ΜΕΘΟΔΟΙ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Εκτίμηση Μέσου Τετραγωνικού Σφάλματος (Mean Square Error Estimation)

Ας επιστρέψουμε και πάλι σε ένα πρόβλημα δυαδικής ταξινόμησης. Στην προηγούμενη θεματική ενότητα είδαμε πως η έξοδος του Perceptron ανήκε στο σύνολο $\{-1, +1\}$ ανάλογα με την πλευρά προέλευσης του διανύσματος χαρακτηριστικών $\underline{x}$. Μάλιστα, υπό την προϋπόθεση της γραμμικής διαχωρισιμότητας των δύο κατηγοριών, η έξοδος του perceptron ταυτίζεται με την σωστή κατηγορία για κάθε διάνυσμα χαρακτηριστικών του συνόλου εκπαίδευσης αφού έχει επιλθει η σύγκλιση του αλγορίθμου.

Στην ενότητα αυτή, θα ασχοληθούμε στην σχεδιασμό ενός γραμμικού ταξινομητή, η επιθυμητή έξοδος του οποίου θα είναι πάλι δυαδική $\{-1, +1\}$ εξαρτώμενη από την πλευρά προέλευσης του διανύσματος εισόδου. Ωστόσο, στην συγκεκριμένη περίπτωση θα επιτρέψουμε την ύπαρξη ταξινομητικών σφαλμάτων. Περιπτώσεις, δηλαδή, όπου οι έξοδος του ταξινομητή θα είναι διαφορετική από την επιθυμητή έξοδο.

Αν επιθυμητή έξοδος του ταξινομητή συμβολισθεί με $y(\underline{x}) = y \in \{-1, +1\}$ (ω_2: -1, ω_1: $+1$) όπου $\underline{x}$ είναι το διάνυσμα των χαρακτηριστικών που τροφοδοτείται ως είσοδος στον ταξινομητή, η έξοδος αυτού θα δίνεται από την σχέση:

$$y(\underline{x}) = \underline{w}^T \underline{x} \quad [1]$$

⊛ Η ύπαρξη σταθερού όρου στην εξίσωση (1) αντιστοιχεί στην θεώρηση επεκτεταμένων διανυσμάτων βαρών και διανυσμάτων χαρακτηριστικών.

Το διάνυσμα των βαρών θα υπολογισθεί έτσι ώστε να ελαχιστοποιείται το μέσο τετραγωνικό σφάλμα ανάμεσα στην επιθυμητή και την πραγματική έξοδο του ταξινομητή. Επομένως, θα πρέπει να ορίσουμε το συναρτησοειδές κόστους ως εξής:

$$J(\underline{w}) = E\left[|y - \underline{x}^T \underline{w}|^2\right] \quad [2]$$

Με βάση τον παραπάνω ορισμό του συναρτησοειδούς κόστους, η βέλτιστη τιμή του διανύσματος των βαρών προκύπτει ως:

$$\underline{w}^* = \arg\min_{\underline{w}} J(\underline{w}) \quad [3]$$

Εύκολα μπορεί να συμπεράνει κανείς πως: $J(\underline{w}) = E\left[(y - \underline{w}^T x)^2\right] \Leftrightarrow$

$$J(\underline{w}) = \int_{-\infty}^{+\infty} p(\underline{x}) \cdot (y(\underline{x}) - \underline{w}^T \underline{x})^2 dx \quad (*)$$

. Ωστόσο, ισχύει ότι:

(i): $p(\underline{x}) = \begin{cases} P(\omega_1), & \underline{x} \in \omega_1; \\ P(\omega_2), & \underline{x} \in \omega_2. \end{cases}$ και (ii) $y(\underline{x}) = \begin{cases} +1, & \underline{x} \in \omega_1; \\ -1, & x \in \omega_2. \end{cases}$

που οδηγεί στην παρακάτω σχέση:

$$J(\underline{w}) = P(\omega_1) \int_{\underline{x} \in \omega_1} (1 - \underline{w}^T \underline{x})^2 d\underline{x} + P(\omega_2) \int_{\underline{x} \in \omega_2} (1 + \underline{w}^T \underline{x})^2 dx \quad [5]$$

Η ελαχιστοποίηση της συνάρτησης κόστους που εμφανίζεται στην εξίσωση (1) πραγματοποιείται κατά τα γνωστά απαιτώντας την συνθήκη πρώτης τάξης, ως εξής:

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \underline{0} \quad [6]$$

Η περαιτέρω ανάλυση της εξίσωσης (6) δίνει:

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \underline{0} \Leftrightarrow \frac{\partial}{\partial \underline{w}} E[(y - \underline{x}^T\underline{w})^2] = \underline{0} \Leftrightarrow E\left[\frac{\partial}{\partial \underline{w}}(y - \underline{x}^T\underline{w})^2\right] = \underline{0} \Leftrightarrow$$

$$E\left[2(y - \underline{x}^T\underline{w})\frac{\partial}{\partial \underline{w}}(y - \underline{x}^T\underline{w})\right] = \underline{0} \Leftrightarrow -2E[\underline{x}(y - \underline{x}^T\underline{w})] = \underline{0} \Leftrightarrow$$

$$E[y \cdot \underline{x} - \underline{x}\underline{x}^T\underline{w}] = \underline{0} \Leftrightarrow E[y \cdot \underline{x}] - E[\underline{x}\underline{x}^T\underline{w}] = \underline{0} \Leftrightarrow$$

$$E[y \cdot \underline{x}] - E[\underline{x}\underline{x}^T]\underline{w} = \underline{0} \Leftrightarrow E[\underline{x}\underline{x}^T] \cdot \underline{w} = E[y \cdot \underline{x}] \Leftrightarrow$$

$$\boxed{\underline{w}^* = \underline{\underline{R}}_x^{-1} \cdot E[y \cdot \underline{x}]} \quad (7), \text{ όπου } \underline{\underline{R}}_x = E[\underline{x}\underline{x}^T] \quad (8)$$

με $\underline{\underline{R}}_x \in \mathbb{R}^{\ell \times \ell}$.

(*) Για τον πίνακα $\underline{\underline{R}}_x$ έχουμε ότι:

$$\underline{\underline{R}}_x = E\left[\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{bmatrix} [x_1\ x_2 \dots x_\ell]\right] = E\begin{bmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_\ell \\ x_\ell x_1 & x_{\ell 2}x_2 & \dots & x_2x_\ell \\ \vdots & \vdots & & \vdots \\ x_\ell x_1 & x_\ell x_2 & \dots & x_\ell x_\ell \end{bmatrix} \Leftrightarrow$$

$$\underline{\underline{R}}_x = \begin{bmatrix} E[x_1x_1] & \dots & E[x_1x_\ell] \\ \vdots & & \\ E[x_\ell x_1] & \dots & E[x_\ell x_\ell] \end{bmatrix} \quad (9)$$

(*) Ο πίνακας $\underline{\underline{R}}_x$ ονομάζεται πίνακας συσχέτισης ή αυτοσυσχέτισης (correlation or autocorrelation matrix) και ταυτίζεται με τον πίνακα συμμεταβλητότητας στην περίπτωση κατά την οποία το μέσο διάνυσμα είναι μηδενικό $\underline{\mu} = \underline{0}$.

(*) Αντίστοιχα, το διάνυσμα $E[y \cdot \underline{x}] = E\left[\begin{bmatrix} yx_1 \\ \vdots \\ yx_\ell \end{bmatrix}\right]$ (10) ποσοτικοποιεί την συσχέτιση (cross-correlation) (ετερο συσχέτιση) ανάμεσα στην επιθυμητή έξοδο και στα διανύσματα χαρακτηριστικών που ο ταξινομητής λαμβάνει στην είσοδο.

Κατά συνέπεια, το βέλτιστο διάνυσμα βαρών στο οποίο οδηγεί η ελαχιστοποίηση της συνάρτησης κόστους που χρησιμοποιείται από την μέθοδο της εκτίμησης του μέσου τετραγωνικού σφάλματος αντιστοιχεί στην επίλυση ενός συνόλου γραμμικών εξισώσεων υπό την προϋπόθεση ότι ο πίνακας αυτοσυσχέτισης R_x είναι αντιστρέψιμος.

Είναι αξιοσημείωτο το γεγονός πως η βέλτιστη λύση για το διάνυσμα των βαρών που συνδέεται με την εκτίμηση του μέσου τετραγωνικού σφάλματος επιδέχεται γεωμετρικής ερμηνείας. Συγκεκριμένα, γνωρίζουμε πως οι τυχαίες μεταβλητές μπορούν να θεωρηθούν ως σημεία σε ένα πολυδιάστατο διανυσματικό χώρο. Μάλιστα, γίνεται εύκολα αντιληπτό πως ο τελεστής της αναμενόμενης τιμής $E[\cdot,\cdot]$ μεταξύ δύο τυχαίων μεταβλητών ικανοποιεί τις ιδιότητες του εσωτερικού γινομένου. Συγκεκριμένα, έχουμε ότι:

$$E[x\cdot y] = Cov(x,y) + E[x]E[y] \quad (12) \qquad \boxed{\varphi(x,y) = E[x\cdot y]} \quad (11)$$

INNER PRODUCT

(*) $Cov(x,y) = E[(x-E[x])(y-E[y])] \Leftrightarrow$

$Cov(x,y) = E[xy - E[y]\cdot x - E[x]\cdot y + E[x]\cdot E[y]] \Leftrightarrow$

$Cov(x,y) = E[x\cdot y] - E[y]E[x] - E[x]E[y] + E[E[x]\cdot E[y]] \Leftrightarrow$

$Cov(x,y) = E[x\cdot y] - E[x]E[y] - E[x]E[y] + E[x]E[y] \Leftrightarrow$

$Cov(x,y) = E[x\cdot y] - E[x]E[y] \Rightarrow \boxed{E[x\cdot y] = Cov(x,y) + E[x]E[y]}$

(i): $\varphi(x,x) \geq 0$, γιατί $E[x^2] = Cov(x,x) + E[x]E[x] \Rightarrow$

$E[x^2] = Var(x) + E^2[x] \Rightarrow$

$$\boxed{E[x^2] = \sigma^2(x) + \mu_x^2} \quad (13)$$

(ii): $\varphi(x,y) = \varphi(y,x)$, γιατί $E[x\cdot y] = Cov(x,y) + E[x]E[y]$ και

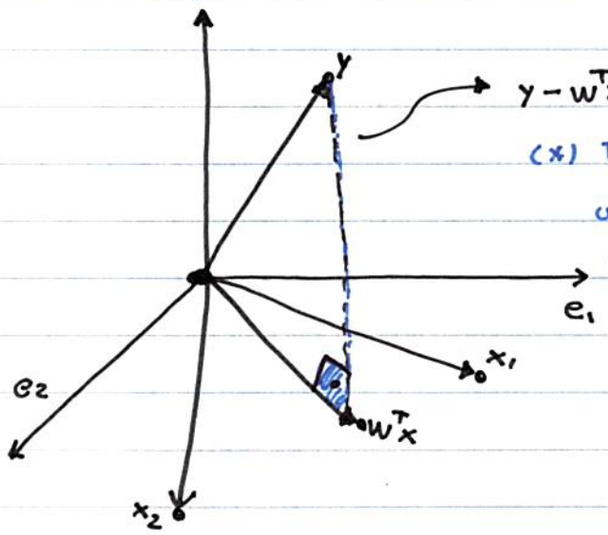
$E[y\cdot x] = Cov(y,x) + E[x]E[y]$ και

$E[xy] = E[y\cdot x]$ καθώς $Cov(x,y) = Cov(y,x)$. (14)

(iii): $\varphi(x,\lambda y+\mu z) = \lambda\varphi(x,y) + \mu\varphi(x,z)$, γιατί $\varphi(x,\lambda y+\mu z) = E[x\cdot(\lambda y+\mu z)] =$

$= E[\lambda xy + \mu xz] = \lambda E[x\cdot y] + \mu E[x\cdot z] = \lambda\varphi(x,y) + \mu\varphi(x,z)$. (15)

Στον συγκεκριμένο διανυσματικό χώρο η ποσότητα $\underline{w}^T\underline{x} = w_1x_1 + \cdots + w_px_p$ υποδηλώνει έναν γραμμικό συνδυασμό διανυσμάτων ο οποίος θα βρίσκεται εντός του υποχώρου που ορίζεται από τα x_i.



(*) Το σφάλμα προσέγγισης του (y) από τον υποχώρο που ορίζουν τα (x_1) και (x_2) γίνεται ελάχιστο όταν γίνεται ορθογώνιο στον εν λόγω υποχώρο.

(*) Η παραπάνω ανάλυση παρουσιάζεται γραφικά στο προηγούμενο σχήμα. Η προσέγγιση του (y) από τον γραμμικό συνδυασμό $w_1x_1 + \cdots + w_px_p$ επιφέρει ένα σφάλμα της μορφής $y - w^Tx$.

(*) Η συνθήκη πρώτης τάξης, που εκφράζεται μέσω της σχέσης (6) $\{ E[\underline{x} \cdot (y - \underline{w}^T\underline{x})] = \underline{0} \}$, υποδηλώνει πως η βέλτιστη τιμή του διανύσματος των βαρών προκύπτει όταν το διάνυσμα του σφάλματος είναι ορθογώνιο σε καθένα από τα x_i. $(1 \le i \le p)$. Με άλλα λόγια, η βέλτιστη εκτίμηση του (y) προκύπτει από μια ορθογώνια προβολή του στον διανυσματικό χώρο που ορίζεται από τα x_i.

Στοχαστική Προσέγγιση και ο Αλγόριθμος Μέσου Τετραγωνικού Σφάλματος

Η επίλυση της εξίσωσης (7) απαιτεί τον υπολογισμό του πίνακα συσχέτισης (correlation matrix) $\underline{R}_x$ και του διανύσματος ετεροσυσχέτισης $E[\underline{x} \cdot y]$ (cross-correlation vector). Ο υπολογισμός, ωστόσο, των εν λόγω διανυσμάτων προϋποθέτει γνώση των υποκείμενων κατανομών που στην γενική περίπτωση είναι άγνωστες. Εξάλλου, αν είχαμε γνώση των κατανομών αυτών για ποιο λόγο να μην χρησιμοποιήσουμε έναν Μαθηματικό ταξινομητή.

Το πρόβλημα, ενομίσως, του υπολογισμού του εκτιμητή των ελαχίστων τετραγώνων αναφέρεται στην επίλυση της εξίσωσης (7) χωρίς την γνώση της απαραίτητης στατιστικής πληροφορίας.

Robbins - Monro Algorithm

Το συγκεκριμένο πρόβλημα λύθηκε από τον αλγόριθμο που πρότειναν οι Herbert Robbins και Sutton Monro το 1951. Η μεθοδολογία που προτάθηκε από τους Robbins και Monro αποτελεί στην πραγματικότητα μια υπολογιστική διαδικασία προσδιορισμού των ριζών μιας συνάρτησης $f(\theta)$, η οποία όπως είναι στην ουσία η αναμενόμενη τιμή μιας συνάρτησης $F(\theta, \xi)$ η οποία εξαρτάται από την τυχαία μεταβλητή ξ. Ισχύει, δηλαδή, ότι:

$$f(\theta) = E_{\xi}[F(\theta, \xi)] \quad (16)$$

Οι αλγόριθμοι στοχαστικής προσέγγισης (Stochastic Approximation Algorithms) στοχεύουν στην ανάλυση ιδιοτήτων της συνάρτησης $f(\cdot)$ παρακάμπτοντας την ανάγκη άμεσης αποτίμησής της. Παραδείγματα, των ενδιαφερόντων ιδιοτήτων της $f(\cdot)$ είναι οι ρίζες της ή τα ακρότατά της.

Ο Robbins και Monro θεώρησαν το πρόβλημα της επίλυσης της εξίσωσης

$$f(\theta) = a \quad (17)$$

όπου a είναι μια σταθερά υπό την προϋπόθεση ότι η εξίσωση:

$$f(\theta) - a = 0 \quad (18)$$

έχει μια μοναδική ρίζα για $\theta = \theta^*$.

Η κεντρική ιδέα των Robbins και Monro βασίστηκε στην υπόθεση πως ενώ η συνάρτηση $f(\theta)$ δεν είναι άμεσα παρατηρήσιμη, είναι δυνατή η απόκτηση δειγμάτων της τυχαίας μεταβλητής $N(\theta)$ για την οποία ισχύει ότι:

$$E[N(\theta)] = f(\theta) \quad (19)$$

Η βασική δομή του αλγορίθμου συνίσταται στην δημιουργία μιας ακολουθίας εκτιμητών για την ζητούμενη ρίζα του θ, η οποία έχει την ακόλουθη μορφή:

$$\theta_{n+1} = \theta_n + a_{n+1}\left(N(\theta_n) - a\right) \quad (20)$$
$$n \geq 0$$

Η ακολουθία $a_1, a_2, \ldots$; είναι μια ακολουθία θετικών βημάτων που χρησιμοποιούνται για την επανεύρεση της ρίζας του θ.

Οι Robbins και Monro απέδειξαν ότι η ακολουθία $(\theta_n)_{n\geq 0}$ συγκλίνει στην ζητούμενη ρίζα θ^* με βάση την μετρική L^2.

$$\lim_{n\to\infty} \|\theta_n\|^2 = \theta^* \quad (21)$$

Η εν λόγω σύγκλιση συνεπάγεται και σύγκλιση της ακολουθίας $(\theta_n)_{n\geq 0}$ σε πιθανότητα:

$$\lim_{n\to\infty} \text{Prob}\left[\theta_n = \theta^*\right] = 1 \quad (22)$$

Τα δύο παραπάνω είδη σύγκλισης εξασφαλίζονται από την προϋπόθεση ότι:

$$\text{(i)}: \sum_{n=1}^{\infty} a_n = \infty \quad (23)$$
$$\text{(ii)}: \sum_{n=1}^{\infty} a_n^2 < \infty \quad (24)$$
$$\Rightarrow \lim_{n\to\infty} a_n = 0 \quad (25)$$

(*) Μια ειδική ακολουθία βημάτων που ικανοποιεί τις παραπάνω συνθήκες είναι η εξής:

$$a_n = \frac{a}{n}, \quad n \geq 1 \text{ και } a > 0 \quad (26)$$

(*) Στην περίπτωση του ενιαίου του μίσου τετραγωνικού σφάλματος η άγνωστη συνάρτηση $F(\cdot,\cdot)$ έχει την ειδική μορφή $F(\underline{x},\underline{w}) = (y - \underline{w}^T\underline{x})^2$, ενώ η συνάρτηση $f(\cdot)$ έχει την ειδική μορφή $f(w) = E_x[F(\underline{x},\underline{w})]$.

(*) Από την σχέση που αναζητούμε την ρίζα της εξίσωσης:

$$E[F(\underline{x},\underline{w})] = 0 \quad (27)$$

η εφαρμογή της μεθοδολογίας των Robbins και Monro οδηγεί στην υιοθεσία του επαναληπτικού σχήματος ενημέρωσης που βασίζεται σε μια ακολουθία τυχαίων διανυσμάτων $\underline{x}_k$ με $k = 1, 2, \ldots$ τα οποία έχουν ληφθεί από την ίδια κατανομή με την συνάρτηση $F(\cdot,\cdot)$, όπου $\underline{w}$ είναι το ζητούμενο διάνυσμα των άγνωστων παραμέτρων:

$$\hat{\underline{w}}(k+1) = \hat{\underline{w}}(k) + a_{k+1} \cdot F(\underline{x}_{k+1}, \hat{\underline{w}}(k)) \quad (28)$$

με $a_k = \frac{a}{k}$ και $k \geq 1$.

(*) Η σύγκλιση της ακολουθίας $(a_k)_{k\geq 1}$ στο μηδέν, $a_k \longrightarrow 0$, εξασφαλίζει πως η διαδικασία ενημέρωσης των βαρών "προχωράει" για μεγάλες τιμές του k, ιδανικά στο άπειρο.

(*) Ωστόσο, η πρώτη συνθήκη πάνω στην ακολουθία των βημάτων ($\sum_{k=1}^{\infty} a_k = \infty$) εξασφαλίζει πως η σύγκλιση δεν επέρχεται πρόωρα προκειμένου να εξασφαλίζεται πως η επαναληπτική διαδικασία δεν "παγώνει" μακριά από την ιδανική λύση.

(*) Η δεύτερη συνθήκη ($\sum_{k=1}^{\infty} a_k^2 < \infty$) εξασφαλίζει πως η συσώρευση του θορύβου που υπεισέρχεται λόγω της στοχαστικής φύσης των μεταβλητών είναι διαχειρίσιμος.

Ας θεωρήσουμε για παράδειγμα την εφαρμογή του αλγοριθμικού πλαισίου των Robbins και Monro στην επίλυση της παρακάτω απλής εξίσωσης:

$$E[\underline{x}-\underline{w}] = \underline{0} \quad (29a) \qquad \text{με} \qquad F(\underline{x},\underline{w}) = \underline{x}-\underline{w} \quad (30)$$

πάνω στη βάση ενός συνόλου τυχαίων δειγμάτων $\underline{x}_k$ για $k=1,2,\ldots$.

Η εφαρμογή της αναδρομικής σχέσης για τον υπολογισμό του $\underline{w}^*$ δίνει:

$$\hat{\underline{w}}(n+1) = \hat{\underline{w}}(n) + a_{n+1} \cdot F(\underline{x}_{n+1}, \hat{\underline{w}}(n)) \Leftrightarrow$$

$$\hat{\underline{w}}(n+1) = \hat{\underline{w}}(n) + \frac{1}{n+1} \cdot (\underline{x}_{n+1} - \hat{\underline{w}}(n)) \Leftrightarrow$$

$$\hat{\underline{w}}(n+1) = \hat{\underline{w}}(n) - \frac{1}{n+1}\hat{\underline{w}}(n) + \frac{1}{n+1}\underline{x}_{n+1} \Leftrightarrow$$

$$\boxed{\hat{\underline{w}}(n+1) = \frac{n}{n+1}\hat{\underline{w}}(n) + \frac{1}{n+1}\underline{x}_{n+1} \quad (31)} \qquad \text{για } n \geq 0$$

Η σχέση (31) γράφεται ως εξής:

$$\hat{\underline{w}}(n+1) = \frac{n}{n+1}\left[\frac{n-1}{n}\hat{\underline{w}}(n-1) + \frac{1}{n}\underline{x}_n\right] + \frac{1}{n+1}\underline{x}_{n+1} \Rightarrow$$

$$\boxed{\hat{\underline{w}}(n+1) = \frac{n-1}{n+1}\hat{\underline{w}}(n-1) + \frac{\underline{x}_n + \underline{x}_{n+1}}{n+1} \quad (32)}$$

Παρατηρούμε επίσης την γενική μορφή της αναδρομικής σχέσης:

$$\hat{\underline{w}}(k) = \frac{k-1}{k}\hat{\underline{w}}(k-1) + \frac{1}{k}\underline{x}_k$$

Η σχέση (32) γράφεται:

$$\hat{W}(n+1) = \frac{n-1}{n+1} \cdot \left[\frac{n-2}{n-1} \cdot \hat{W}(n-1) + \frac{1}{n-1} \cdot x_{n-1} \right] + \frac{x_n + x_{n+1}}{n+1} \Leftrightarrow$$

$$\hat{W}(n+1) = \frac{n-2}{n+1} \cdot \hat{W}(n-1) + \frac{x_{n-1} + x_n + x_{n+1}}{n+1} \quad (33)$$

Από την σχέση (33) μπορούμε να παραγωγίσουμε την παρακάτω γενική μορφή της σχέσης:

$$\hat{W}(n+1) = \frac{n-k}{n+1} \cdot \hat{W}(n+1-k) + \frac{x_{n+1-k} + \cdots + x_{n+1}}{n+1} \Rightarrow$$

$$\hat{W}(n+1) = \frac{n-k}{n+1} \cdot \hat{W}(n+1-k) + \frac{1}{n+1} \cdot \sum_{r=n+1-k}^{n+1} x_r \quad (34)$$

Η σχέση (34) ισχύει προφανώς για $0 \le k \le n$ (k διαφορετικά των δειγμάτων)

Επομένως, για $k=n$, θα έχουμε ότι:

$$\hat{W}(n+1) = \frac{1}{n+1} * \sum_{r=1}^{n+1} x_r \quad (35)$$

⊛ Η σχέση (35) υποδηλώνει πως η βέλτιστη εκτίμηση για την παράμετρο $\underline{W}$ καθώς αυξάνεται το πλήθος των δειγμάτων σύμφωνα με την εκτίμηση του μίγου τετραγωνικού σφάλματος είναι ίση με τον αριθμητικό μίγο των παρατηρήσεων. Η συγκεκριμένη εκτίμηση αποτελεί την πλέον φυσική εκτίμηση.

⊛ Για την περίπτωση της εκτίμησης του μέσου τετραγωνικού σφάλματος με $F(x,w) = x \cdot (y - w^T x)$ θα έχουμε ότι:

$$\hat{\underline{W}}(n+1) = \hat{\underline{W}}(n) + a_{n+1} \cdot F(\underline{x}_{n+1}, \hat{\underline{W}}(n)) \Rightarrow$$

$$\hat{W}(n+1) = \hat{W}(n) + a_{n+1} \cdot x_{n+1} \cdot (y_{n+1} - \hat{W}^T(n) x_{n+1}) \quad (36)$$