Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 05

Avinash Sharma
CVIT, IIIT Hyderabad

Lecture 05: Plan

- Recap
- Minimum Squared Error Procedures
- The Widrow-Hoff /LMS Procedure
- The Ho-Kashyap Procedure

Non Separable Behavior

- Error Correcting Procedures
- Generalization to unseen test data not guaranteed
- Fails to handle non-separable case
- Many Heuristic exists to handle non-separable cases:
  - Forced termination of loop
  - Annealing of η with increasing k

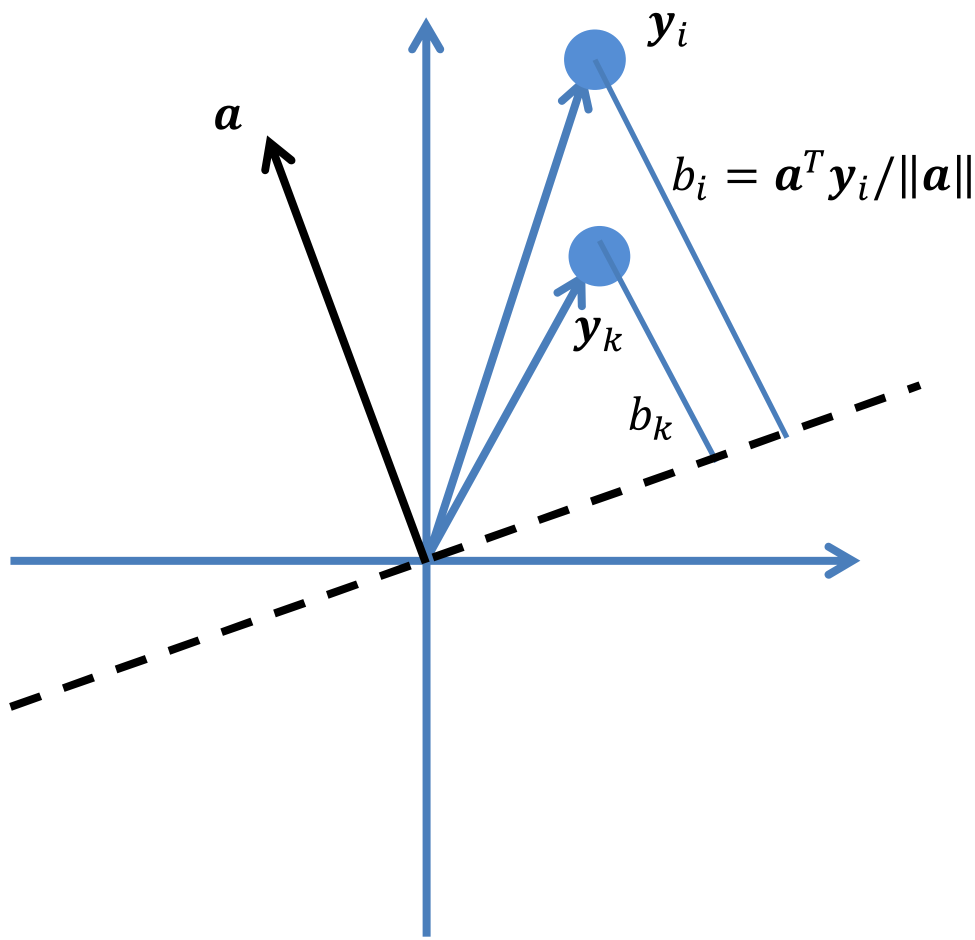
Minimum Squared Error Procedures

- MSE consider all samples instead of just miss-classified ones.

$\boldsymbol{a}^T \boldsymbol{y}_i > 0$ for all samples $\boldsymbol{y}_i$

⇓

$\boldsymbol{a}^T \boldsymbol{y}_i = b_i$ for all samples $\boldsymbol{y}_i$ where $b_i > 0$



Minimum Squared Error Procedures

- $\boldsymbol{e} = \boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}$ (Error definition)

- $J(\boldsymbol{a}) = \|\boldsymbol{e}\|^2 = \|\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}\|^2$

- $\nabla J(\boldsymbol{a}) = 2\boldsymbol{Y}^T(\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b})$=0

- $\boldsymbol{Y}^T\boldsymbol{Y}\boldsymbol{a} = \boldsymbol{Y}^T\boldsymbol{b}$

- $\boldsymbol{a} = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\boldsymbol{b} = \boldsymbol{Y}^\dagger\boldsymbol{b}$

Iterative Gradient Descent for MSE

- $(\boldsymbol{Y}^T\boldsymbol{Y})$ is a square matrix and **often(?)** non-singular and hence invertible.
  - Singular if data points are highly correlated (rows $\boldsymbol{Y}$ are almost linear combination of each other).
- Computing inverse of $(\boldsymbol{Y}^T\boldsymbol{Y})$ can be too expensive for high dimensional data (i.e., large matrix inversion).
- These issues can be avoided by adapting an iterative gradient descent solution.

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \eta(k)\boldsymbol{\nabla}J(\boldsymbol{a}(k)) \text{ or}$$

$$\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \eta(k)\,\boldsymbol{Y}^T(\boldsymbol{Y}\boldsymbol{a}(k) - \boldsymbol{b})$$

The Widrow-Hoff /LMS Procedure

- Reduce storage requirements significantly by considering single sample sequentially, i.e., $\widehat{\boldsymbol{d}}$-dimensional data point ($\boldsymbol{y}_i$) separately.
- Prof. Bernard Widrow and Dr. Ted Hoff proved the convergence of individual "***Least Mean Square (LMS)*** Error" minimization in the stochastic sense.

1. Initialize $\boldsymbol{a}, \boldsymbol{b}, \theta\ (threshold), \eta(\cdot), k = 0, i = 0$
2. do $i = mod(i + 1, n)$
3. $\boldsymbol{a}(k+1) = \boldsymbol{a}(k) - \eta(i)\boldsymbol{y}_i({\boldsymbol{y}_i}^T\boldsymbol{a}(k) - b_i)$
4. until $|\eta(k)\boldsymbol{y}_i({\boldsymbol{y}_i}^T\boldsymbol{a}(k) - b_i)| < \theta$
5. return $\boldsymbol{a}$

*Use Annealing for learning rate $\eta(k)$=$\eta(1)/k$

Single Sample Relaxation with Margin

- Single Sample relaxation with margin

1. Initialize $\boldsymbol{a}, \eta(\cdot), k = 0$

2. do $k = mod(k+1, n)$

   if $\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y}^k \leq b$ then $\boldsymbol{a} = \boldsymbol{a} + \eta(k)\frac{(b-\boldsymbol{a}^T\boldsymbol{y}^k)^2}{\|\boldsymbol{y}^k\|^2}\boldsymbol{y}^k$

3. untill $\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y}^k > b$ for all $\boldsymbol{y}^k$

4. return $\boldsymbol{a}$

Relaxation Procedures (Lecture 04)

The Ho-Kashyap Procedure

- In the MSE procedure, for linearly separable case one can find a separable hyperplane ($\boldsymbol{a}$) iff $\boldsymbol{b} > \mathbf{0}$, which however is chosen arbitrarily.

- Here, the idea is to find both $\boldsymbol{a}$ and $\boldsymbol{b}$, simultaneously using the modified gradient descent procedure that minimizes $J(\boldsymbol{a}, \boldsymbol{b})$.
  - ❖ $J(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{e}\|^2 = \|\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}\|^2$
  - ❖ Fix $\boldsymbol{b}$ and minimize $J(\boldsymbol{a}, \boldsymbol{b})$ w.r.t. to $\boldsymbol{a}$
  - ❖ Fix $\boldsymbol{a}$ and minimize $J(\boldsymbol{a}, \boldsymbol{b})$ w.r.t. to $\boldsymbol{b}$

- The partial derivatives of $J(\boldsymbol{a}, \boldsymbol{b})$ will be
  - ❖ $\nabla_{\boldsymbol{a}} J(\boldsymbol{a}, \boldsymbol{b}) = 2\boldsymbol{Y}^T(\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b})$
  - ❖ $\nabla_{\boldsymbol{b}} J(\boldsymbol{a}, \boldsymbol{b}) = -2\,(\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b})$ or $(\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{b}) = -\frac{1}{2}\nabla_{\boldsymbol{b}} J(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{e}$

The Ho-Kashyap Procedure

- The second step will look like

$$\boldsymbol{b}(k+1) = \boldsymbol{b}(k) - \eta(k)\nabla_{\boldsymbol{b}} J(\boldsymbol{a}(k), \boldsymbol{b}(k))$$

- However, during the iterative gradient descent we need to always ensure that $\boldsymbol{b} > \boldsymbol{0}$.

- This can be enforced by reducing the positive elements of $\nabla_{\boldsymbol{b}} J$ to zero.

$$\boldsymbol{b}(k+1) = \boldsymbol{b}(k) - \eta(k)\tfrac{1}{2}(\nabla_{\boldsymbol{b}} J - |\nabla_{\boldsymbol{b}} J|) \text{ ** or}$$

$$\boldsymbol{b}(k+1) = \boldsymbol{b}(k) + \eta(k)\left(-\tfrac{1}{2}\nabla_{\boldsymbol{b}} J + \tfrac{1}{2}|\nabla_{\boldsymbol{b}} J|\right) \text{ or}$$

**|v| is a vector with all components having absolute values of vector v

$$\boldsymbol{b}(k+1) = \boldsymbol{b}(k) + \eta(k)(\boldsymbol{e}_k + |\boldsymbol{e}_k|)$$

The Ho-Kashyap Procedure

1. Initialize $\boldsymbol{a}, \boldsymbol{b}, \eta(\cdot) < 1, k = 0, threshold(b_{min}, k_{max})$
2. do $k = k + 1$
3. $\boldsymbol{e}_k = (\boldsymbol{Y}\boldsymbol{a}(k) - \boldsymbol{b}(k))$
4. $\boldsymbol{b}(k+1) = \boldsymbol{b}(k) + \eta(k)(\boldsymbol{e}_k + |\boldsymbol{e}_k|)$
5. $\boldsymbol{a}(k+1) = \boldsymbol{Y}^{\dagger}\boldsymbol{b}(k+1)$
6. if $|\boldsymbol{e}_k| \leq b_{min}$ then return $\boldsymbol{a}, \boldsymbol{b}$ and exit
7. until $k < k_{max}$
8. Print “NO SOLUTION”

The Ho-Kashyap Procedure

- Since $\boldsymbol{a}$ is determined by $\boldsymbol{b}$ only (step 5), the procedure can be interpreted as the one which sequentially produce margin vectors.
- Not doing steepest descent anymore, but we are still doing descent and ensure that $\boldsymbol{b}$ is positive.
- In case when none of the element of $\boldsymbol{e}_k$ vector is positive then
  - Either they all are (close to) zero which means we have solution
  - Or there is no solution
- In Linearly Separable case: solution will be found in finite step when we reach to $\boldsymbol{e}_k = \boldsymbol{0}$.
- In the Non Separable case: $\boldsymbol{e}_k$ will have only negative components.
  - No bound on number of iterations needed to prove the non-separability.