EE 6885 Statistical Pattern Recognition

Fall 2005

Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

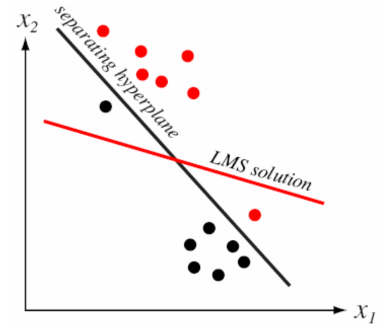
Lecture 12 (10/19/05)

- Reading
  - Linear Discriminant Functions
    - DHS Chap. 5.3-5.6
- Midterm Exam
  - Oct. 24th 2005 Monday 1pm-2:30pm (90mins)
    - Main Material, Textbook Chap. 1 – 5.3
    - Open books/notes, no computer
- Review Class
  - Oct. 21st Friday 4pm. EE Conf. Room (Mudd Rm 1312)

Discriminant Functions (Chap. 5)

- Define discriminant functions, e.g., linear functions

$$\boxed{g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \quad ,\mathbf{w}\text{: weight vector, } w_0 : \text{bias}}$$



- Two-Category Case

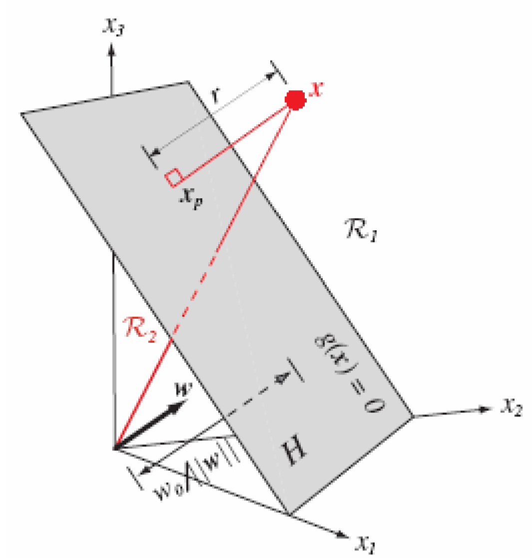
$$\boxed{\text{map } \mathbf{x} \text{ to class } \omega_1 \text{ if } g(\mathbf{x})>0, \text{ otherwise class } \omega_2}$$

$$\text{Decision surface } H: g(\mathbf{x}) = 0$$

$$\boxed{\text{distance from } x \text{ to } H: \; r = g(\mathbf{x}) / \|\mathbf{w}\|}$$

$$\mathbf{x} = \mathbf{x}_p + r \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$\mathbf{x}_p : \text{ projection of } \mathbf{x} \text{ onto } H, \; g(\mathbf{x}_p) = 0$$



Method for searching decision boundaries

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \quad \Rightarrow \text{find weight } \mathbf{w} \text{ and bias } w_o$$

- Augmented Vector

$$\mathbf{y} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \quad \mathbf{a} = \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \quad \Rightarrow g(\mathbf{x}) = g(\mathbf{y}) = \mathbf{a}^t\mathbf{y}$$

- Decision Boundary

$$H: \ (\mathbf{w})^t\mathbf{x} + w_0 = 0 \quad \Rightarrow H: \ (\mathbf{a})^t\mathbf{y} = 0$$

- A hyperplane in augmented y space, with normal vector **a**

all sample points reside in the $y_1 = 1$ subspace

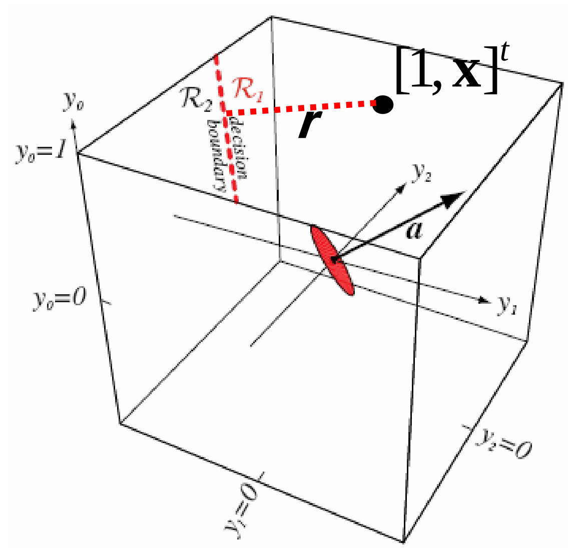
distance from $\mathbf{x}$ to boundary in $\mathbf{x}$ space: $r = \dfrac{g(\mathbf{x})}{\|\mathbf{w}\|}$

distance from $\mathbf{x}$ to boundary in $\mathbf{y}$ space:

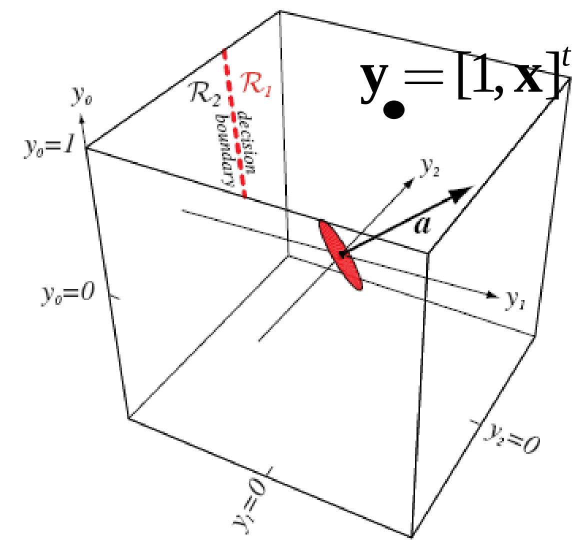
$$r' = \left|\mathbf{a}^t\mathbf{y}\right| / \|\mathbf{a}\| \ \le r$$

i.e., r' and r same signs,

if $r' \ge b$ then $r \ge b$



Search Method for Linear Discriminant



- Design Objective for finding **a**
  - Find **a** that correctly classify each sample data
  - Assume data are separable

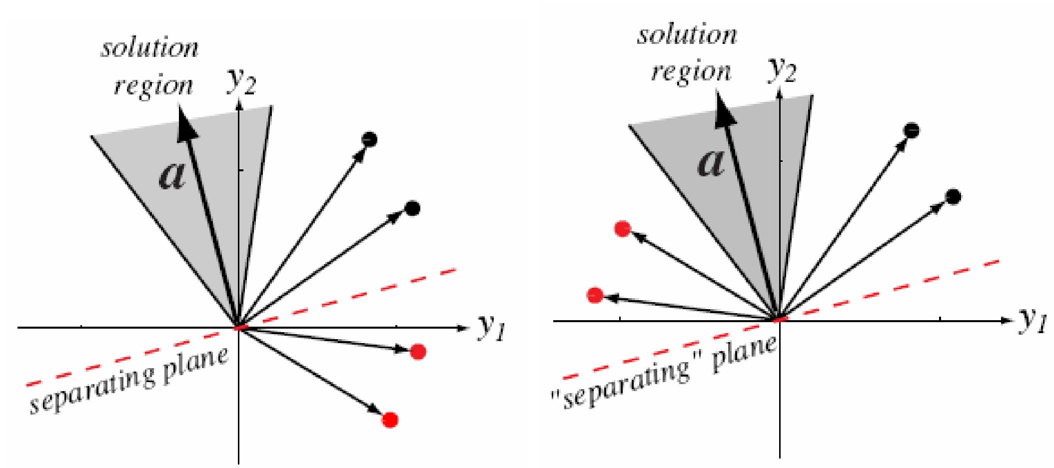
$$\forall \mathbf{y}_i \text{ in class } \omega_1, \mathbf{a}^t\mathbf{y}_i > 0 \quad \forall \mathbf{y}_i \text{ in class } \omega_2, \mathbf{a}^t\mathbf{y}_i < 0$$

- Normalization $\forall \mathbf{y}_i \text{ in class } \omega_2, \mathbf{y}_i \leftarrow -(\mathbf{y}_i)$

- New Design Objective $\mathbf{a}^t\mathbf{y}_i > 0, \ \forall \mathbf{y}_i$

solution $\mathbf{a}$ should be on the positive side of every plane $\mathbf{a}^t\mathbf{y}_i = 0$

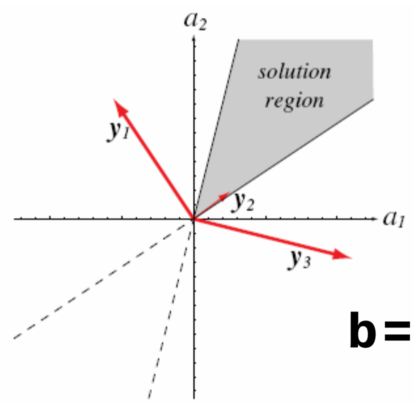
- Solution region
  - Intersection of positive sides of all hyperplanes



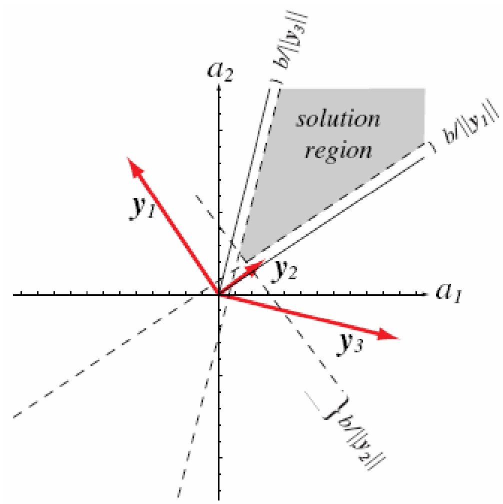
Searching Linear Discriminant Solutions

- Stricter criterion: Solution region with margin b
  - Ensure each sample unambiguously classified

$$\forall \mathbf{y}_i \text{ in class } \omega_1 \text{ or } \omega_2,\ \mathbf{a}^t\mathbf{y}_i > b$$



b=0



b>0

- Search Approaches
  - Gradient decent methods to find a solution in the solution region
  - Maximize margin

Gradient Decent (GD)

- Choose criterion function $\mathbf{J(a)}$
  - $\mathbf{J(a)}$ is minimized when **a** is in the solution region
  - Examples of criterion function
    - # of samples misclassified $\quad$ $\#\text{ of } y \in Y: \text{misclassified samples}$
    - Sum of distances from misclassified samples to H → **perceptron distance**

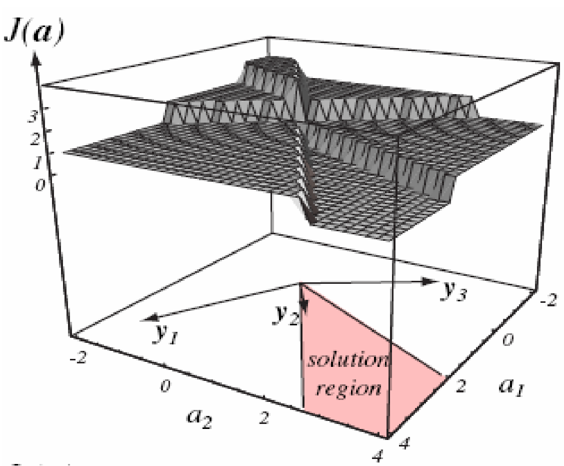
$$\mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y} (-\mathbf{a}^t\mathbf{y}), \quad \text{where } Y \text{ is the set of misclassified samples}$$

    - Quadratic error $\quad \mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y} (\mathbf{a}^t\mathbf{y})^2$
    - Quadratic error with margin (Relaxation Criterion)

$$\mathbf{J}_q(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y} \frac{\left(\mathbf{a}^t\mathbf{y} - b\right)^2}{\|\mathbf{y}\|^2}, \quad \text{where } Y{:}\{\mathbf{y} \mid \mathbf{a}^t\mathbf{y} < b\}$$

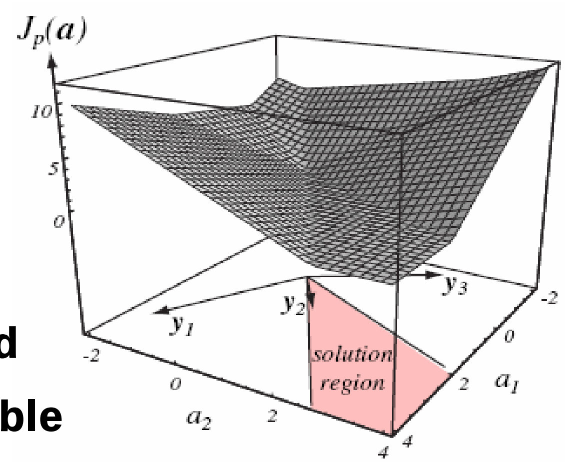
$$\text{Repeat } \ \mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\nabla\mathbf{J}(\mathbf{a}(k)) \qquad \eta(k): \ \text{learning rate}$$

Different Criterion Functions



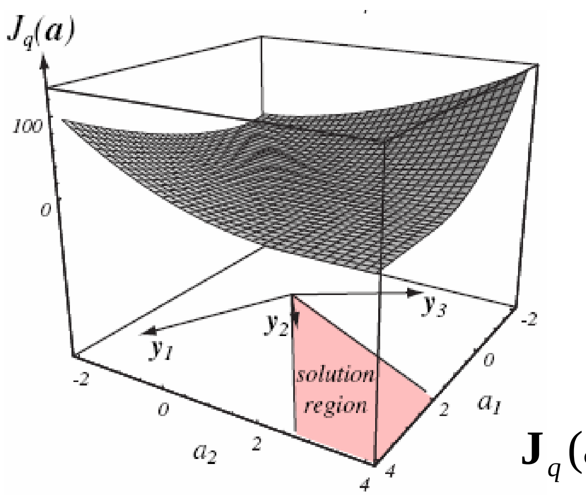
 # misclassified

GD not applicable



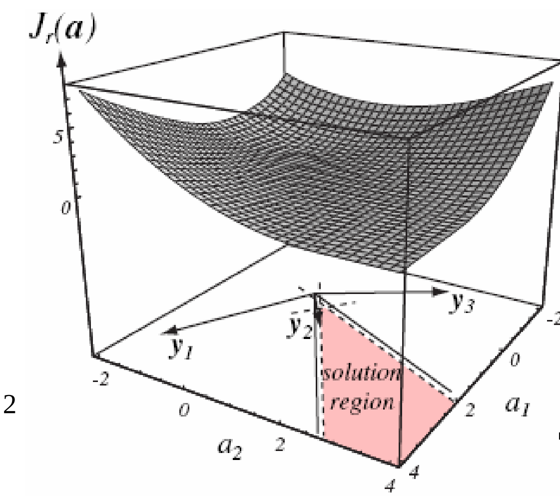
$$\mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(-\mathbf{a}^t\mathbf{y})$$

Not differentiable



$$\mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(\mathbf{a}^t\mathbf{y})^2$$

Smooth, but solutions may be trapped to boundaries



$$\mathbf{J}_q(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y}\frac{\left(\mathbf{a}^t\mathbf{y}-b\right)^2}{\|\mathbf{y}\|^2}$$

Solutions moved away from boundaries

Example: GD based on perceptron criterion

$$\mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(-\mathbf{a}^t\mathbf{y}), \quad \text{where } Y \text{ is the set of misclassified samples}$$

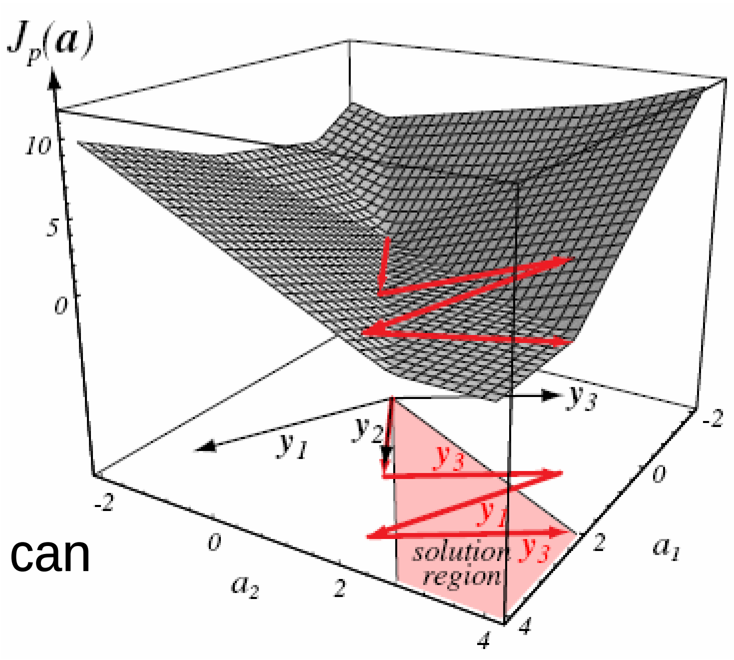
$$\nabla \mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(-\mathbf{y}) \qquad \text{GD: } \mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\nabla\mathbf{J(a}(k))$$

- Batch Perceptron Update

initialize $\mathbf{a}(1)$, choose rate $\eta(.)$, and stop criterion θ

$$\text{Loop} \quad \mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\sum_{\mathbf{y}\in Y}\mathbf{y}$$

$$\text{until} \quad \left|\eta(k)\sum_{\mathbf{y}\in Y}\mathbf{y}\right| < \theta$$

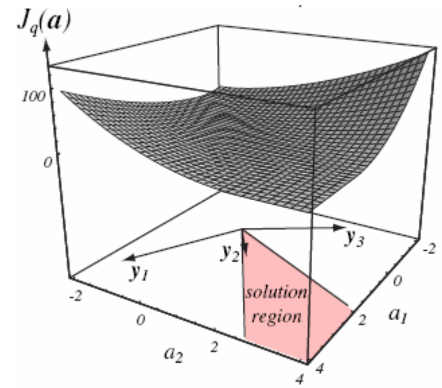


- Example $\mathbf{a}(1) = 0, \ \eta(\text{k})=1$
  - Add sum of misclassified samples

- Theorem:

If samples are separable, then a solution can always be found within finite steps.

Relaxation Procedure



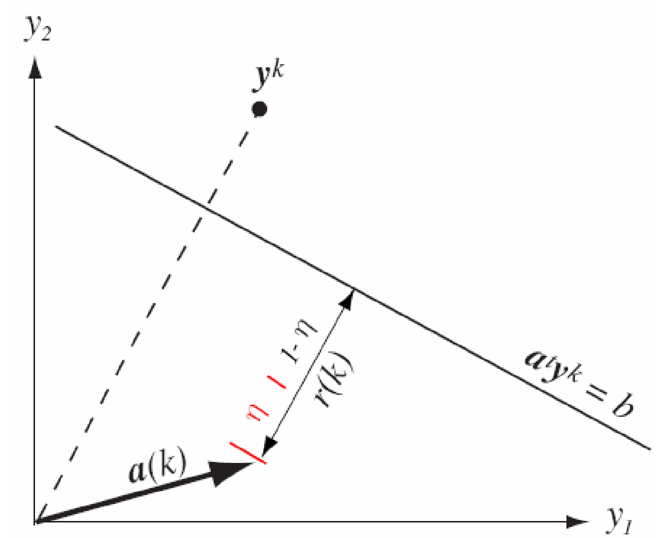
- Problems with Quadratic Criterion $\quad \mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y} (\mathbf{a}^t\mathbf{y})^2$
  - Too smooth, solution trapped at boundaries
  - Dominated by large mis-classified sample
- Relaxation Criterion

$$\mathbf{J}_q(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y} \frac{\left(\mathbf{a}^t\mathbf{y} - b\right)^2}{\|\mathbf{y}\|^2} \qquad \nabla \mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y} \frac{\left(\mathbf{a}^t\mathbf{y} - b\right)}{\|\mathbf{y}\|^2}\mathbf{y}$$

- Gradient Decent with single sample $\mathbf{y}^k$

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

$$= \mathbf{a}(k) + \eta(k)\overset{r(k)}{\boxed{\frac{\left(b - \mathbf{a}^t(k)\mathbf{y}^k\right)}{\|\mathbf{y}^k\|}}}\frac{\mathbf{y}^k}{\|\mathbf{y}^k\|}$$



- Move **a** towards boundary

$$\mathbf{a}^t(k+1)\mathbf{y}^k - b = (1-\eta(k))\left(\mathbf{a}^t(k)\mathbf{y}^k - b\right)$$

$$0 < \eta < 2 \quad , \; \eta < 1: \text{ underrelaxation}, \;\; \eta > 1: \text{ overrelaxation}$$