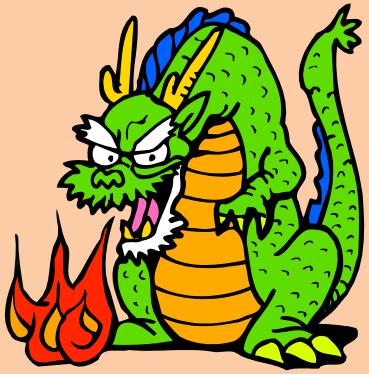
情報学習理論

渡辺澄夫
東京工業大学



深層学習

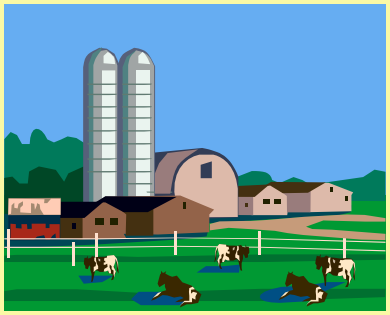
ボルツマンマシン

ニューラルネットワーク

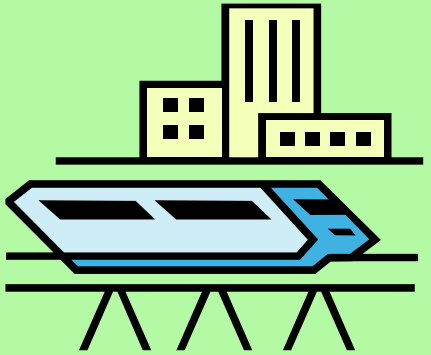
教師なし学習

教師あり学習

自己組織化

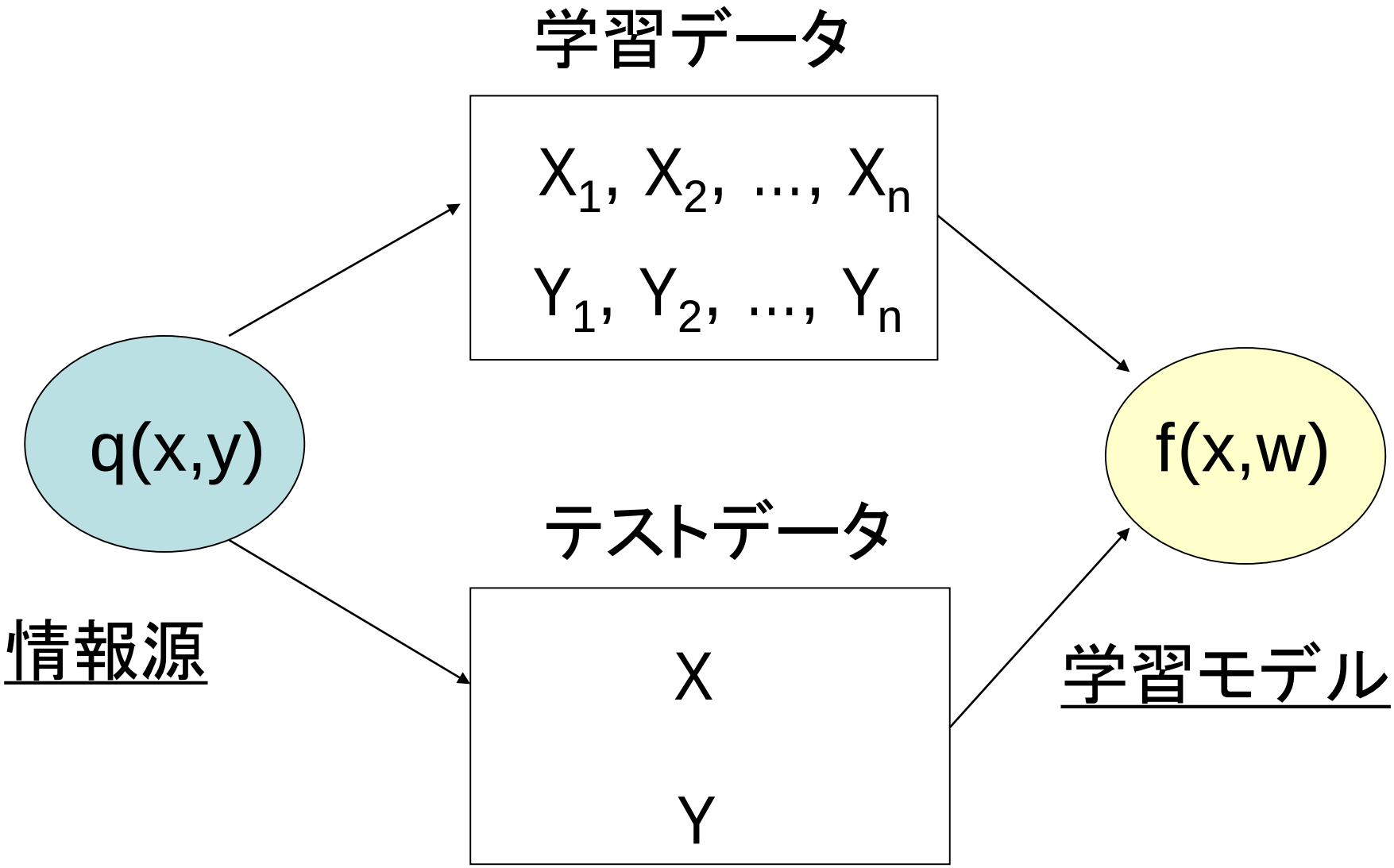


競合学習



サポートベクタマシン

教師あり学習の枠組み



汎化誤差を小さくしたい

学習誤差　$E(w) = (1/n) \sum_{i=1}^{n} (Y_i - f(X_i, w))^2$

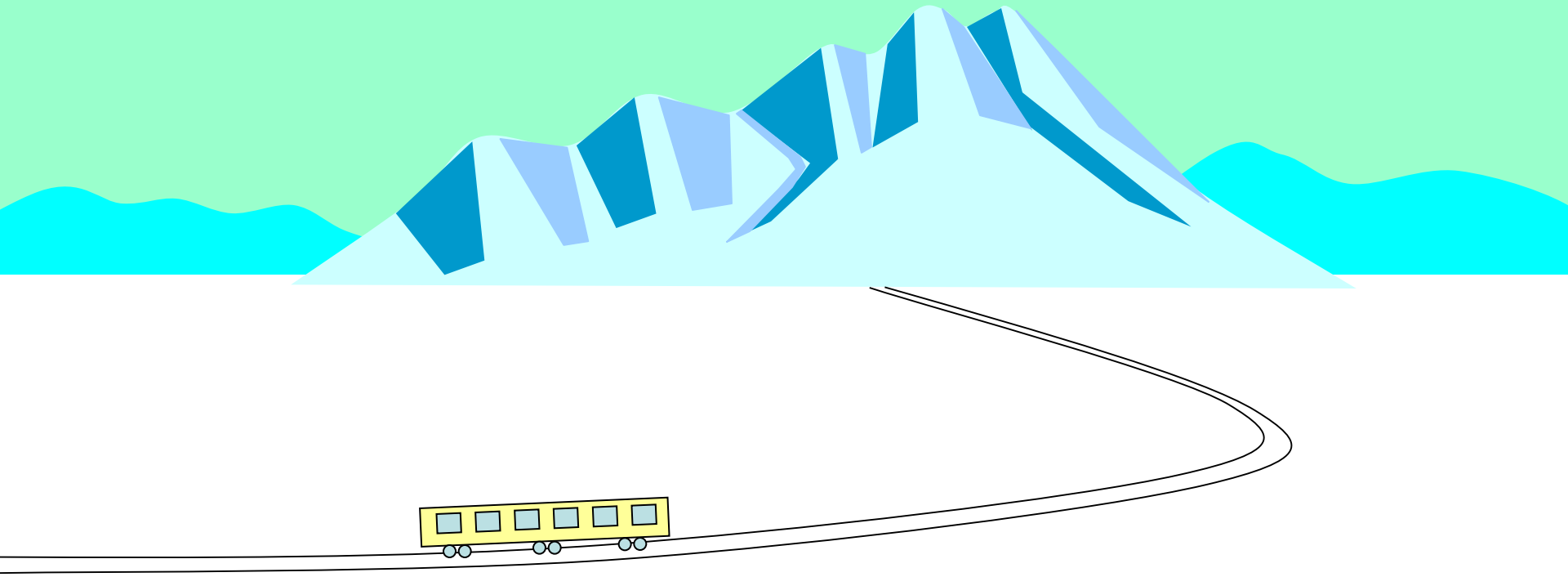
◎　汎化誤差を小さくすること、すなわち予測を正確にすることが学習の目的である。

○　学習誤差を小さくしても汎化誤差は小さくならなかった。

どうすれば汎化誤差を小さくできるだろうか。
汎化誤差を小さくする方法を求めて様々な工夫がなされてきた。

マージン最大化とカーネル法の組み合わせにより成功したのがSVM（サポートベクタマシン）である。1995年から2005年の間にSVM は多くの実問題で使われるようになった。

大都会　カーネルマシンへ



識別問題

カテゴリを二つに分ける問題を考える

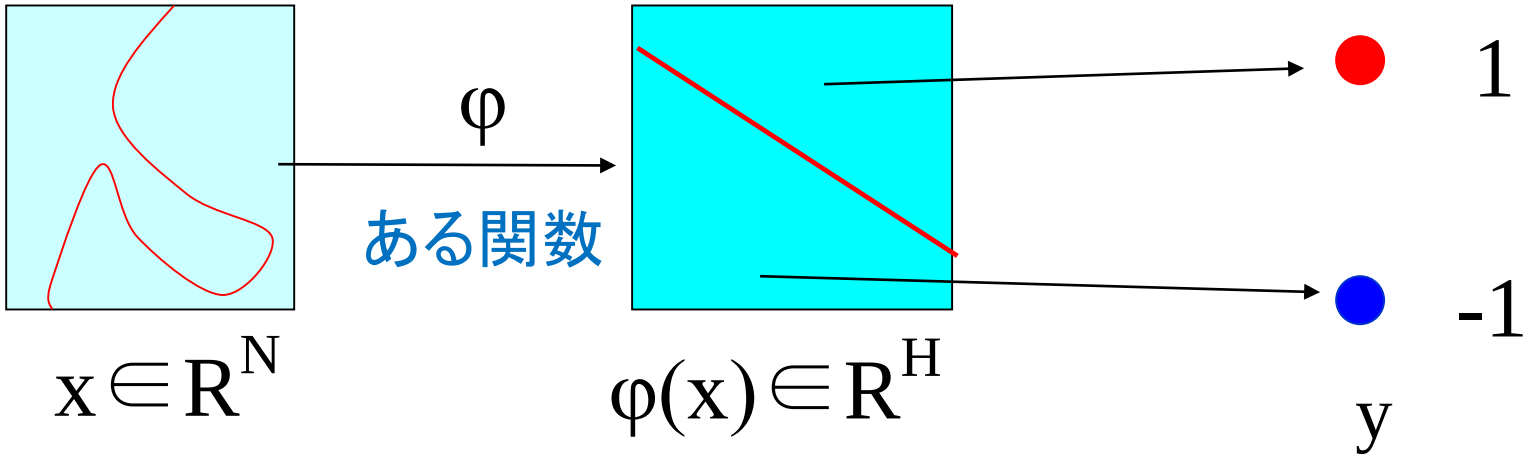
学習用データ　$\{ (x_i, y_i)\ \ i = 1,2,\ldots,n \}$

$$x_i \in R^N$$

$y_i = 1$ or -1　　（正例，負例）

「x から y 」を定めているルールを推定する。

サポートベクタマシンの仕組み

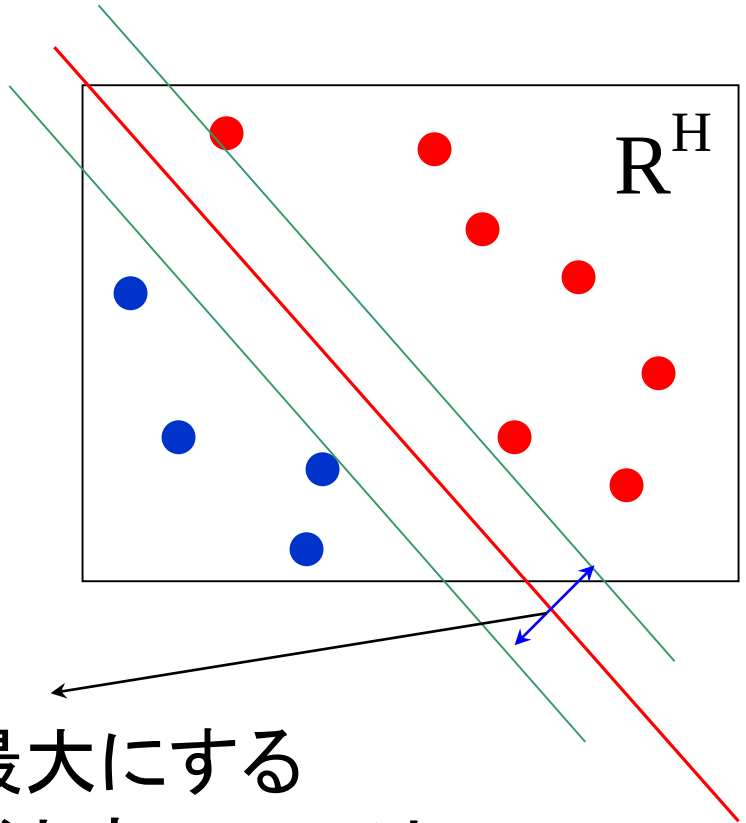


$$sgn(t) = \begin{cases} 1 & (t \geqq 0) \\ -1 & (t<0) \end{cases}$$

$$y = sgn(w \cdot \varphi(x)+b)$$

w, b, φ(x) をどのように決めたらよいか？

マージン最大化



マージンとは
正例と負例の
間の帯の太さ

マージンを最大にする
(w, b) を選ぶと良いのでは

$y=w \cdot \varphi(x)+b$

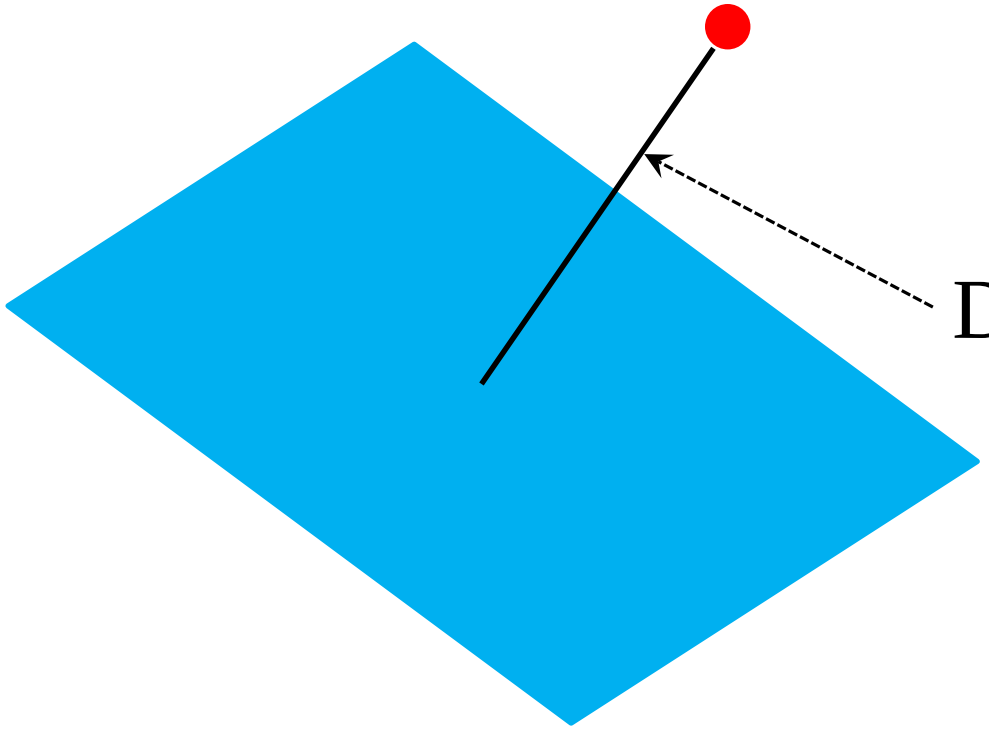
(1) なぜ　マージンを最大化するのか

(2) マージンを最大化するにはどうしたらよいか

(3) 関数 $\varphi(x)$ の作り方は次回。今回は(1)と(2)を説明します。

復習：点から平面に降ろした垂線の長さ

点：(x_1,y_1,z_1)

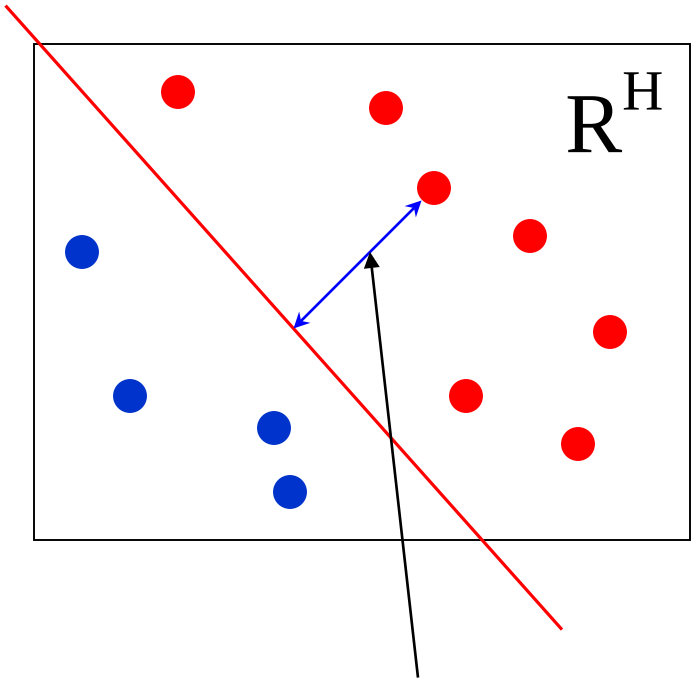


$$D = \frac{|ax_1+by_1+cz_1+d|}{(a^2+b^2+c^2)^{1/2}}$$

平面：$ax+by+cz+d=0$

高次元空間でも
同様の公式が成立

マージンを数式に



正例も負例も同じ条件で書ける

$$y_i\,(w \cdot \varphi(x_i)+b) > 0 \qquad (i=1,2,..,n)$$

この条件のもとで

$$\text{Margin} = \min_{i=1,2,\ldots,n} \frac{|\, w \cdot \varphi(x_i)+b \,|}{\|w\|}$$

点$\varphi(x_i)$から
平面 $w \cdot v+b=0$
に降ろした
垂線の長さ

を最大にする (w,b) を求めればよい。

簡単な式に変形する

Marigin は w, b を定数倍しても変わらないので

$$\mathrm{Min}\,|\, w \cdot \varphi(x_i)+b \,|\, = 1$$ と仮定してよい

条件 $y_i\,(w \cdot \varphi(x_i)+b) \geqq 1 \quad (i=1,2,..,n)$ のもとで

$$\mathrm{Margin}= \frac{1}{\|w\|}$$ を最大にする

$\Leftrightarrow$ $\|w\|$を最小にする

最適化問題に帰着

マージン最大化問題

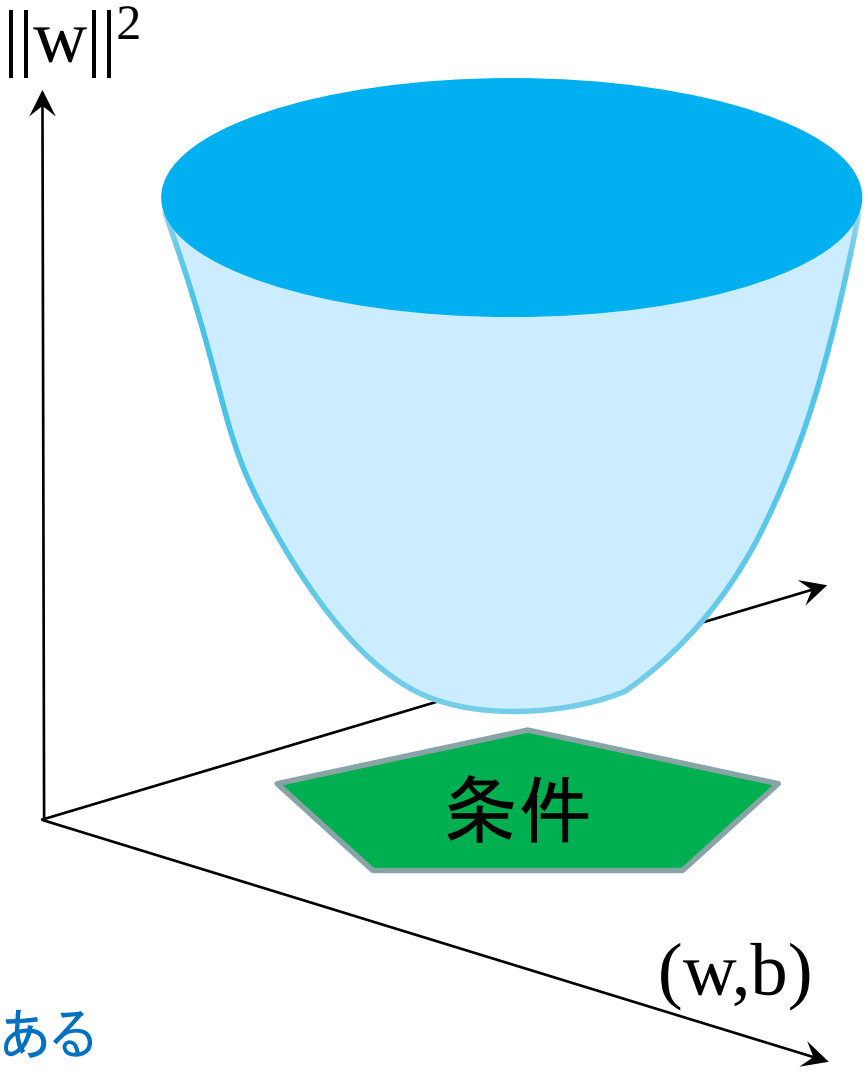
条件

$y_i\,(w \cdot \varphi(x_i)+b) \geqq 1$
$(i=1,2,..,n)$

のもとで

$\|w\|^2$

を最小にしたい



(注) 条件を満たす{(w,b)}は凸集合である

線形制約2次最適化問題

マージン最大化問題

条件　$y_i\,(w \cdot \varphi(x_i)+b) \geqq 1 \quad (i=1,2,..,n)$

のもとで　$\|w\|^2$　を最小にしたい

複数の1次不等式で定義された凸集合の中で
2次式を最小にする最適化問題

高次元空間上なので直接解くことは難しいが、
効率的な近似最適化法が研究されている

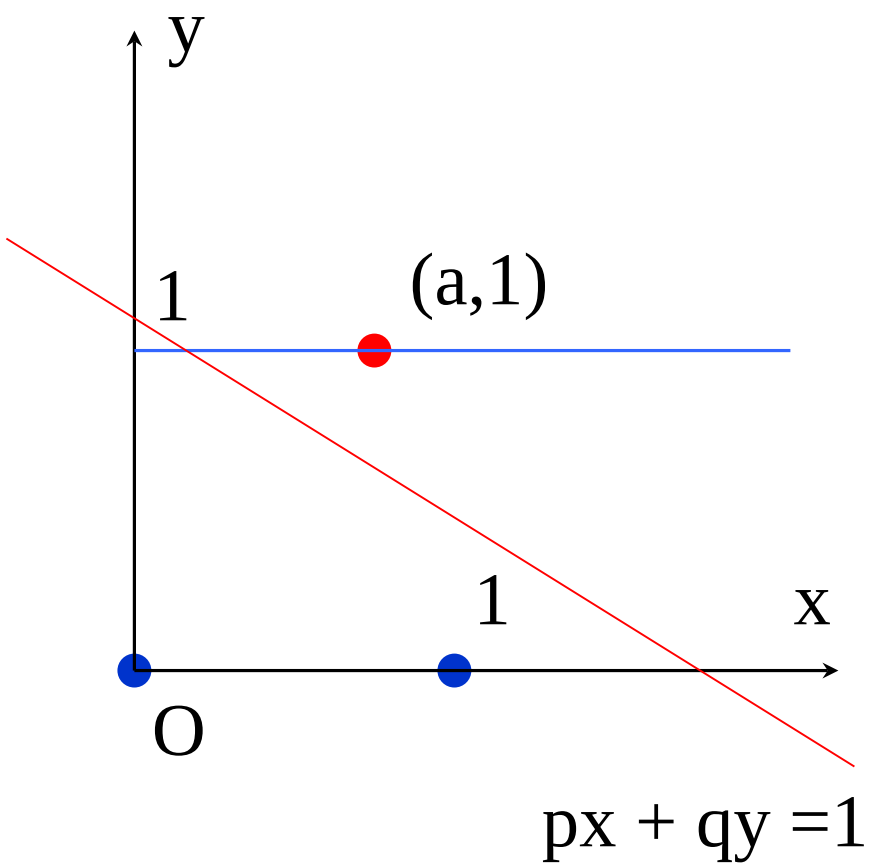
（注意） SVMの問題は最適化問題に帰着するが、
学習理論の中には最適化問題に帰着しないものも多い。

問1

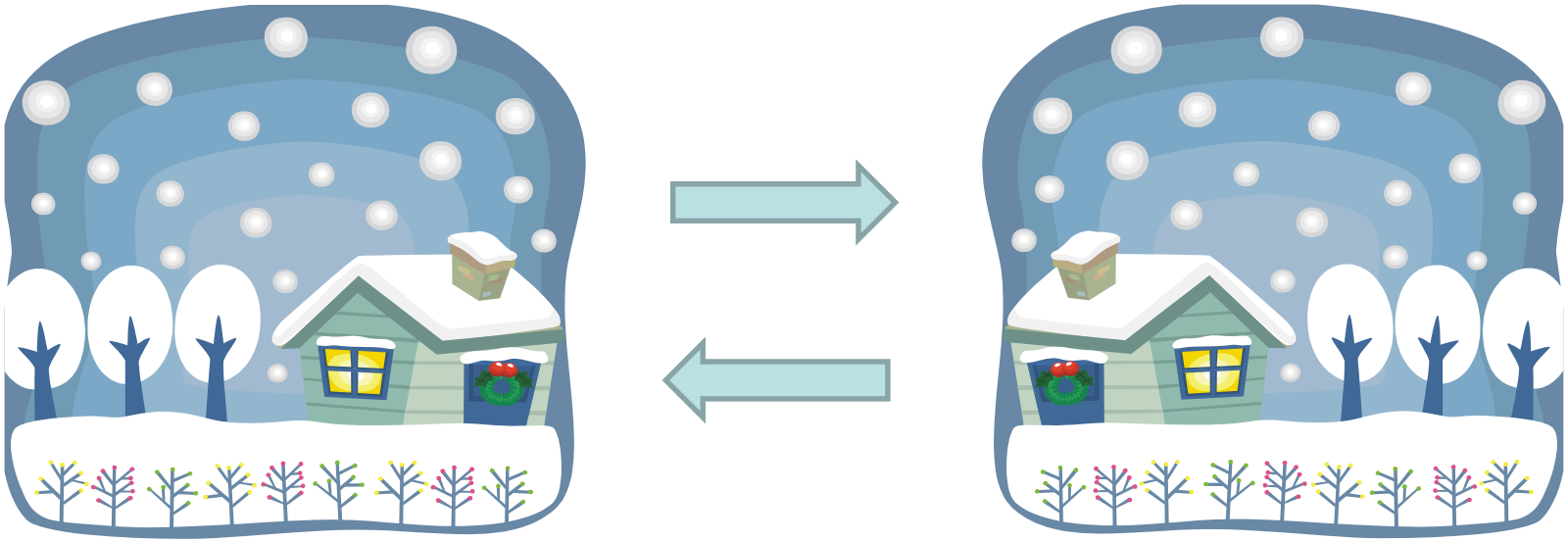
(1) 赤い点 (a,1) が $a \geqq 0$
を動くとき、

マージンを最大にする直線
$px + qy = 1$
について p, q をa で表せ。

(2) (p,q) は a の連続関数か
微分可能関数か。



双対問題とは



復習： 凸最適化問題

◎ $f(x), g_1(x),\ldots,g_K(x)$ を凸関数（下に凸な関数）とする。

元問題　$S=\{x\,;\, g_k(x)\leqq 0\ (k=1,2,\ldots,K)\}$　とする。

$f(x)$ を最小化する $x \in S$ を見つけたい。

◎ $A=\{(a_1,a_2,\ldots,a_K)\,;\, a_k \geqq 0\}$　とする。ラグランジュアン $L(a,x)$ を次の式で定義する。

$$L(a, x) = f(x) + \sum_{k=1}^{K} a_k\, g_k(x)$$

復習： Kuhn-Tacker の定理

元問題の任意の最適解を x^* とする。

定理. $S^o=\{x;\ g_k(x)<0\ (k=1,2,\ldots,K)\}$ が空集合でないとする。
このとき次の $a^*\in A$ が存在する。

(1) $a_k^*\, g_k(x^*)=0 \quad (k=1,2,\ldots,K)$.

(2) $\min_{x\in S} L(a^*, x) = L(a^*, x^*) = \max_{a\in A} L(a, x^*)$.

上記の (a^*,x^*) を見つければよい。集合Sの定義から

$$\min_{x\in S} L(a^*, x) = \min_{x:\text{自由}} L(a^*, x)$$

が成り立つので x の制約条件を除いて探すことができる。

双対問題　「$\partial L/\partial x=0$ を満たす x に対して $L(a, x)$ を最大化する $a^*\in A$ を見つけてから $L(a^*,x)$ を最小化」を解けばよい。

（参考） マージン最大化に Kuhn-Tacker の定理を適用する手順

問題

$(1/2)\|w\|^2$最小化, 条件 $y_i\,(w\cdot\varphi(x_i)+b) \geqq 1 \quad (i=1,2,..,n)$

↓

ラグランジュアン

$$L(a, w, b) = 1/2\|w\|^2 + \sum_{i=1}^{n} a_i\,(1-y_i\,(w\cdot\varphi(x_i)+b))$$

↓

$$\begin{cases} \partial L/\partial w = w - \sum a_i\, y_i\, \varphi(x_i) = 0 \;\rightarrow\; w = \sum a_j\, y_j\, \varphi(x_j) \\ \partial L/\partial b = - \sum a_i\, y_i = 0 \;\rightarrow\; \sum a_i\, y_i = 0 \end{cases}$$

双対問題

$\Sigma\, a_i - \frac{1}{2} \Sigma\, a_i\, a_j\, y_i\, y_j\, \varphi(x_j)\cdot\varphi(x_i)$最大化, 条件 $a_i \geqq 0, \Sigma\, a_i\, y_i = 0$

↓

この問題の解 $\{a^*_i\}$に対して $L(a^*,w,b)$ を最小化する(w,b)は

答

$$w = \sum a^*_i\, y_i\, \varphi(x_i), \quad b = y_i - w\cdot\varphi(x_i) \quad (a^*_i>0\text{のとき})$$

等価な双対問題

マージン最大化問題は次の双対問題と等価になる

双対問題

変数　$a_1,a_2,\ldots,a_n$ に関する最大化問題

条件 $a_i \geqq 0$ かつ $\Sigma\, a_i y_i = 0$ のもとで下記を最大化

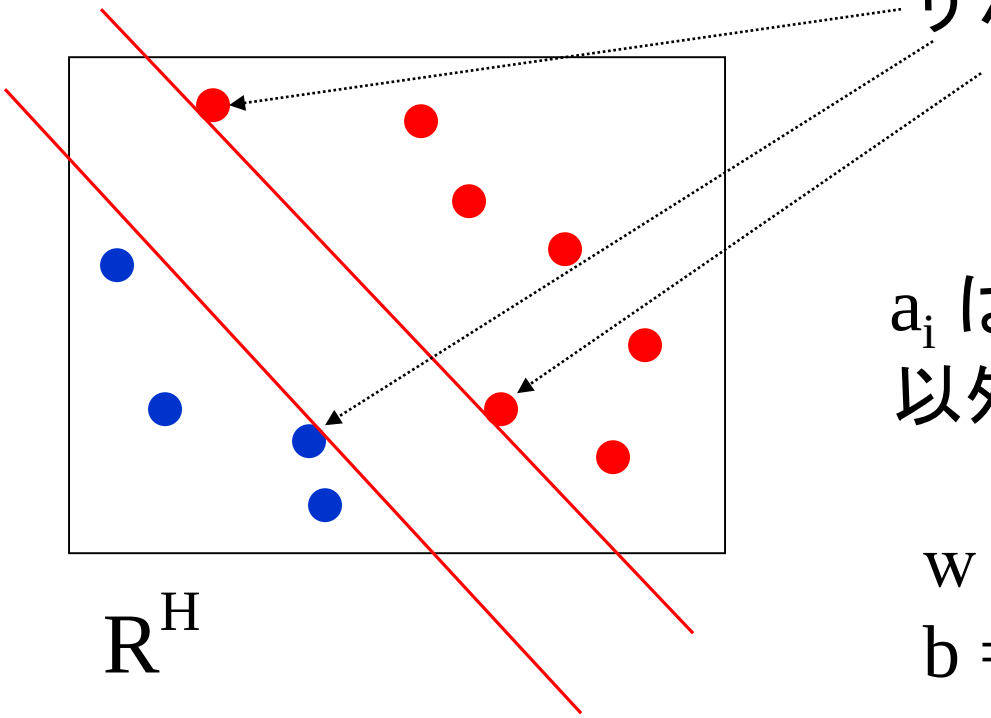
$$E(a_1,a_2,\ldots,a_n) = \Sigma\, a_i - (1/2)\, \Sigma\, a_i a_j y_i y_j (\varphi(x_i) \cdot \varphi(x_j))$$

答　双対問題が解けると、マージン最大化の解 (w,b) は

$$w = \Sigma a_i y_i \varphi(x_i)$$

$$b = y_i - w \cdot \varphi(x_i) \quad (a_i > 0 \text{となる } i \text{ について})$$

解の関係



a_i はサポートベクトル
以外では0になる。

$$w = \Sigma a_i y_i \varphi(x_i)$$
$$b = y_i - w \cdot \varphi(x_i) \quad (a_i > 0)$$

(w,b)はサポートベクトルで定まる。つまり

マージン最大化問題 ⇔ サポートベクトルを見つける問題

マージン最大化はなぜ汎化誤差を小さくするのか

真の分布は不明でありサンプルの出方についてもばらつきがあるが、ともかく、

確率 $(1-\delta)$ 以上で

未知データに対する誤り確率は

$$Pr(err) \leqq (4/n) \{ V_c(1+\log(2n/V_c)) - \log \delta / 4 \}$$

となる。ここで

$$V_c \leqq \min \{ [D^2/Margin^2], H \} +1$$

$$D = \max||\varphi(x_i)||$$

おおよそ「Margin：大 ⇔ Pr(err)：小」

マージン最大化は、予測精度を最良にするものではないが、
実用上うまく行くことが多いので広く使われている。

（注）V_C はVC次元と呼ばれ学習モデルの複雑さを表す量

（注意）双対問題の数値的な解き方の例

変数 $a_1,a_2,\ldots,a_n$ について

条件 $a_i \geqq 0$ かつ $\Sigma\, a_i y_i = 0$ のもとで下記を最大化

$$E(a_1,a_2,\ldots,a_n) = \Sigma\, a_i - (1/2)\, \Sigma\, a_i a_j y_i y_j (\varphi(x_i) \cdot \varphi(x_j))$$

双対問題も変数の個数が多いと完全に解くことは難しい

配布プログラムでは $0 \leqq a_i \leqq C$ で定数$(L>0)$を用いて

$$H(a_1,a_2,\ldots,a_n) = E(a_1,a_2,\ldots,a_n) - L(\Sigma a_i y_i)^2$$

の最大化問題を最急上昇法で解いている.
C の制限をつけることをソフトマージンという.

サポートベクタマシンのまとめ

(1) サポートベクタマシンでは、与えられた例を
説明できるものの中でマージンを最大化する。

(2) マージンの最大化は、汎化誤差を最小にするものではないが、応用上、うまくいくことが多い。

(3) マージンの最大化問題は双対問題に帰着させて解ける。
マージンの最大化はサポートベクトルを見つけること。

(4) 関数 φ(x) の作り方は次週。

(注) サポートベクタマシンの成功により，統計的学習の問題に対して最適化問題を定めて解くというアプローチが広く行われるようになった。

問2

$\varphi(x)=x$ のときを考える。データ発生時に
真の識別ルールに雑音が加わるとき
サポートベクトルの数と汎化誤差が
どのように変化するかを調べてみよう。

雑音の大きさ	0.0	0.1	0.2	0.3
サポートベクトル数				
汎化誤差				