INTRODUCTION TO MACHINE LEARNING

Perceptron, Multilayer Perceptron (ANN), SVM

Hien Tran

Department of Mathematics
Center for Research in Scientific Computation
North Carolina State University

July 31, 2019

WHAT IS MACHINE LEARNING?

- Machine learning is a term invented by *Arthur Samuel*, an American pioneer in computer gaming and artificial intelligence, in 1959 while at IBM

"Machine learning teaches computers to do what comes naturally to humans and animals, learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases"

MathWorks, Natick, MA

MACHINE LEARNING

- The *essence* of machine learning
  - A *pattern* exists
    - How a bank gives credit approval to a customer depending on the customer's income relative to its debt, age, employment history, etc.
  - We *cannot* describe it mathematically
  - We have *data* on it (from previous borrowers)

DATA

- *Labeled* data

$$(x(1), y(1)), (x(2), y(2)), \ldots, (x(N), y(N))$$

$$x(i) = (\text{Gender}, \text{Age}, \text{Debt}, \text{MaritalStatus}, \text{BankCustomer}, \ldots) \in R^d$$

$$= \text{features}$$

$$y(i) = \begin{cases} \text{Approved} \\ \text{Denied} \end{cases}$$

Labels/outcomes/endpoints

- *Unlabeled* data

$$(x(1), x(2), \ldots, x(N))$$

MACHINE LEARNING

Labeled data (Supervised Learning)

Classification (discrete responses)

Nearest Neighbors (kNN)
Naïve Bayes Classifier
Perceptron (Neural Networks)
Support Vector Machines
Decision Trees
Discriminant Analysis

Regression (continuous responses)

Linear Regression
Generalized Linear Models
Perceptron (Neural Networks)
Support Vector Regression
Decision Trees
Ensemble Methods

Unlabeled data (Unsupervised Learning)

Clustering

K-Means
Gaussian Mixtures
Neural Networks
Hidden Markov Model

PERCEPTRON

- *Frank Rosenblatt* published the first model of a perceptron in 1958

Input (*features*)

$$x(i) = \begin{cases} \text{Gender} \\ \text{Age} \\ \text{Debt} \\ \text{MaritalStatus} \\ \vdots \\ \text{Income} \end{cases}$$

Output (*labels*)

$$y(i) = \begin{cases} 1 & \text{Approved} \\ -1 & \text{Denied} \end{cases}$$

Target Function ("ideal" credit approval formula)

$$f : \mathcal{X}(\text{input}) \rightarrow \mathcal{Y}(\text{output})$$

This function is unknown. ML "learns" the target function from the data

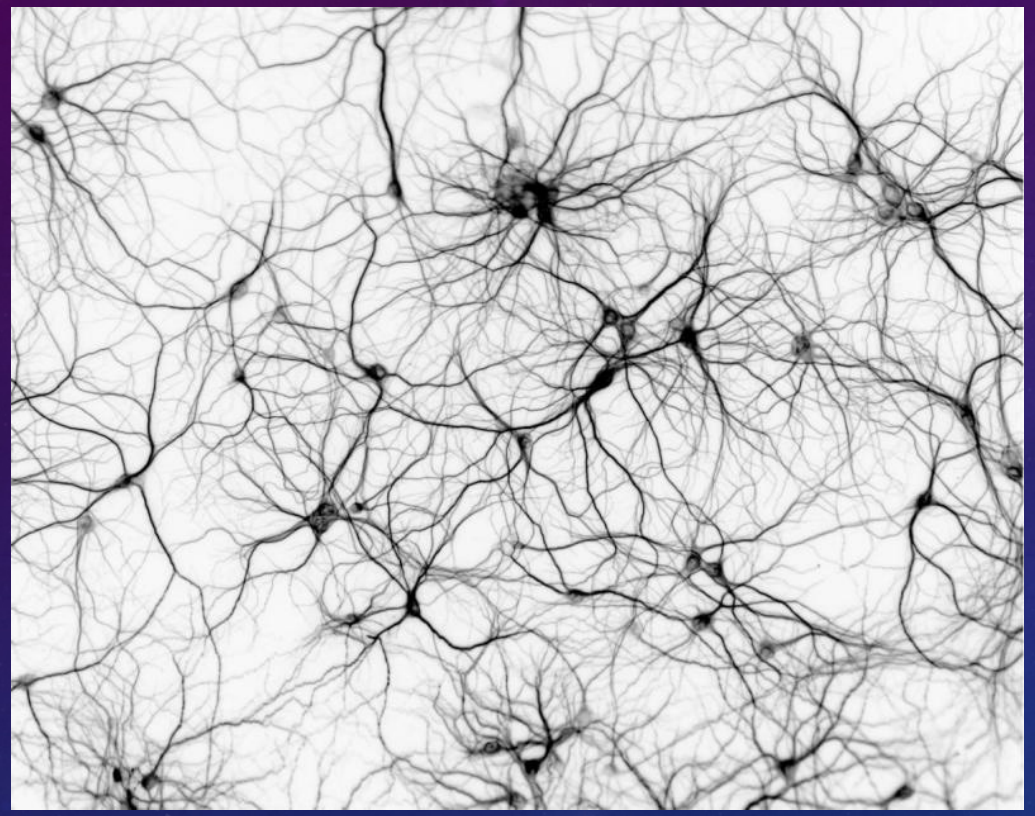
PERCEPTRON MODEL

- Some features are important for credit approval (e.g., income), while outstanding debt is not good for credit approval
- Idea of Perceptron is to weight the features according to their importance or not
- *Approve* credit if

$$\sum_{j=1}^{d} w_j x_j > \text{threshold}$$

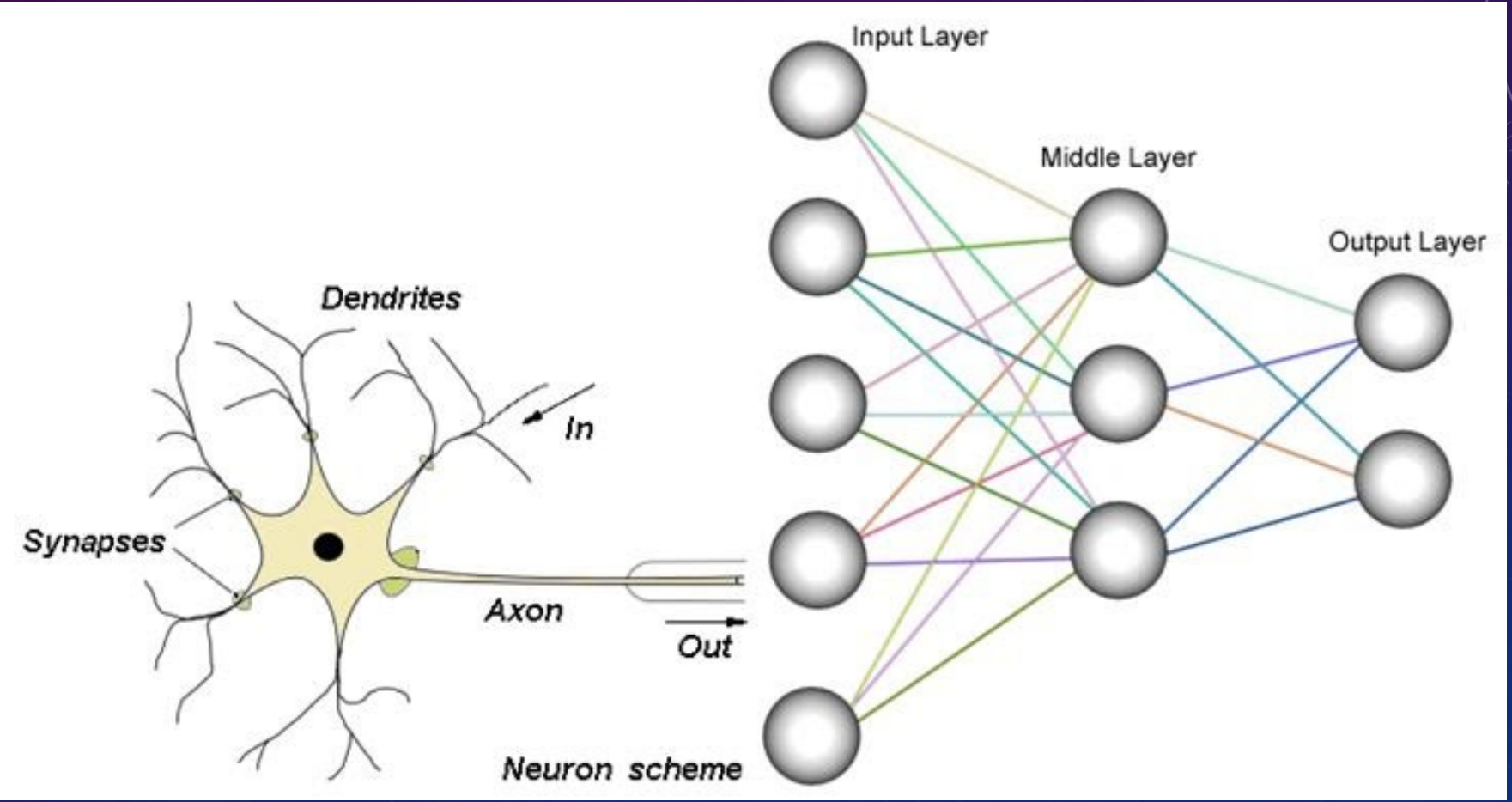
w_j and threshold are unknown and are "learned" from data

MULTILAYER PERCEPTRON (ARTIFICIAL NEURON NETWORK)

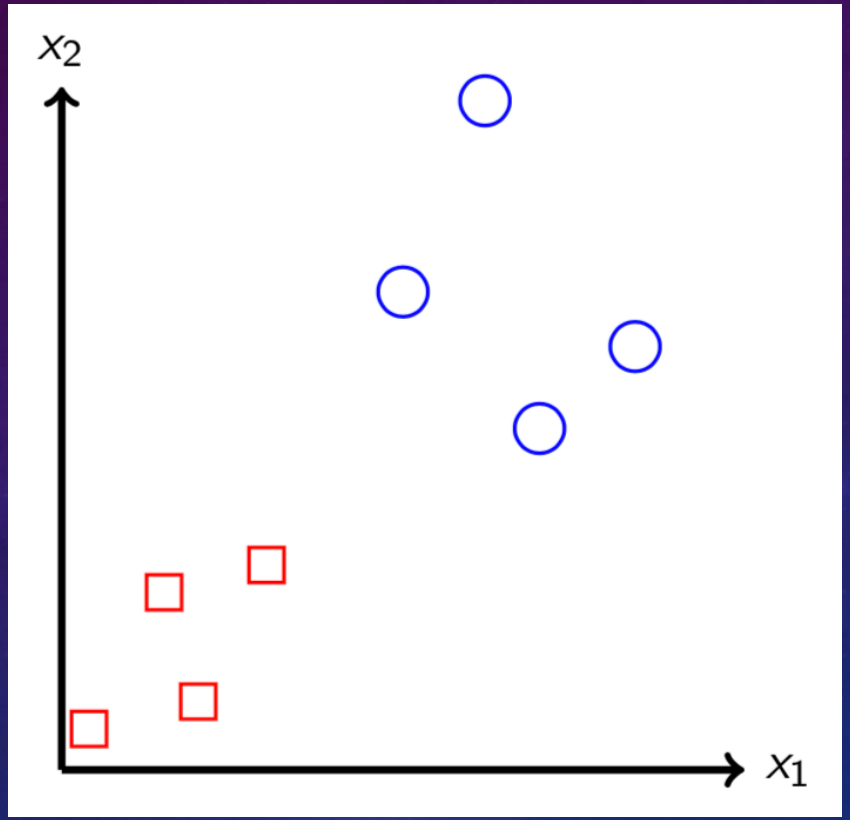


- *Biological* neural network (brain) has about 100 billion neurons, which are connected through synapses
- Each neuron receives thousands of connections with other neurons
- If the *resulting sum of the signals surpasses a certain threshold*, a signal is sent through the axon
  - *A perceptron can be thought of as a single neuron*

BIOLOGICAL VS. ARTIFICIAL NEURON NETWORK

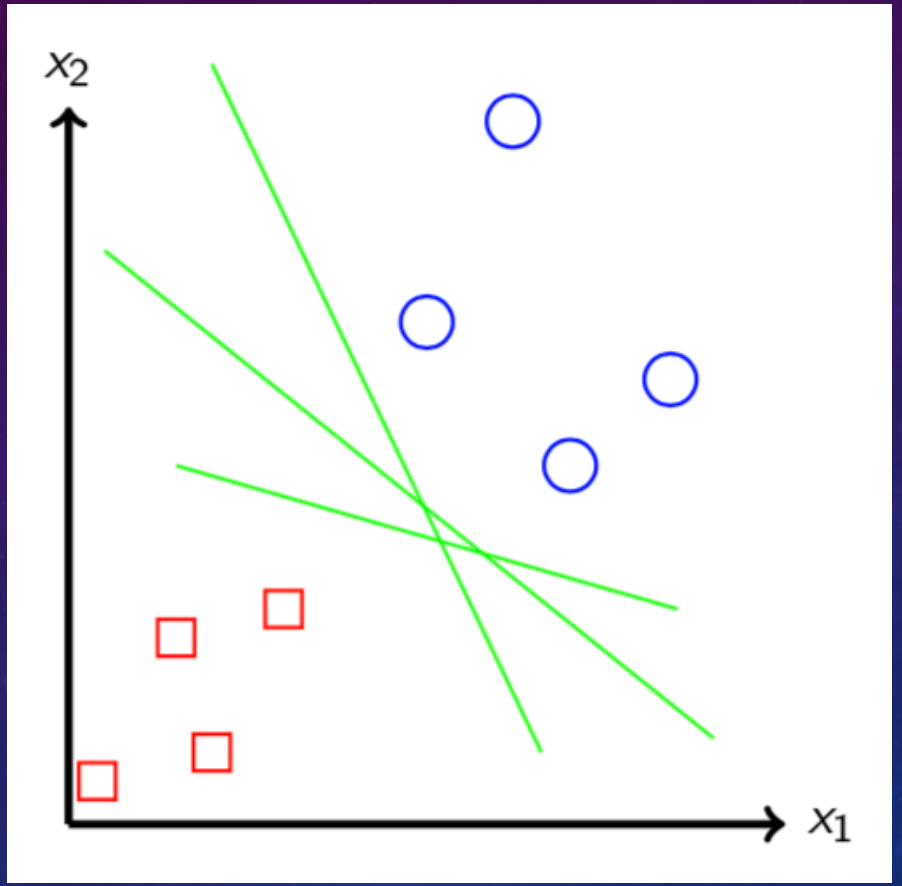


SUPPORT VECTOR MACHINES (SVM)



- Given two classes of linearly separable data, y_0 and y_1 for a given feature vector $\vec{x}$

SUPPORT VECTOR MACHINES (SVM)



- The goal is to not just find any decision plane $d(x) = \vec{w} \cdot \vec{x} + b$ that separates the two

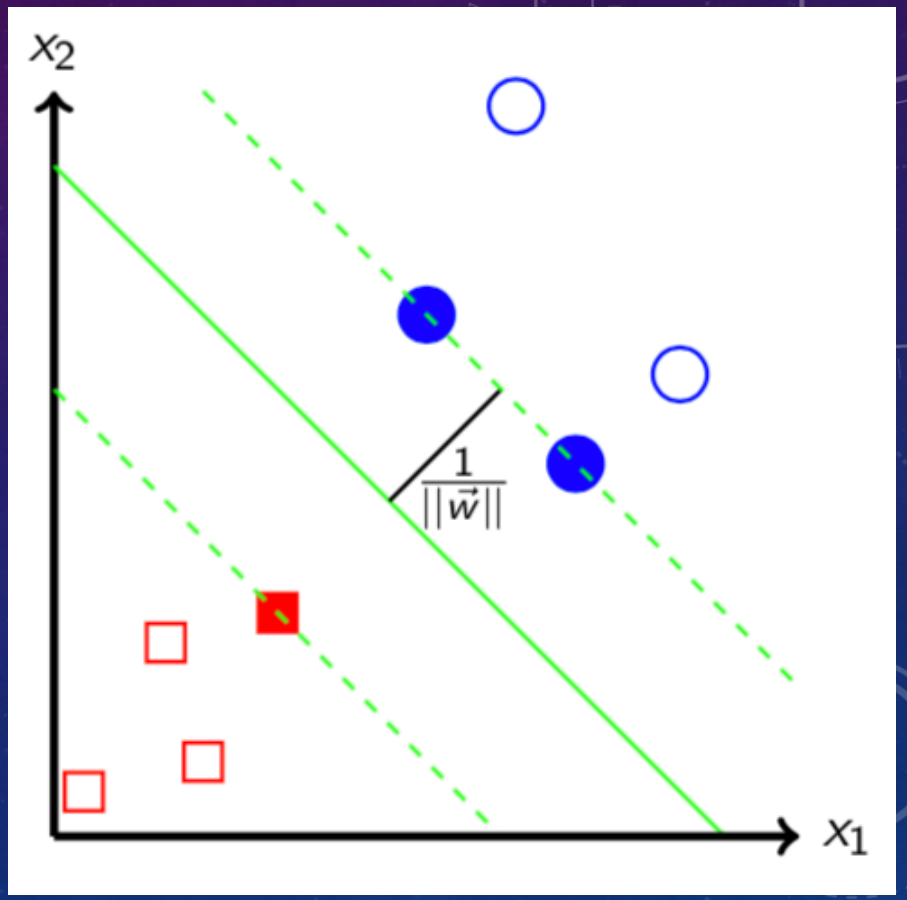
SUPPORT VECTOR MACHINES (SVM)

- But to find the best hyperplane that **maximizes the distance between them**

Constrained Quadratic Optimization Problem

Min $\frac{1}{2}||\vec{w}||^2$

subject to $y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1$



LINEARLY NONSEPARABLE (SOFT MARGIN SVM)

- Introduce a slack variable ξ_i

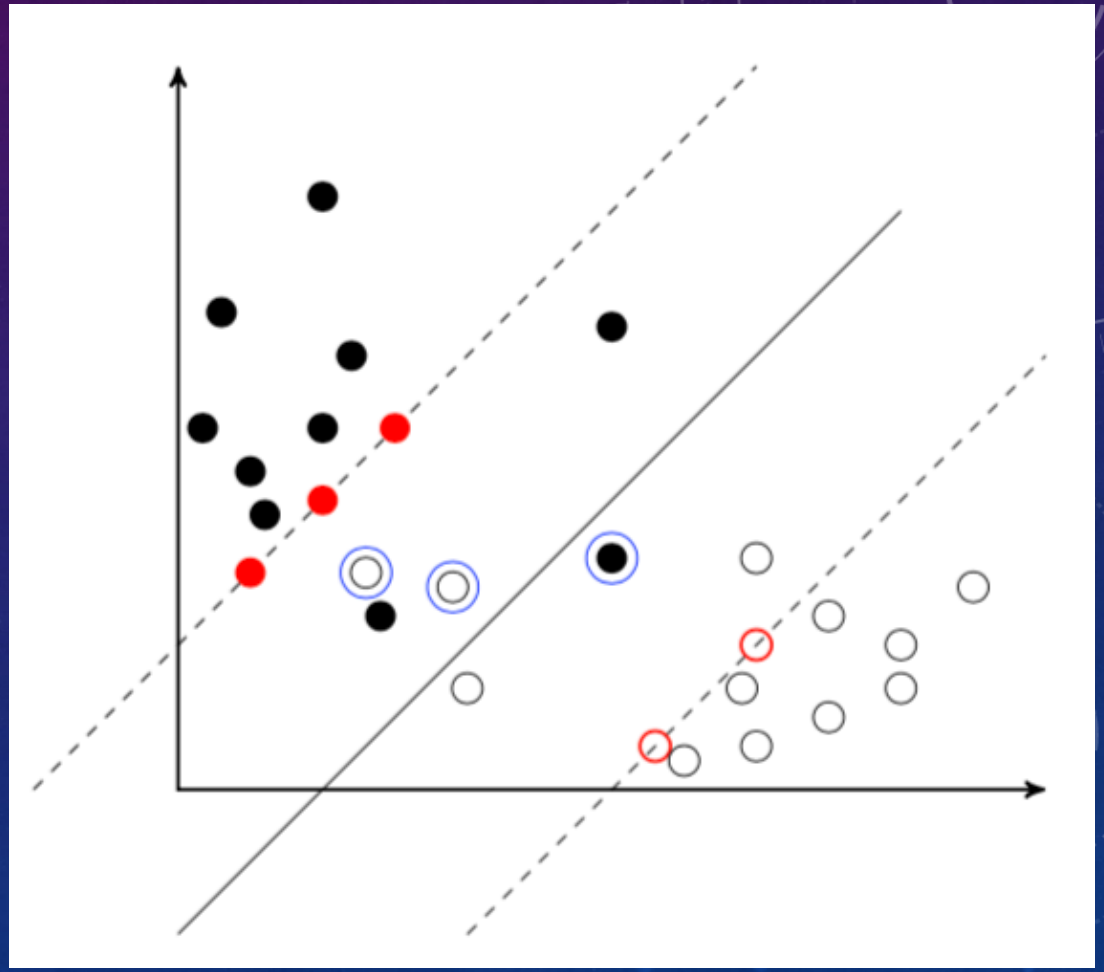
→ **A New Optimization Problem**

$$\text{Min} \quad \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{l}\xi_i$$

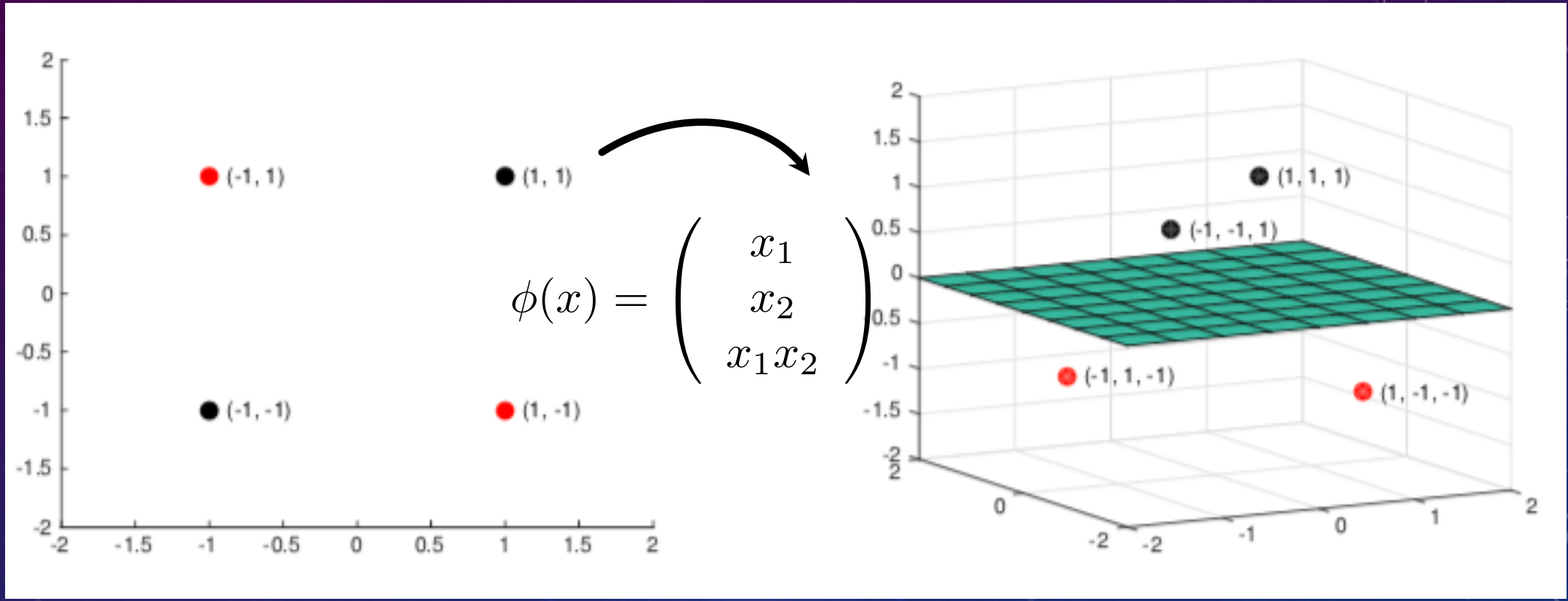
$$\text{subject to} \quad y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 - \xi_i$$

$$0 \leq \xi_i$$

- *C* controls the trade-off between
  - Maximizing margin
  - Having fewer of misclassification



NONLINEARLY SEPARABLE (KERNEL SVM)



- Use the Gaussian Kernel $K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2}$ to map data into a separable domain