Northeastern University

Introduction to Machine Learning

Roi Yehoshua

Agenda

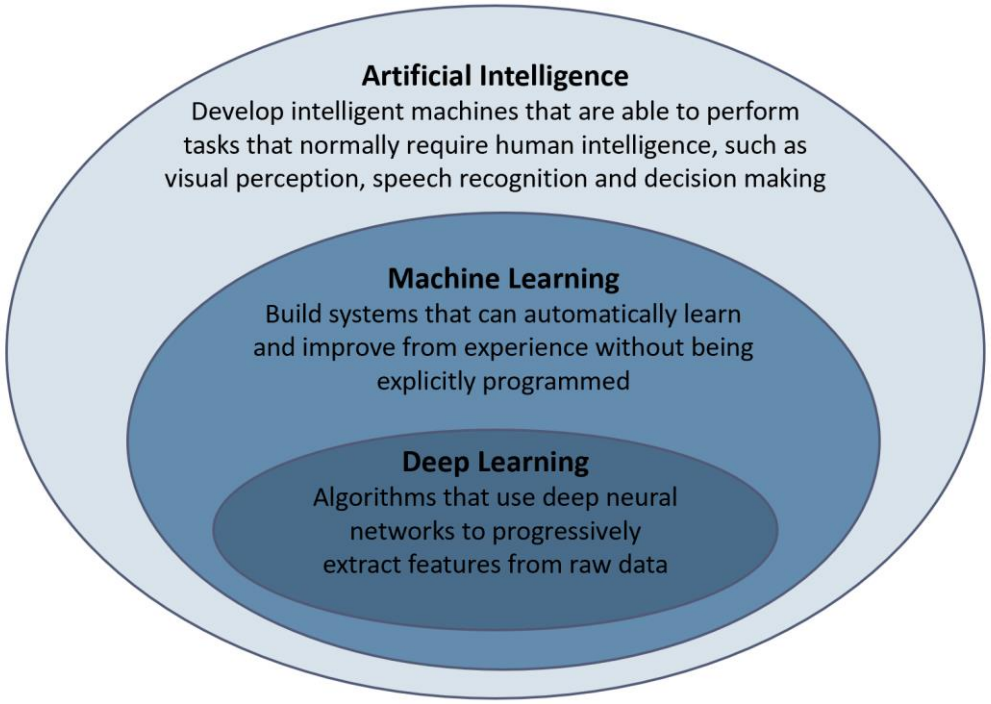
Northeastern University

- What is machine learning?
- Types of machine learning
- Limitations of machine learning
- Machine learning models
- Python ML ecosystem
- Course objectives and overview

What is Machine Learning (ML)?

Northeastern University

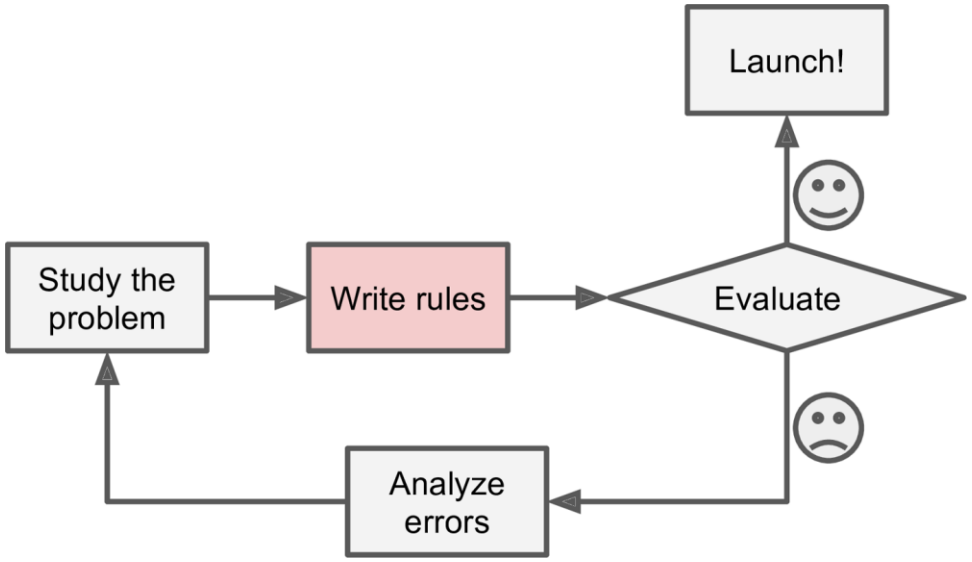
- Subfield of artificial intelligence (AI)
- Design algorithms that learn patterns from data without being explicitly programmed



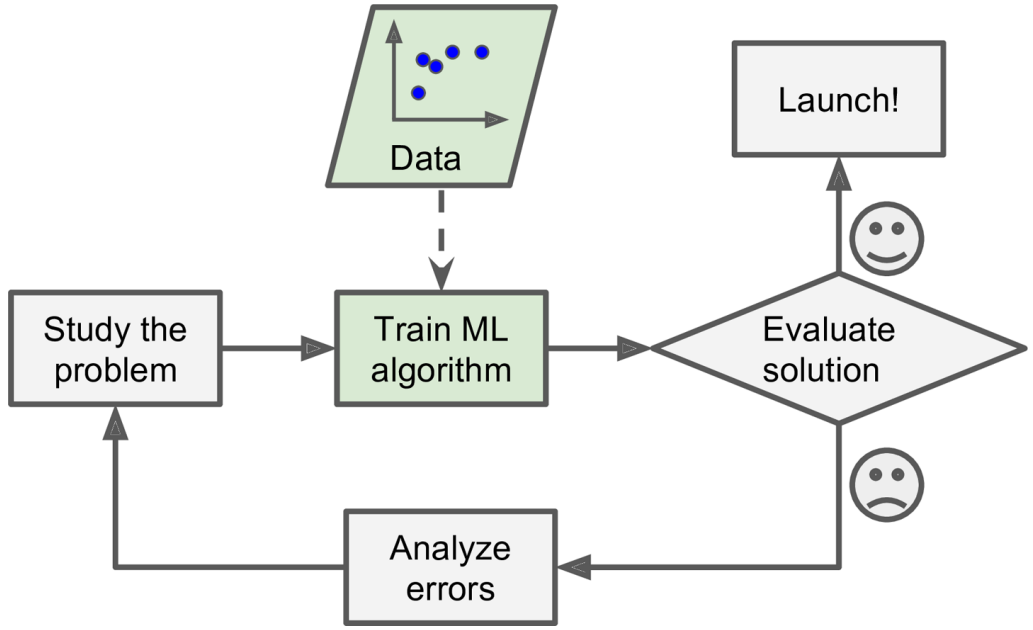
What is Machine Learning (ML)?

Northeastern University

- Solving problems with ML vs. the traditional programming tools:



Traditional approach



The ML approach

Learning Problems

Northeastern University

- Which types of problems can be solved with machine learning?
  - When the task can be expressed as a learning task
    - e.g., sorting a list of numbers can be solved without learning
  - When high-quality data is available
    - GIGO (Garbage In – Garbage Out)
  - When some error is acceptable
    - Can rarely achieve 100% accuracy

Short History

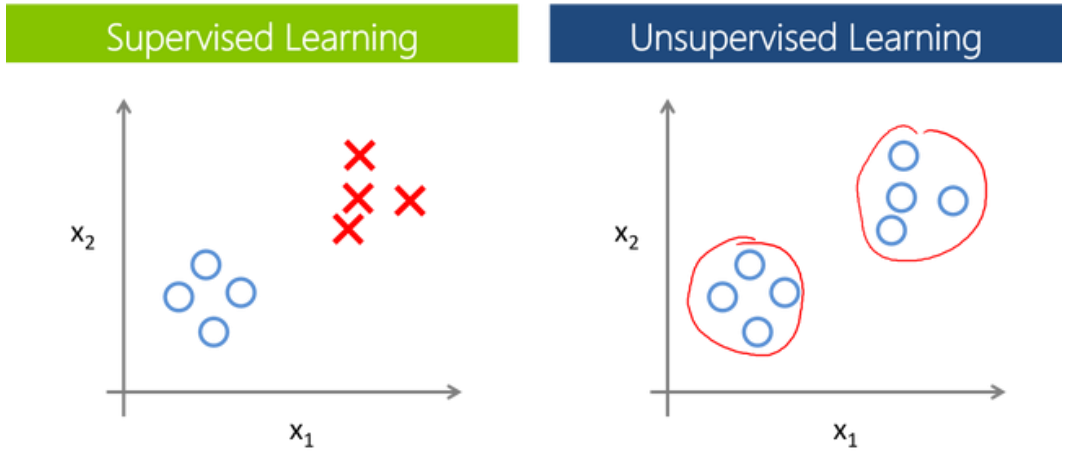
Northeastern University

- 1763 – Thomas Bayes, Bayes theorem
- 1805 – Legendre and Gauss used linear regression to predict planetary movement
- 1943 – McCulloch-Pitts neurons, first mathematical model of biological neurons
- 1952 – Arthur Samuel's checkers program, the first computer learning program
- 1957 – Frank Rosenblatt, the perceptron model
- 1967 – The nearest neighbors algorithm
- 1970 – "AI winter", limitations of perceptron
- 1979 – The Stanford Cart, a robot that was able to navigate obstacles in a room
- 1986 – Quinlan, ID3, decision tree learning
- 1990 – Cortes and Vapnik, SVM with kernels
- 1997 – IBM Deep blue defeats Kasparov at chess
- 2006 – Geoffrey Hinton, deep learning
- 2011 – IBM's Watson beats its human competitors at Jeopardy
- 2014 – Facebook develops DeepFace that is able to recognize faces at the same level as humans
- 2016 – Google's AlphaGo defeats the 18-time Go world champion Lee Sedol
- 2022 – OpenAI released ChatGPT, a powerful chatbot that can assist on many tasks

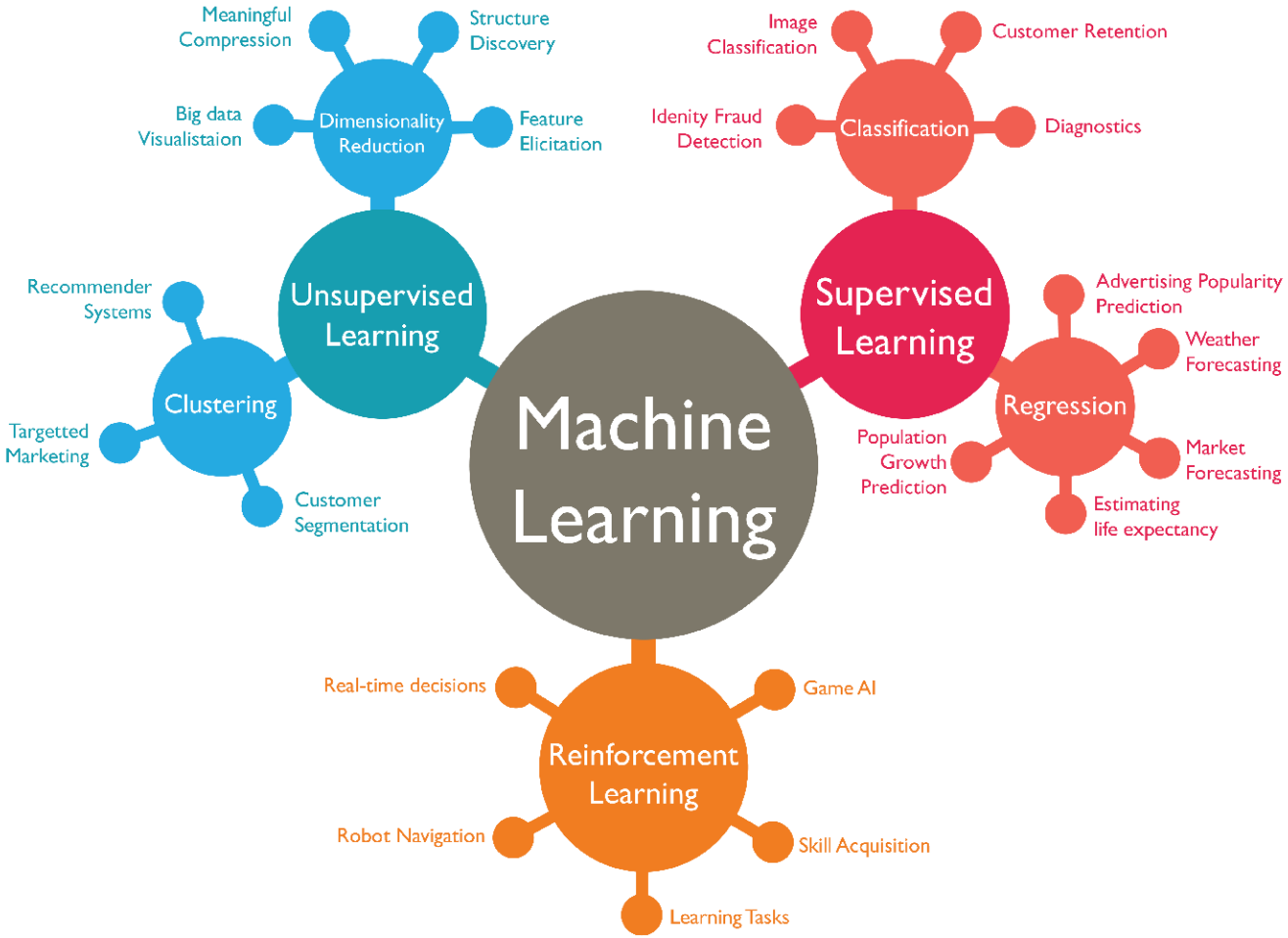
Types of Machine Learning

Northeastern University

- **Supervised learning** – learn from **labeled data**
  - Goal is to learn a function that maps an input to an output
- **Unsupervised learning** – learn from **unlabeled data**
  - Goal is to extract useful patterns from the data
- **Reinforcement learning** – learn to take the best actions in an environment in order to maximize a cumulative reward



Types of Machine Learning



Machine Learning Model Components

Northeastern University

- **Dataset**: Collection of examples used to train and evaluate a model
  - **Training set**: Subset of data used to fit the model's parameters
  - **Test set**: Held-out data used to assess the model's final generalization performance
- **Parameters**: Internal values learned from the training data
- **Hyperparameters**: Configuration settings defined before training
  - e.g., learning rate, tree depth
- **Loss function**: Quantifies the model's error, guides the learning process
- **Optimization method**: Algorithm used to minimize the loss function
  - e.g., gradient descent
- **Evaluation metrics**: Measures used to judge model performance
  - e.g., accuracy, RMSE, AUC

The Learning Process

Northeastern University

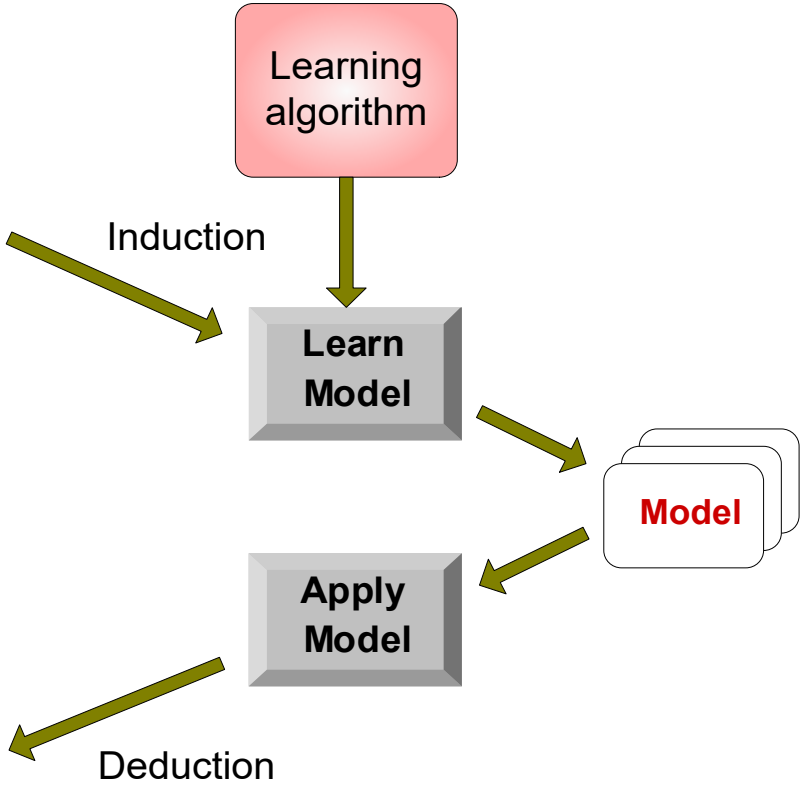
- Define the problem space
- Collect data
- Extract/select features
- Pick a model (hypothesis)
- Develop a learning algorithm
  - Train and learn model parameters
- Make predictions on new data
  - Testing phase
- In practice, usually re-train when new data is available and use feedback from deployment

The Learning Process

Northeastern University

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

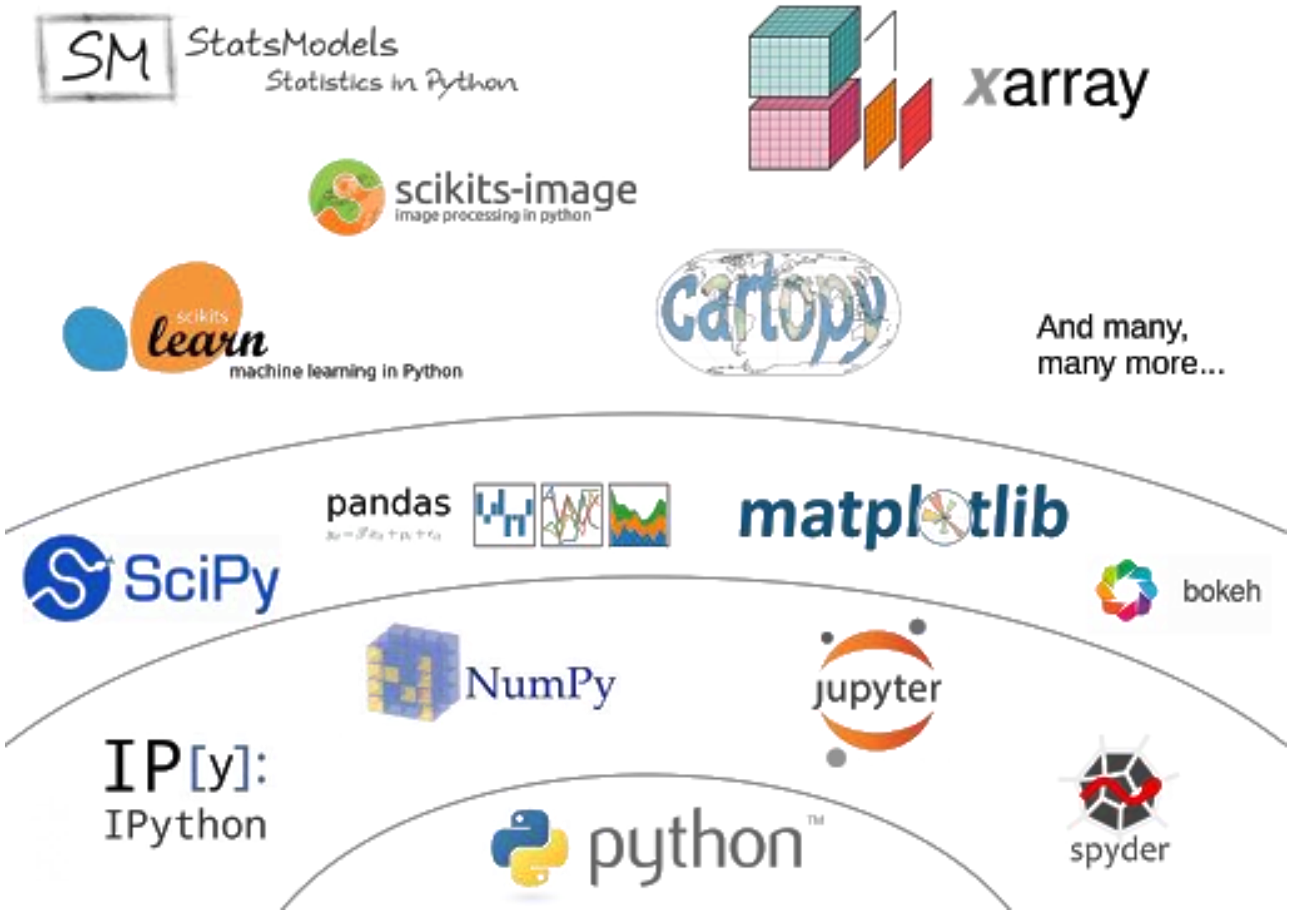
Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Python ML Ecosystem

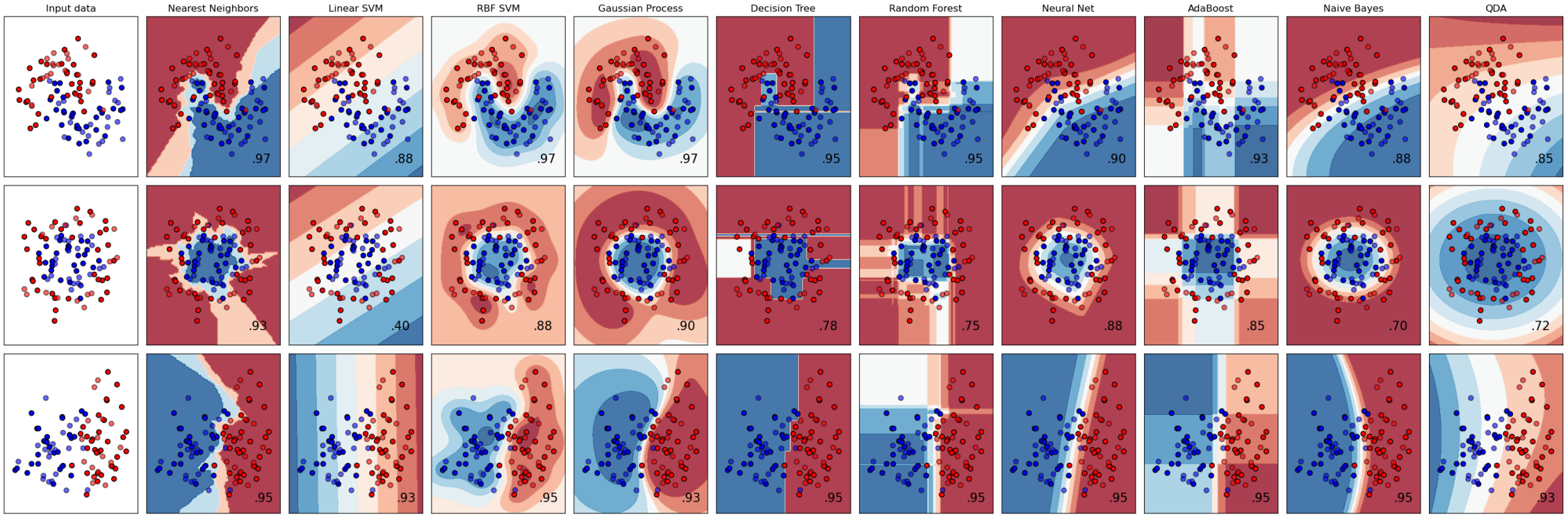


Scikit-Learn

Northeastern University

- scikit-learn is a free machine learning library for Python
- Provides efficient versions of a large number of common ML algorithms, including various classification, regression and clustering algorithms
- Characterized by a clean, uniform, and streamlined API, as well as by very useful and complete online documentation
- Designed to interoperate with Python scientific libraries NumPy and SciPy
- Largely written in Python, with some core algorithms written in Cython to achieve fast performance
- http://scikit-learn.org/

Classification Algorithms



https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Course Objectives

Northeastern University

- Become familiar with main machine learning tasks and concepts
  - Model capacity, generalization
  - Model selection, regularization, cross-validation
  - Bias vs. variance tradeoff
- Study most well-known algorithms
  - Regression (linear regression, polynomial regression)
  - Classification (decision trees, Naïve Bayes, KNN, SVMs, ensembles, etc.)
  - Neural networks (+ intro to deep learning)
  - Clustering (K-means, GMM, hierarchical clustering)
  - Dimensionality reduction (PCA, t-SNE)
- Learn the theory and foundation behind ML algorithms and how to apply them to real-world data sets

Math Background

Northeastern University

- Linear algebra
  - Vectors, norms of vectors, dot product, projections
  - Matrices, basic properties (transpose, trace, determinant, norm)
  - Systems of linear equations
  - Eigenvectors and eigenvalues
- Calculus
  - Limits, derivatives, integrals
  - The chain rule of derivatives
  - Multivariable functions, gradients, Hessian, Jacobian matrix
  - Minimization/maximization of a function, Lagrange multipliers
- Probability theory
  - Random variables, expectation, variance
  - Common probability distributions (Bernoulli, Binomial, Poisson, Uniform, Normal)
  - Covariance, correlation

Assignments

Northeastern University

- Weekly problem sets including both theoretical and programming exercises
  - Assignments are due at 11:59PM on the specified date
  - Up to one day late is allowed, penalized by 10%
  - **Assignments must be done independently**
    - Can use tools like ChatGPT only for clarifying concepts, not for generating code or answers
- Final Project
  - Can choose any topic related to supervised machine learning
  - Team size: 1-3 members
  - Project proposal: mid class
  - Presentation: on the last lecture of the semester
  - Final report: when semester ends

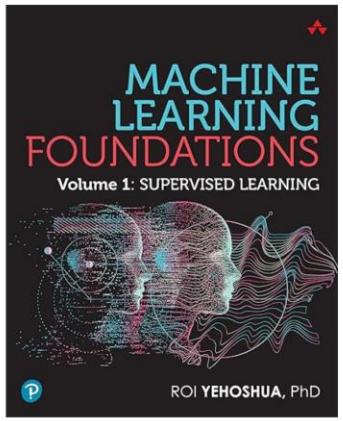
Grading

Northeastern University

- Assignments – **20%**
- Midterm exam – **20%**
- Final exam – **20%**
- Final project – **40%**
- Class participation is used to adjust grade upwards (at the discretion of the instructor)

Course Textbook

- **Machine Learning Foundations (Volume 1)**, Roi Yehoshua, Pearson, 2026
- Textbook will be published on February 2026
  - Can be purchased from Northeastern's bookstore
  - Or directly from Amazon
  - https://www.amazon.com/Machine-Learning-Foundations-Roi-Yehoshua/dp/0135337860



Follow the author

Roi Yehoshua Follow

Machine Learning Foundations: Volume 1: Supervised Learning 1st Edition

by Roi Yehoshua (Author)

See all formats and editions

Savings Pre-order Price Guarantee. Terms

The Essential Guide to Machine Learning in the Age of AI

Machine learning stands at the heart of today's most transformative technologies: advancing scientific discovery, reshaping industries, and transforming everyday life. From large language models to medical diagnosis and autonomous vehicles, the demand for robust, principled machine learning models has never been greater.

Machine Learning Foundations, Volume 1: Supervised Learning, offers a comprehensive and accessible roadmap to the core algorithms and concepts behind modern AI systems. Balancing mathematical rigor with hands-on implementation, this book not only teaches how machine learning works, but why it works.

As part of a three-volume series, Volume 1 lays the foundation for mastering the full landscape of modern machine learning, including deep learning, large language models, and cutting-edge research. Whether you are a student starting out, a researcher seeking a reliable reference, or a practitioner looking to sharpen your

Read more

Report an issue with this product or seller