EE 6885 Statistical Pattern Recognition

Fall 2005
Prof. Shih-Fu Chang
http://www.ee.columbia.edu/~sfchang

Lecture 13 (10/26/05)

- Reading
  - Linear Discriminant Functions
    - DHS Chap. 5.5-5.8
  - Review of vector derivative and chain rule
  - Discriminant Functions with Higher Dimensions
    - DHS Chap. 5.3
- Grading options
  - Option A: complete HW#5-8, no project required
  - Option B: complete a project on image classification, no more HWs
  - Final exam required for either option
- Class schedules
  - No classes on
    - 10/31 (M), 11/7 (M, Uni. Holiday), 11/9 (W), 11/14 (M)
  - Long lectures (start at 12 noon)
    - 11/2 (W), 11/16 (W), 11/21 (M)

Linear Discriminant Classifiers

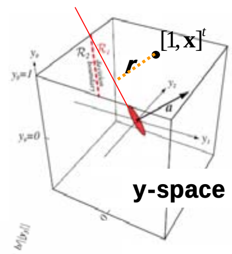
$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 \quad \Rightarrow \text{find weight } \mathbf{w} \text{ and bias } w_o$$

- Augmented Vector

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

map $\mathbf{y}$ to class ω_1 if $g(\mathbf{y})>0$, otherwise class ω_2

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

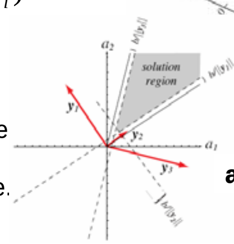


y-space

- Normalization $\quad \forall \mathbf{y}_i$ in class ω_2, $\mathbf{y}_i \leftarrow -(\mathbf{y}_i)$
- Design Objective

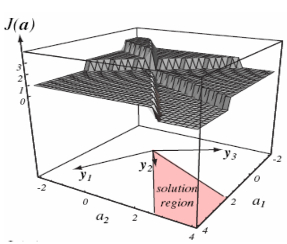
$$\mathbf{a}^t\mathbf{y}_i > b, \ \forall \mathbf{y}_i$$

  - Each y_i defines a half plane in the weight space (a).
  - Note we search weight solutions in the a-space.

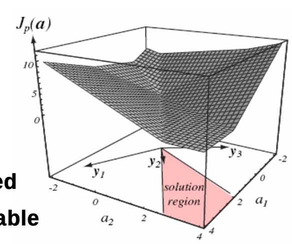


a-space

Gradient Decent Search with Different Criterion Functions

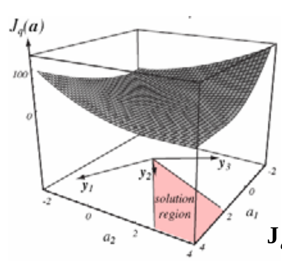


 # misclassified
GD not applicable



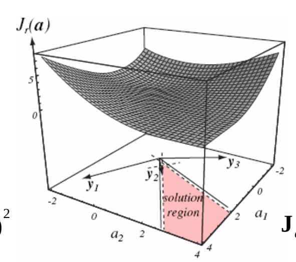
$$\mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y} \in Y} (-\mathbf{a}^t\mathbf{y})$$

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



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

Solutions moved away from boundaries

Example: GD based on perceptron criterion

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

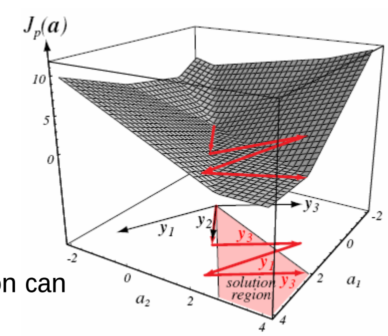
$$\nabla\mathbf{J}_p(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(-\mathbf{y}) \qquad \text{GD: } \mathbf{a}(k+1) = \mathbf{a}(k) - \eta(k)\nabla\mathbf{J}(\mathbf{a}(k))$$

- Batch Perceptron Update

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

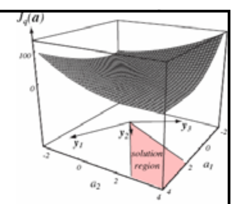
Loop $\mathbf{a}(k+1) = \mathbf{a}(k) + \eta(k)\sum_{\mathbf{y}\in Y}\mathbf{y}$

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



- Example $\mathbf{a}(1) = 0,\ \eta(k)=1$
  - Add sum of misclassified samples
- Theorem:
  If samples are separable, then a solution can always be found within finite steps.

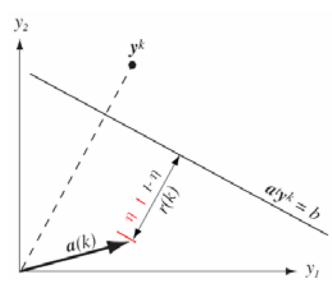
Relaxation Procedure



- Problems with Quadratic Criterion $\mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y}(\mathbf{a}^t\mathbf{y})^2$
  - Too smooth, solution trapped at boundaries
  - Dominated by large mis-classified sample
- Relaxation Criterion $\mathbf{J}_q(\mathbf{a}) = \frac{1}{2}\sum_{\mathbf{y}\in Y}\frac{(\mathbf{a}^t\mathbf{y}-b)^2}{\|\mathbf{y}\|^2} \qquad \nabla\mathbf{J}_q(\mathbf{a}) = \sum_{\mathbf{y}\in Y}\frac{(\mathbf{a}^t\mathbf{y}-b)}{\|\mathbf{y}\|^2}\mathbf{y}$
- Gradient Decent with single sample $\mathbf{y}^k$

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

$$= \mathbf{a}(k) + \eta(k)\underbrace{\frac{(b-\mathbf{a}^t(k)\mathbf{y}^k)}{\|\mathbf{y}^k\|}}_{r(k)}\frac{\mathbf{y}^k}{\|\mathbf{y}^k\|}$$



- Move **a** towards boundary $\mathbf{a}^t(k)\mathbf{y}^k = b$

$0 < \eta < 1$: underrelaxation

$1 < \eta < 2$: overrelaxation

Vector Derivative (Gradient) and Chain Rule

Consider scalar function of vector input: $J(\mathbf{x})$

- **Vector derivative (gradient)** $\nabla_{\mathbf{x}} J(\mathbf{x}) = [\partial J/\partial x_1, \partial J/\partial x_2, \cdots, \partial J/\partial x_d]^t$
- **inner product** $J = \mathbf{a}^t\mathbf{b} = \sum_k a_k b_k \qquad \partial J/\partial a_i = b_i$

$$\Rightarrow \nabla_{\mathbf{a}} \mathbf{a}^t\mathbf{b} = \mathbf{b} \qquad \nabla_{\mathbf{b}} \mathbf{a}^t\mathbf{b} = \nabla_{\mathbf{b}} \mathbf{b}^t\mathbf{a} = \mathbf{a}$$

- **Hermitian** $J = \mathbf{x}^t A\mathbf{x} = \sum_i \sum_j x_i A_{ij} x_j$ $\boxed{\Rightarrow \nabla_{\mathbf{x}} \mathbf{x}^t A\mathbf{x} = A\mathbf{x} + A^t\mathbf{x}}$

  if A is symmetric, then $\nabla_{\mathbf{x}} J = 2A\mathbf{x}$

  if $A = I$, then $\nabla_{\mathbf{x}} J = 2\mathbf{x}$

- **Generalized chain rule**

  now consider $\mathbf{x} = A\mathbf{x}'$, *i.e.* $x_i = \sum_j A_{ij} x_j' \quad \Rightarrow \quad \delta x_i / \delta x_j' = A_{ij}$

$$\nabla_{\mathbf{x}'} J = \left(\frac{\delta x_i}{\delta x_j'}\right)^t \nabla_{\mathbf{x}} J \qquad \boxed{\Rightarrow \nabla_{\mathbf{x}'} J = A^t \nabla_{\mathbf{x}} J}$$

Example of gradient chain rule

if $\mathbf{x} = A\mathbf{x}'$ then $\nabla_{\mathbf{x}'} J = A^t \nabla_{\mathbf{x}} J$

example (mean squared error) $J = \|Y\mathbf{a} - \mathbf{b}\|^2 = (Y\mathbf{a} - \mathbf{b})^t (Y\mathbf{a} - \mathbf{b})$

Let $\mathbf{x} = Y\mathbf{a} - \mathbf{b}, \quad \mathbf{x}' = \mathbf{a}$

$\Rightarrow \quad \mathbf{x} = Y\mathbf{x}' - \mathbf{b}, \quad \nabla_{\mathbf{x}'} J = Y^t \nabla_{\mathbf{x}} J$ **chain rule of gradient**

note $J_{\mathbf{x}} = \mathbf{x}^t\mathbf{x} \quad \Rightarrow \quad \nabla_{\mathbf{x}} J = 2\mathbf{x} = 2(Y\mathbf{a} - \mathbf{b})$

$$\Rightarrow \nabla_{\mathbf{x}'} J = Y^t \nabla_{\mathbf{x}} J = 2Y^t (Y\mathbf{a} - \mathbf{b})$$

$$\boxed{\therefore \nabla_{\mathbf{a}} J = 2Y^t (Y\mathbf{a} - \mathbf{b})}$$

Minimal Squared-Error Solution

$$Y = \begin{bmatrix} \mathbf{y}_1^t \\ \mathbf{y}_2^t \\ \vdots \\ \mathbf{y}_n^t \end{bmatrix}$$

Training sample matrix

dimension: n x (d+1)

Objective: $\mathbf{a}^t\mathbf{y}_i = b, \ \forall \mathbf{y}_i$

$$\Rightarrow \text{define } J_s = \sum_{i=1}^{n} (\mathbf{a}^t\mathbf{y}_i - b_i)^2$$

$$= \|Y\mathbf{a} - \mathbf{b}\|^2 = (Y\mathbf{a} - \mathbf{b})^t (Y\mathbf{a} - \mathbf{b})$$

$$\nabla_{\mathbf{a}} J_s = 2Y^t(Y\mathbf{a} - \mathbf{b}) = 0 \quad \Rightarrow Y^tY\mathbf{a} = Y^t\mathbf{b}$$

$$\text{if } Y^tY \text{ is nonsingular} \quad \Rightarrow \mathbf{a} = \left(Y^tY\right)^{-1} Y^t\mathbf{b} = Y^\dagger \mathbf{b}$$

$$Y^\dagger = \left(Y^tY\right)^{-1} Y^t$$

pseudo-inverse : (d+1) x n

- **Example**

  training samples: *class* $\omega_1 : (1,2)^t, (2,0)^t$ *class* $\omega_2 : (3,1)^t, (2,3)^t$

$$Y = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ -1 & -3 & -1 \\ -1 & -2 & -3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

find $Y^\dagger$, then compute $\mathbf{a}^* = Y^\dagger \mathbf{b}$

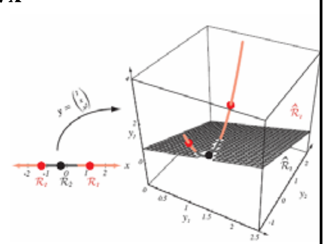
(see figure in textbook)

Generalized Linear Discriminant Functions

- **Include more than just the linear terms**

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij} x_i x_j = w_0 + \mathbf{w}^t\mathbf{x} + \mathbf{x}^t\mathbf{W}\mathbf{x}$$

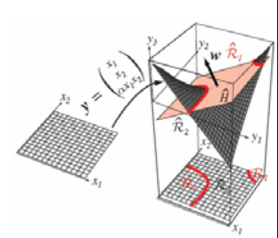
- **Shape of decision boundary**
  - **ellipsoid, hyperhyperboloid, lines etc**
- **In general** $g(\mathbf{x}) = \sum_{i=1}^{\hat{d}} a_i y_i(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$



- **Example**

$$g(x) = a_1 + a_2 x + a_3 x^2$$
$$= [a_1 \quad a_2 \quad a_3][1 \quad x \quad x^2]^t$$

$$g(x) = a_1 x_1 + a_2 x_2 + a_3 x_1 x_2$$
$$= [a_1 \quad a_2 \quad a_3][1 \quad x_1 \quad x_1 x_2]^t$$



- **SVM**
  - **learning all the parameters is hard (curse of dim.)**
  - **instead, try to maximize margins**