COMPUTER SCIENCE

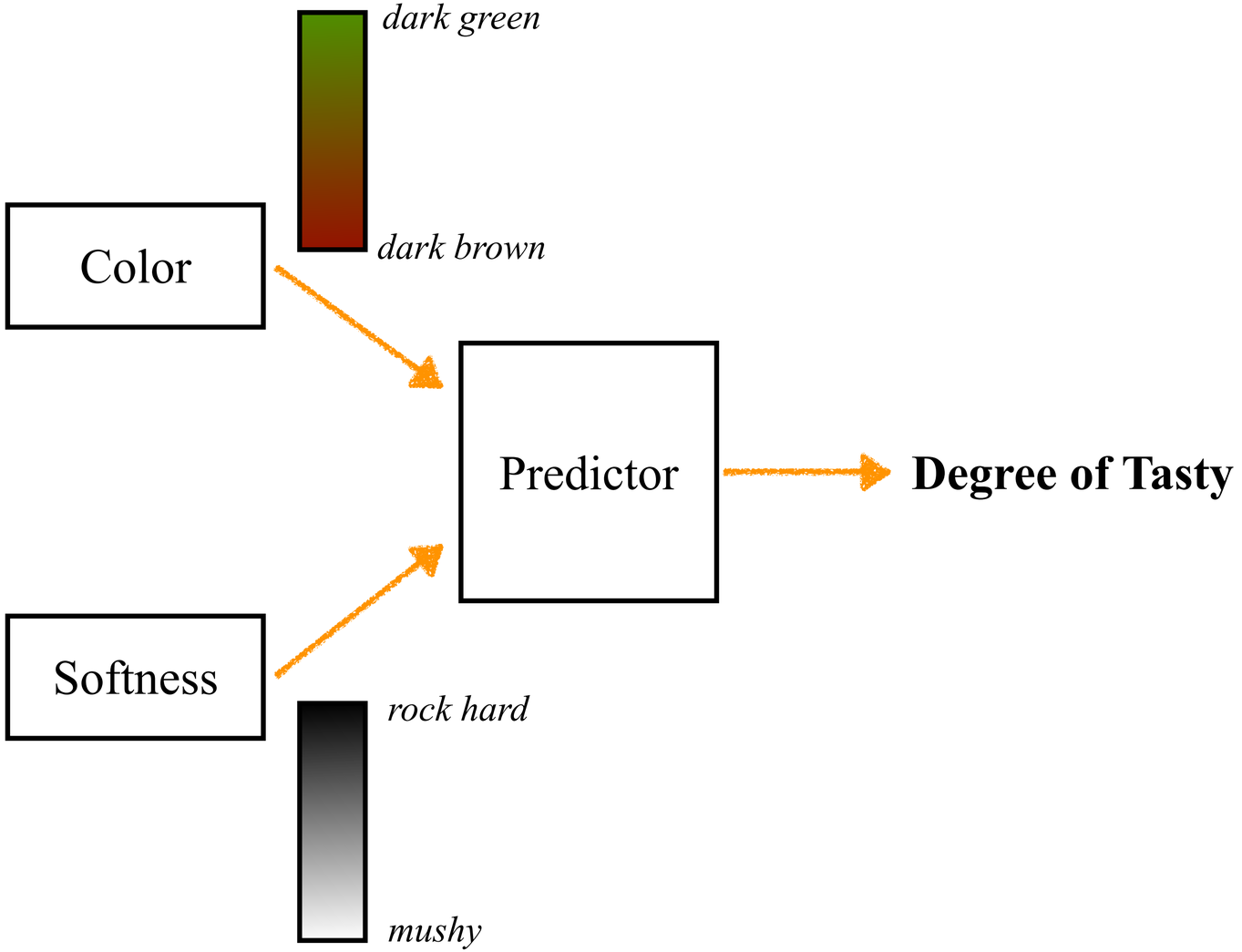
Lecture 2: Supervised Learning

Xin Chen
Assistant Professor
University of New Mexico

Outline

- Problem Formulation.
- Classification and Regression.
- Basic Terminology and Notations.
- Data and Data Representation.
- Building Machine Learning Systems.
- Python Programming.

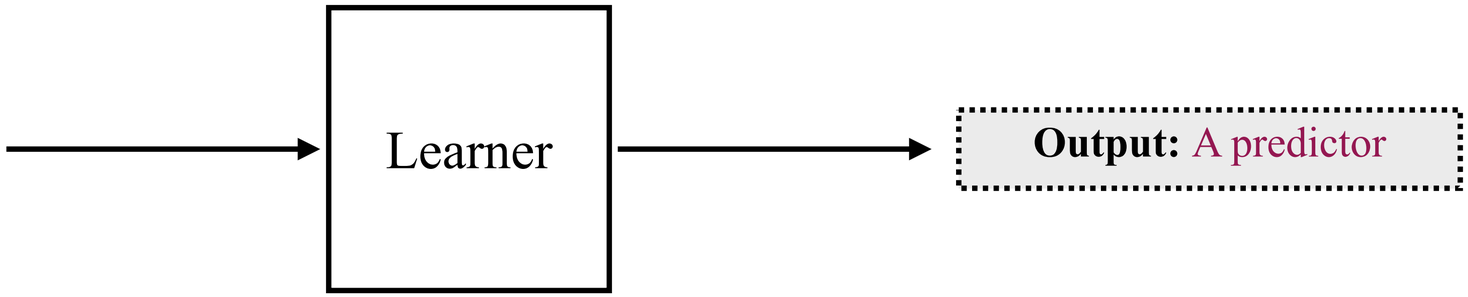
Which Papaya is Tasty?



Statistical Learning Framework

Input

- **Domain set** - The set of objects that we may wish to label.
- **Label set** - The set of labels applied to the objects.
- **Training data** - A finite set of pairs in the form of (X, Y) where X is an object and Y is the label of X.



Output: A predictor

Supervised Learning

Problem Definition:

Given a finite set of data such that $X_1, \ldots, X_n$ are the inputs and $Y_1, \ldots, Y_n$ are the corresponding outputs. How can we find a model f such that the **distance** between $f(X_i)$ and Y_i is **minimized** for all $i = 1, \ldots, n$.

The two basic ML tasks in supervised learning are **classification** and **regression**.

Problem Modeling

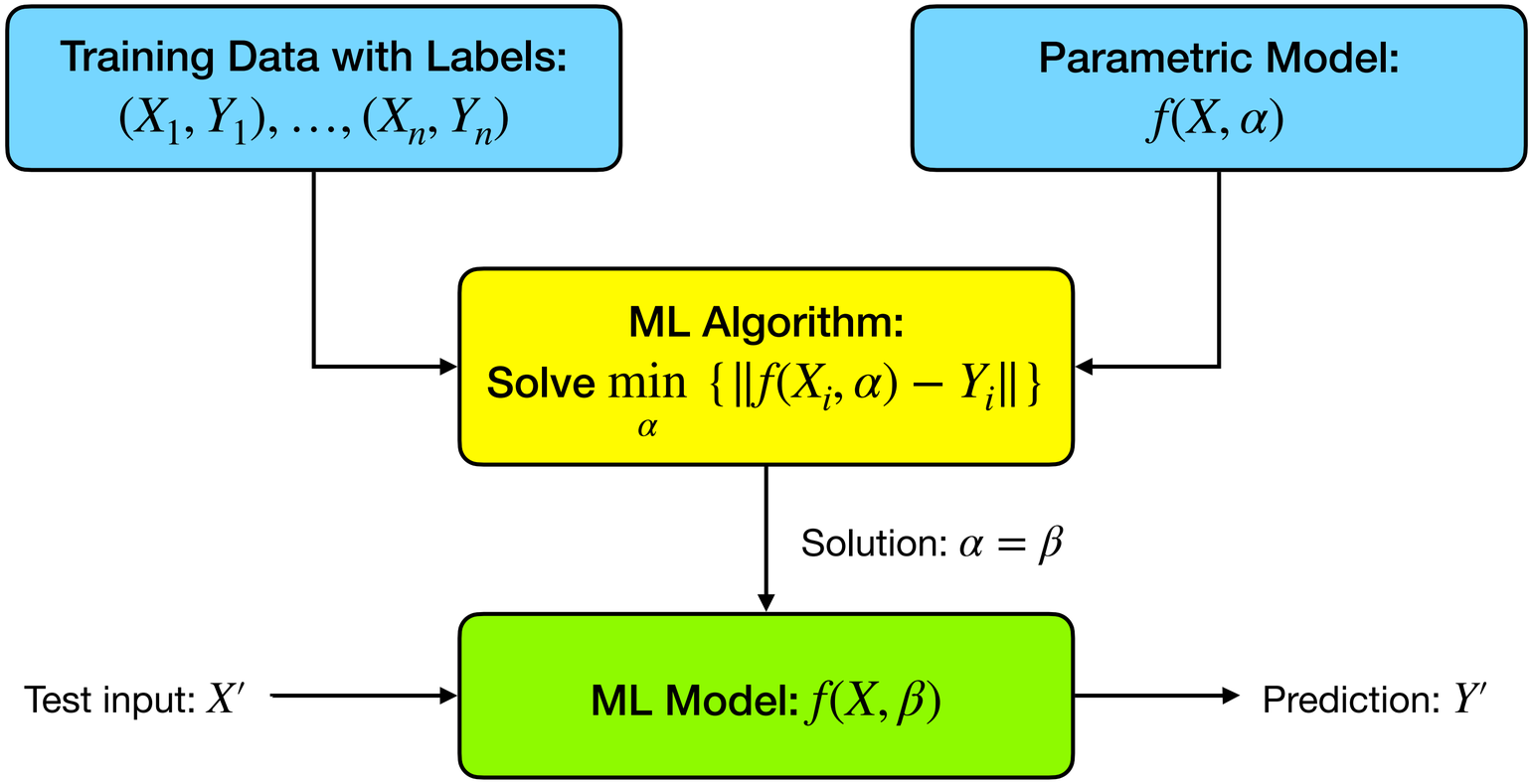
A supervised learning task can be formally described as

$$\min_{\alpha} \{ \|f(X_i, \alpha) - Y_i\| \} \text{ for all } i = 1, \ldots, n$$

such that $f(X, \alpha)$ is the parametric model and α are the unknown parameters.

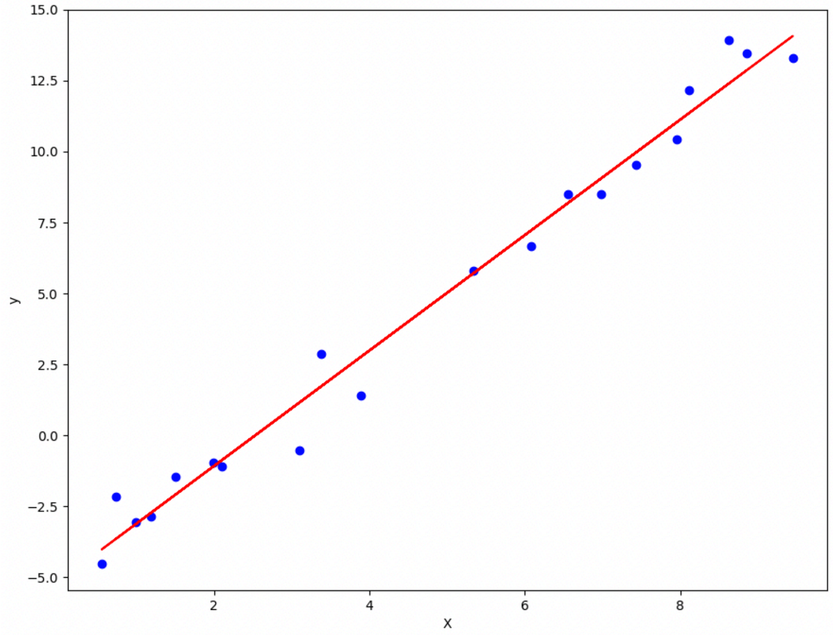
Example: A linear parametric model can be defined as $f(x, a, b) = ax + b$ wherein x is the variable for the input data and a, b are the unknown parameters.

Supervised Learning Process



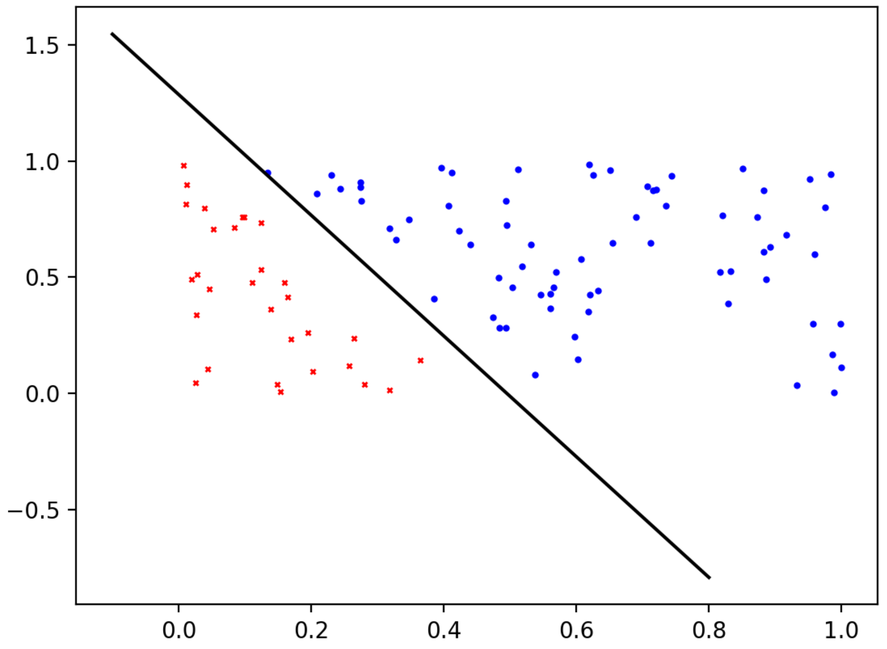
Example: Linear Regression

- Every datum is represented as a pair (x_i, y_i) such that x_i is the input and y_i is the output.
- The parametric model is given as $f(x, a, b) = ax + b$.
- The ML task tries to find the unknown parameters a, b such that the distance $f(x_i, a, b) - y_i$ is minimized.
- The red line indicates the result $f(x, a, b)$ with $a = 2.03342064,\ b = -5.1457367.$



Example: Linear Classification

- Every datum is represented as a pair $((x_i, y_i), c_i)$ such that (x_i, y_i) is the input and c_i is the output (color).
- The goal is to find a separation function $f(x, y, \alpha)$ such that $f(x_i, y_i, \alpha) > 0$ when (x_i, y_i) is blue, $f(x_i, y_i, \alpha) < 0$ when (x_i, y_i) is red.
- We use the linear parametric model $f(x, y, a, b, c) = ax + by + c$.
- The black line indicates the hyperplane $ax + by + c = 0$ wherein $a = 0.73373,\ b = 0.30899,\ c = -0.32596.$



Notations

- The set of real numbers is denoted as $\mathbb{R}$.
- The set of all real-valued matrices of the size $n \times m$ is denoted as $\mathbb{R}^{n \times m}$.
- Formally (mathematically), a datum is represented as a column vector $\bar{x} = (x_1, x_2, \ldots, x_n)^T$. We also use a vector to represent a set of ordered variables, e.g., $\bar{x} = x_1, \ldots, x_n$.
- In this course, we use $\bar{x}[j]$ or x_j to indicate the j-th component of the vector $\bar{x}$.
- The set of integers is denoted as $\mathbb{I}$.
- We use capital letters to denote matrices.

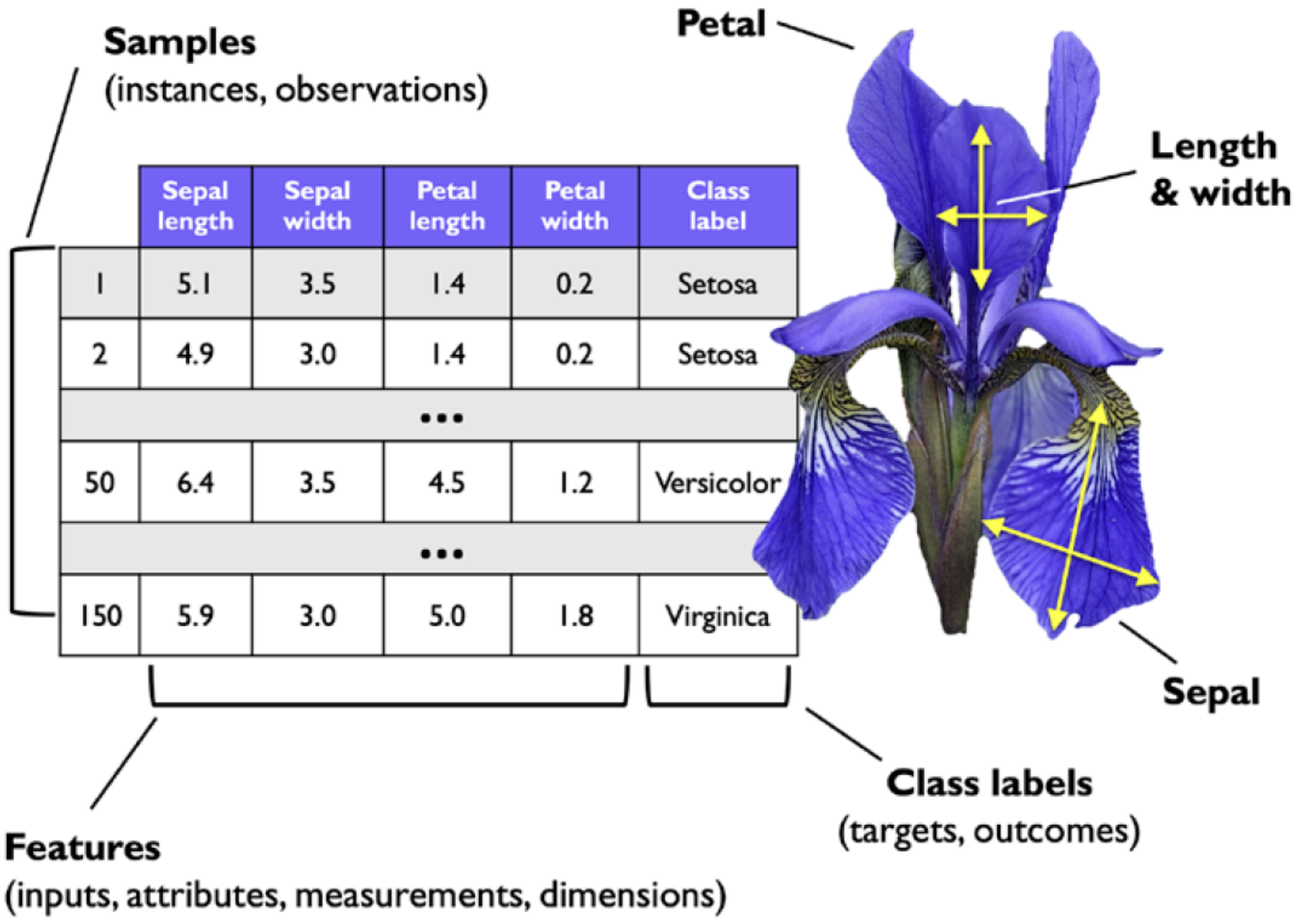
Data Representation in Python

Although a sample (datum) is represented as a column vector in math, it is more convenient to represent it as a **row vector (1-D array)** in Python.

Therefore, a set of data should be represented as a **matrix (2-D array)** in Python such that every row is a sample.

Data are not always 1-D in ML. For example, a black-white image is represented as a matrix, while an RGB color image is a 3-D matrix.

Iris Flower Dataset



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

[Textbook, Page 10]

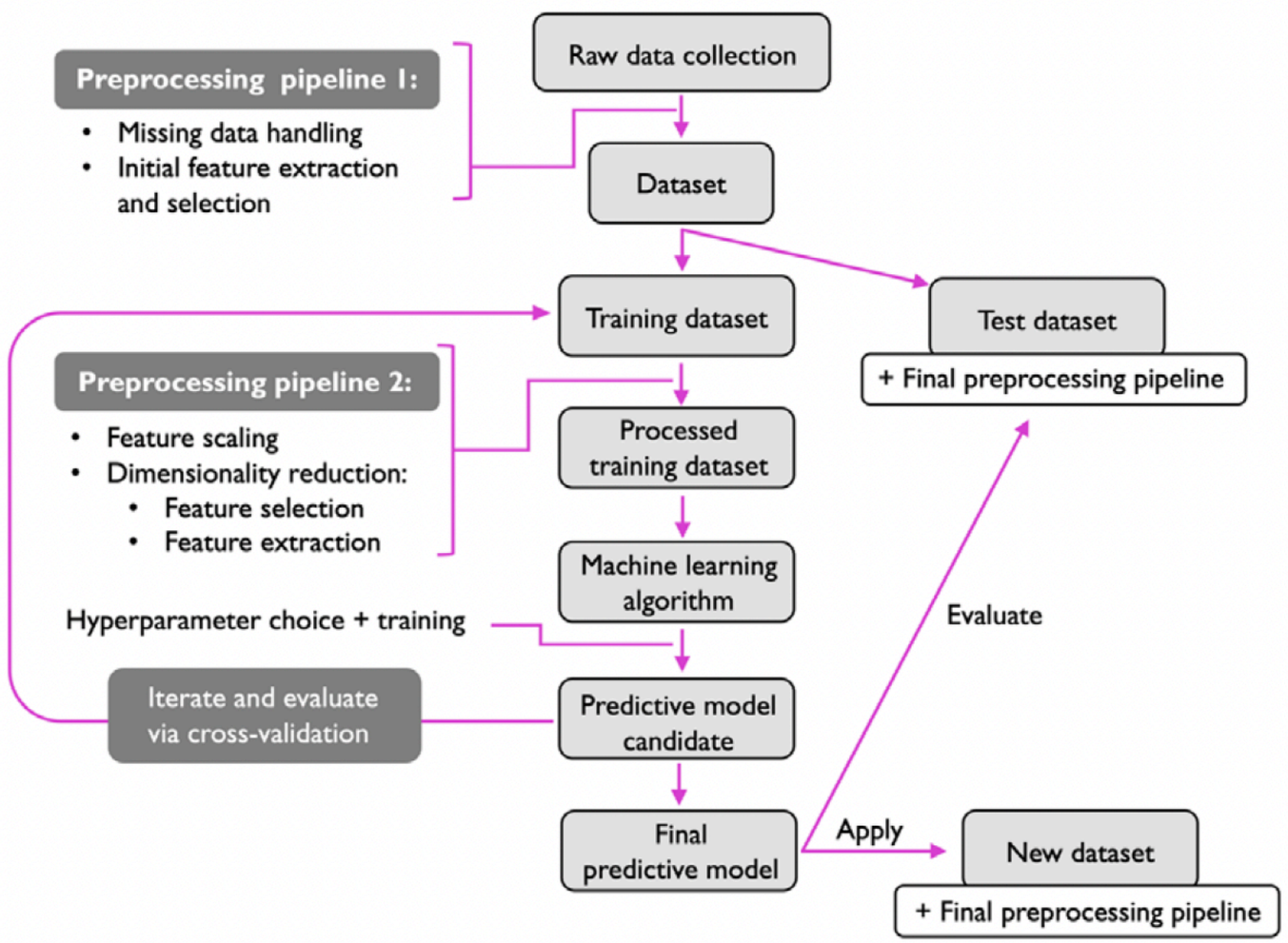
Reading Assignment

- **Python syntax**: assignments, conditions, loops, functions, classes, ...
- **NumPy**: arrays and their operations. *https://numpy.org/*
- **Matplotlib** - *https://matplotlib.org/*
  - Creating new figures.
  - Figure settings.
  - Plotting points, lines, shapes, functions, ...
  - Plot settings.
  - Reading the examples on the official website.

Machine Learning Terminology

- **Training:** Model fitting; Finding the unknown parameters for the ML model.
- **Training example/sample/datum:** Samples used for training an ML model.
- **Testing example/sample/datum:** Samples used for testing an ML model.
- **Feature:** The input of a sample.
- **Target:** The output of a sample; ground truth.
- **Prediction:** Output of the ML model.
- **Loss/cost function:** Error measurement between targets and predictions.

Roadmap for Building Machine Learning Systems



[Textbook, Page 12]

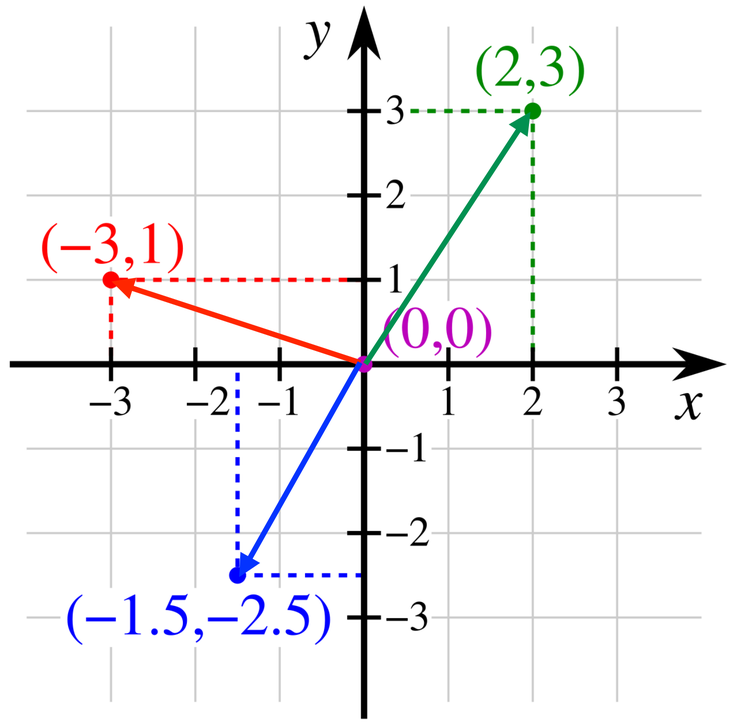
Basics of Analytic Geometry

Analytic Geometry in Machine Learning

- Multidimensional vector.
- Hyperplane and halfspace.
- Surface.
- Convex set.
- Convex/concave function.

Multidimensional Vector

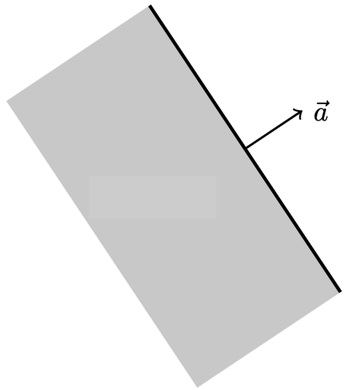
- An n-dimensional vector is denoted as a tuple $\bar{x} = (x_1, \ldots, x_n)^T$ of n components each of which indicates the coordinate in the corresponding dimension.
- Vectors sometimes are also treated as points.
- In Python, we represent $\bar{x} = (x_1, \ldots, x_n)^T$ as a 1-D NumPy array $[x_1, \ldots, x_n]$.

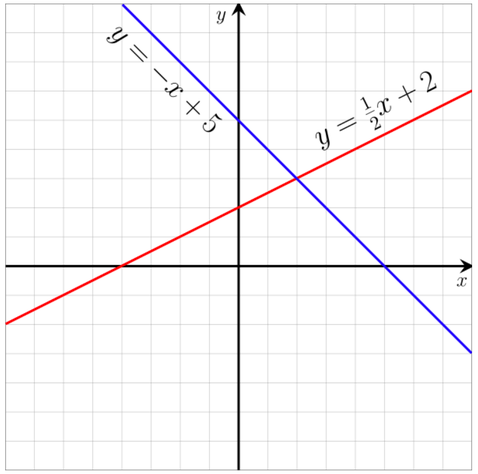


Hyperplane and Halfspace

- An n-dimensional **hyperplane** is the set of all points/vectors $(x_1, \ldots, x_n)^T$ that satisfy a linear equation of the form $a_1 x_1 + \cdots + a_n x_n = b$.
- The vector of the coefficients $(a_1, \ldots, a_n)^T$ is called the **normal vector** of the hyperplane. It is orthogonal to the hyperplane.
- The corresponding **halfspace** is the set of points satisfying the linear inequality $a_1 x_1 + \cdots + a_n x_n \leq b$.

$a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \leq b$



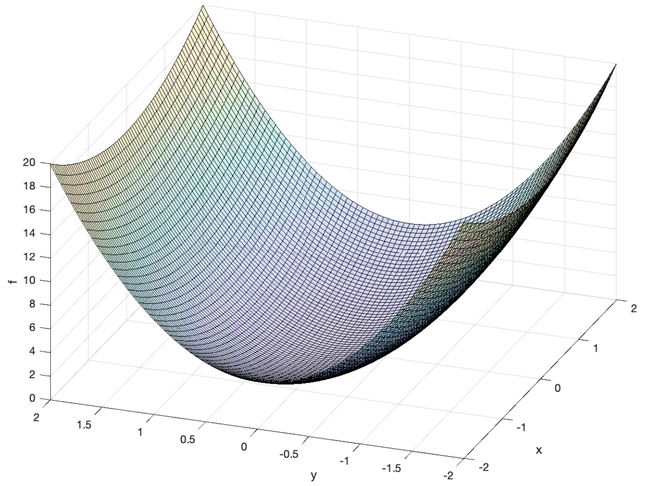


Nonlinear Surface

A nonlinear function $f(\bar{x}) = f(x_1, \ldots, x_{n-1})$ defines a **surface** in the Euclidean space $\mathbb{R}^n$.

Example: The equation $f(x_1, x_2) = x_1^2 + 4x_2^2$ defines the following surface in the space $\mathbb{R}^3$.

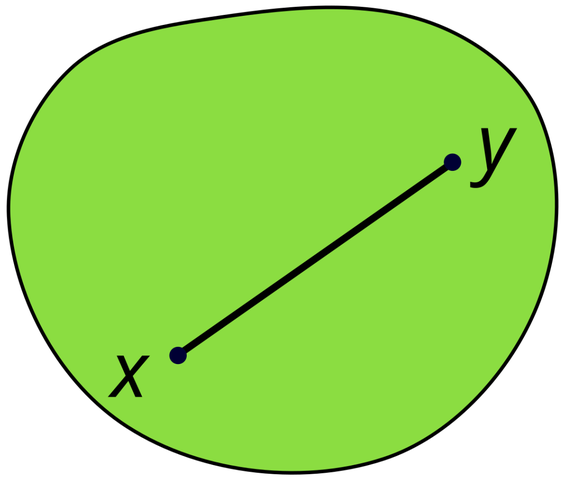
Question: What is the surface defined by $f(x_1, x_2) = x_1 + 4x_2$?



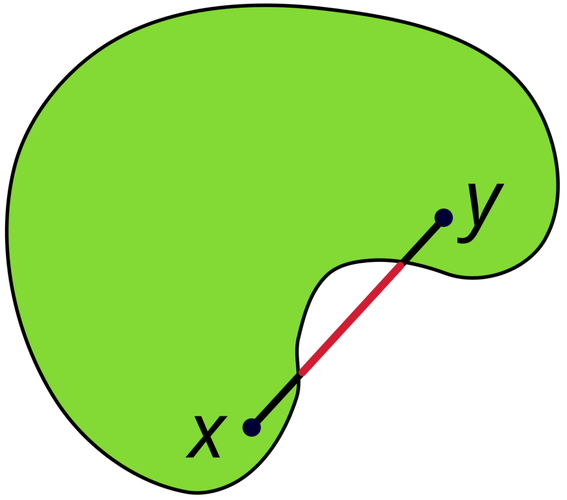
Convex Set

A set $S \subseteq \mathbb{R}^n$ is **convex**, if for all $\lambda \in [0,1]$ and $x, y \in S$, the point $\lambda x + (1-\lambda)y$ is still in the set.

In other words, the line segment $\{\lambda x + (1-\lambda)y \mid \lambda \in [0,1], x, y \in S\}$ is contained in S.



convex set

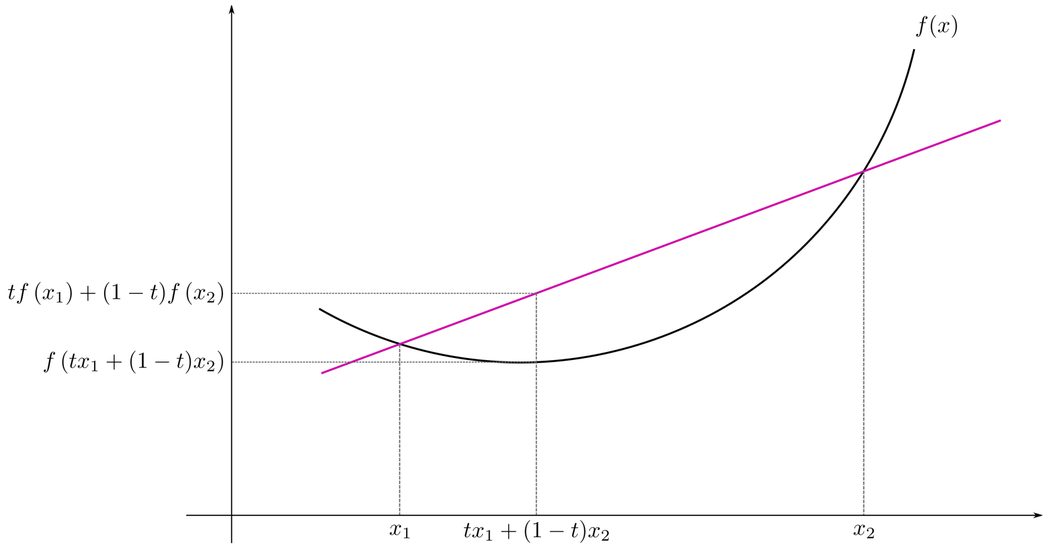


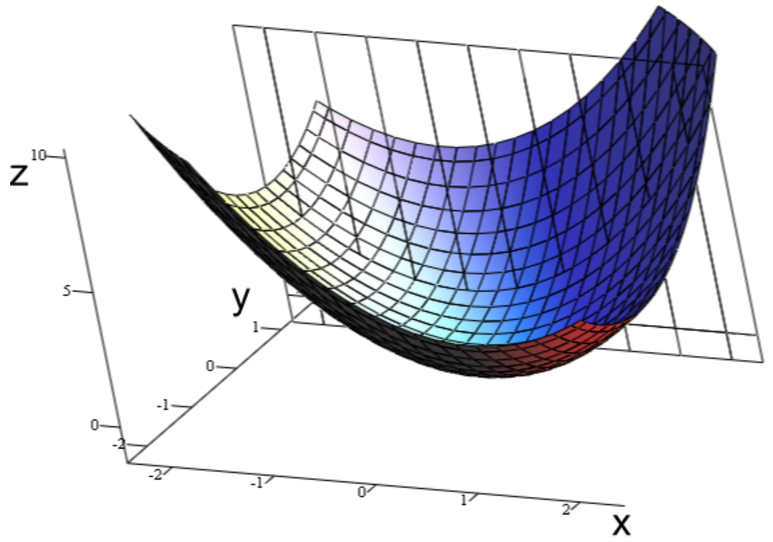
nonconvex set

Convex Function

A function $f(x)$ is **convex** over a convex set X, if for all $\lambda \in [0,1]$ and $x_1, x_2 \in X$, we have that

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$





Assignments and Projects

- All of the assignments and projects require you to perform the following three steps:
  - **Read:** Review the previous classes and read the given materials.
  - **Think:** Discuss the work plan with your teammates to find a solution to the problems/questions.
  - **Write:** Implement your solution, do experiments, write a report.