Chapter 5:
Linear Discriminant Functions
(Sections 5.8-5.9-5.11-5.12)

q MSE and Pseudoinverse

q LMS procedure

q Ho-Kashyap Procedure

q SVM

q Multicategory cases

MSE and the Pseudoinverse

- Until now the criterion functions depend on misclassified y`s
- Now, we use all samples and can specify a margin for each sample $a^t y_i = b_i > 0$
- Remember: y=-y if y belongs to the second category
- So, we obtain a system of n equations with d+1 unknown (Y=[1,x], y`s are in rows) in the form of Ya=b.

Pseudoinverse

- In general n>>d, (Y is nx(d+1)) and the system has no solution
- We search a solution which minimizes the error:

$$e=Ya-b$$

- Or in the sense of MSE

$$J_s(a)=||Ya-b||^2$$

- Gradient $\nabla J_s=2Y^t(Ya-b)$
- So $Y^tYa=Y^tb$
- Or (if Y^tY is nonsingular) $a=(Y^tY)^{-1}Y^tb$
- Pseudoinverse matrix of Y: $(Y^tY)^{-1}Y^t$

Widrow-Hoff or LMS (least mean squared)

q Problems with MSE:

  q Y^tY may be singular

  q d may be very large and we must work with large matrices

q Solution: Gradient descent procedure

q $\nabla J_s = 2Y^t(Ya-b)$ so

  $a(k+1)=a(k)-\beta(k)Y^t(Ya(k)-b)$

  (Y=y`s in rows)

q Single-Sample:

  $a(k+1)=a(k)+\beta(k)(b_k-a(k)^t y^k)y^k$

  (y^k in column form)

q Stop when $|\beta(k)(b_k-a^t y^k)y^k| <\theta$

Similarity with Relaxation?

- Relaxation

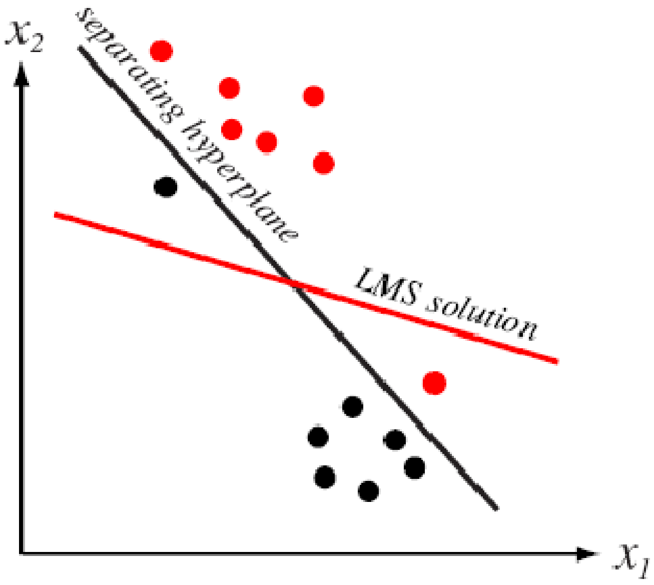
  If $a^t y^k \le b$ then correcting the error

$$a \longleftarrow a + b(k) \frac{b - a^t y^k}{\left\| y^k \right\|^2} y^k$$

- LMS

  If $|\beta(k)(b_k - a^t y^k) y^k| \ge \theta$ then

  $a(k+1) = a(k) + \beta(k)(b_k - a(k)^t y^k) y^k$



Ho-Kashyap Procedure

- LMS gives always a solution
- For an arbitrary b, there is no guarantee for separating the linearly separatable samples
- How can we find a with a margin for b
- Ho-Kashyap searches a as well as b
- $J_s(a)=||Ya-b||^2$
- $\nabla_a J_s=2Y^t(Ya-b)$ and $\nabla_b J_s=-2(Ya-b)$
- $a=(Y^tY)^{-1}Y^tb$
- Gradient descent for b: $b(k+1)=b(k)-\beta(k)\nabla_b J_s(b(k))$

Ho-Kashyap Procedure

- We must respect b>0, so

  (refuse to reduce b (it is a vector) when the initial b is positive)

$$b(k+1)=b(k)-b\,(k)\frac{\nabla_b J_s(b(k)-abs(\nabla_b J_s(b(k))}{2}$$

<u>Algorithm:</u>

- Begin initialize a and b>0
- $$b(k+1)=b(k)-b\,(k)\frac{\nabla_b J_s(b(k)-abs(\nabla_b J_s(b(k))}{2}$$
- a(k+1)=(Y^{t}Y)$^{-1}$Y^{t}b(k+1)
- Convergence if 0<β<1 and linearly separable samples

Summery

- Perceptron (consider only misclassified samples)

$$a(k+1) = a(k) + y^k \qquad a(k+1) = a(k) + b(k) y^k$$

- Relaxation (+margin, $0<\beta<2$)

$$a(k+1) = a(k) + b(k) \frac{b - a^t y^k}{\left\| y^k \right\|^2} y^k$$

- LMS

$$a(k+1) = a(k) + b(k)(b - a^t y) y$$

- Pseudoinverse $a=(Y^tY)^{-1}Y^tb$
- Ho-Kashyap

$$b(k+1) = b(k) - b(k) \frac{\nabla_b J_s (b(k) - abs(\nabla_b J_s (b(k))}{2}$$

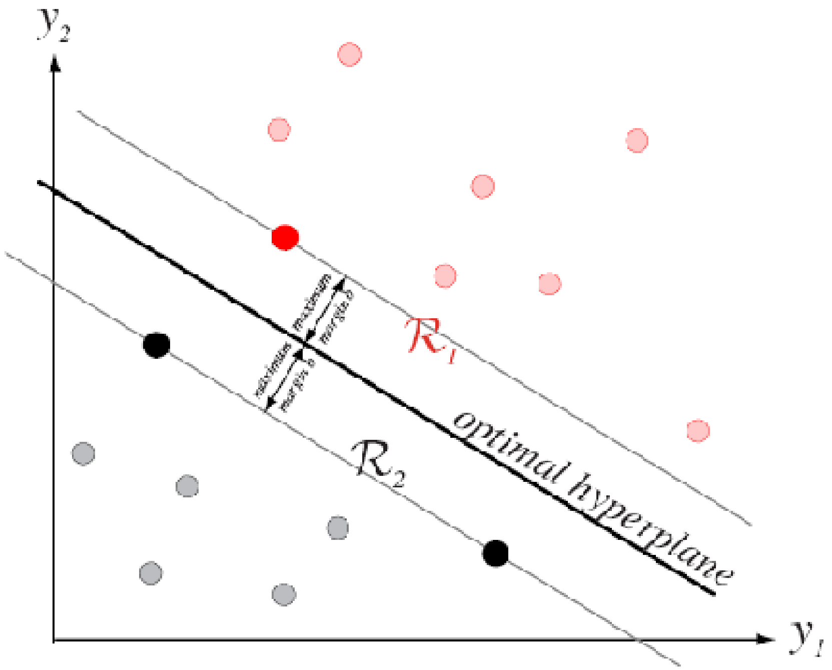
$$a(k+1)=(Y^tY)^{-1}Y^tb(k+1)$$

Support Vector Machine (SVM)

- q a relatively straightforward engineering solution for classification tasks
- q has nice computational properties
- q key ideas in brief
  - q constructs a separating hyperplane in a high-dimensional feature space
  - q maximizes separability
  - q expresses the hyperplane in the original space using a small set of training vectors, the “support vectors”
  - q is nonlinear in the input space

SVM

- Goal: Construct a separating hyperplane that maximizes the margin of separation
- Support vectors are the vectors with b distance from hyperplane



Why a large margin is good?

q It can be shown (Vapnik) that the capacity of a classifier (expressed as Vapnik-Chervonenkis-dimension h) is bounded by a term that decreases as margin b increases.

q Structural risk minimization: For fixed N the total generalization error is equal to:

training error + confidence interval

q where the confidence interval increases as h increases.

q Here training error = 0, and hence we should use the hyperplane for which the margin b is maximal.

Hyperplane decision

q Distance y^k from the hyperplane

$$\frac{|g(y^k)|}{\|a\|}$$

q So the condition to verify:

$$\frac{|g(y^k)|}{\|a\|} \ge b$$

q We impose:

$$\begin{aligned} &z = +1 \quad if \quad w_1 \\ &z = -1 \quad if \quad w_2 \\ &|g(y^k)| = zg(y^k) \end{aligned}$$

q We search *a* so that *b* is maximal

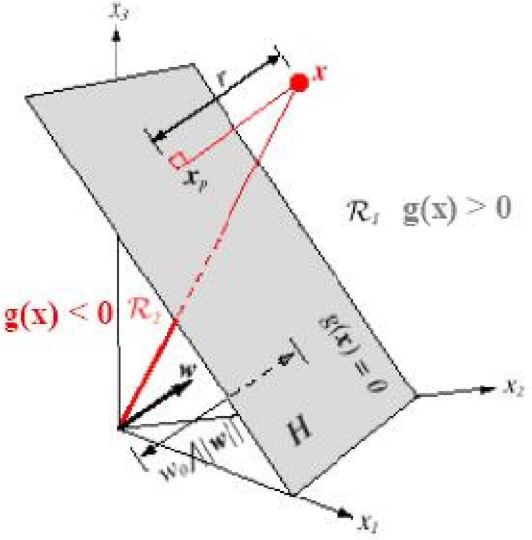
$$\frac{z^k g(y^k)}{\|a\|} \ge b$$

q For having one solution, we impose *b||a||=1*

q Maximizing *b* equals minimizing ||a||

q So min $||a||^2$

q with $z^k g(y^k) \geq 1$



How to find the hyperplane

q Goal: find the weight vector a having the smallest norm and fulfilling the following constraint for each sample

q $z^k a^t y^k \geq 1$

q Use Lagrange multiplier $\alpha^k > 0$ and minimize with respect to a and maximize it with respect to α

$$L(a,a) = \frac{1}{2}\|a\|^2 - \sum_{k=1}^{n} a^k \left[z^k g(y^k) - 1\right]$$

q Assume $a^t y^k = w^t x^k + w_0$

q So

$$L(w, w_0, a) = \frac{1}{2} w^t w - \sum_{k=1}^{n} a^k \left[z^k (w^t x^k + w_0) - 1\right]$$

Solution

$$\frac{\partial L(w, w_0, a)}{\partial w} = 0 \qquad \Rightarrow w = \sum_{k=1}^{n} a^k z^k x^k$$

$$\frac{\partial L(w, w_0, a)}{\partial w_0} = 0 \qquad \Rightarrow 0 = \sum_{k=1}^{n} a^k z^k$$

- Insert this to the L, we obtain:

$$L(a) = \sum_{k=1}^{n} a^k - \frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} a^k a^j z^k z^j x_k^t x_j$$

- The L should be maximized with respect to α, subject to the constraints:

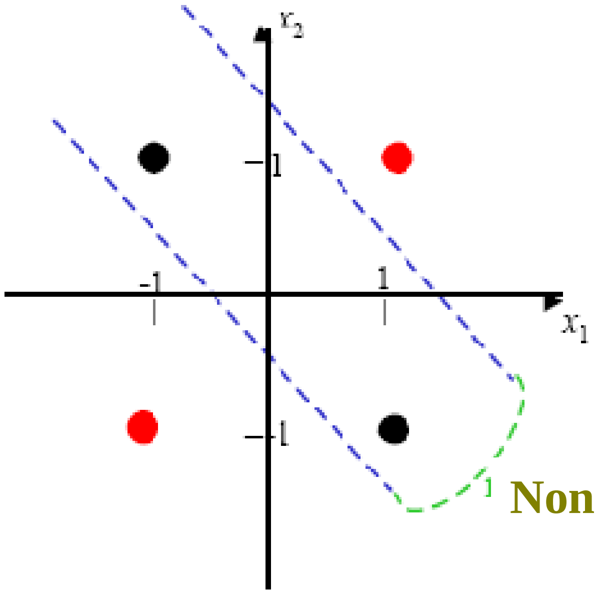
$$\sum_{k=1}^{n} a^k z^k = 0$$

$$a^k \geq 0 \qquad k = 1, \ldots, n$$

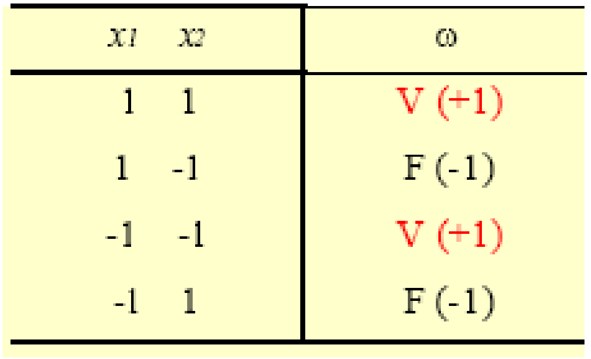
- quadratic optimization

Example: XOR

- XOR is a the simplest problem of nonlinearly separable case



Non-linear



x_1	x_2	ω
1	1	V (+1)
1	-1	F (-1)
-1	-1	V (+1)
-1	1	F (-1)

Example: XOR

q Non-linear

$$\mathbf{y}(\mathbf{x}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1{}^2, x_2{}^2)^t$$

q Using 4 prototypes

$$\begin{aligned}
\mathbf{x}_1 &= (1,1), \quad z_1 = 1 \textbf{ for } \omega_1 \\
\mathbf{x}_2 &= (1,-1), \quad z_2 = -1 \textbf{ for } \omega_2 \\
\mathbf{x}_3 &= (-1,-1), \quad z_3 = 1 \textbf{ for } \omega_1 \\
\mathbf{x}_4 &= (-1,1), \quad z_4 = -1 \textbf{ for } \omega_2
\end{aligned}$$

Example: XOR

$$L(\alpha) = \Sigma\, \alpha_i - \tfrac{1}{2}\, \Sigma\, \Sigma\, \alpha_i\, \alpha_j\, z_i\, z_j\, y(\underline{x}_i)^T y(\underline{x}_j)$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \tfrac{1}{2}\,(1+2+2+2+1+1)\, \alpha_1\, \alpha_1$$

$$+ \tfrac{1}{2}\,(1+2 -2 -2 +1+1)\, \alpha_1\, \alpha_2 + \ldots$$

$$= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \tfrac{1}{2}(9\, \alpha_1 \alpha_1 - 2\, \alpha_1 \alpha_2 - 2\, \alpha_1 \alpha_3 + 2\, \alpha_1 \alpha_4$$

$$+ 9\, \alpha_2 \alpha_2 + 2\, \alpha_2 \alpha_3 - 2\, \alpha_2 \alpha_4 + 9\, \alpha_3 \alpha_3 - 2\, \alpha_3 \alpha_4 + 9\, \alpha_4 \alpha_4)$$

With

$$\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0$$

and

$$\alpha_i \geq 0$$

Example: XOR

$$\partial L / \partial \alpha - 0$$

$$1 = 9\,\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4$$

$$1 = -\alpha_1 + 9\,\alpha_2 + \alpha_3 - \alpha_4$$

$$1 = -\alpha_1 + \alpha_2 + 9\,\alpha_3 - \alpha_4$$

$$1 = \alpha_1 - \alpha_2 - \alpha_3 + 9\,\alpha_4$$

$$\Rightarrow \boxed{\alpha_i = 1/8}$$

With

$$\alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 - 0$$

and

$$\alpha_i \geq 0$$

Example: XOR

$$\mathbf{a} = \sum_{k=1}^{n} \alpha_k z_k \mathbf{y}_k$$

$$\alpha_i = 1/8$$

$$\mathbf{a} = \frac{1}{8}\left[+ \begin{pmatrix} 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ \sqrt{2} \\ -\sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ \sqrt{2} \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ -\sqrt{2} \\ \sqrt{2} \\ -\sqrt{2} \\ 1 \\ 1 \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1/\sqrt{2} \\ 0 \\ 0 \end{pmatrix}$$

$$g(\mathbf{y}) = \mathbf{a}^t\mathbf{y} = (0,0,0,\tfrac{1}{\sqrt{2}},0,0) \cdot (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2) = x_1x_2$$

$$b = 1/\|\mathbf{a}\| = \sqrt{2}$$

$$z_k g(\mathbf{y}_k) \geq 1$$

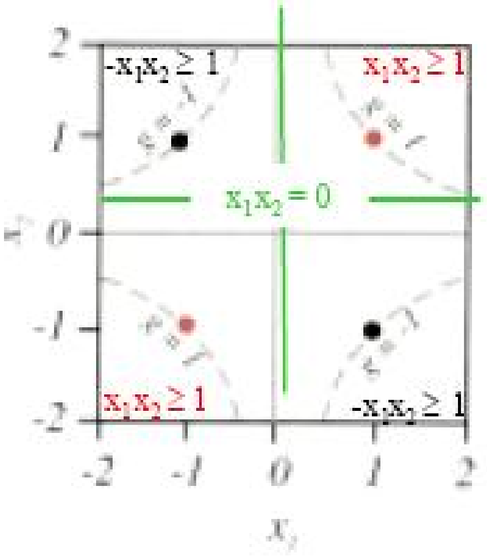
Example: XOR

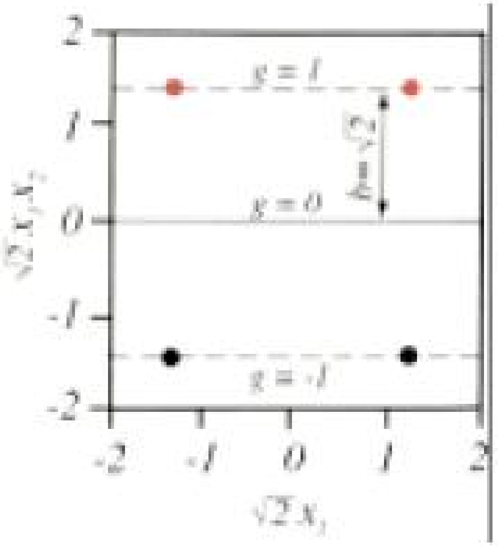
$$z_k g(\mathbf{y}_k) \geq 1$$

$$z x_1 x_2 \geq 1$$

z=1 or -1 depends on w1 or w2

$$\mathbf{x}_1 = (1,1),\quad z_1 = 1 \text{ for } \omega_1$$
$$\mathbf{x}_2 = (1,-1),\quad z_2 = -1 \text{ for } \omega_2$$
$$\mathbf{x}_3 = (-1,-1),\quad z_3 = 1 \text{ for } \omega_1$$
$$\mathbf{x}_4 = (-1,1),\quad z_4 = -1 \text{ for } \omega_2$$





Here, all four prototypes are support vectors

Multicategory Generalization

q $g_i(x) > g_j(x) \Leftrightarrow \mathbf{a}_i^t\mathbf{y} > \mathbf{a}_j^t\mathbf{y}$ " $j \neq i$ if $\mathbf{y} \in \omega_i$

q if $\mathbf{y} \in \omega_1$ $\mathbf{a}_1^t\mathbf{y} - \mathbf{a}_j^t\mathbf{y} > 0$ **for j=2,…,c**

$$a = \begin{bmatrix} a_1 \\ a_2 \\ M \\ a_c \end{bmatrix} \qquad h_{12} = \begin{bmatrix} y \\ -y \\ 0 \\ M \\ 0 \end{bmatrix} \qquad h_{13} = \begin{bmatrix} y \\ 0 \\ -y \\ M \\ 0 \end{bmatrix} \qquad \mathbf{L} \qquad h_{1c} = \begin{bmatrix} y \\ 0 \\ 0 \\ M \\ -y \end{bmatrix}$$

q **So** $a^t h_{1j} > 0$

q **In general, we search** a **so that** $a^t h_{ij} > 0$, **i ¹ j**

q **Using this construction (Kesler), the problem with c classes becomes a problem with 2 classes (d dimensions become cd dimensions-n samples become n(c-1) samples-theoretically applicable!)**