

# Lecture 1: An Overview of Machine Learning

Beidi Qiang

SIUE

# What is Machine learning?

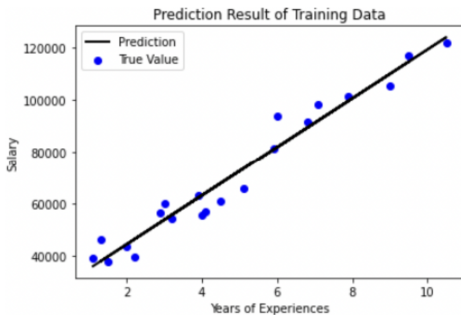
Machine learning (ML) is a category of an algorithm that enables computer to learn and adapt from data, draw inference and become more accurate in predicting outcomes. It refers to a vast set of tools used for understanding data.

Machine learning can be classified into 2 types of algorithms.

- ▶ Supervised Learning: building a statistical model for predicting, or estimating, an output (label) based on one or more inputs (features). Linear regression is a simple example of supervised learning.
- ▶ Unsupervised Learning (Chapter 12 ISLR): there are inputs but no supervising output (label); nevertheless we can learn relationships and structure from such data, such as grouping similar data together (clustering).

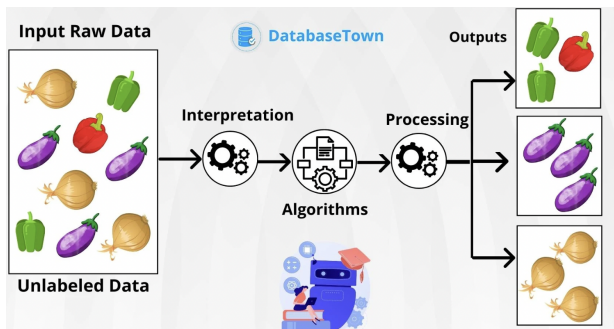
# Examples of Supervised Learning

- ▶ Predict the salary by years of expert. (for example using linear regression)
- ▶ Recognize images of handwritten digits (classify an image as 0,1,2,... or 9).
- ▶ Financial industry and trading: companies use ML in fraud investigations (classify a transaction as fraud or not)



# Example of Unsupervised Learning

- ▶ Clustering: you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior, or grouping images by their characters.



# A Quick History of Machine Learning

- ▶ Linear discriminant analysis was proposed in 1936 and in the 1940s, logistic regression was developed.
- ▶ In 1970s, many more non-linear techniques for learning from data were available and the computation become feasible in 1980s by the improvement of computing technology. that non-linear methods were no longer computationally prohibitive
- ▶ In the mid 1980s, classification and regression trees were developed. Neural networks gained popularity in the 1980s, and support vector machines arose in the 1990s.
- ▶ Machine learning became very famous in the 1990s. The intersection of computer science and statistics gave birth to probabilistic approaches in AI. This shifted the field further toward data-driven approaches.

- ▶ Labels: The thing we're predicting. Other names: output, response, target. Example: the  $y$  variable in simple linear regression.
- ▶ Features: The input variables. Other names: input, predictors, independent variables. Example: the  $x$  variable in simple linear regression.
- ▶ Examples: An example is a particular instance of data, labeled or unlabeled.

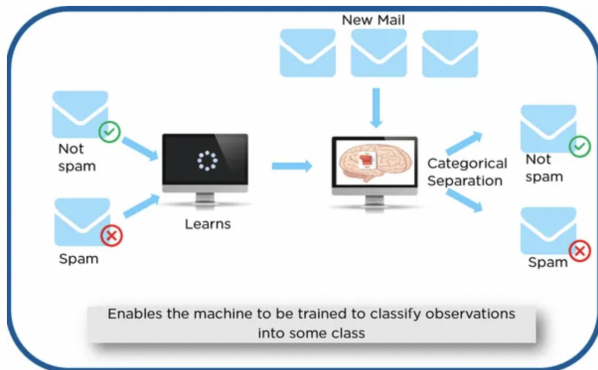
**Example:** Email spam filtering

Labels: spam or not spam.

Features: words in the email text, sender's address, time of day the email was sent, etc...

# Key ML Terminology, cont.

- ▶ **Model:** A model defines the relationship between features and label.
- ▶ **Model Training:** Means creating or learning the model. That is, you show the model examples and enable the model to gradually learn the relationships between features and label.
- ▶ **Inference** means applying the trained model to unlabeled examples.





We tend to select statistical learning methods on the basis of whether the label is quantitative (numerical) or qualitative (categorical).

- ▶ A **regression model** predicts continuous values. For example, regression models make predictions that answer questions like the following:
  - ▶ What is the value of a house in Edwardsville, IL?
  - ▶ What is the probability that a user will click on this ad?
- ▶ A **classification model** predicts discrete values or classes. For example, classification models make predictions that answer questions like the following:
  - ▶ Is a given email message spam or not spam?
  - ▶ Is this an image of a dog, a cat, or a hamster?