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 ‘s from ‘s
• or ‘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” intimitely.

• 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) \( h(x) = y \in \{1, 2, 3, \ldots, n\} \) (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)
$$h(x) = w^T x + b \quad h(\bullet) \leq -1 \quad h(\bigcirc) \geq +1$$

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

\[
\begin{align*}
\text{minimize} & \quad ||w||^2 / 2 \\
\text{w,b} & \\
\text{s.t.} & \quad y_i(w^T x_i + b) \geq 1, \; i=1,\ldots,m
\end{align*}
\]
Non-separable training sets

Separable:
\[
\begin{align*}
&\text{minimize } \frac{||w||^2}{2} \\
&w, b \\
&\text{s.t. } y_i(w^T x_i + b) \geq 1, \ i=1,\ldots,m
\end{align*}
\]

Non separable: Add slack variables $\varepsilon_i$:
\[
\begin{align*}
&\text{minimize } \frac{||w||^2}{2} + c\sum \varepsilon_i \\
&w, b, \varepsilon \\
&\text{s.t. } y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \ i=1,\ldots,m
\end{align*}
\]

Non separable: Non-linearly distort space:
\[
\begin{align*}
&\text{minimize } \frac{||w||^2}{2} \\
&w, b \\
&\text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1, \ i=1,\ldots,m
\end{align*}
\]
The dual problem

Separable: minimize $\frac{\|w\|^2}{2}$
subject to $y_i(w^T x_i + b) \geq 1, i=1,..,m$

Introduce Lagrange Multipliers: $\alpha_i$

maximize $\alpha^T 1 - \frac{\alpha^T H\alpha}{2}$
subject to $y^T \alpha = 0, \alpha \geq 0$

At optimality $w = \sum y_i \alpha_i x_i$

Distorted space version: $H_{ij} = y_i (\phi(x_i)^T \phi(x_j)) y_j$
Slack Variable version: $y^T \alpha = 0, c \geq \alpha \geq 0$
Solving the dual problem

maximize $\alpha^T 1 - \alpha^T H \alpha / 2$

$s.t. \ y^T \alpha = 0, \ \alpha \geq 0$

Quadratic programming problem!
Tony has a projected conjugate gradient approach!