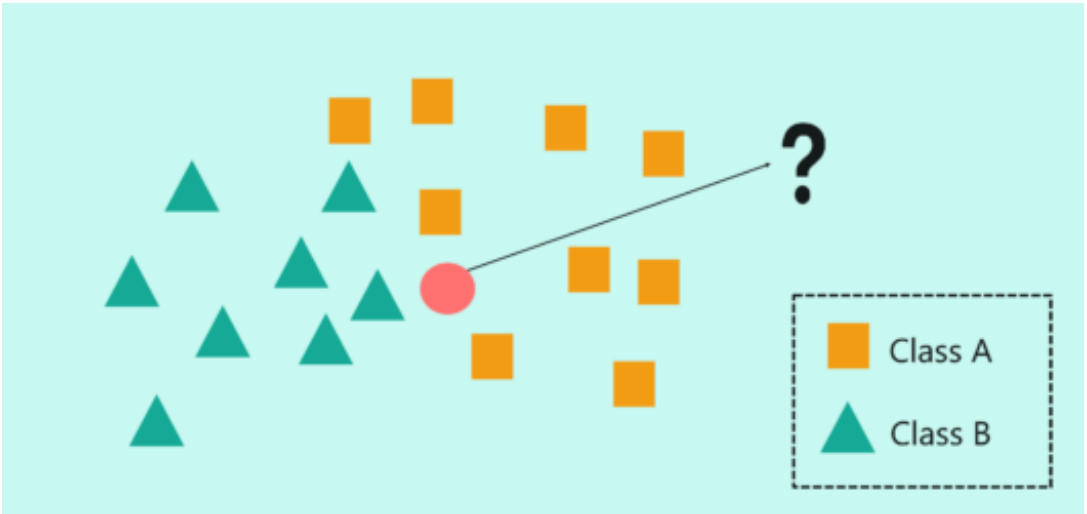
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

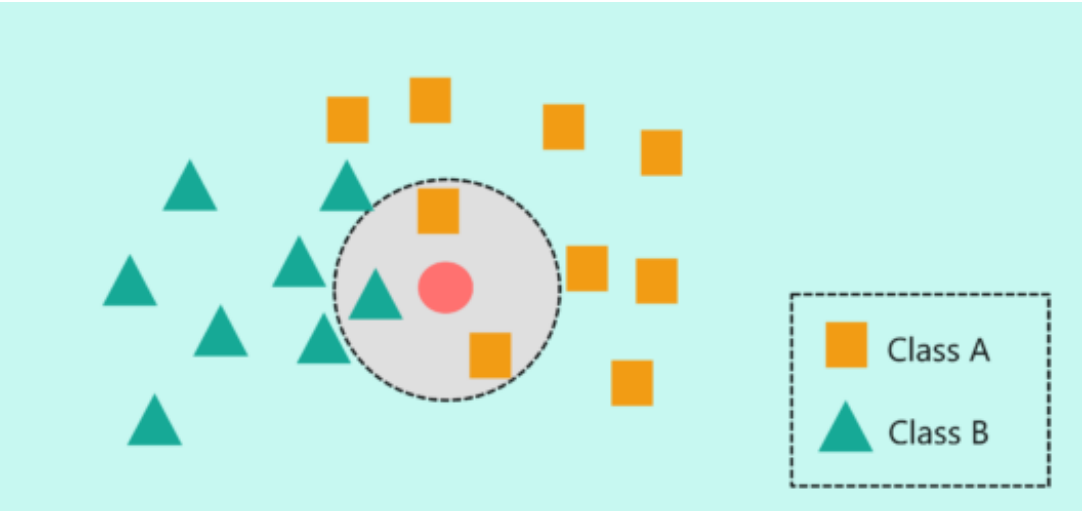
Boosting Methods





- In the above image, we have two classes of data, namely class A (squares) and Class B (triangles)
- The problem statement is to assign the new input data point to one of the two classes by using the KNN algorithm
- The first step in the KNN algorithm is to define the value of 'K'. But what does the 'K' in the KNN algorithm stand for?
- 'K' stands for the number of Nearest Neighbors and hence the name K Nearest Neighbors (KNN).

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods
Support Vector Machine (SVM)
Artificial Neural Networks
Missing Data



- In the above image, I've defined the value of 'K' as 3. This means that the algorithm will consider the three neighbors that are the closest to the new data point in order to decide the class of this new data point.
- The closeness between the data points is calculated by using measures such as Euclidean and Manhattan distance.
- At 'K' = 3, the neighbors include two squares and 1 triangle. So, if I were to classify the new data point based on 'K' = 3, then it would be assigned to Class A (squares).

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

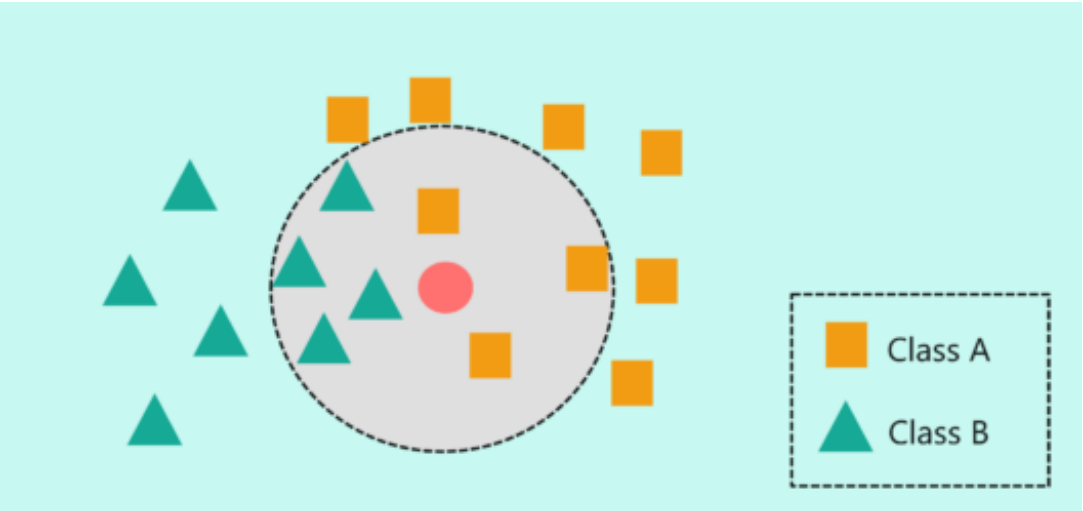
Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data



- But what if the 'K' value is set to 7? Here, I'm basically telling my algorithm to look for the seven nearest neighbors and classify the new data point into the class it is most similar to.
- At 'K' = 7, the neighbors include three squares and four triangles. So, if I were to classify the new data point based on 'K' = 7, then it would be assigned to Class B (triangles) since the majority of its neighbors were of class B.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

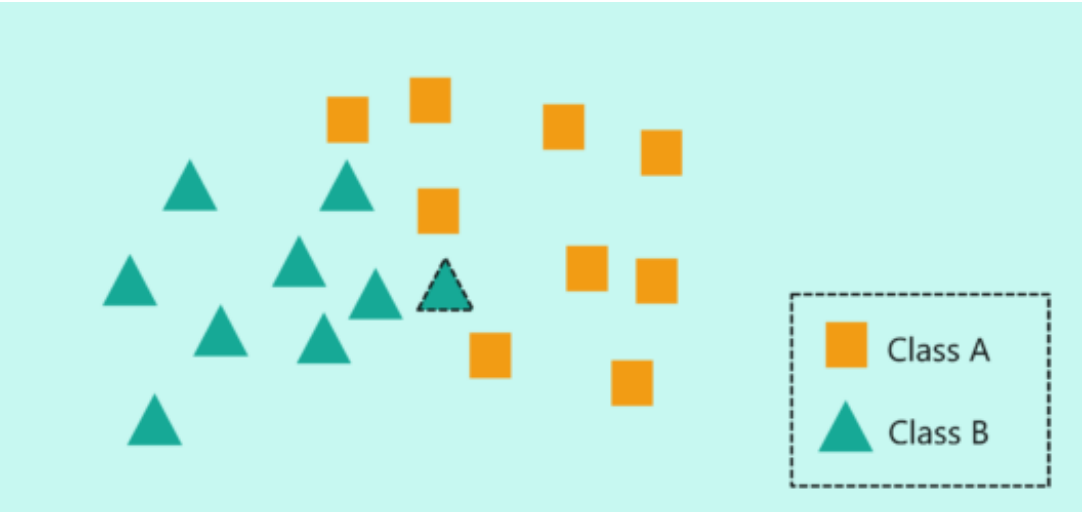
Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data



It is as simple as that! KNN makes use of simple measure in order to solve complex problems, this is one of the reasons why KNN is such a commonly used algorithm.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

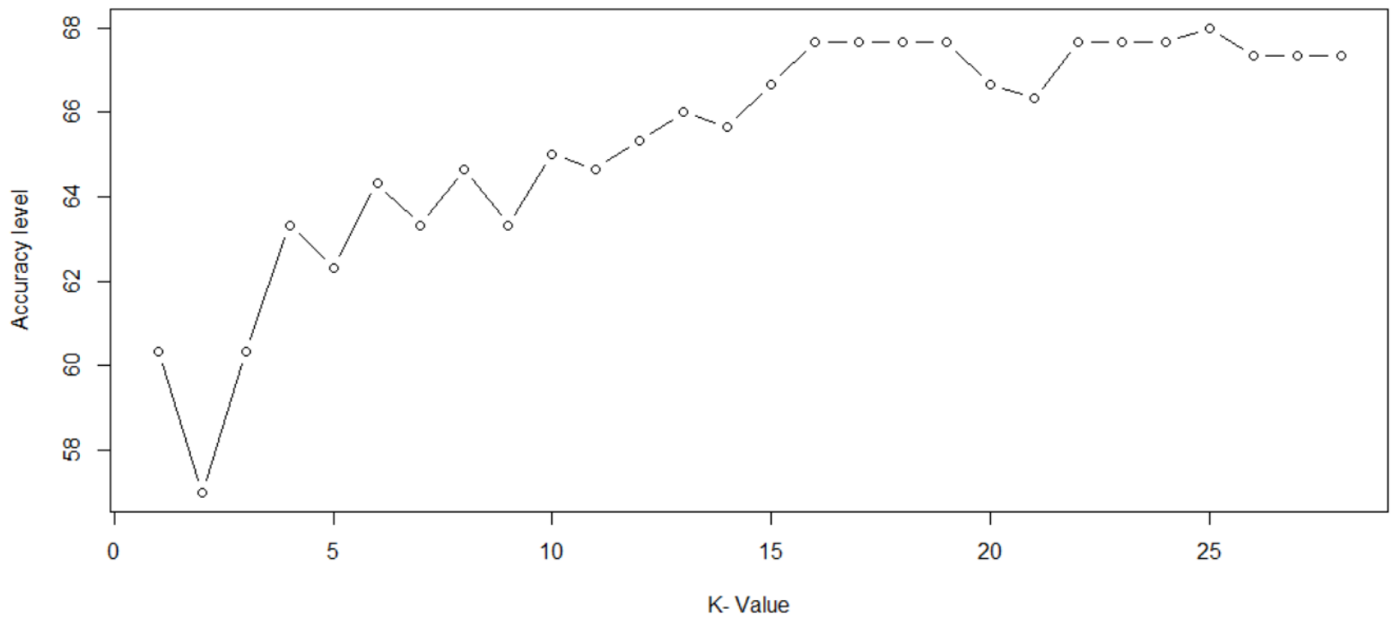
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Optimization: In order to improve the accuracy of the model, we can use n number of techniques such as the below method and maximum percentage accuracy graph. This way we can check which 'K' value will result in the most accurate model.



The above graph shows that for 'K' value of 25 we get the maximum accuracy.

Decision Trees

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

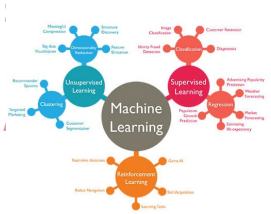
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Tree-based methods for classification and regression

Decision Trees

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

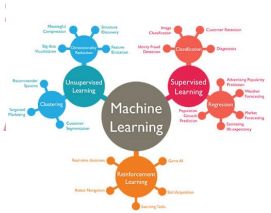
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Decision Trees:**
  - **What are they?** A graphical model that helps make decisions based on conditions.
  - **How do they work?** They split data into subsets based on features to make decisions in a step-by-step manner.

Example 1: Predicting Plant Species

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

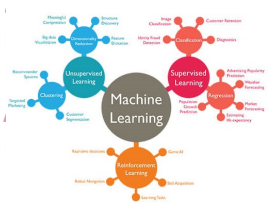
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Scenario:** Scientists want to predict the species of a plant in a new ecosystem.
- **Decision Tree Approach:** They create a decision tree based on features like leaf shape, size, and flower color. By following the branches of the tree, they can classify the plant into a specific species.

Example 2: Identifying Animal Species

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

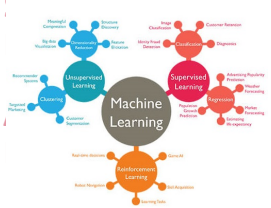
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Scenario:** Researchers aim to identify an unknown animal species based on observable traits.
- **Decision Tree Approach:** They construct a decision tree using features like fur color, size, and tail length. Following the branches of the tree, they can determine the likely species of the animal.

Example: simple classification tree

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

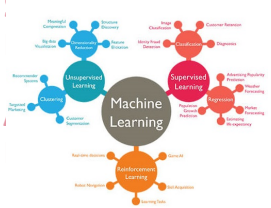
Discrimination Analysis

K-Nearest Neighborhood (KNN)

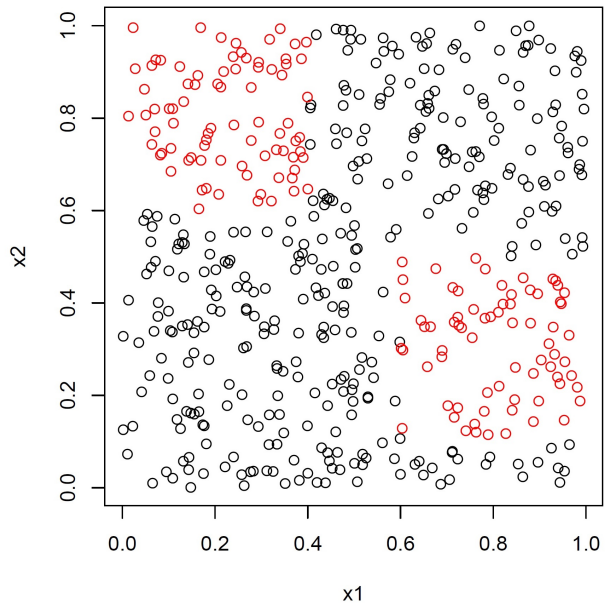
Decision Trees

Random Forests

Boosting Methods



Example: n = 500 points in p = 2 dimensions, falling into classes 0 and 1, as marked by colors



Does dividing up the feature space into rectangles look like it would work here?

Example: simple classification tree

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

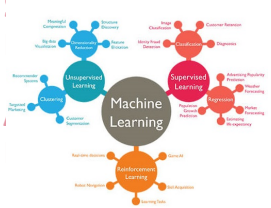
Discrimination Analysis

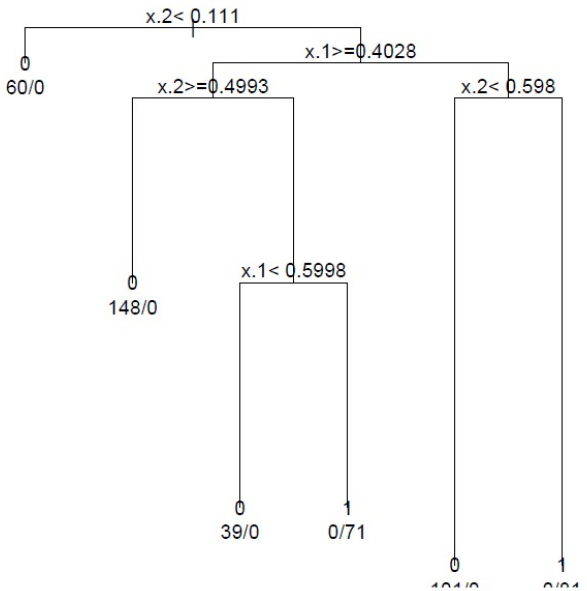
K-Nearest Neighborhood (KNN)

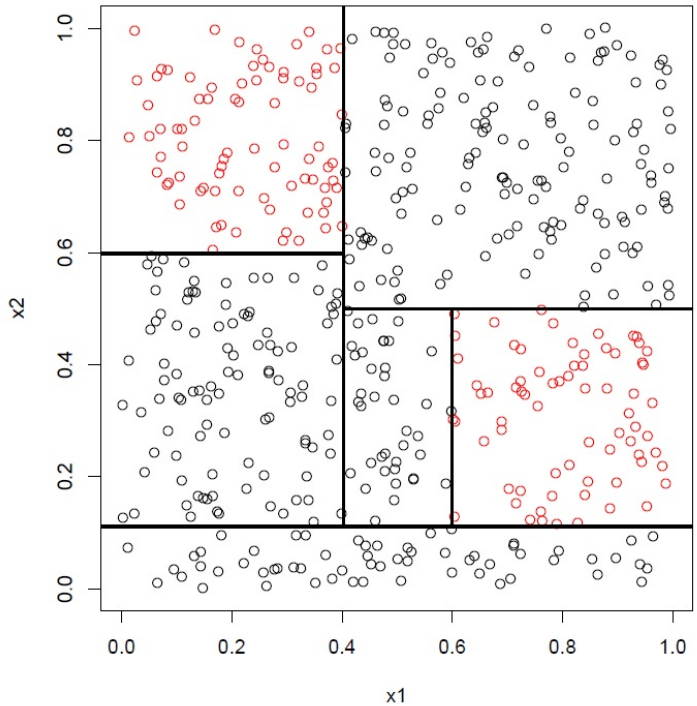
Decision Trees

Random Forests

Boosting Methods







Example: regions defined by a tree

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

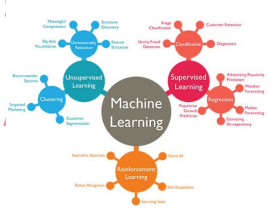
Discrimination Analysis

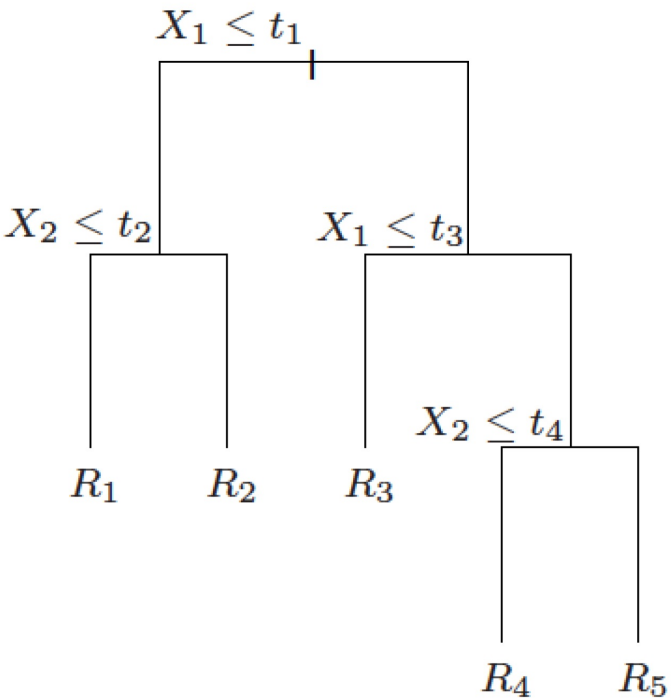
K-Nearest Neighborhood (KNN)

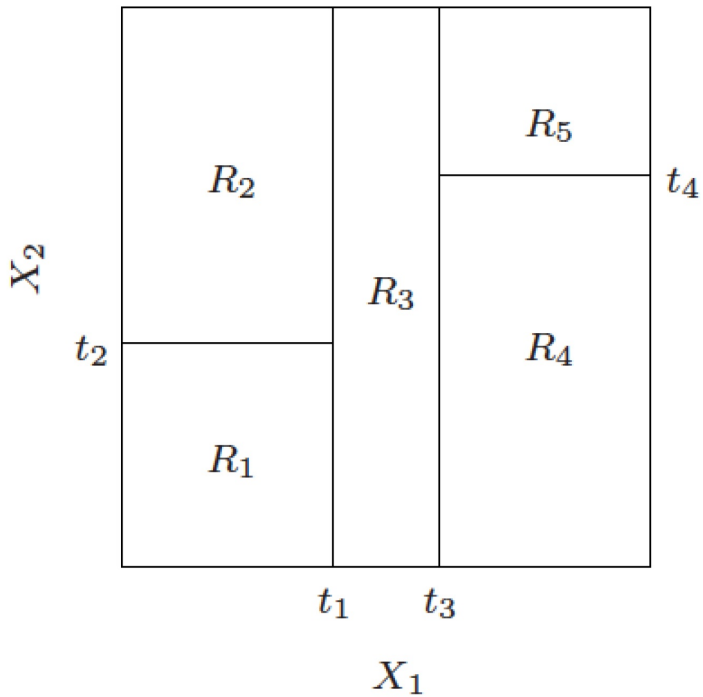
Decision Trees

Random Forests

Boosting Methods







How to build trees?

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

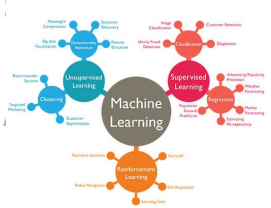
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



There are two main issues to consider in building a tree:

1. How to choose the splits?
2. How big to grow the tree?

Think first about varying the depth of tree ... which is more complex, a big tree or a small tree? What trade off is at play here?

- A very large tree might over fit the data
- A small tree might not capture the important structure

- Tree size is a tuning parameter governing the model's complexity
- The optimal tree size should be adaptively chosen from the data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

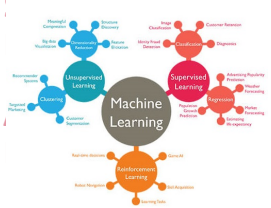
Artificial Neural Networks

Missing Data

Random Forests

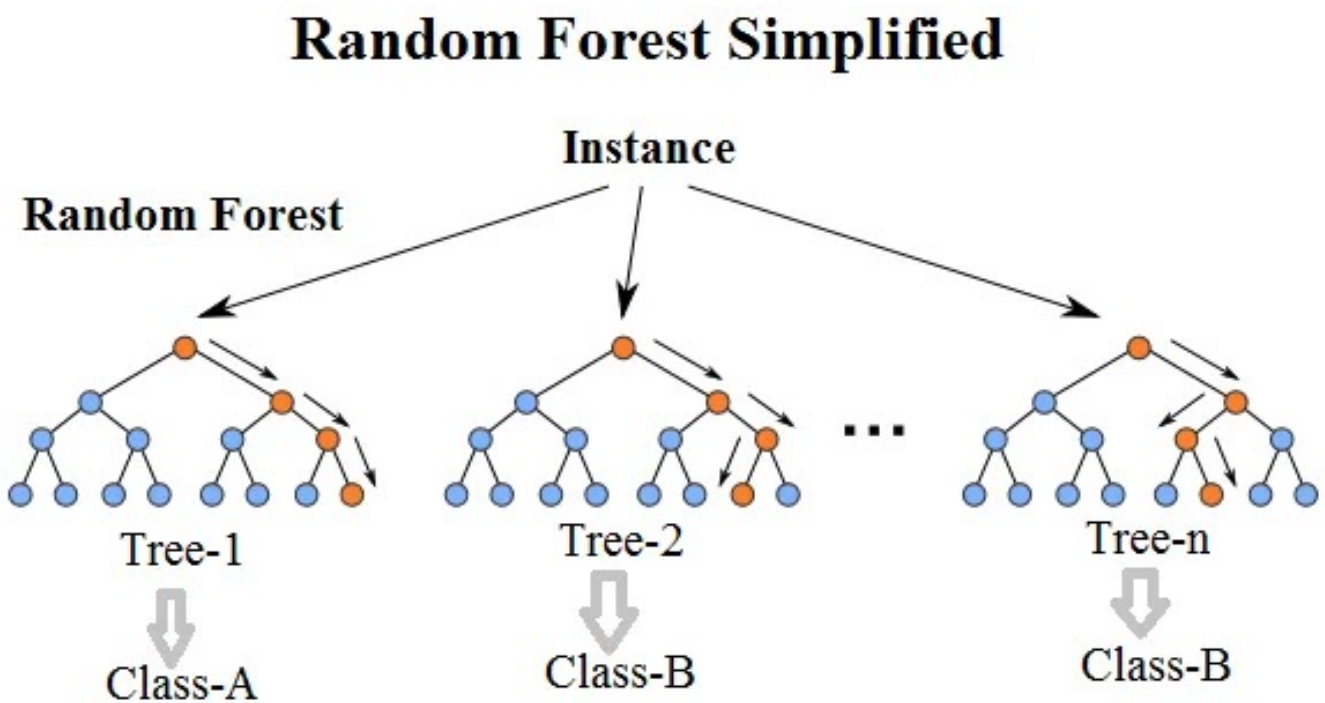
The Random Forest Classifier

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.

Put simply: random forest builds multiple decision trees and merges them together to get a **more accurate** and **stable prediction**.



Random Forest Algorithm

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

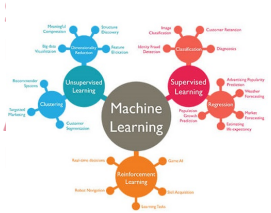
Discrimination Analysis

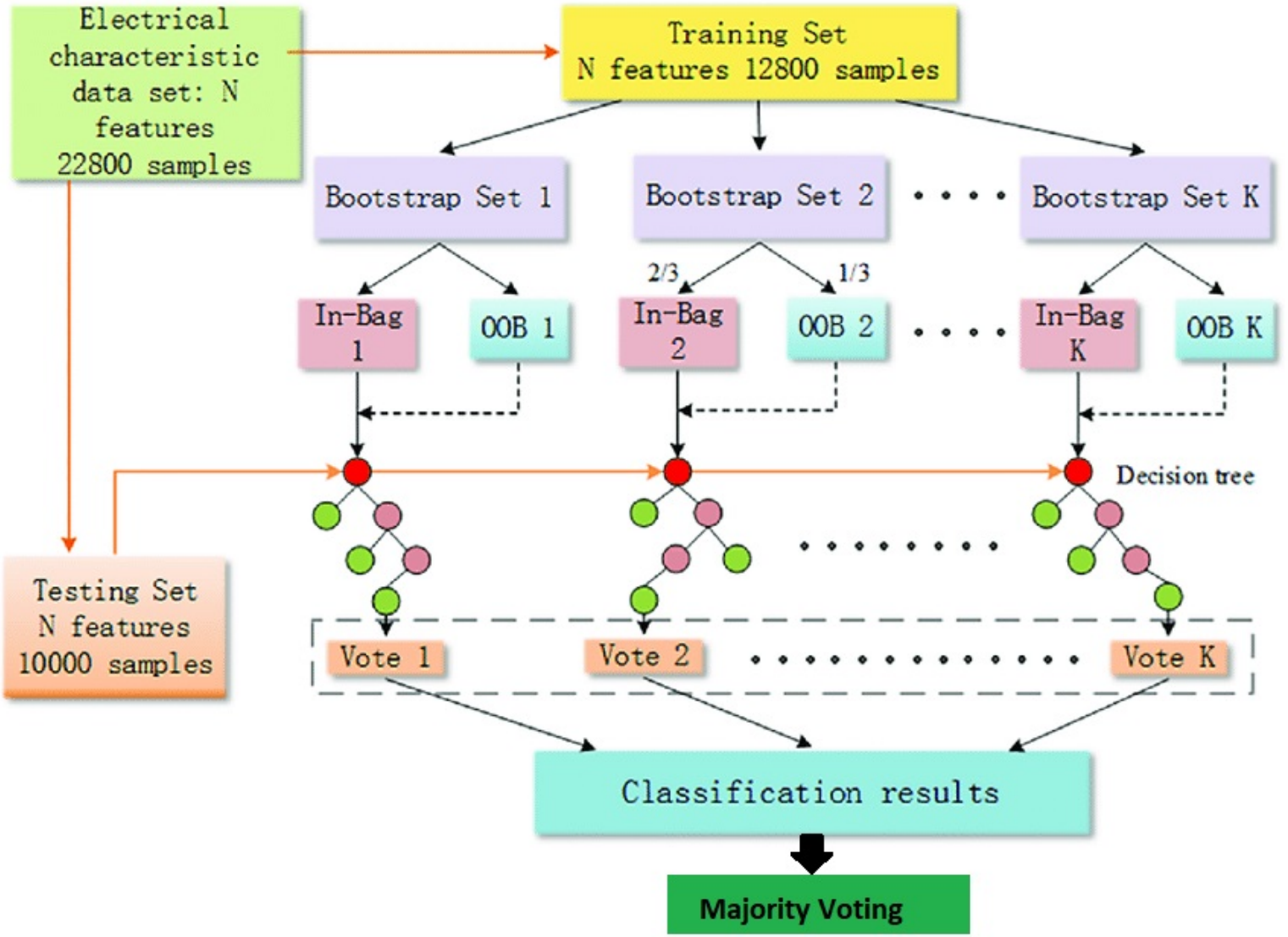
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods





Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Boosting Methods

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

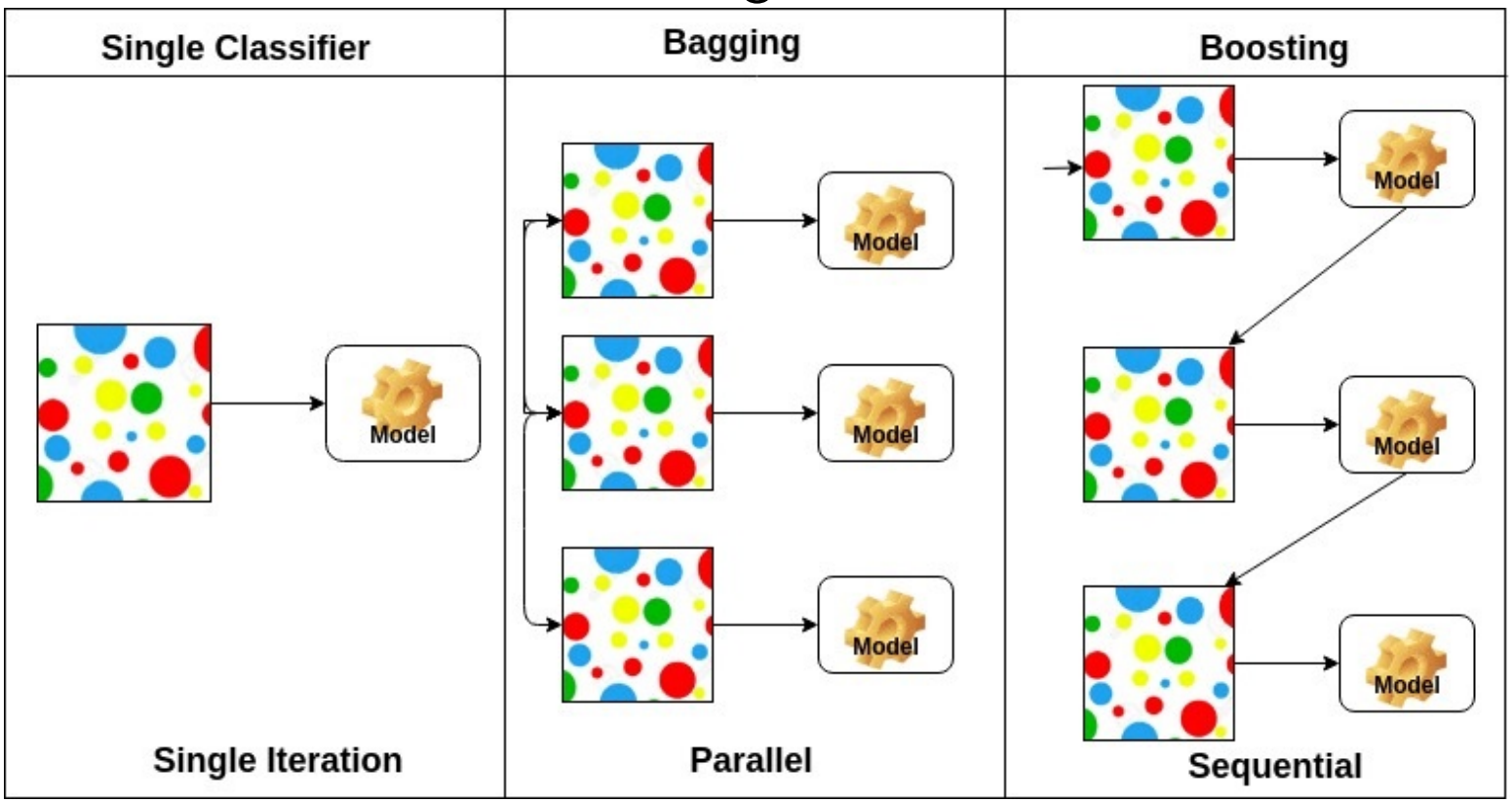
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

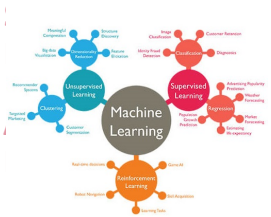
Missing Data

The term **'Boosting'** refers to a family of algorithms which converts weak learner to strong learners.



Types of Boosting Algorithms

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



Bagging: It is an approach where you take random samples of data, build learning algorithms and take simple means to find bagging probabilities.

Boosting: Boosting is similar, however the selection of sample is made more intelligently. We subsequently give more and more weight to hard to classify observations.

Types of Boosting Algorithms

- AdaBoost (Adaptive Boosting)
- Gradient Tree Boosting
- XGBoost

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Support Vector Machine (SVM)

Support Vector Machine (SVM)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

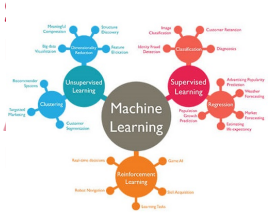
Discrimination Analysis

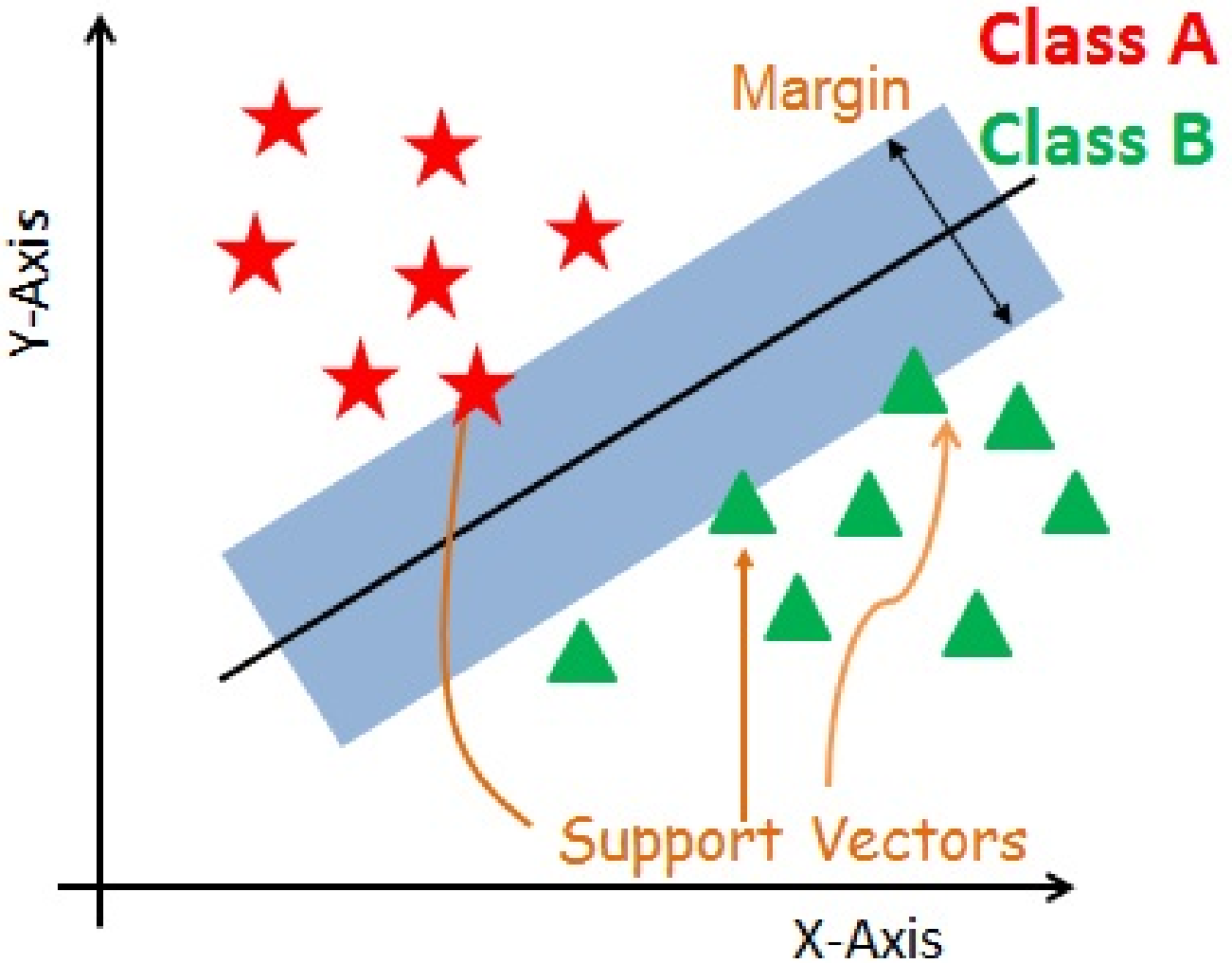
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods





An Introduction

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

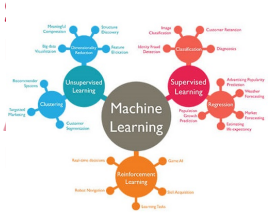
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **What is SVM?**
  - A supervised machine learning algorithm used for classification and regression tasks.
  - It works by finding the optimal hyperplane that separates different classes or predicts numerical values.
- **Key Concepts:**
  - **Hyperplane:** Decision boundary that maximally separates classes.
  - **Support Vectors:** Data points crucial for determining the hyperplane.

Disease Prediction in Livestock

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

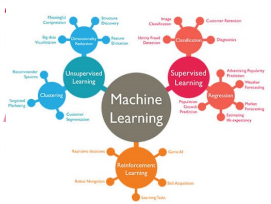
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Scenario:**
  - Predicting the likelihood of a disease (e.g., mastitis) in dairy cattle.
- **Data Input:**
  - Features like milk production, udder health, and environmental conditions.
- **SVM Use:**
  - SVM can effectively classify animals into healthy and diseased categories based on various features, aiding in early detection.

Biodiversity Classification

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

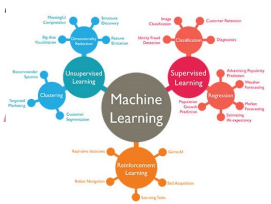
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Scenario:**
  - Assessing biodiversity in a specific ecosystem.
- **Data Input:**
  - Ecological features such as vegetation types, climate, and soil composition.
- **SVM Use:**
  - SVM can classify different species or biodiversity levels by creating a hyperplane that separates distinct ecological classes.

Forest Cover Classification

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

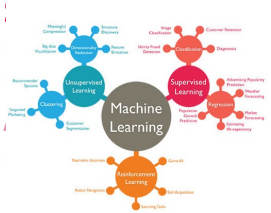
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Scenario:**
  - Classifying types of vegetation or determining forest cover in a region.
- **Data Input:**
  - Remote sensing data, including spectral information from satellite imagery.
- **SVM Use:**
  - SVM excels in image classification tasks, making it suitable for mapping and monitoring forest cover by creating a hyperplane to distinguish different vegetation types.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

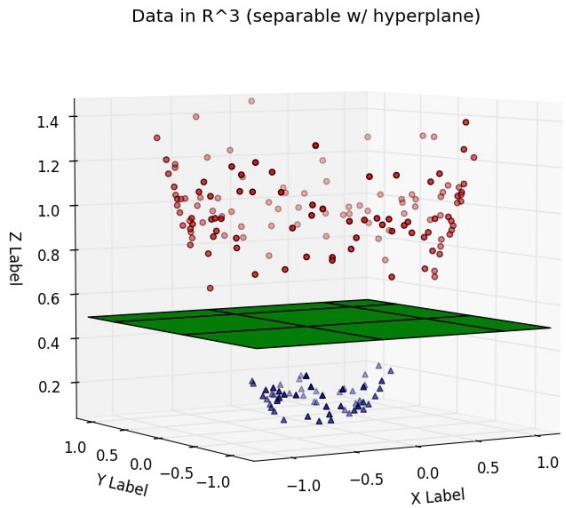
Random Forests

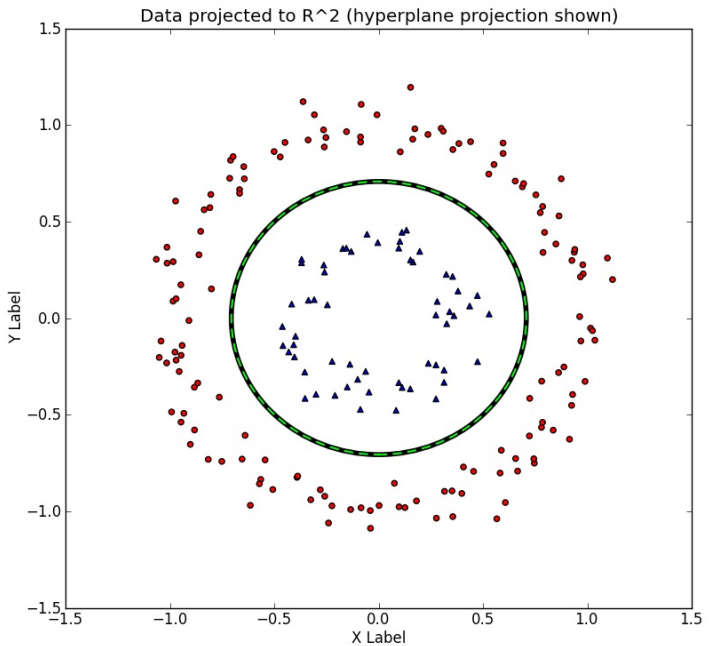
Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data





Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Lecture 3

Artificial Neural Networks (ANN)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

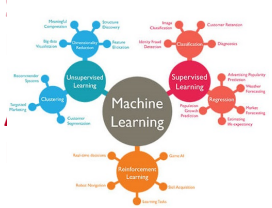
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **What is an ANN?**
  - A computational model inspired by the structure and function of the human brain.
  - Composed of interconnected nodes (neurons) organized in layers that process information.
- **Key Components:**
  - **Input Layer:** Receives input features.
  - **Hidden Layers:** Process information.
  - **Output Layer:** Produces the final result.

Simplest Artificial Neural Network

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

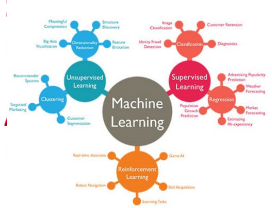
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



The simplest form of an Artificial Neural Network (ANN) consists of:

- An input layer ($\mathbf{X}$) with n neurons.
- A single neuron in the hidden layer with weights ($\mathbf{W}$) and bias (b).
- An output layer with a single neuron representing the predicted value ($\hat{y}$).

The output $\hat{y}$ is calculated using the sigmoid activation function:

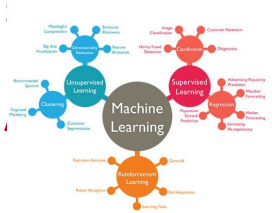
$$\hat{y} = \sigma(\mathbf{W}^T\mathbf{X} + b)$$

where $\sigma(z)$ is the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

One-Layer Artificial Neural Network (ANN)

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



The output a of a single-layer perceptron is calculated using the weighted sum z and the activation function σ:

$$z = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b$$

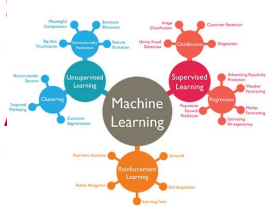
$$a = \sigma(z) = \frac{1}{1 + e^{-z}}$$

where:

- $x_1, x_2, \ldots, x_n$: Input features
- $w_1, w_2, \ldots, w_n$: Weights associated with each input feature
- b: Bias term
- z: Weighted sum of inputs plus bias
- σ: Activation function (e.g., sigmoid function)

Two-Layer Artificial Neural Network (ANN)

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



Mathematical Expression: The output a of a two-layer ANN with one hidden layer is calculated using the following steps:

1. Calculate the activations of the hidden layer:

$$z^{(1)} = W^{(1)}X + b^{(1)}$$

$$a^{(1)} = \sigma(z^{(1)})$$

2. Calculate the output layer activations:

$$z^{(2)} = W^{(2)}a^{(1)} + b^{(2)}$$

$$a^{(2)} = \sigma(z^{(2)})$$

where:

- X: Input features
- $W^{(1)}, W^{(2)}$: Weights matrices for the hidden and output layers, respectively
- $b^{(1)}, b^{(2)}$: Bias vectors for the hidden and output layers, respectively

Two-Layer Artificial Neural Network (ANN)

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

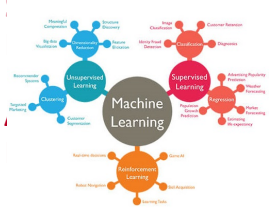
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- $z^{(1)}, z^{(2)}$: Weighted sums of inputs plus biases for the hidden and output layers, respectively
- $a^{(1)}, a^{(2)}$: Activations of the hidden and output layers, respectively
- σ: Activation function (e.g., sigmoid function)

Common Activation Functions in ANNs

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

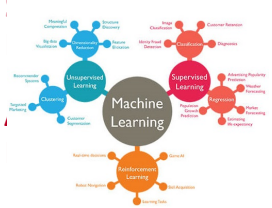
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- Sigmoid Function (Logistic):

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- Hyperbolic Tangent (Tanh):

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

- Rectified Linear Unit (ReLU):

$$\text{ReLU}(x) = \max(0, x)$$

- Leaky ReLU:

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{otherwise} \end{cases}$$

- Softmax Function:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Activation functions

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

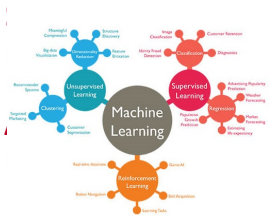
Discrimination Analysis

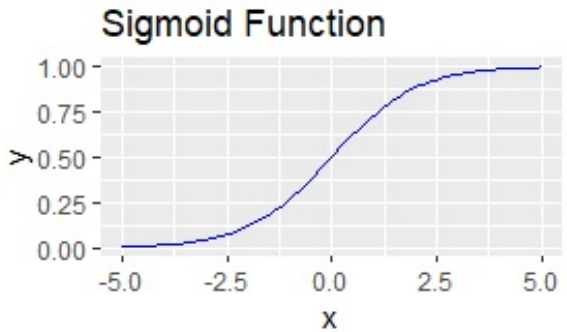
K-Nearest Neighborhood (KNN)

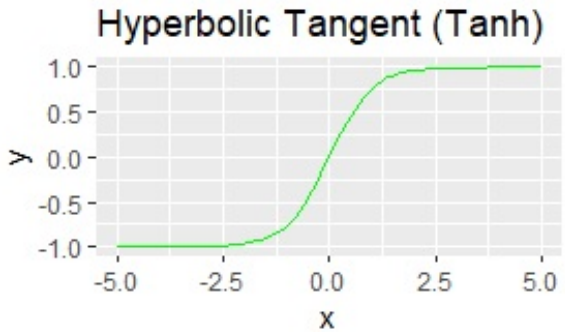
Decision Trees

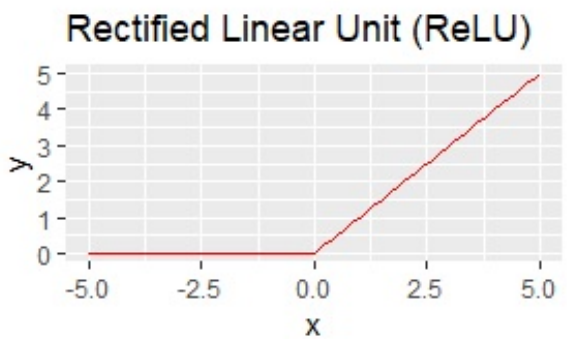
Random Forests

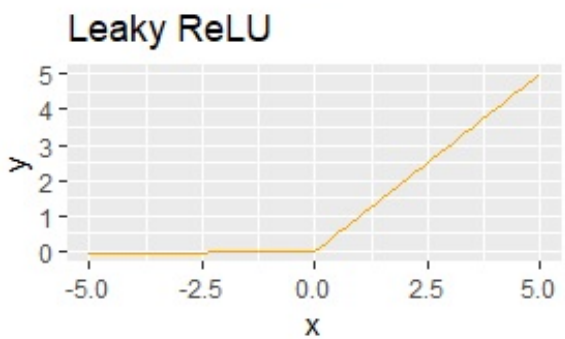
Boosting Methods











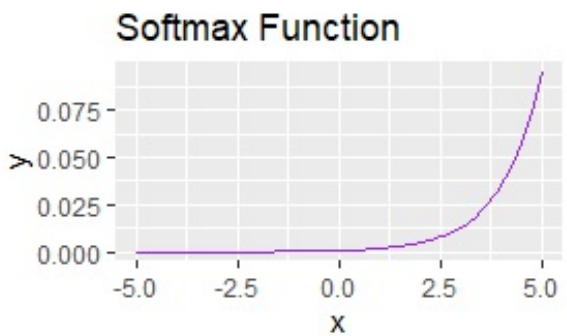


Figure: Activation functions

Activation Functions

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

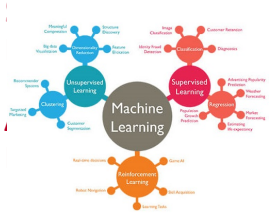
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- Sigmoid: Historically used in binary classification problems.
- Tanh: Similar to sigmoid but with a range between -1 and 1.
- ReLU: Becoming the most popular activation function (for spars problems).
- Leaky ReLU: Addresses the "dying ReLU" problem.
- Softmax: Primarily used in multi-class classification.

Choosing the Best Activation Function

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

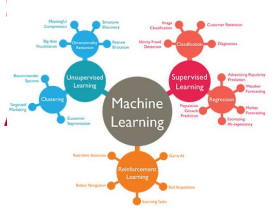
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



To choose the best activation function for your specific problem:

- Experiment with different functions.
- Evaluate their performance on a validation dataset.
- Consider factors such as network architecture, data nature, and computational efficiency.

Activation Functions

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

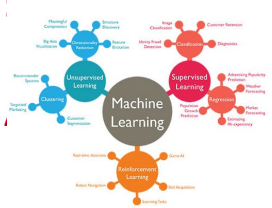
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



A good reference for understanding and selecting activation functions is the book "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville. In Chapter 6, they discuss different activation functions and provide insights into when to use each one based on empirical observations and theoretical considerations.

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data

Lecture 4

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

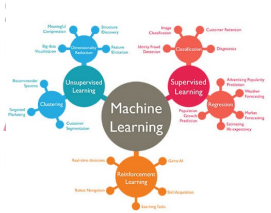
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Missing Data

Missing Data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

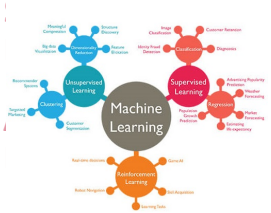
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Missing Data:** Occurs when no data is stored for certain observations within a variable.
- **Common Causes:** Survey non-response, equipment failure, data entry errors, etc.
- **Importance:** Can impact the quality and validity of analyses and results.

Types of Missing Data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

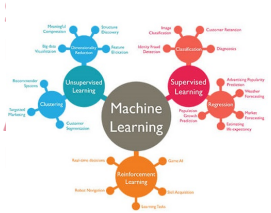
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Missing Completely at Random (MCAR):** No systematic pattern in the missing data.
- **Missing at Random (MAR):** The likelihood of missing data depends on observed data.
- **Missing Not at Random (MNAR):** The missing data pattern is related to the unobserved data.

Dealing with Missing Data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

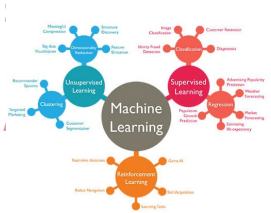
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



1. **Removal:** Remove rows or columns with missing data.
2. **Imputation:** Fill in missing values using various techniques.
3. **Model-Based Methods:** Utilize statistical models to predict missing values.
4. **Multiple Imputation:** Generate multiple datasets with different imputed values.

Impact of Missing Data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

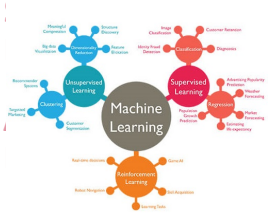
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Bias:** Missing data can introduce bias in statistical analyses.
- **Reduced Power:** Missing data may lead to reduced statistical power.
- **Invalid Conclusions:** Conclusions drawn from incomplete data may not be valid.
- **Misleading Results:** Inaccurate representations of relationships and patterns in the data.

Conclusion about Missing data

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

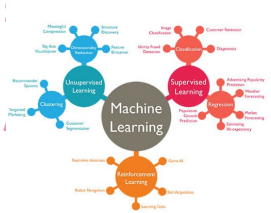
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Addressing missing data is crucial** for robust and reliable data analysis.
- **Choose appropriate methods** based on the nature and extent of missing data.
- **Transparent reporting:** Clearly document how missing data were handled in your analyses.

Imbalanced Classes

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

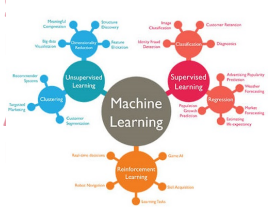
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Imbalanced Classes:** Occurs when one class in the target variable has significantly fewer instances than the others.
- **Common in Real-world Data:** Fraud detection, disease diagnosis, etc.
- **Challenges:** Traditional machine learning models may perform poorly on minority classes.

Impacts of Imbalanced Classes

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

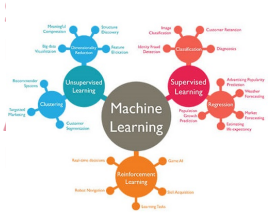
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

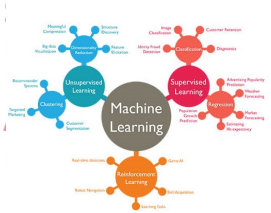
Boosting Methods



- **Bias in Model:** The model may be biased towards the majority class.
- **Poor Generalization:** The model may not generalize well to minority classes.
- **Misleading Accuracy:** High accuracy may not reflect actual model performance.

Treatments for Imbalanced Classes

Introduction to Machine Learning
Logistic Regression
Penalized Logistic Regression (LASSO, Ridge, Elastic Net)
Discrimination Analysis
K-Nearest Neighborhood (KNN)
Decision Trees
Random Forests
Boosting Methods



1. **Resampling Techniques:**
   - **Oversampling:** Increase the number of instances in the minority class.
   - **Undersampling:** Reduce the number of instances in the majority class.
2. **Synthetic Data Generation:** Create synthetic instances for the minority class.
3. **Cost-sensitive Learning:** Assign different misclassification costs to different classes.
4. **Ensemble Methods:** Combine multiple models to improve performance on minority classes.
5. **Evaluation Metrics:** Use metrics like precision, recall, F1 score, and area under the ROC curve for better assessment.

Conclusion

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

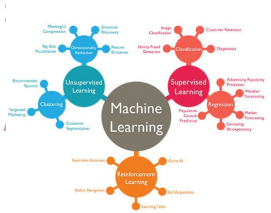
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Addressing imbalanced classes is crucial** for accurate and meaningful model predictions.
- **Choose appropriate treatment methods** based on the characteristics of your dataset.
- **Regularly evaluate model performance** using relevant metrics.

Variable Selection

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

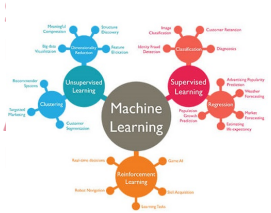
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Variable Selection

Introduction

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

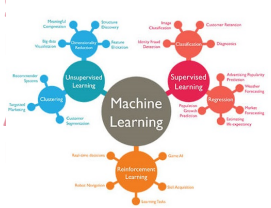
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Variable Selection:** Choosing the most relevant features for a machine learning model.
- **Importance:** Reduces overfitting, improves model interpretability, and enhances model performance.
- **Challenges:** High-dimensional datasets may contain irrelevant or redundant features.

Methods of Variable Selection

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

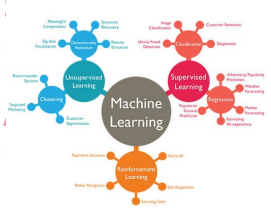
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



1. **Filter Methods:** Evaluate features independently of the learning algorithm.
2. **Wrapper Methods:** Use the learning algorithm's performance to evaluate subsets of features.
3. **Embedded Methods:** Perform variable selection as part of the model training process.

Boruta Algorithm

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

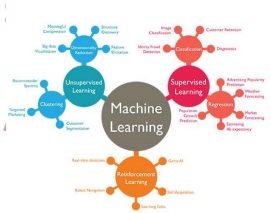
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Purpose:** Specifically designed for random forest models.
- **How it Works:**
  - Boruta compares the importance of each feature in the actual dataset with the importance of features obtained from a shadow dataset (where the class labels are shuffled).
  - Features with significantly higher importance in the actual dataset are considered relevant.
- **Output:** Each feature is labeled as "important," "unimportant," or "tentative."

Using Boruta in Practice

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

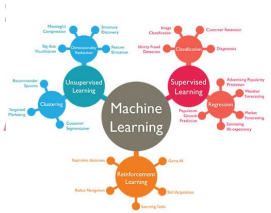
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



1. **Data Preprocessing:** Handle missing values and ensure data compatibility with random forest.
2. **Run Boruta:** Apply Boruta algorithm to rank and select features.
3. **Evaluate Results:** Analyze the output of Boruta to identify important features.
4. **Integrate with Model:** Train the final model using the selected features.

Advantages and Considerations

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

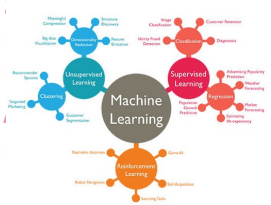
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Advantages:**
  - Handles non-linear relationships.
  - Accounts for interaction effects.
  - Works well with high-dimensional datasets.
- **Considerations:**
  - Computationally intensive for large datasets.
  - Sensitivity to hyperparameters.
  - Interpretability challenges with a large number of features.

Conclusion about variable selection

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

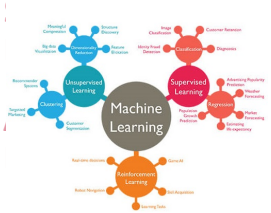
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Variable selection is a critical step** in building effective machine learning models.
- **Boruta offers a powerful approach** for selecting features, especially in the context of random forest models.
- **Consider the trade-offs** in terms of computational cost and interpretability when using Boruta.

Comparison of ML algorithms

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

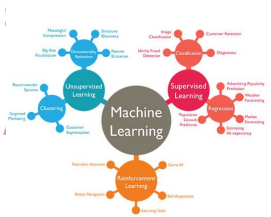


Table: Comparison of Machine Learning Models

Model	Speed	Interpretability	Scalability	User Adoption	Training Time
LR	Fast	High	High	Widely Accepted	5
DT	Moderate	Medium	Moderate	Moderate	10
RF	Moderate	Low	High	Moderate	20
GB	Slow	Low	High	Moderate	30
SVM	Moderate	Medium	Moderate	Low	15
ANN	Slow	Low	High	Low	50

Some Considerations

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

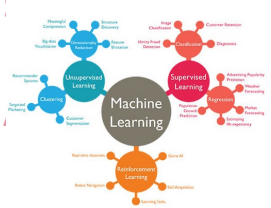
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Sample size**
  - Determining the minimum sample size is crucial in machine learning for building robust and reliable models.
- **Factors Affecting Sample Size:**
  - Complexity of the problem.
  - Dimensionality of the data.
  - Algorithm sensitivity.
  - Desired model performance.
- **Rule of Thumb:**
  - Several hundred samples are often a starting point, but the actual requirement varies.

Issues and Considerations

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

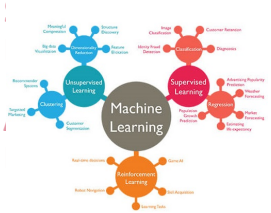
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Overfitting:**
  - Small datasets are prone to overfitting, where the model learns noise instead of the underlying patterns.
- **Generalization:**
  - Models trained on small samples might struggle to generalize well to new, unseen data.
- **Statistical Power:**
  - Low statistical power can lead to unreliable results and challenges in drawing meaningful conclusions.

Task-Specific Recommendations

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

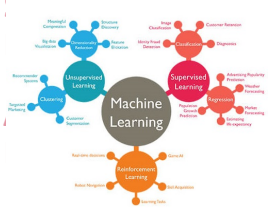
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **Classification Tasks:**
  - For simple classification tasks, a few hundred samples may suffice. For complex tasks or deep learning, larger datasets are often necessary.
- **Regression Tasks:**
  - The required sample size depends on the number of predictors and the desired effect size. Several hundred samples are often recommended.
- **Deep Learning:**
  - Deep learning models, particularly neural networks, may require thousands to millions of samples, especially for image and speech recognition.

Key Takeaways

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

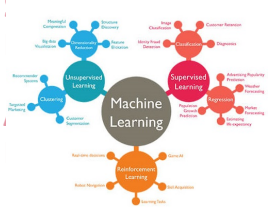
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



- **No One-Size-Fits-All:**
  - The minimum sample size varies based on factors such as task complexity, data dimensionality, and algorithm choice.
- **Quality Matters:**
  - Focus on collecting high-quality, relevant data to enhance model performance.
- **Adaptability:**
  - Continuously assess and adapt the sample size based on ongoing model evaluation and performance.

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

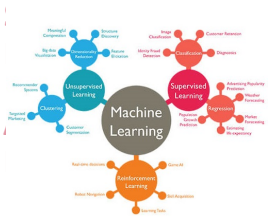
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



ELSEVIER

Check for updates

Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 110 (2019) 12–22

REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

[a]*Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium*

[b]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK*

[c]*Oxford University Hospitals NHS Foundation Trust, Oxford, UK*

[d]*Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands*

[e]*Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium*

[f]*Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK*

Accepted 5 February 2019; Published online 11 February 2019

Abstract

Objectives: The objective of this study was to compare performance of logistic regression (LR) with machine learning (ML) for clinical prediction modeling in the literature.

Study Design and Setting: We conducted a Medline literature search (1/2016 to 8/2017) and extracted comparisons between LR and ML models for binary outcomes.

Results: We included 71 of 927 studies. The median sample size was 1,250 (range 72–3,994,872), with 19 predictors considered (range 5–563) and eight events per predictor (range 0.3–6,697). The most common ML methods were classification trees, random forests, artificial neural networks, and support vector machines. In 48 (68%) studies, we observed potential bias in the validation procedures. Sixty-four (90%) studies used the area under the receiver operating characteristic curve (AUC) to assess discrimination. Calibration was not addressed in 56 (79%) studies. We identified 282 comparisons between an LR and ML model (AUC range, 0.52–0.99). For 145 comparisons at low risk of bias, the difference in logit(AUC) between LR and ML was 0.00 (95% confidence interval, −0.18 to 0.18). For 137 comparisons at high risk of bias, logit(AUC) was 0.34 (0.20–0.47) higher for ML.

Conclusion: We found no evidence of superior performance of ML over LR. Improvements in methodology and reporting are needed for studies that compare modeling algorithms.

Keywords: Clinical prediction models; Logistic regression; Machine learning; AUC; Calibration; Reporting

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

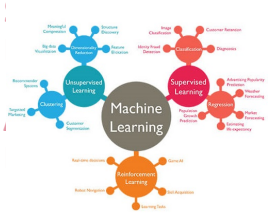
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



RESEARCH

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Marc M J Bonten,[8,10] Darren L Dahly,[11,12] Johanna A A Damen,[8,9] Thomas P A Debray,[8,9] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,13] Paula Dhiman,[4,5] Maria C Haller,[7,14] Michael O Harhay,[15,16] Liesbet Henckaerts,[17,18] Pauline Heus,[8,9] Nina Kreuzberger,[19] Anna Lohmann,[20] Kim Luijken,[20] Jie Ma,[5] Glen P Martin,[21] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[22,23] Chunhu Shi,[24] Nicole Skoetz,[19] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[25] René Spijker,[8,9,26] Ewout W Steyerberg,[3] Toshihiko Takada,[8] Ioanna Tzoulaki,[27,28] Sander M J van Kuijk,[29] Florien S van Royen,[8] Jan Y Verbakel,[30,31] Christine Wallisch,[7,32,33] Jack Wilkinson,[22] Robert Wolff,[34] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

For numbered affiliations see end of the article

Correspondence to: L Wynants laure.wynants@maastrichtuniversity.nl (ORCID 0000-0002-3037-122X)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;369:m1328 http://dx.doi.org/10.1136/bmj.m1328

Originally accepted: 31 March 2020

Final version accepted: 1 July 2020

ABSTRACT

OBJECTIVE

To review and appraise the validity and usefulness of published and preprint reports of prediction models for diagnosing coronavirus disease 2019 (covid-19) in patients with suspected infection, for prognosis of patients with covid-19, and for detecting people in the general population at increased risk of becoming infected with covid-19 or being admitted to hospital with the disease.

DESIGN

Living systematic review and critical appraisal by the COVID-PRECISE (Precise Risk Estimation to optimise covid-19 Care for Infected or Suspected patients in diverse sEttings) group.

DATA SOURCES

PubMed and Embase through Ovid, arXiv, medRxiv, and bioRxiv up to 5 May 2020.

STUDY SELECTION

Studies that developed or validated a multivariable covid-19 related prediction model.

DATA EXTRACTION

At least two authors independently extracted data using the CHARMS (critical appraisal and data extraction for systematic reviews of prediction modelling studies) checklist; risk of bias was assessed using PROBAST (prediction model risk of bias assessment tool).

RESULTS

14 217 titles were screened, and 107 studies describing 145 prediction models were included. The review identified four models for identifying people at risk in the general population; 91 diagnostic models for detecting covid-19 (60 were based on medical imaging, nine to diagnose disease severity); and 50 prognostic models for predicting mortality risk, progression

Best Reference Book for Statistical Learning

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

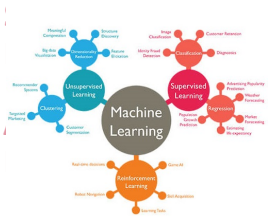
Discrimination Analysis

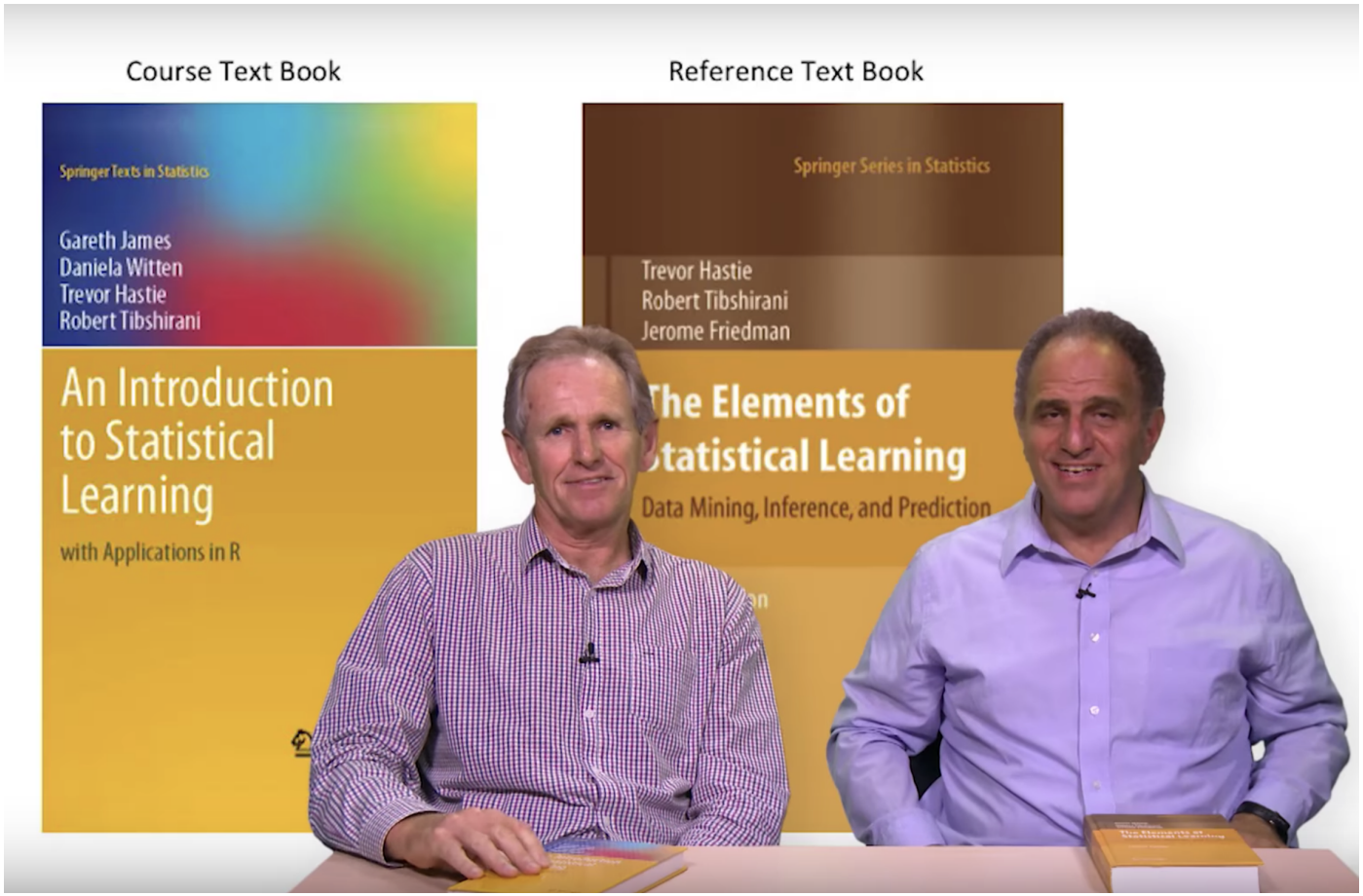
K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods





Best Reference Book for Statistical Learning

Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

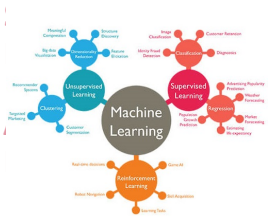
Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods



Introduction to Machine Learning

Logistic Regression

Penalized Logistic Regression (LASSO, Ridge, Elastic Net)

Discrimination Analysis

K-Nearest Neighborhood (KNN)

Decision Trees

Random Forests

Boosting Methods

Support Vector Machine (SVM)

Artificial Neural Networks

Missing Data