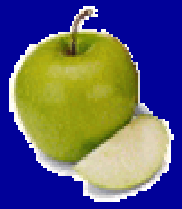
Support Vector Machines:The Latest in Learning Algorithms

- 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 1's from 7'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

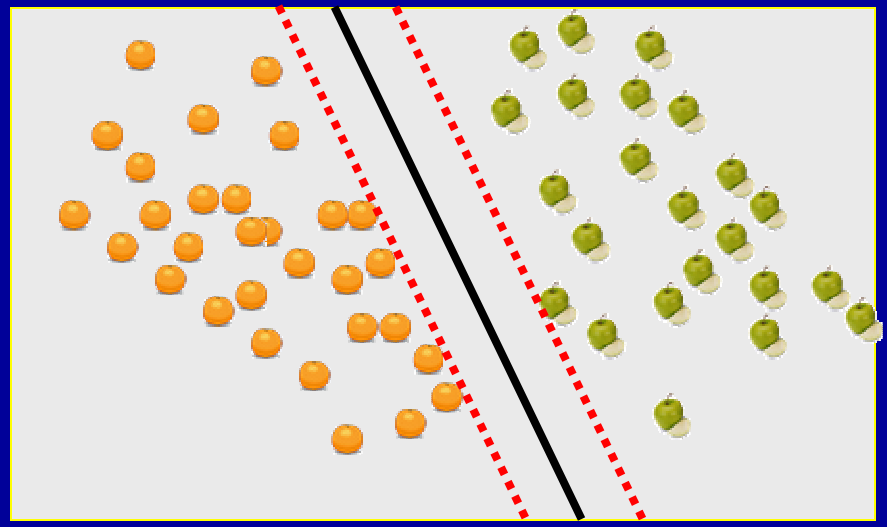
- (x,y) $h(x)=y \in \pm 1$ (Binary Classification)
- (x,y) $h(x)=y \in \{1,2,3,\ldots,n\}$ (Discrete)
- (x,y) $h(x)=y \in R^n$ (Continuous)

  - For example if x is a collection of data points, y could be the slope and intercept of the best fit line → 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.

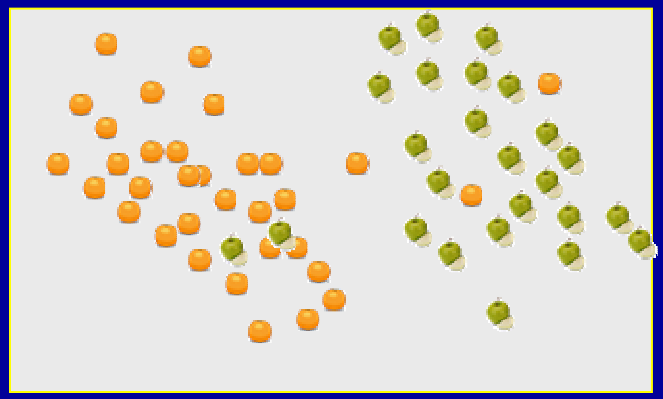
Linear Classification

Examples in R^2:



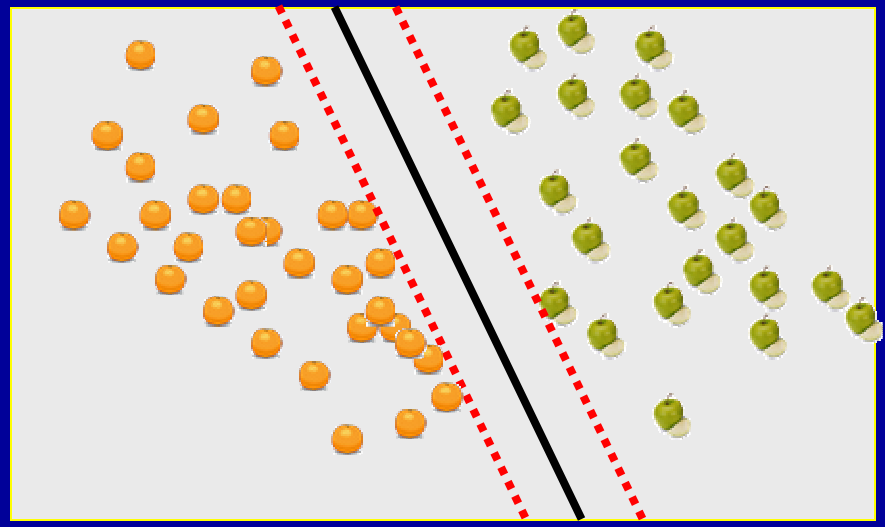
Separating Hyperplanes!

These points are linearly separable (a hyperplane exists)



These points are not,
but all is not lost

Examples in R^2:



Assume Separating Hyperplanes!

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

$h(x)=w^Tx+b$ $h(\text{●})\leq -1$ $h(\text{●}) \geq +1$

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

$$\underset{w,b}{\text{minimize}} \ \|w\|^2/2$$

$$\text{s.t. } y_i(w^Tx_i+b) \geq 1,\ I=1,..,m$$

Non-separable training sets

Separable:

$$\min_{\mathbf{w},b} \|w\|^2/2$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1,\ i=1,..,m$$

Non separable: Add slack variables ε_i :

$$\min_{\mathbf{w},b,\boldsymbol{\varepsilon}} \|w\|^2/2 + c\Sigma\, \varepsilon_i$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \varepsilon_i,\ i=1,..,m$$

Non separable: Non-linearly distort space:

$$\min_{\mathbf{w},b} \|w\|^2/2$$
$$\text{s.t. } y_i(w^T \phi(x_i) + b) \geq 1,\ i=1,..,m$$

The dual problem

Separable:

$$\min_{\mathbf{w},b} \|w\|^2/2 \quad \text{s.t. } y_i(w^T x_i + b) \geq 1,\ i=1,..,m$$

PRIMAL

Introduce Lagrange Multipliers: α_i

$$\max_{\alpha} \alpha^T 1 - \alpha^T H \alpha/2 \quad \text{s.t. } y^T\alpha=0,\ \alpha \geq 0$$

DUAL

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

$$H_{ij} = y_i (x_i^T x_j) y_j$$

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

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

Quadratic programming problem!

Tony has a projected conjugate gradient approach!