Αλγόριθμοι και Πολυπλοκότητα

Ν. Μ. Μισυρλής

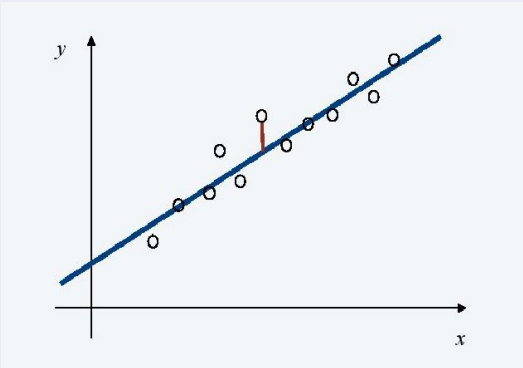
Τμήμα Πληροφορικής και Τηλεπικοινωνιών,
Πανεπιστήμιο Αθηνών

Ελάχιστα τετράγωνα (Least squares)

Ελάχιστα τετράγωνα. Θεμελιώδες πρόβλημα της Στατιστικής.

- Δεδομένου ενός συνόλου P με n σημεία στο επίπεδο: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Να βρεθεί ευθεία L με εξίσωση $y = ax + b$ που ελαχιστοποιεί το άθροισμα του τετραγώνου του σφάλματος:

$$Error(L, P) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$



Ελάχιστα τετράγωνα (Least squares)

Ελάχιστα τετράγωνα. Θεμελιώδες πρόβλημα της Στατιστικής.

- Δεδομένου ενός συνόλου P με n σημεία στο επίπεδο: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Να βρεθεί ευθεία L με εξίσωση $y = ax + b$ που ελαχιστοποιεί το άθροισμα του τετραγώνου του σφάλματος:

$$Error(L, P) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Λύση. Το ελάχιστο σφάλμα επιτυγχάνεται όταν

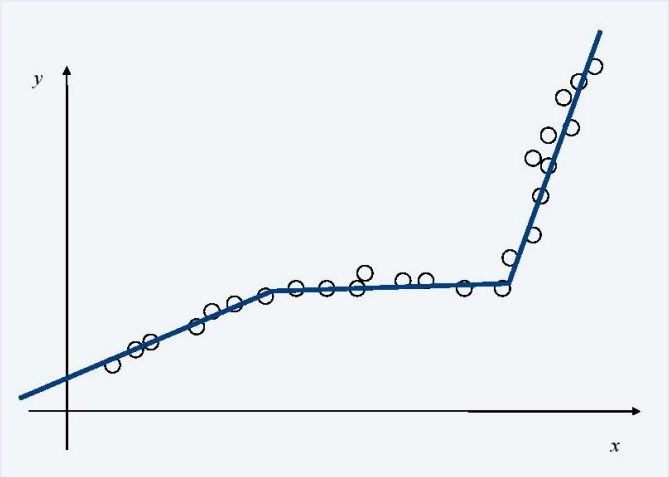
$$a = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i^2 - (\sum_i x_i)^2}, \; b = \frac{\sum_i y_i - a\sum_i x_i}{n}$$

Τμηματοποιημένα ελάχιστα τετράγωνα (Segmented least squares)

Τμηματοποιημένα ελάχιστα τετράγωνα

- Τα σημεία βρίσκονται κοντά σε μία ακολουθία ευθύγραμμων τμημάτων.
- Δεδομένων n σημείων στο επίπεδο: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ με $x_1 < x_2 < \ldots x_n$, να βρεθεί ακολουθία ευθύγραμμων τμημάτων που ελαχιστοποιεί μια συνάρτηση $f(x)$.

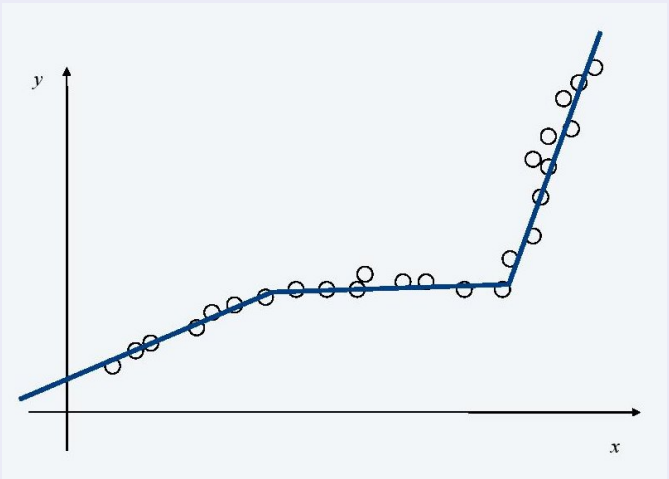
Ερώτηση. Ποιά $f(x)$ να επιλέξουμε ώστε να εξασφαλίσουμε ισορροπία μεταξύ ακρίβειας (καλή προσαρμογή) και οικονομίας (πλήθος τμημάτων);



Τμηματοποιημένα ελάχιστα τετράγωνα (Segmented least squares)

Δεδομένων n σημείων στο επίπεδο: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ με $x_1 < x_2 < \ldots x_n$ και μίας σταθεράς $C > 0$, να βρεθεί ακολουθία ευθειών που ελαχιστοποιούν την $f(E, L) = E + CL$:

1. $E =$ Το άθροισμα των αθροισμάτων των τετραγώνων των σφαλμάτων της βέλτιστης ευθείας που διέρχεται από κάθε τμήμα.
2. $L =$ το πλήθος των ευθύγραμμων τμημάτων στα οποία διαμερίζουμε το σύνολο των σημείων P.



Σχεδιασμός Αλγορίθμου

- Στόχος μας είναι η εύρεση μιας διαμέρισης με

$$min\ f(E, L)$$

- Αύξηση του L σημαίνει μείωση του E
- Μείωση του L σημαίνει αύξηση του E
- Υπάρχουν εκθετικά πολλές διαμερίσεις του P
- Θέλουμε να διαμερίσουμε n αντικείμενα
- Αναλογία με Πολλαπλασιασμό Αλληλουχίας Πινάκων

Συμβολισμοί.

- $p_i = (x_i, y_i)$
- $OPT(j) =$ βέλτιστη λύση για τα σημεία $p_1, p_2, \ldots, p_j$.
- $e_{i,j} =$ ελάχιστο σφάλμα για τα σημεία $p_i, p_{i+1}, \ldots, p_j$.

Σχεδιασμός Αλγορίθμου

Παρατήρηση

- Το τελευταίο σημείο p_n ανήκει σε ένα μόνο τμήμα της βέλτιστης διαμέρισης και αυτό το τμήμα ξεκινά από κάποιο προγενέστερο σημείο p_i
- Αν γνωρίζουμε την ταυτότητα του τελευταίου τμήματος τότε μπορούμε να ανάγουμε το πρόβλημα για τα υπόλοιπα σημεία $p_1, p_2, \ldots, p_{i-1}$.

Υπολογισμός του $OPT(n)$**:**

- Το τελευταίο τμήμα χρησιμοποιεί τα σημεία $p_i, p_{i+1}, \ldots, p_n$.
- $OPT(n) = e_{i,n} + C + OPT(i-1)$. (ιδιότητα βέλτιστης υποδομής)

$$OPT(n) = \begin{cases} 0 & \text{αν } n = 0 \\ e_{i,n} + C + OPT(i-1) & \text{διαφορετικά} \end{cases}$$

Σχεδιασμός Αλγορίθμου

Υπολογισμός του $OPT(j)$:

- Για το υποπρόβλημα των σημείων $p_1, p_2, \ldots, p_j$ ισχύει

$$OPT(j) = \begin{cases} 0 & \text{αν } j = 0 \\ \min\limits_{1 \le i \le j} \{e_{i,j} + C + OPT(i-1)\} & \text{διαφορετικά} \end{cases}$$

και το τμήμα $p_i, p_{i+1}, \ldots, p_j$ χρησιμοποιείται σε μια βέλτιστη λύση για το υποπρόβλημα αν και μόνον αν η ελάχιστη τιμή επιτυγχάνεται με την χρήση του δείκτη i.

Αλγόριθμος τμηματοποιημένων ελαχίστων τετραγώνων

SEGMENTED-LEAST-SQUARES$(n, p_1, \ldots, p_n, C)$

1. **for** $j = 1$ **to** n
2. **for** $i = 1$ **to** j
3. Υπολόγισε το σφάλμα $e[i, j]$ για το τμήμα $p_i, p_{i+1}, \ldots, p_j$.
4.
5. $M[0] \leftarrow 0$
6. **for** $j = 1$ **to** n
7. $M[j] \leftarrow \infty$
8. **for** $i = 1$ **to** j
9. $q \leftarrow \{e[i, j] + C + M[i - 1]\}$.
10. **if** $q < M[j]$
11. **then** $M[j] = q$
12. $s[j] = i$
13. **return** M και s .

Ανάλυση τμηματοποιημένων ελαχίστων τετραγώνων

Θεώρημα. (Bellman 1961) Ο αλγόριθμος δυναμικού προγραμματισμού λύνει το πρόβλημα τμηματοποιημένων ελαχίστων τετραγώνων σε $O(n^3)$ χρόνο και $O(n^2)$ χώρο.

Απόδειξη.

- Σημείο συμφόρησης είναι ο υπολογισμός του $e_{i,j}$ για $O(n^2)$ ζεύγη.
- $O(n)$ για κάθε ζεύγος χρησιμοποιώντας τον τύπο.

$$a = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n\sum_i x_i^2 - (\sum_i x_i)^2}, b = \frac{\sum_i y_i - a\sum_i x_i}{n}$$

- Συνεπώς η πολυπλοκότητα για τον υπολογισμό των $e_{i,j}$ είναι $O(n^3)$
- Υπολογισμός του $M[j]$ απαιτεί χρόνο $O(n)$
- Για $j = 1, 2, \ldots n$ ο συνολικός χρόνος για τον υπολογισμό του $M[j]$ είναι $O(n^2)$.