California State University
Los Angeles

Machine Learning

Linear Regression

Nikhil Dhiman
Computational Molecular Biology (COMB) Lab

Agenda Overview

01 What is Machine Learning?

02 Terminology

03 Machine Learning Settings

04 Supervised Learning

05 Approaches of Supervised Learning

06 Supervised Learning Algorithms

07 Linear Regression

08 Cost Function in Linear Regression

What is Machine Learning?

Definition

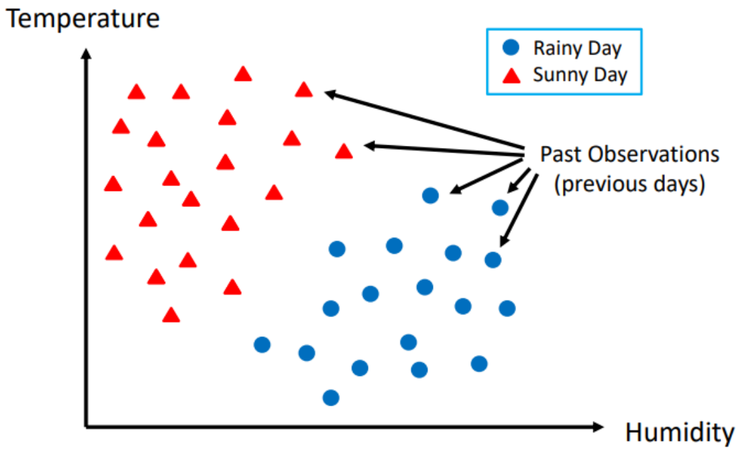
- Designing and constructing algorithms or methods that give computers the ability to learn from past data, without being explicitly programmed, and then make predictions on future data.

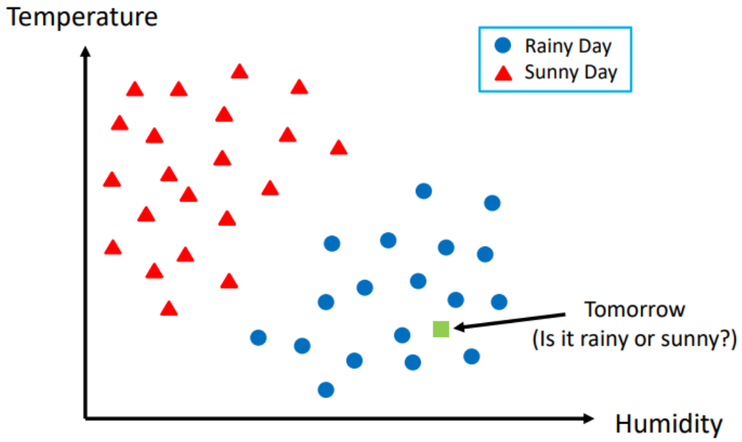
- A set of algorithms that can automatically detect and extract patterns in past data, and then use the extracted patterns to predict future data, or to perform other kinds of decision-making.

Example: Weather Forecasting

- Suppose that we have the Temperature and Humidity of the past 30 days.
- We also know whether those days were Sunny or Rainy.

Questions: Now, If we know the Temperature and Humidity of tomorrow, can we predict tomorrow's outlook (predict whether tomorrow is rainy or sunny)?





Terminology

Observations:

Data Samples (Data Examples)

Features (inputs):

Attributes that represent an observation, e.g., temperature, humidity

Labels (outputs):

Values assigned to observations (also called class, target), e.g., rainy/sunny

Training Data:

Past observations are given to the Machine Learning algorithm to learn from.

Testing or Prediction Data:

Observations given to a “predictive model” for prediction.

Training Stage (Modeling):

Building a predictive model based on the training dataset (past data).

- The model does not have to be perfect. As long as it is close, it is useful.

-

We should tolerate randomness and mistakes.

Testing Stage (Prediction):

Applying the trained model to forecast what is going to happen in future

(on future testing data)

Machine Learning Settings

Supervised learning

Learning from labeled observations

Unsupervised learning

Learning from unlabeled observations

Semi-supervised learning

Labels are provided only for a part of the training data

Reinforcement learning

Learning from an agent taking actions in an environment so as to maximize a long-term reward.

Transfer learning

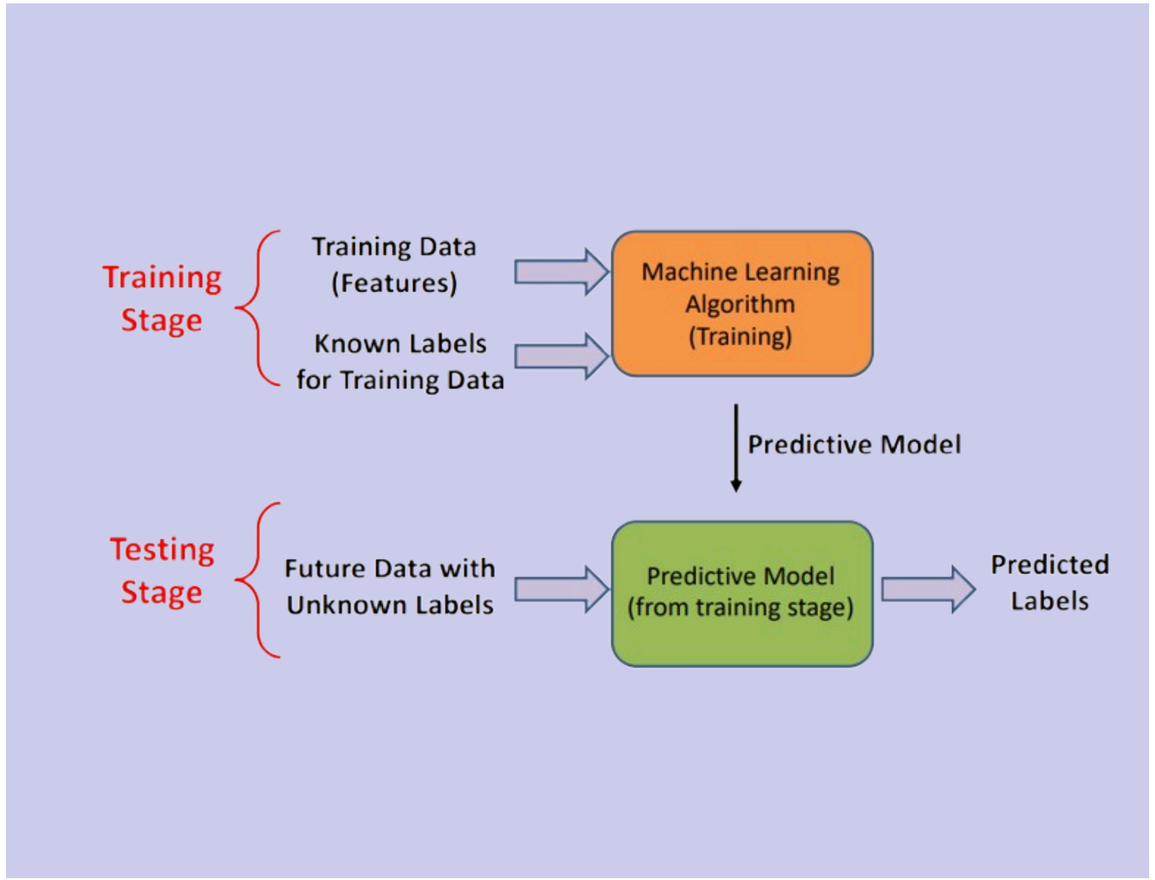
Learning from a dataset while solving a problem, and then applying the extracted knowledge to a different but related dataset/problem.

Active learning

Similar to Semi-Supervised Learning, but the algorithm is able to interactively query the user or some other information source to obtain the labels as needed.

Supervised Learning

- Supervised learning is a type of machine learning where a model is trained on a labeled dataset, meaning each training example is paired with a correct output (label).
- The goal is to learn a mapping from inputs (features) to outputs (labels) so that the model can accurately predict the label for new, unseen data.
- It's widely used in applications such as spam detection (email labeled as spam or not), image classification (e.g., identifying objects in photos), and medical diagnosis (e.g., predicting disease based on symptoms).



Two Important Approaches of Supervised Learning

- **Classification**: Predict a discrete valued output for each observation. - Labels are discrete (categorical) - Labels can be binary (e.g., rainy/sunny, spam/non-spam,) or non-binary (e.g., rainy/sunny/cloudy, object recognition (100classes))

- **Regression**: Predict a continuous valued output for each observation. - Labels are continuous (numeric), e.g., stock price, housing price - Can define 'closeness' when comparing prediction with true values

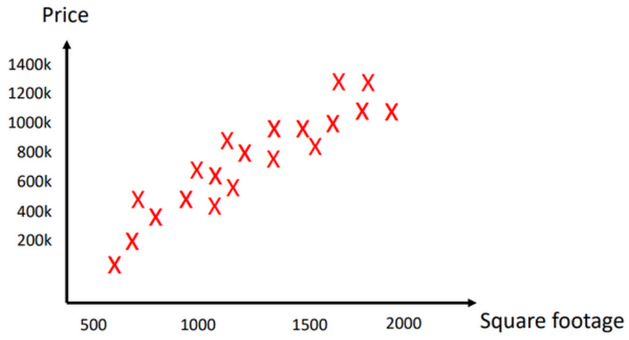
Supervised Learning Algorithms

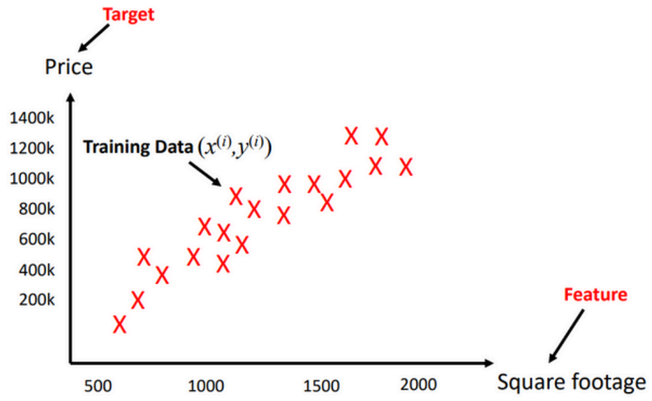
- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVM)
- Neural Networks

Linear Regression

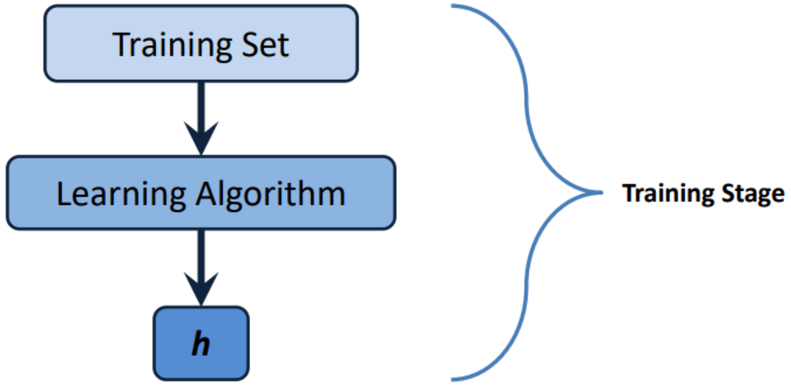
- A statistical method to model the relationship between a dependent variable and one or more independent variables.
- Predicts continuous outcomes.
- Simple yet powerful for many real-world applications.
- Linear Regression with single input variable (one feature). It is also called “Univariate Linear Regression.

Regression Example: Housing Price





Linear Regression Training

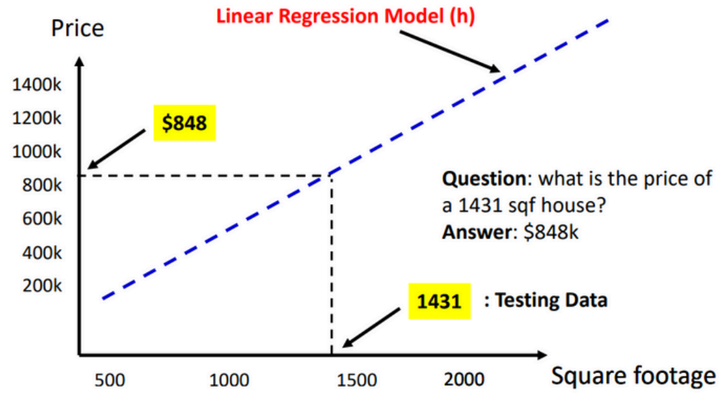


"h" is the Regression Model (Predictive Model) built by machine learning algorithm based on the training set. It will be later used to predict or estimate the "target" (e.g. price) on new observations.

Linear Regression Training

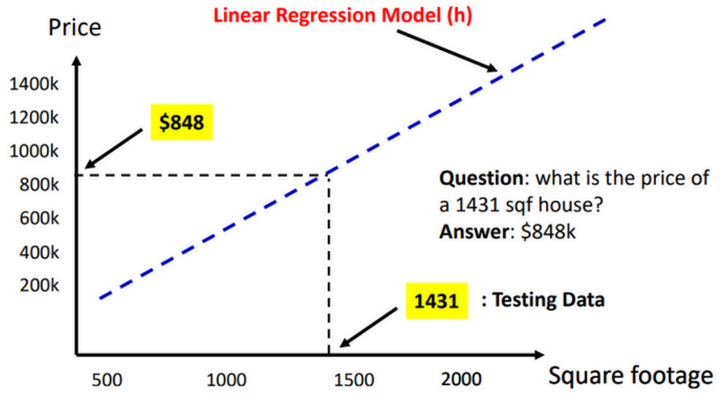
$$h_\theta(x) = \theta_0 + \theta_1 x$$

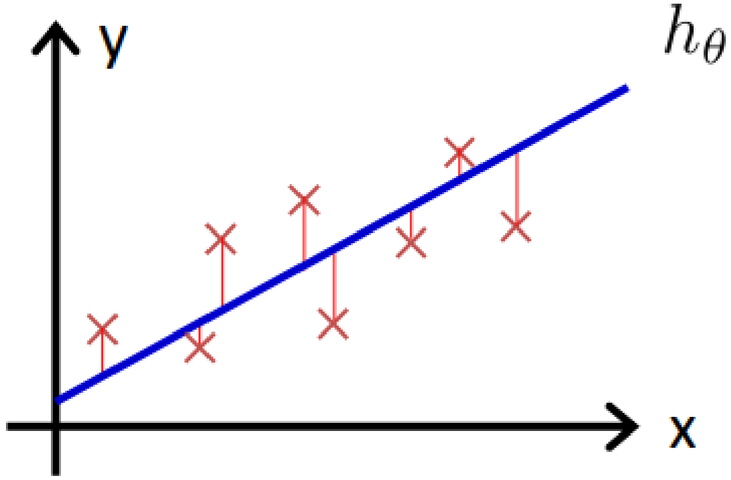
In the case of Linear Regression with One Variable, the Regression model (h) will be a line



How to find the best fit line? In other word, How to find the Parameters that correspond to the best line?

$$h_\theta(x) = \theta_0 + \theta_1 x$$





$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$\theta_0, \theta_1 = ?$$

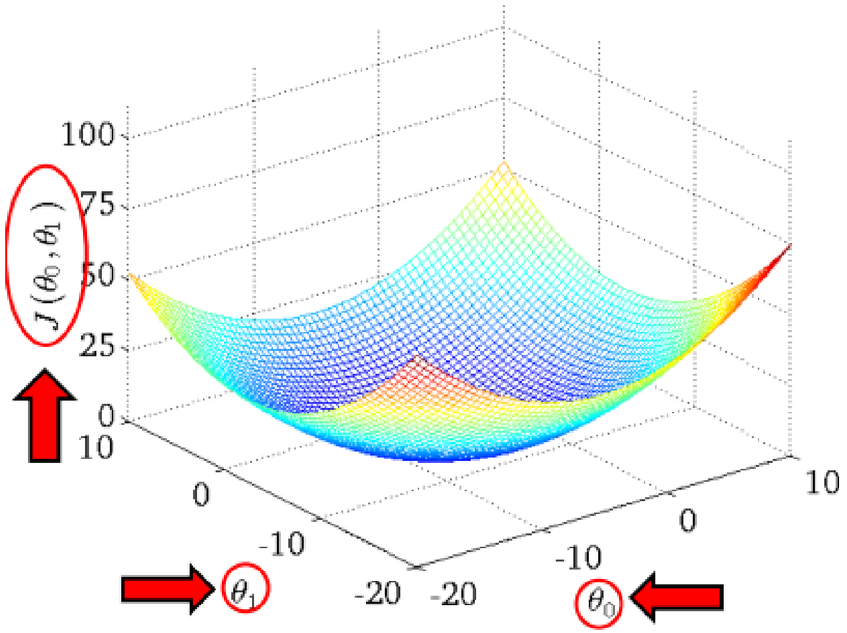
Cost Function in Linear Regression

- The Cost Function, J(θ0,θ1), quantifies the error between the predicted values and actual values in the training data.
- It helps measure how well the model fits the data.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Cost Function in Linear Regression

The linear regression cost function is convex because it is a quadratic function in the model parameters, and its second derivative is non-negative. This property ensures that optimization algorithms like gradient descent will always converge to the global minimum, making the training process stable and predictable.



California State University
Los Angeles

Thank You

Computational Molecular Biology (COMB) Lab