

# The recently announced Earth Simulator Computer:

$n = 1,041,216;$

(8.7 TB of memory)

5.8 hours to run

35 TFlop/s

# Support Vector Machines: The Latest in Learning Algorithms

- Not a computer architecture – an algorithm!
- “3 is prime, 5 is prime, 7 is prime, 9 is experimental error, 11 is prime, 13 is prime, ...

- The latest way to separate
- or 1’s from 7’s



‘s from



‘s

- or



‘s from



‘s

# Supervised Learning

- Traditional programming: Given  $x$  as input, produce  $y=h(x)$  as output. The programmer and then the software know “ $h$ ” intimately.
- Supervised Learning: Given pairs  $(x,y)$  e.g. (picture of tank, “tank”), produce an “ $h$ ” that works well with high probability
- Unsupervised Learning: Given  $x$  figure out the pattern

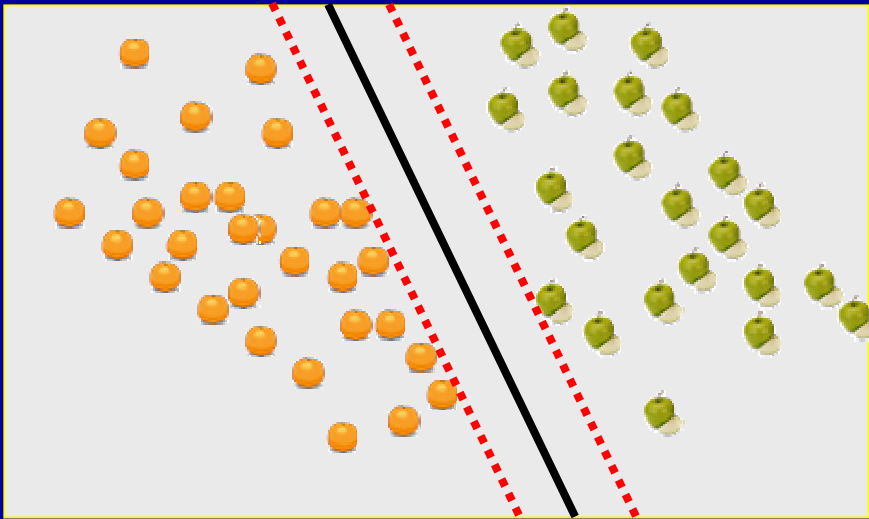
# Binary, Discrete, Continuous learning

- $(x, y)$   $h(x)=y \in \pm 1$  (Binary Classification)
- $(x, y)$   $h(x)=y \in \{1, 2, 3, \dots, n\}$  (Discrete)
- $(x, y)$   $h(x)=y \in \mathbb{R}^n$  (Continuous)
  - For example if  $x$  is a collection of data points,  $y$  could be the slope and intercept of the best fit line  $\rightarrow$  Regression

Issues: Complexity and Accuracy: Should not be so complex that each example is directly built into  $x$ .  
Need not be perfectly accurate since data comes with noise, etc.

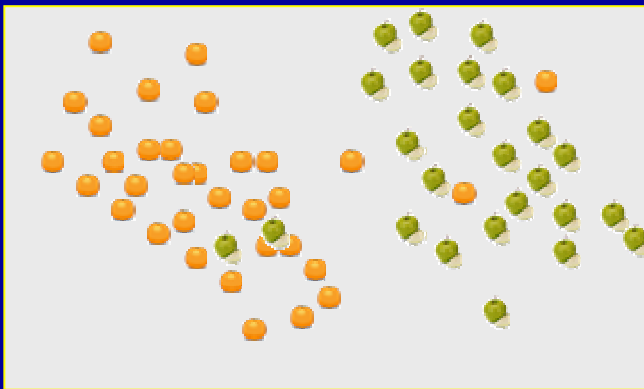
# Linear Classification

Examples in  $\mathbb{R}^2$ :



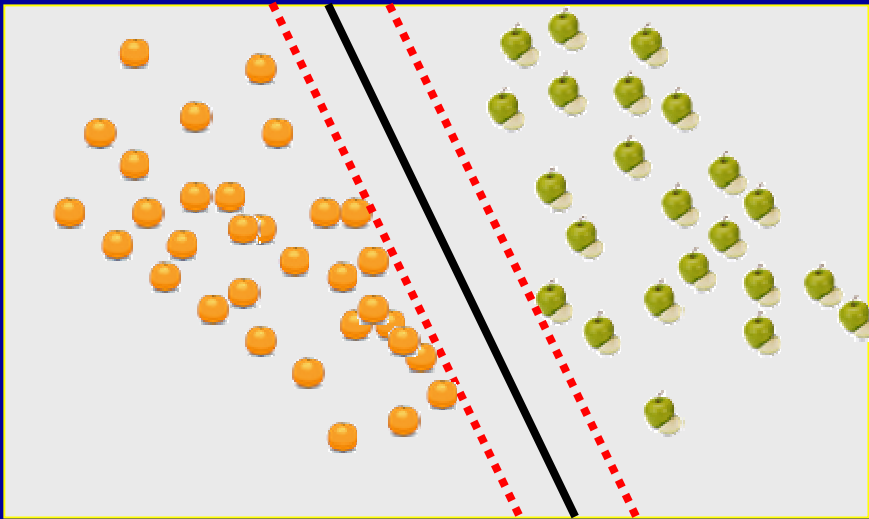
Separating Hyperplanes!

These points are linearly separable (a hyperplane exists)



These points are not,  
but all is not lost

Examples in  $\mathbb{R}^2$ :



Assume Separating Hyperplanes!

Find a separating hyperplane: Rosenblatt's Perceptron (1956)

$$h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad h(\text{orange}) \leq -1 \quad h(\text{green}) \geq +1$$

Find the best separating hyperplane: Maximize  $1/\|\mathbf{w}\|$

$$\begin{aligned} & \underset{\mathbf{w}, b}{\text{minimize}} \quad \|\mathbf{w}\|^2/2 \\ & \text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m \end{aligned}$$

# Non-separable training sets

Separable:

$$\begin{aligned} & \underset{\mathbf{w}, \mathbf{b}}{\text{minimize}} \quad \|\mathbf{w}\|^2/2 \\ & \text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + \mathbf{b}) \geq 1, \quad i=1, \dots, m \end{aligned}$$

Non separable: Add slack variables  $\varepsilon_i$  :

$$\begin{aligned} & \underset{\mathbf{w}, \mathbf{b}, \varepsilon}{\text{minimize}} \quad \|\mathbf{w}\|^2/2 + c \sum \varepsilon_i \\ & \text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + \mathbf{b}) \geq 1 - \varepsilon_i, \quad i=1, \dots, m \end{aligned}$$

Non separable: Non-linearly distort space:

$$\begin{aligned} & \underset{\mathbf{w}, \mathbf{b}}{\text{minimize}} \quad \|\mathbf{w}\|^2/2 \\ & \text{s.t.} \quad y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + \mathbf{b}) \geq 1, \quad i=1, \dots, m \end{aligned}$$

# The dual problem

Separable:

$$\begin{aligned} & \underset{\mathbf{w}, \mathbf{b}}{\text{minimize}} \quad \|\mathbf{w}\|^2/2 \\ & \text{s.t.} \quad y_i(\mathbf{w}^T \mathbf{x}_i + \mathbf{b}) \geq 1, \quad i=1, \dots, m \end{aligned}$$

PRIMAL

Introduce Lagrange Multipliers:  $\alpha_i$

$$\begin{aligned} & \underset{\alpha}{\text{maximize}} \quad \alpha^T \mathbf{1} - \alpha^T \mathbf{H} \alpha / 2 \\ & \text{s.t.} \quad \mathbf{y}^T \alpha = 0, \quad \alpha \geq 0 \end{aligned}$$

DUAL

At optimality  $\mathbf{w} = \sum y_i \alpha_i \mathbf{x}_i$

$$H_{ij} = y_i (\mathbf{x}_i^T \mathbf{x}_j) y_j$$

Distorted space version:  $H_{ij} = y_i (\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)) y_j$

Slack Variable version:  $\mathbf{y}^T \alpha = 0, \quad c \geq \alpha \geq 0$

# Solving the dual problem

$$\begin{array}{ll} \text{maximize} & \alpha^T 1 - \alpha^T H \alpha / 2 \\ & \alpha \\ \text{s.t.} & y^T \alpha = 0, \alpha \geq 0 \end{array}$$

Quadratic programming problem!

Tony has a projected conjugate gradient approach!