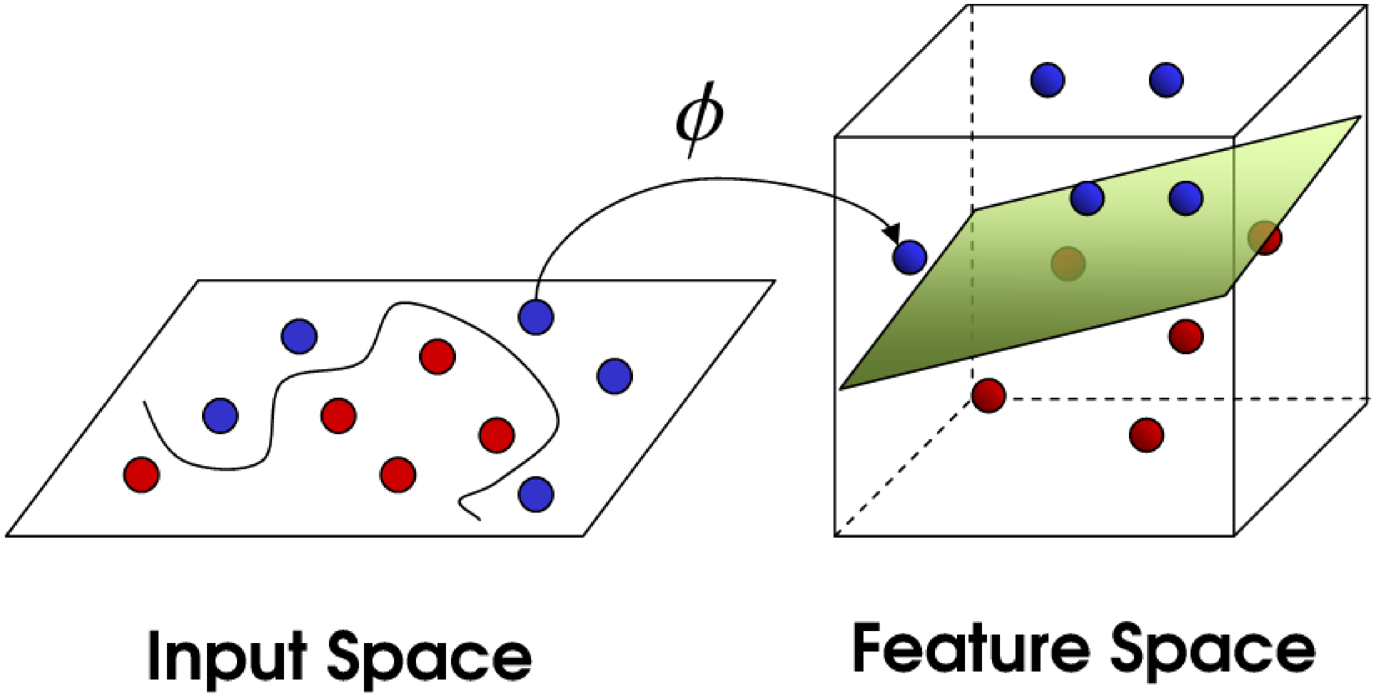
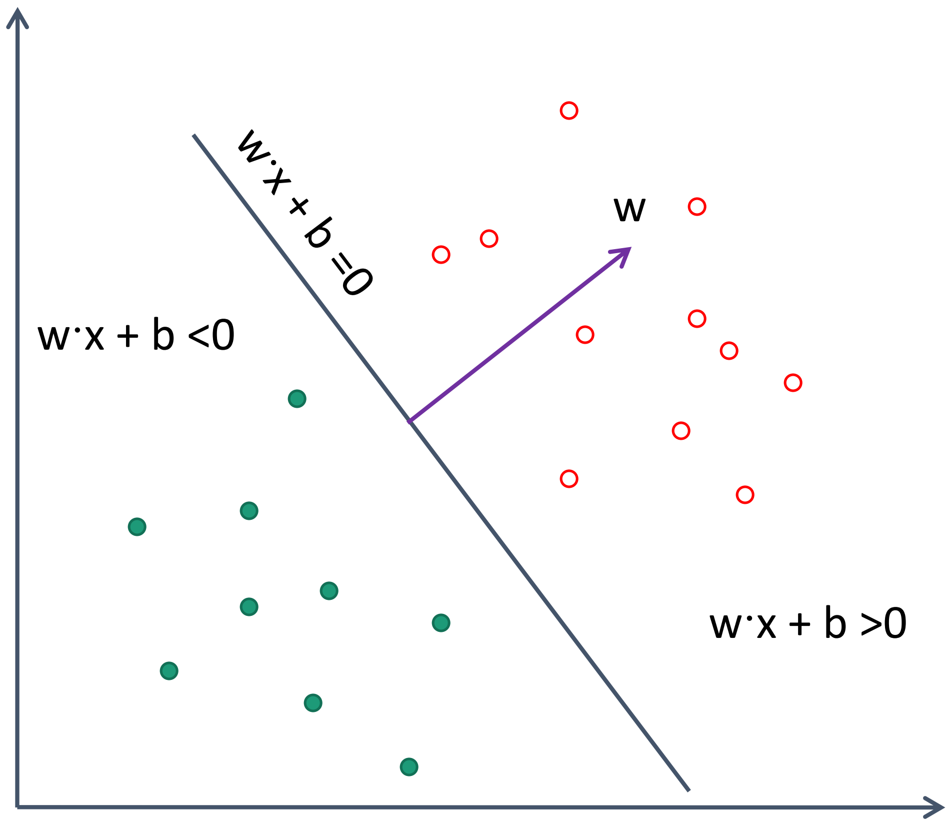
大数据科学与应用技术

Lecture：焦在滨（ Zaibin JIAO ）
Email：jiaozaibin@mail.xjtu.edu.cn

Overview



Linear Classifier



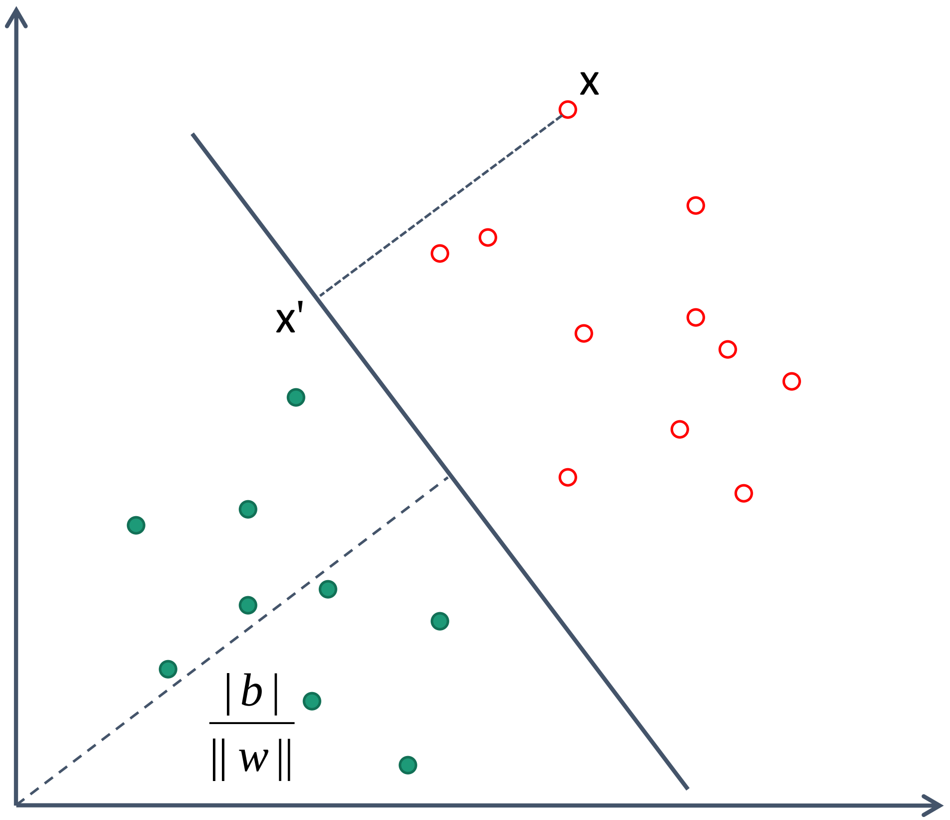
$$f(x, w, b) = sign(g(x))$$
$$= sign(w \cdot x + b)$$

Just in case ...

$$w \cdot x = \sum_{i=1}^{n} w_i x_i$$

$$w \cdot x_1 + b = w \cdot x_2 + b$$
$$w(x_1 - x_2) = 0$$

Distance to Hyperplane



$$g(x) = w \cdot x + b$$

$$x = x' + \lambda w$$

$$g(x) = w(x' + \lambda \cdot w) + b$$

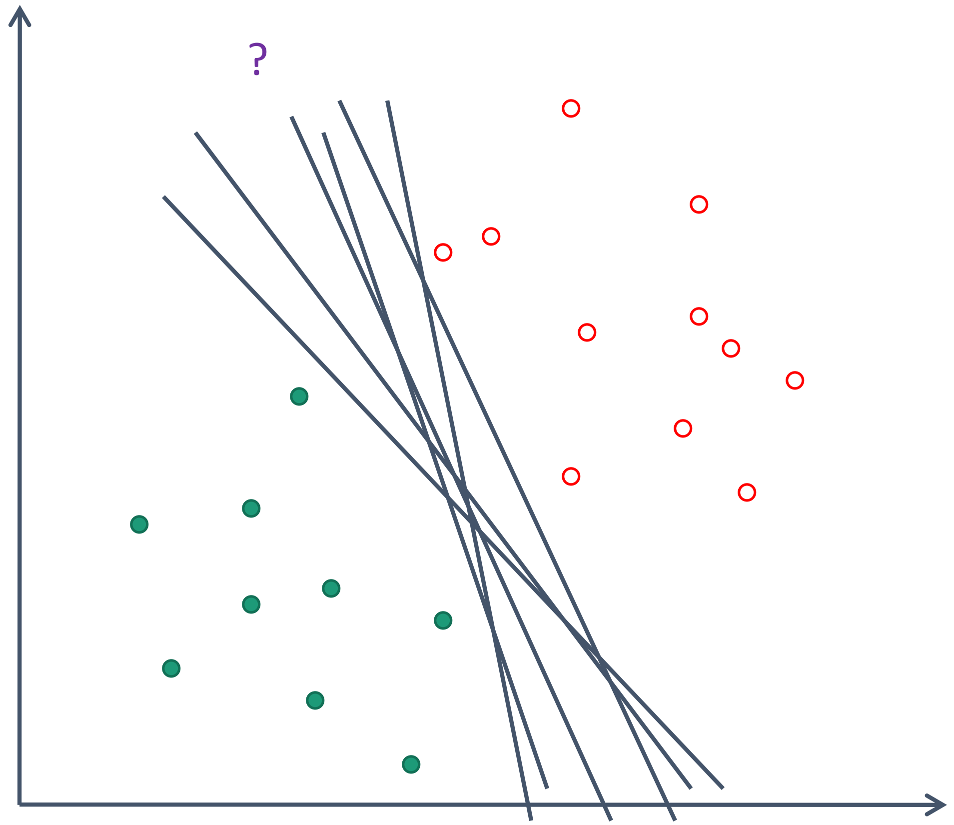
$$= w \cdot x' + b + \lambda w \cdot w$$

$$= \lambda w \cdot w$$

$$M = \| x - x' \| = \| \lambda w \|$$

$$= \frac{|g(x)| \times \| w \|}{w \cdot w} = \frac{|g(x)|}{\| w \|}$$

Selection of Classifiers

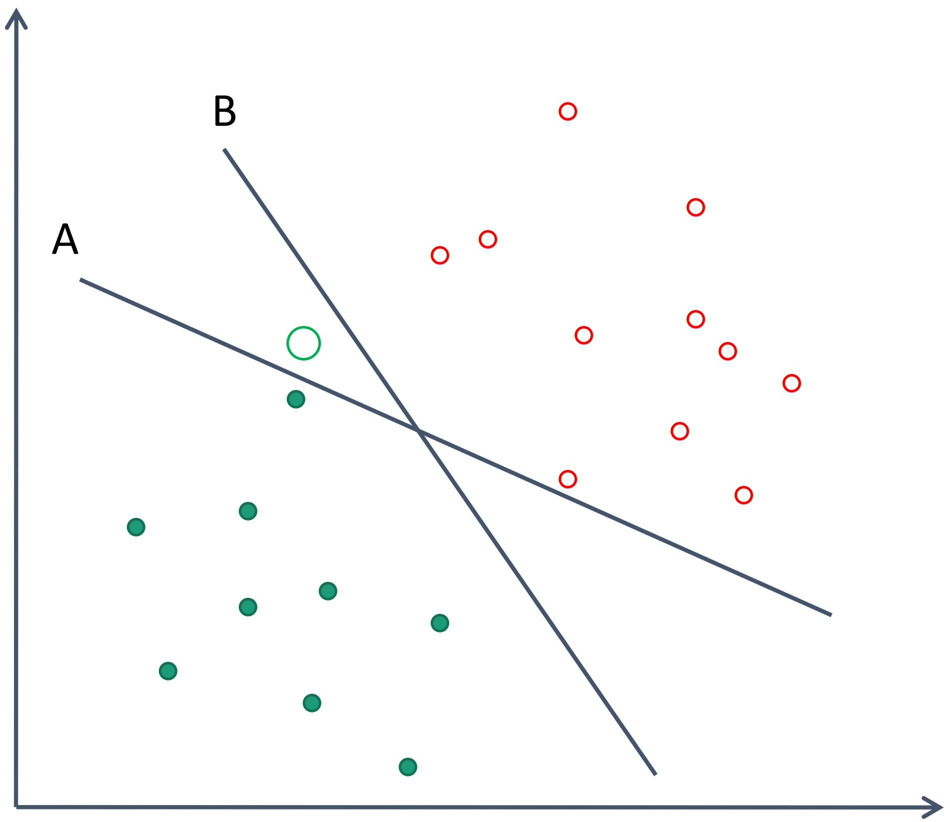


Which classifier is the best?

All have the same training error.

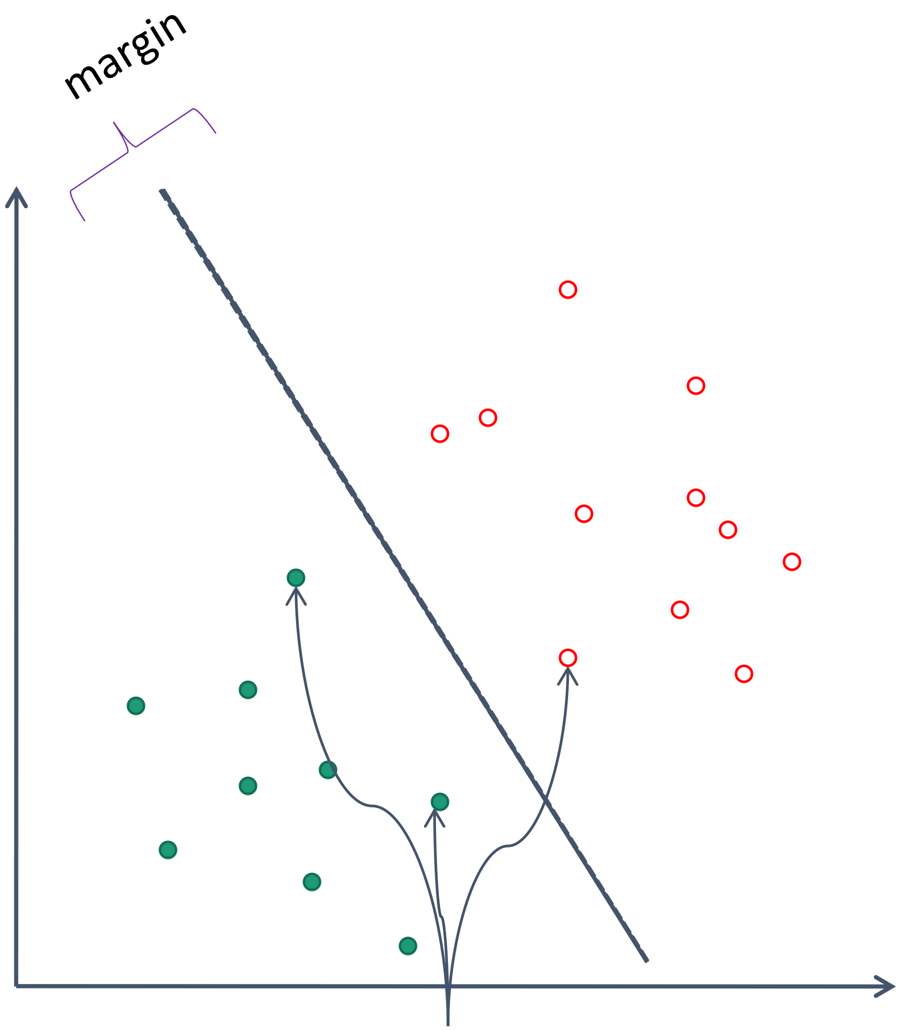
How about generalization?

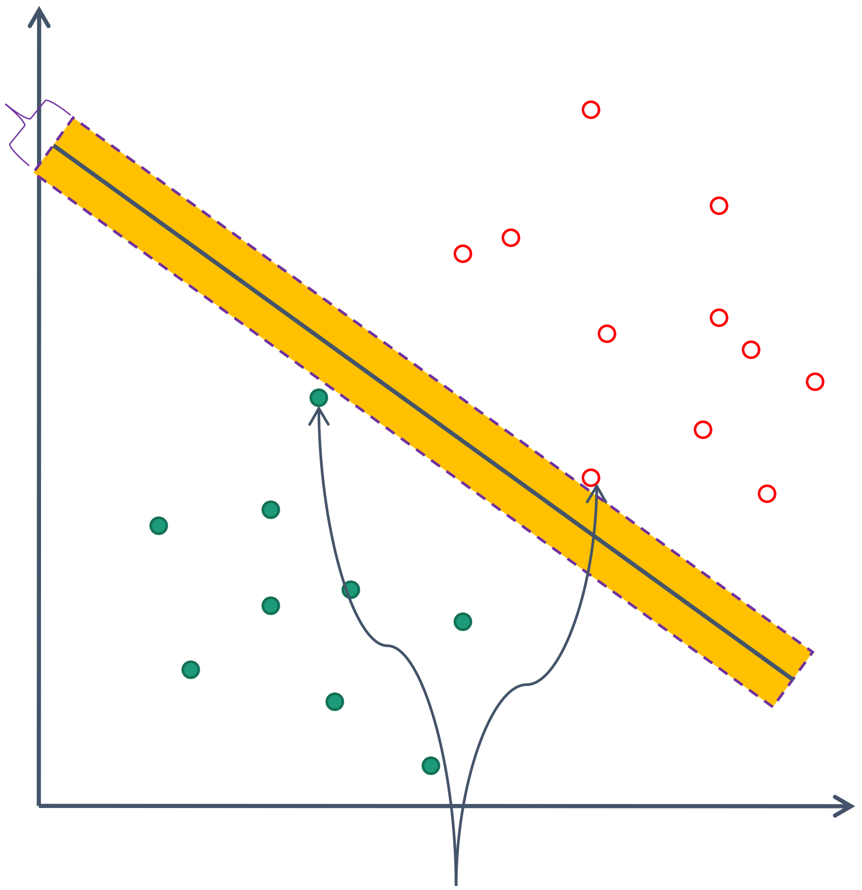
Unknown Samples



Classifier B divides the space more consistently (unbiased).

Margins

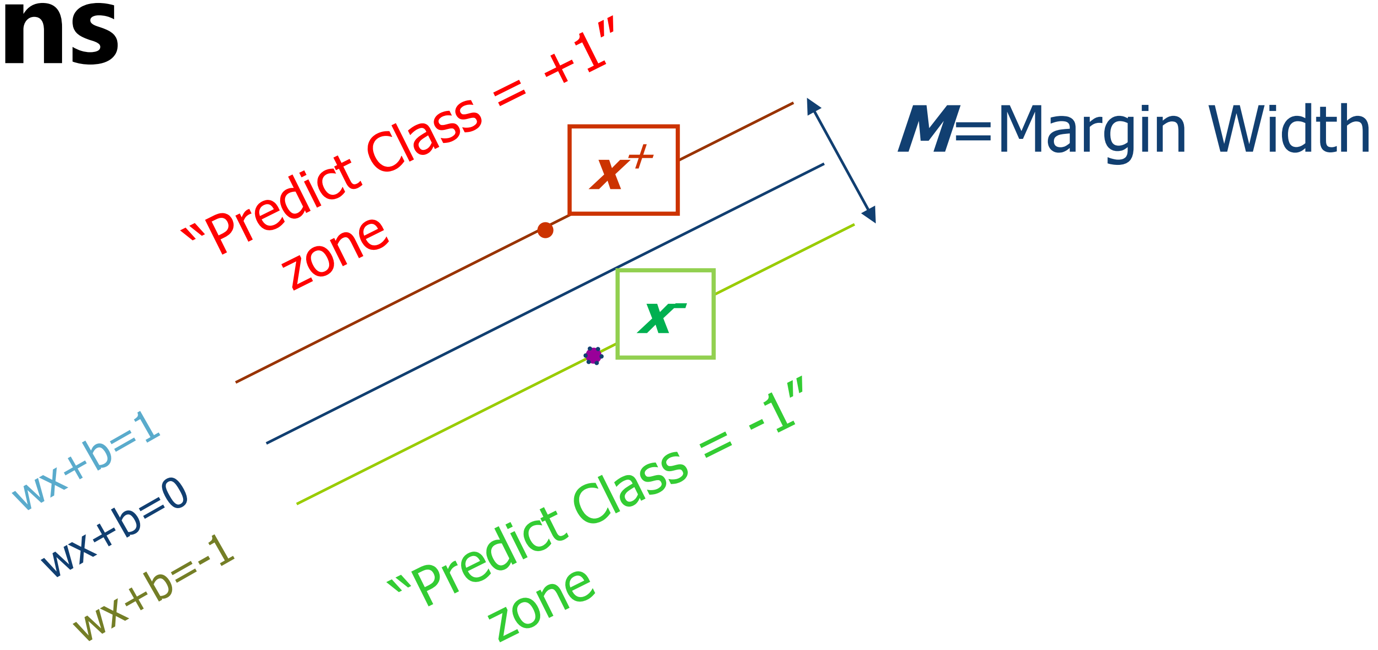




Margins

- The margin of a linear classifier is defined as the width that the boundary could be increased by before hitting a data point.
- Intuitively, it is safer to choose a classifier with a larger margin.
- Wider buffer zone for mistakes
- The hyperplane is decided by only a few data points.
  - Support Vectors
  - Others can be discarded!
- Select the classifier with the maximum margin.
  - Linear Support Vector Machines (LSVM)
- How to specify the margin formally?

Margins



$$M = \boxed{\frac{2}{\|w\|}}$$

Objective Function

- Correctly classify all data points:

$$w \cdot x_i + b \geq 1 \qquad if \;\; y_i = +1$$

$$w \cdot x_i + b \leq -1 \qquad if \;\; y_i = -1$$

$$y_i(w \cdot x_i + b) - 1 \geq 0$$

- Maximize the margin: $\max M = \dfrac{2}{\|w\|} \Rightarrow \min \dfrac{1}{2} w^T w$

- Quadratic Optimization Problem

  - Minimize $\Phi(w) = \dfrac{1}{2} w^t w$

  - Subject to $y_i(w \cdot x_i + b) \geq 1$

Lagrange Multipliers

$$y_i(w\cdot x_i+b)-1\geq 0$$
$$\alpha_i \geq 0$$
$$\alpha_i\left[y_i(w\cdot x_i+b)-1\right]=0$$

$$L_P \equiv \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{l}\alpha_i$$

$$\frac{\partial L_p}{\partial w}=0 \Rightarrow w=\sum_{i=1}^{l}\alpha_i y_i x_i$$

$$\frac{\partial L_p}{\partial b}=0 \Rightarrow \sum_{i=1}^{l}\alpha_i y_i=0$$

Dual Problem

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i\cdot x_j$$

$$\equiv \sum_i \alpha_i - \frac{1}{2}\alpha^T H\alpha \; where \; H_{ij}=y_i y_j x_i\cdot x_j$$

Quadratic problem again!

$$subject\ to: \sum_i \alpha_i y_i = 0 \;\; \& \;\; \alpha_i \geq 0$$

Solutions of w & b

$Support\ Vectors: Samples\ with\ \ positive\ \alpha$

$$g(x) = \sum_{i=1}^{l} \alpha_i y_i x_i \cdot x + b$$

inner product

$$y_s(x_s \cdot w + b) = 1$$

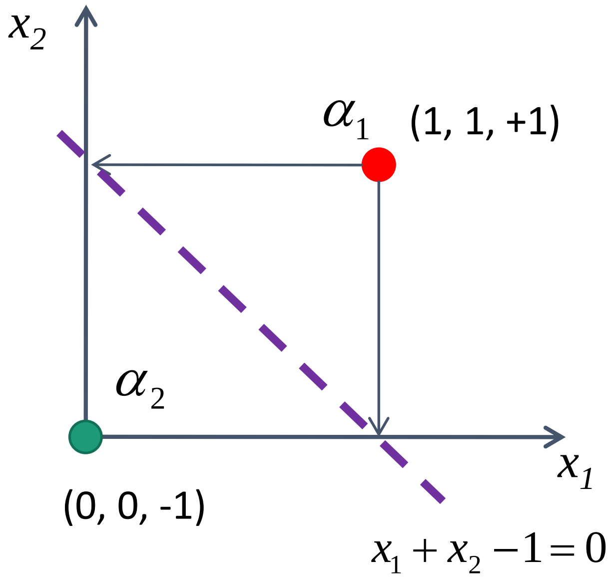
$$y_s\left(\sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b\right) = 1$$

$$y_s^2\left(\sum_{m \in S} \alpha_m y_m x_m \cdot x_s + b\right) = y_s$$

$$b = y_s - \sum_{m \in S} \alpha_m y_m x_m \cdot x_s$$

$$b = \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{m \in S} \alpha_m y_m x_m \cdot x_s\right)$$

An Example



$$\sum_{i=1}^{2} \alpha_i y_i = 0 \Rightarrow \alpha_1 - \alpha_2 = 0 \Rightarrow \alpha_1 = \alpha_2$$

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} y_1 y_1 x_1 \cdot x_1 & y_1 y_2 x_1 \cdot x_2 \\ y_2 y_1 x_2 \cdot x_1 & y_2 y_2 x_2 \cdot x_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$L_D \equiv \sum_{i=1}^{2} \alpha_i - \frac{1}{2} [\alpha_1, \alpha_2] H \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = 2\alpha_1 - \alpha_1^2$$

$$w = \sum_{i=1}^{2} \alpha_i y_i x_i = 1 \times 1 \times [1,1] + 1 \times (-1) \times [0,0] = [1,1]$$

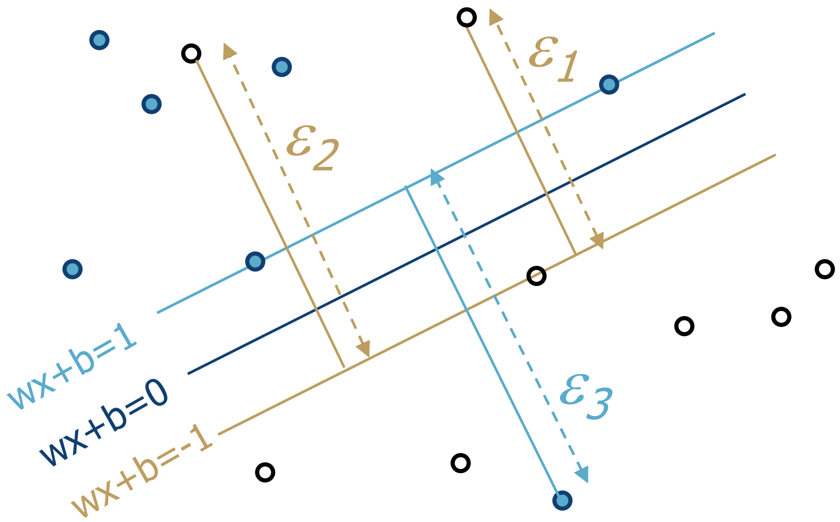
$$\boxed{\alpha_1 = 1; \ \alpha_2 = 1}$$

$$b = -w x_1 + 1 = -2 + 1 = -1$$

$$g(x) = wx + b = x_1 + x_2 - 1$$

$$M = \frac{2}{\|w\|} = \frac{2}{\sqrt{2}} = \sqrt{2}$$

Soft Margin



$$y_i(wx_i + b) - 1 + \xi_i \geq 0$$

$$\Phi(w) = \frac{1}{2} w^t w + C \sum_i \xi_i$$

$$\xi_i \geq 0$$

$$L_P \equiv \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{l} \mu_i \xi_i$$

Soft Margin

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l} \alpha_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l} \alpha_i y_i = 0$$

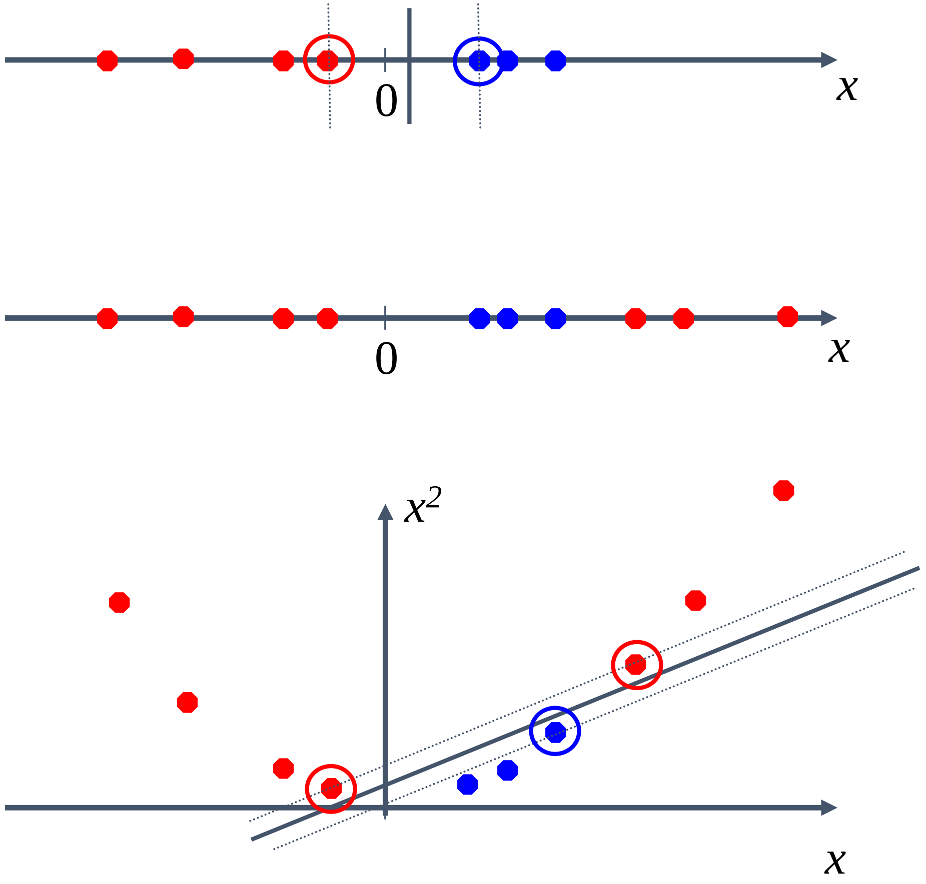
Same as before

$$\frac{\partial L_p}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i$$

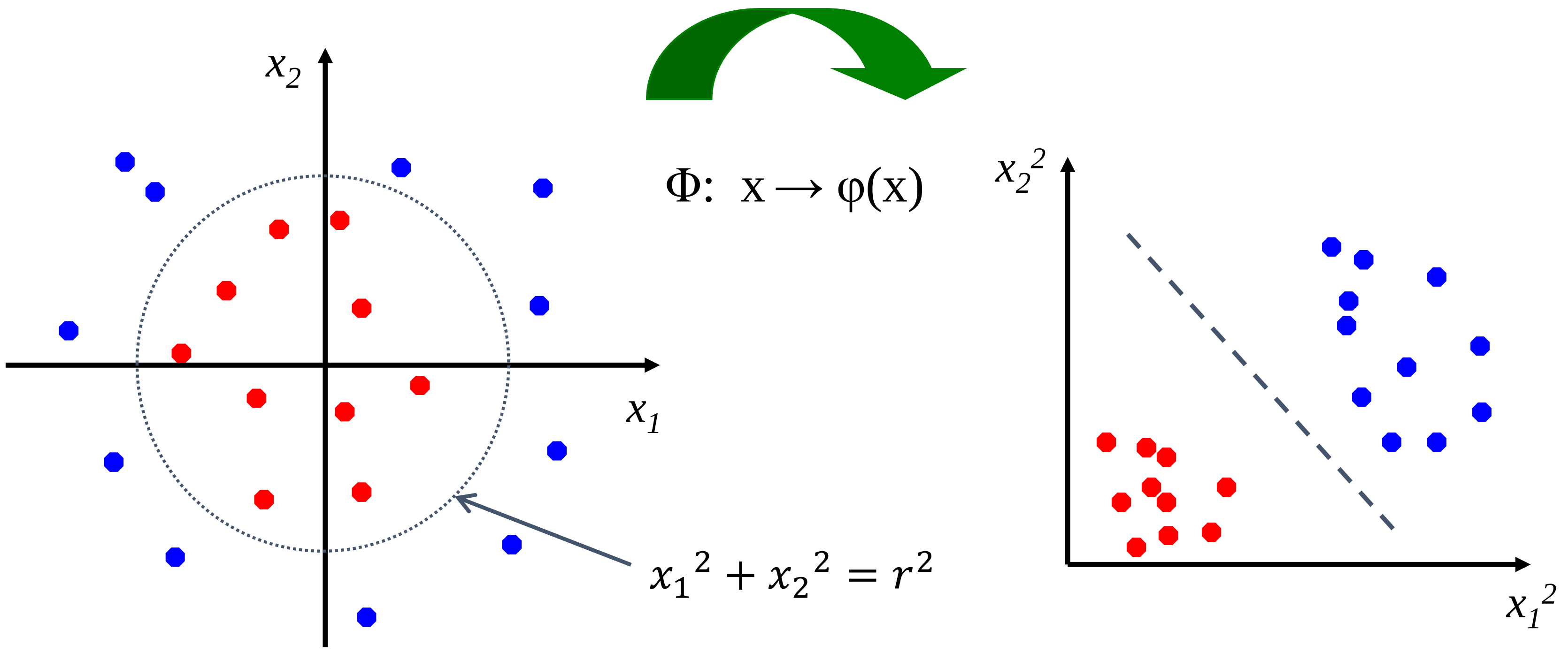
$$L_P \equiv \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i - \sum_{i=1}^{l} \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^{l} \mu_i \xi_i$$

$$\boxed{L_D \equiv \sum_i \alpha_i - \frac{1}{2}\alpha^T H \alpha \quad s.t.\, 0 \le \alpha_i \le C \quad and \ \sum_i \alpha_i y_i = 0}$$

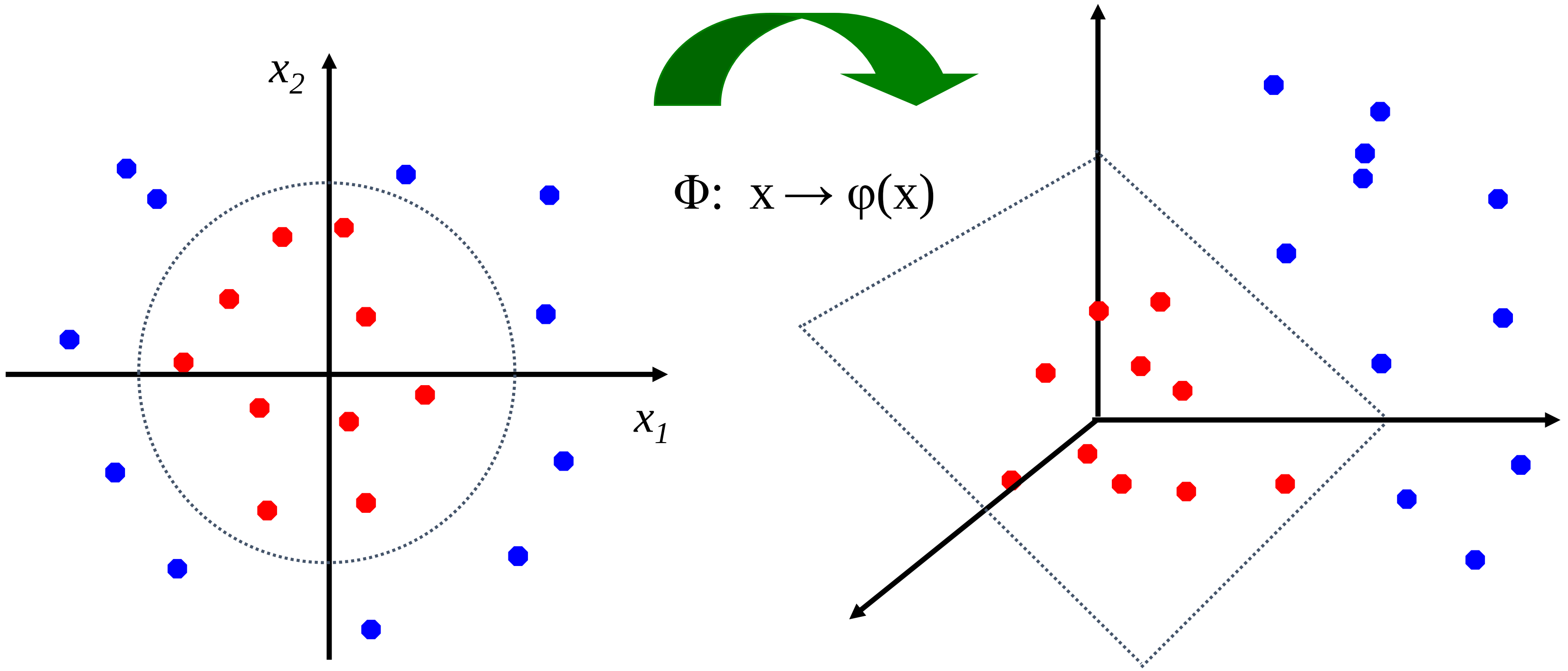
Non-linear SVMs



Feature Space



Feature Space



Quadratic Basis Functions

$$\Phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \vdots \\ \sqrt{2}x_m \\ x_1^2 \\ x_2^2 \\ \vdots \\ x_m^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1x_3 \\ \vdots \\ \sqrt{2}x_1x_m \\ \sqrt{2}x_2x_3 \\ \vdots \\ \sqrt{2}x_2x_m \\ \vdots \\ \sqrt{2}x_{m-1}x_m \end{bmatrix}$$

Constant Terms

Linear Terms

Pure Quadratic Terms

Quadratic Cross-Terms

Number of terms

$$C_{m+2}^2 = \frac{(m+2)(m+1)}{2} \approx \frac{m^2}{2}$$

Calculation of Φ(xi)·Φ(xj)

$$\Phi(a)\cdot\Phi(b)=\begin{bmatrix}1\\ \sqrt{2}a_1\\ \sqrt{2}a_2\\ \vdots\\ \sqrt{2}a_m\\ a_1^2\\ a_2^2\\ \vdots\\ a_m^2\\ \sqrt{2}a_1a_2\\ \sqrt{2}a_1a_3\\ \vdots\\ \sqrt{2}a_1a_m\\ \sqrt{2}a_2a_3\\ \vdots\\ \sqrt{2}a_2a_m\\ \vdots\\ \sqrt{2}a_{m-1}a_m\end{bmatrix}\cdot\begin{bmatrix}1\\ \sqrt{2}b_1\\ \sqrt{2}b_2\\ \vdots\\ \sqrt{2}b_m\\ b_1^2\\ b_2^2\\ \vdots\\ b_m^2\\ \sqrt{2}b_1b_2\\ \sqrt{2}b_1b_3\\ \vdots\\ \sqrt{2}b_1b_m\\ \sqrt{2}b_2b_3\\ \vdots\\ \sqrt{2}b_2b_m\\ \vdots\\ \sqrt{2}b_{m-1}b_m\end{bmatrix}$$

$$\mathbf{1}$$

$$\sum_{i=1}^{m}2a_ib_i$$

$$\sum_{i=1}^{m}a_i^2b_i^2$$

$$\boxed{x_i\cdot x_j\Rightarrow\Phi(x_i)\cdot\Phi(x_j)}$$

$$\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}2a_ia_jb_ib_j$$

It turns out …

$$\Phi(a)\cdot\Phi(b) = \underline{1 + 2\sum_{i=1}^{m} a_i b_i + \sum_{i=1}^{m} a_i^2 b_i^2 + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} 2a_i a_j b_i b_j}$$

$$(a\cdot b+1)^2 = (a\cdot b)^2 + 2a\cdot b + 1 = (\sum_{i=1}^{m} a_i b_i)^2 + 2\sum_{i=1}^{m} a_i b_i + 1$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m} a_i b_i a_j b_j + 2\sum_{i=1}^{m} a_i b_i + 1$$

$$= \underline{\sum_{i=1}^{m} (a_i b_i)^2 + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} a_i b_i a_j b_j + 2\sum_{i=1}^{m} a_i b_i + 1}$$

$$K(a,b) = (a\cdot b+1)^2 = \Phi(a)\cdot\Phi(b)$$

$O(m)$ $O(m^2)$

Kernel Trick

- The linear classifier relies on dot products $x_i \cdot x_j$ between vectors.
- If every data point is mapped into a high-dimensional space via some transformation Φ: $x \rightarrow \varphi(x)$, the dot product becomes: $\varphi(x_i) \cdot \varphi(x_j)$
- A *kernel function* is some function that corresponds to an inner product in some expanded feature space: $K(x_i,x_j) = \varphi(x_i) \cdot \varphi(x_j)$
- Example: $x=[x_1,x_2]$; $K(\mathrm{x_i},\mathrm{x_j}) = (1+x_i \cdot x_j)^2$

$$K(x_i,x_j) = (1+x_i \cdot x_j)^2 = 1 + x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}$$

$$= [1, x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}] \cdot [1, x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2, \sqrt{2}x_{j1}, \sqrt{2}x_{j2}]$$

$$= \Phi(\mathrm{x_i}) \cdot \Phi(\mathrm{x_j}), \quad \text{where } \Phi(x) = [1, x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]$$

Kernels

$$Polynomial: K(x_i, x_j) = (x_i \cdot x_j + 1)^d$$

$$Gaussian: K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)$$

$$Hyperbolic\ Tangent: K(x_i, x_j) = \tanh(\kappa x_i \cdot x_j + c)$$

Solutions of w & b

$$w = \sum_{i=1}^{l} \alpha_i y_i \Phi(x_i)$$

$$w \cdot \Phi(x_j) = \sum_{i=1}^{l} \alpha_i y_i \Phi(x_i) \cdot \Phi(x_j) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j)$$

$$b = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} \alpha_m y_m \Phi(x_m) \cdot \Phi(x_s)) = \frac{1}{N_s} \sum_{s \in S} (y_s - \sum_{m \in S} \alpha_m y_m K(x_m, x_s))$$

$$\boxed{g(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b} \Longleftrightarrow g(x) = w \cdot x + b = \sum_{i=1}^{l} \alpha_i y_i x_i \cdot x + b$$

SVM Roadmap

Linear Classifier

Maximum Margin

↓

Linear SVM

Noise

↓

Soft Margin

Nonlinear Problem

↓

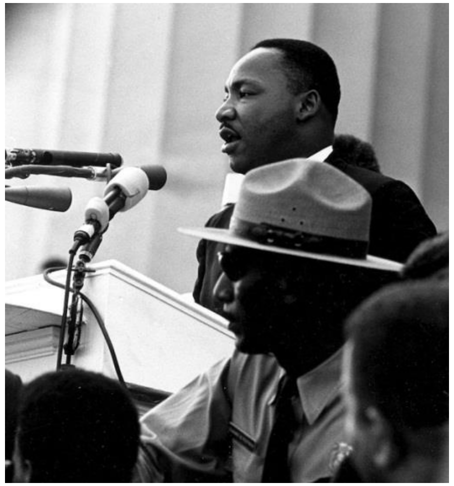
$a \cdot b \rightarrow \Phi(a) \cdot \Phi(b)$

High Computational Cost

↓

Kernel Trick

$K(a,b)=\Phi(a) \cdot \Phi(b)$



“I have a dream — one day there will be a classifier that can handle nonlinear problems …”

Reading Materials

- N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

- C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

- H. Lodhi et al., "Text Classification Using String Kernels". *The Journal of Machine Learning Research*, vol. 2, pp. 419-444, 2002.

- Online Resources
  - http://www.kernel-machines.org/
  - http://www.support-vector-machines.org/
  - http://www.tristanfletcher.co.uk/SVM%20Explained.pdf
  - http://www.csie.ntu.edu.tw/~cjlin/libsvm/


ALSO
BIG
DATA
TOOLS
DISK
TARGET
APPLIED
SENSOR
SHARED
EXAMPLES
EVERY
DEFINITION
CURRENT
MAY
MOVING
MANAGE
WITHIN
ARCHIVES
PRACTITIONERS
CAPTURE
BUSINESS
TECHNOLOGIES
USING
INTERNET
SETS
DISTRIBUTED
GENOMICS
NEEDED
TIME
ONE
LARGE
MANAGEMENT
TERABYTES
CASE
ABILITY
SOFTWARE
INFORMATION
PROCESS
PETABYTES
INCLUDE
PERFORMANCE
STORE
ELAPSED
SYSTEMS
TOLERABLE
SIZE
RESEARCH
STORAGE
MPP
LOGS
COST
PARALLEL
SAN
MASSIVELY
GROW
DIFFICULTY
SET
RELATIONAL
SOCIAL
USE
LARGER
CONTINUES
CITATION
TENS
COMPLEX
ANALYTICS
BIOLOGICAL
PROCESSING
CAPACITY
HUNDREDS
NETWORKS
DATABASES
RECORDS
SEARCH
GARTNER
WORKING