



Αναγνώριση Προτύπων - Μηχανική Μάθηση

Παλινδρόμηση (Regression)

Γιάννης Παναγάκης

Artificial intelligence

```
graph TD; AI[Artificial intelligence] --> ML[Machine learning]; ML --> SL[Supervised learning]; ML --> UL[Unsupervised learning]; ML --> RL[Reinforcement learning]; SL --> DL[Deep learning]; UL --> DL; RL --> DL;
```

Machine learning

Supervised
learning

Unsupervised
learning

Reinforcement
learning

Deep learning

Αναγνώριση Προτύπων - Μηχανική Μάθηση

Η **αναγνώριση προτύπων** (pattern recognition) είναι η διαδικασία προσδιορισμού μοτίβων σε δεδομένα (εικόνες, ήχο, κείμενο, κτλ) χρησιμοποιώντας κάποια μέθοδο **μηχανικής μάθησης** (machine learning). Διακρίνεται σε:

- **Supervised pattern recognition:** χρησιμοποιεί μεθόδους μηχανικής μάθησης με επίβλεψη (supervised learning) ώστε να αναγνωρίσει δεδομένα ως μέλη κάποιων προκαθορισμένων κατηγοριών ή να αντιστοιχίσει σε αυτά κάποιες τιμές που ανήκουν σε κάποιο προκαθορισμένο σύνολο τιμών.
- **Unsupervised pattern recognition:** χρησιμοποιεί μεθόδους μηχανικής μάθησης χωρίς επίβλεψη (unsupervised learning) ώστε να ομαδοποιήσει τα δεδομένα με βάση κάποια λανθάνοντα μοτίβα. Η ομαδοποίηση των δεδομένων δημιουργεί διαμέριση των δεδομένων με βάση τα λανθανόντα μοτίβα, γεγονός που βοηθά σε προβλήματα εξαγωγής γνώσης από δεδομένα, συμπίεση δεδομένων κτλ.

Supervised learning



Bicycle



Apple



Aardvark

Unsupervised learning



Αναπαράσταση δεδομένων

— Στη μηχανική μάθηση, χρησιμοποιούμε διανύσματα για να αναπαραστήσουμε δεδομένα, δηλαδή μετρήσεις ή χαρακτηριστικά (*features*) που αφορούν κάποιο “φαινόμενο”.

		Feature 1	Feature 2	Feature 3	Labels	
		↓	↓	↓	↓	
		fruit	length	width	weight	label
Data sample 1	→	fruit 1	165	38	172	Banana
Data sample 2	→	fruit 2	218	39	230	Banana
		fruit 3	76	80	145	Orange
		fruit 4	145	35	150	Banana
		fruit 5	90	88	160	Orange
		...				
Data sample n	→	fruit n

Supervised vs Unsupervised Learning

- Supervised learning: μάθηση ενός στατιστικού μοντέλου από **επισημασμένα δεδομένα** (labeled data).

example $x_1 \rightarrow$	x_{11}	x_{12}	\dots	x_{1d}	$y_1 \leftarrow$ label
\dots	\dots	\dots	\dots	\dots	\dots
example $x_i \rightarrow$	x_{i1}	x_{i2}	\dots	x_{id}	$y_i \leftarrow$ label
\dots	\dots	\dots	\dots	\dots	\dots
example $x_n \rightarrow$	x_{n1}	x_{n2}	\dots	x_{nd}	$y_n \leftarrow$ label

- Unsupervised learning: μάθηση ενός στατιστικού μοντέλου από **μή-επισημασμένα δεδομένα** (unlabeled data).

example $x_1 \rightarrow$	x_{11}	x_{12}	\dots	x_{1d}
\dots	\dots	\dots	\dots	\dots
example $x_i \rightarrow$	x_{i1}	x_{i2}	\dots	x_{id}
\dots	\dots	\dots	\dots	\dots
example $x_n \rightarrow$	x_{n1}	x_{n2}	\dots	x_{nd}

Διανύσματα

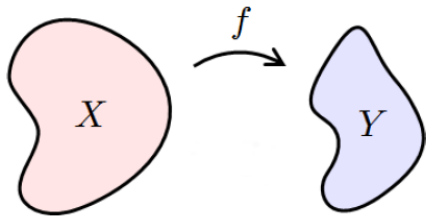
— **Διάνυσμα (vector)**: μια λίστα D βαθμωτών, τα οποία παρατάσσονται σε μία στήλη.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \in \mathbb{R}^D \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \in \mathbb{R}_+^D \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \in [0,1]^D$$

— Το πλήθος των στοιχείων ενός διανύσματος, δηλαδή ο αριθμός D , ονομάζεται **διάσταση** του διανύσματος.

Σύνολα δεδομένων

Supervised learning



— Σύνολο εκπαίδευσης (training set):

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \sim p$$

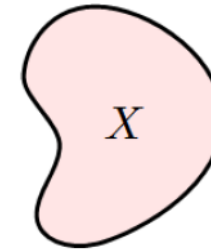
— **Validation set:** χρησιμοποιείται για την επιλογή των υπέρ-παραμέτρων

$$\mathcal{V} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^V \sim p$$

— **Σύνολο ελέγχου (test set):** χρησιμοποιείται για την αξιολόγηση της επίδοσης των αλγορίθμων μηχανικής μάθησης

$$\mathcal{T} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^T \sim p$$

Unsupervised learning



$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N \sim p$$

$$\mathcal{V} = \{\mathbf{x}_i\}_{i=1}^V \sim p$$

$$\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^T \sim p$$

Σύνολα δεδομένων

— Τα παραπάνω υποσύνολα δεδομένων συγκροτούν το **σύνολο δεδομένων (dataset)**:

$$\mathcal{D} \cup \mathcal{V} \cup \mathcal{T} \sim p$$



Αναγνώριση προτύπων σε δεδομένα

1. Διατύπωση του προβλήματος ως πρόβλημα μηχανικής μάθησης.
2. Συλλογή δεδομένων.
3. Προεπεξεργασία δεδομένων.
4. Επιλογή του στατιστικού μοντέλου που θα χρησιμοποιηθεί με βάση κάποιες υποθέσεις για τα δεδομένα.
5. Εκπαίδευση (learning/training/estimation/fitting) του στατιστικού μοντέλου χρησιμοποιώντας δεδομένα εκπαίδευσης (training set).
6. Αξιολόγηση της επίδοσης του μοντέλου σε δεδομένα τα οποία δεν ανήκουν στο σύνολο δεδομένων εκπαίδευσης (test set).

Μάθηση με επίβλεψη

— Σύνολο εκπαίδευσης: $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \sim p$

Δεδομένα εισόδου

Μεταβλητές στόχου

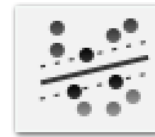
$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in X \subseteq \mathbb{R}^D$

f

$\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}, \mathbf{y}_i \in Y \subseteq \mathbb{R}^K$

— Δεδομένου ενός συνόλου εκπαίδευσης $\{(\mathbf{x}_i, \mathbf{y}_i = f(\mathbf{x}_i))\}_{i=1}^N$ ο στόχος της **μηχανικής μάθησης με επίβλεψη (supervised learning)** είναι να εκτιμήσουμε τις τιμές της συνάρτησης $f(\cdot)$ εκτός του συνόλου εκπαίδευσης, ώστε να μπορούμε να κάνουμε προβλέψεις. Διακρίνεται σε:


- Ταξινόμηση: $\mathcal{Y} = \{c_1, c_2, \dots, c_C\} \subseteq \mathbb{Z}$



- Παλινδρόμηση: $\mathcal{Y} = \mathbb{R}^K$



Μάθηση με επίβλεψη: Ταξινόμηση και Παλινδρόμηση

- Ταξινόμηση: $\mathcal{Y} = \{c_1, c_2, \dots, c_C\} \subseteq \mathbb{Z}$ 

- κατηγορικές μεταβλητές (categorical variables):

1 = brown hair, 2 = red hair, 3 = blonde hair

1 = Adenine, 2 = Thymine, 3 = Cytosine, 4 = Guanine

- Παλινδρόμηση: $\mathcal{Y} = \mathbb{R}^K$ 

- Αριθμητικές μεταβλητές (numerical variables):

17.31 kg, 22.37 kg, 51.34 kg

250 EUR, 1007.5 EUR, 350.98 EUR

Σε αυτή τη διάλεξη: Παλινδρόμηση (Regression)

Στατιστικά μοντέλα:

- Γραμμική παλινδρόμηση
- Πολυωνυμική παλινδρόμηση
- Γραμμικά μοντέλα συναρτήσεων βάσης

Μέθοδοι εκτίμησης παραμέτρων:

- Least squares estimation

Γραμμική παλινδρόμηση

Real world input

Model
input

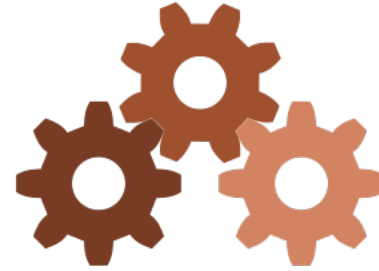
Model

Model
output

Real world output

6000 square feet,
4 bedrooms,
previously sold for
\$235K in 2005,
1 parking spot.

$\begin{bmatrix} 6000 \\ 4 \\ 235 \\ 2005 \\ 1 \end{bmatrix}$



Supervised learning
model

$[340]$

Predicted price
is \$340k

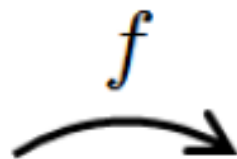
Γραμμική παλινδρόμηση

Σύνολο εκπαίδευσης: N ζεύγη βαθμωτών $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \sim p$

i	x_i	y_i
1	18.5451	27.5695
2	13.0981	58.1471
3	10.1800	62.6110
4	14.1688	48.9667
5	11.4464	48.5689
6	13.6516	49.6105
7	13.7936	44.2449
8	14.1050	41.9451
9	29.0000	50.8000
10	30.0000	53.8000
11	31.0000	55.8000

Δεδομένα εισόδου

$\{x_1, x_2, \dots, x_N\}, x_i \in \mathbb{R}$

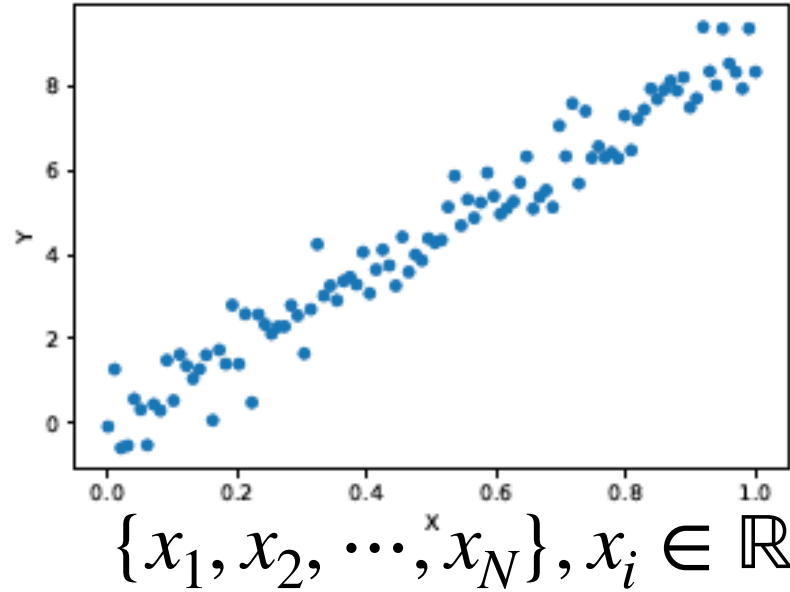


Μεταβλητές στόχου

$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$

Γραμμική παλινδρόμηση

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$



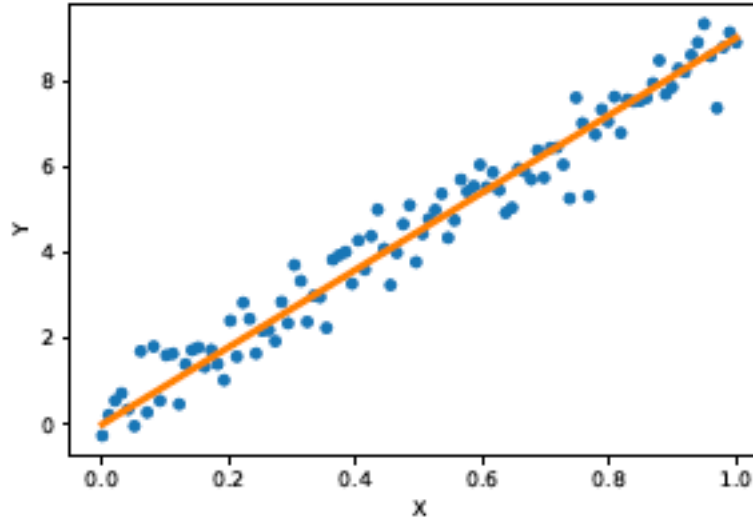
$$\{x_1, x_2, \dots, x_N\}, x_i \in \mathbb{R}$$

— Για κάθε δεδομένο x_i γνωρίζουμε μια συνεχή τιμή y_i , δηλαδή γνωρίζουμε ότι $y_i = f(x_i)$.

— **Στόχος μας** είναι να μοντελοποιήσουμε τη σχέση ανάμεσα στις μεταβλητές x_i και y_i . Με άλλα λόγια, στόχος είναι να μάθουμε την $f(\cdot)$ από τα N σημεία, δεδομένα εκπαίδευσης.

Γραμμική παλινδρόμηση

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$

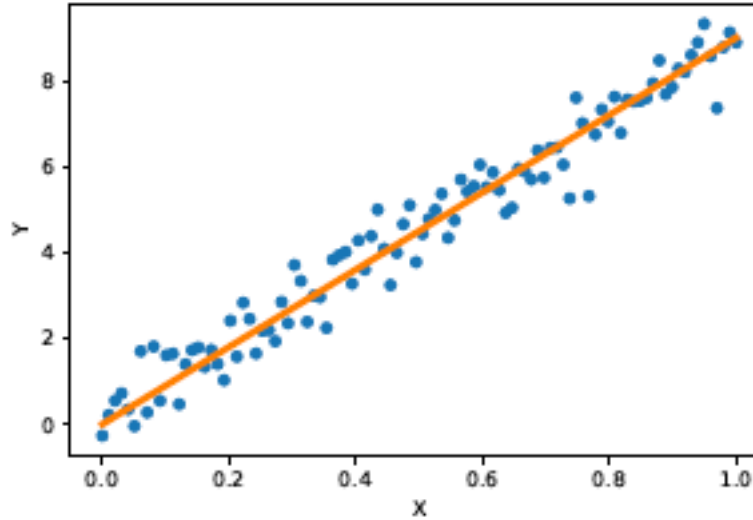


$$\{x_1, x_2, \dots, x_N\}, x_i \in \mathbb{R}$$

- Η ευθεία γραμμή (δηλ. μια γραμμική συνάρτηση) προσεγγίζει την σχέση ανάμεσα στις μεταβλητές x και y .
- Συνεπώς, στόχος είναι να μάθουμε μια γραμμική συνάρτηση μέσω της οποίας να προβλέψουμε την τιμή y από τη μεταβλητή x .

Γραμμική παλινδρόμηση

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$

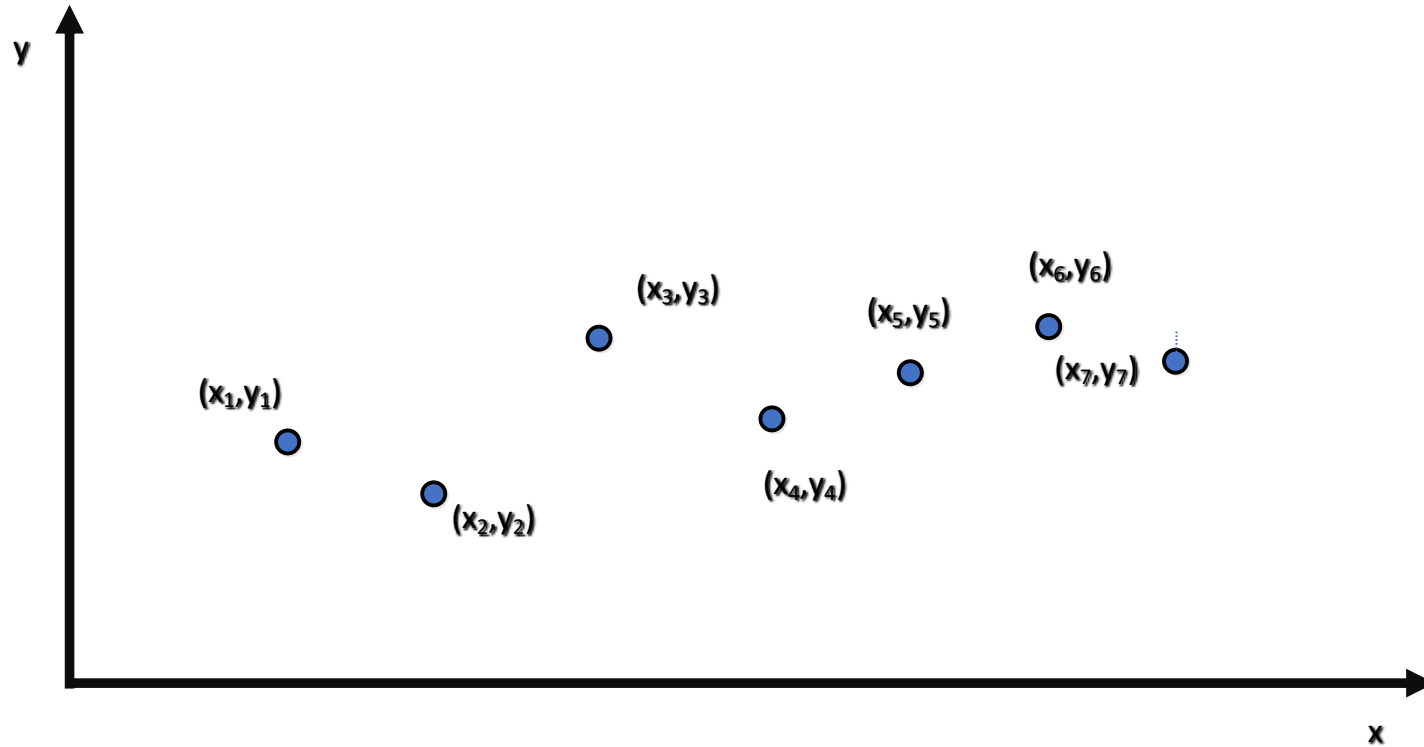


$$\{x_1, x_2, \dots, x_N\}, x_i \in \mathbb{R}$$

- Η γραμμική συνάρτηση καλείται μοντέλο γραμμικής παλινδρόμησης (linear regression).

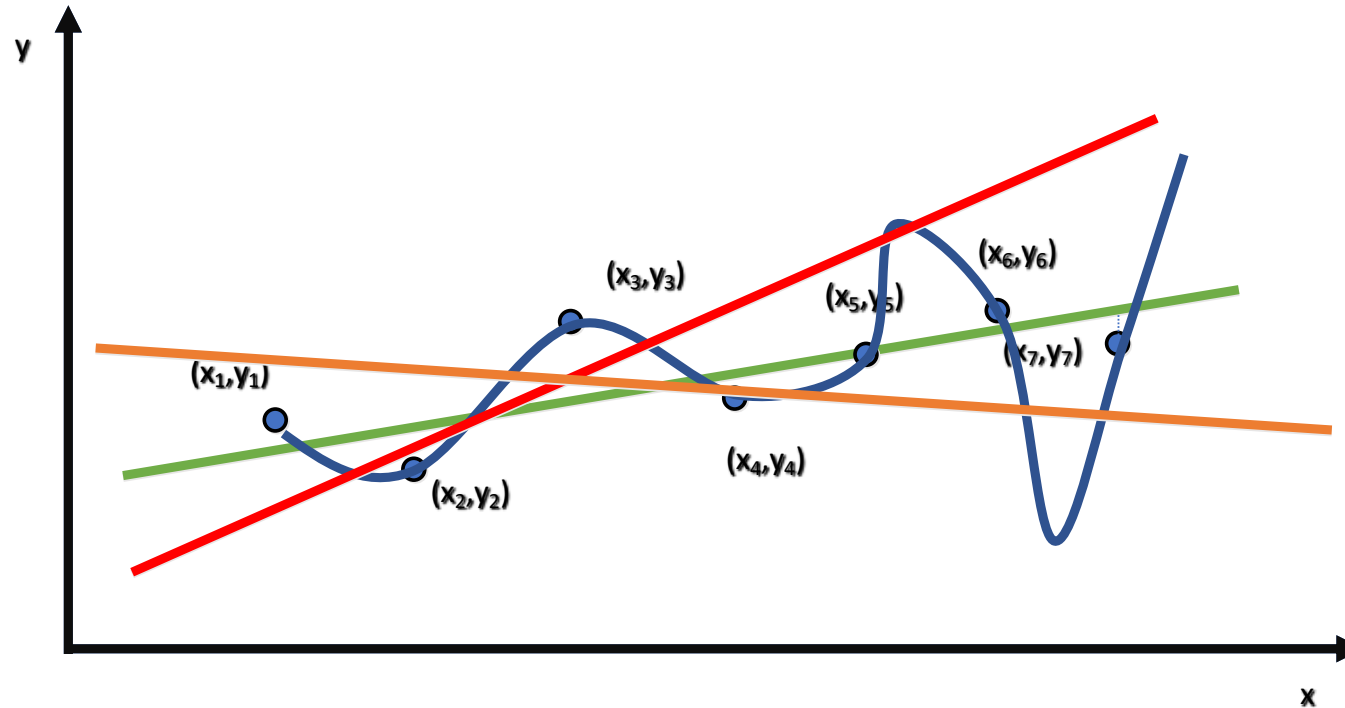
Ποιο είναι ένα καλό μοντέλο;

Σύνολο εκπαίδευσης: $\{(x_i, y_i)\}_{i=1}^7$



Ποιο είναι ένα καλό μοντέλο;

Σύνολο εκπαίδευσης: $\{(x_i, y_i)\}_{i=1}^7$



— Ποια συνάρτηση από όλες θα επιλέξουμε; Ο χώρος υποθέσεων, δηλ. το σύνολο συναρτήσεων που περιγράφουν τα δεδομένα, είναι πολύ μεγάλος.

Γραμμική παλινδρόμηση (linear regression)

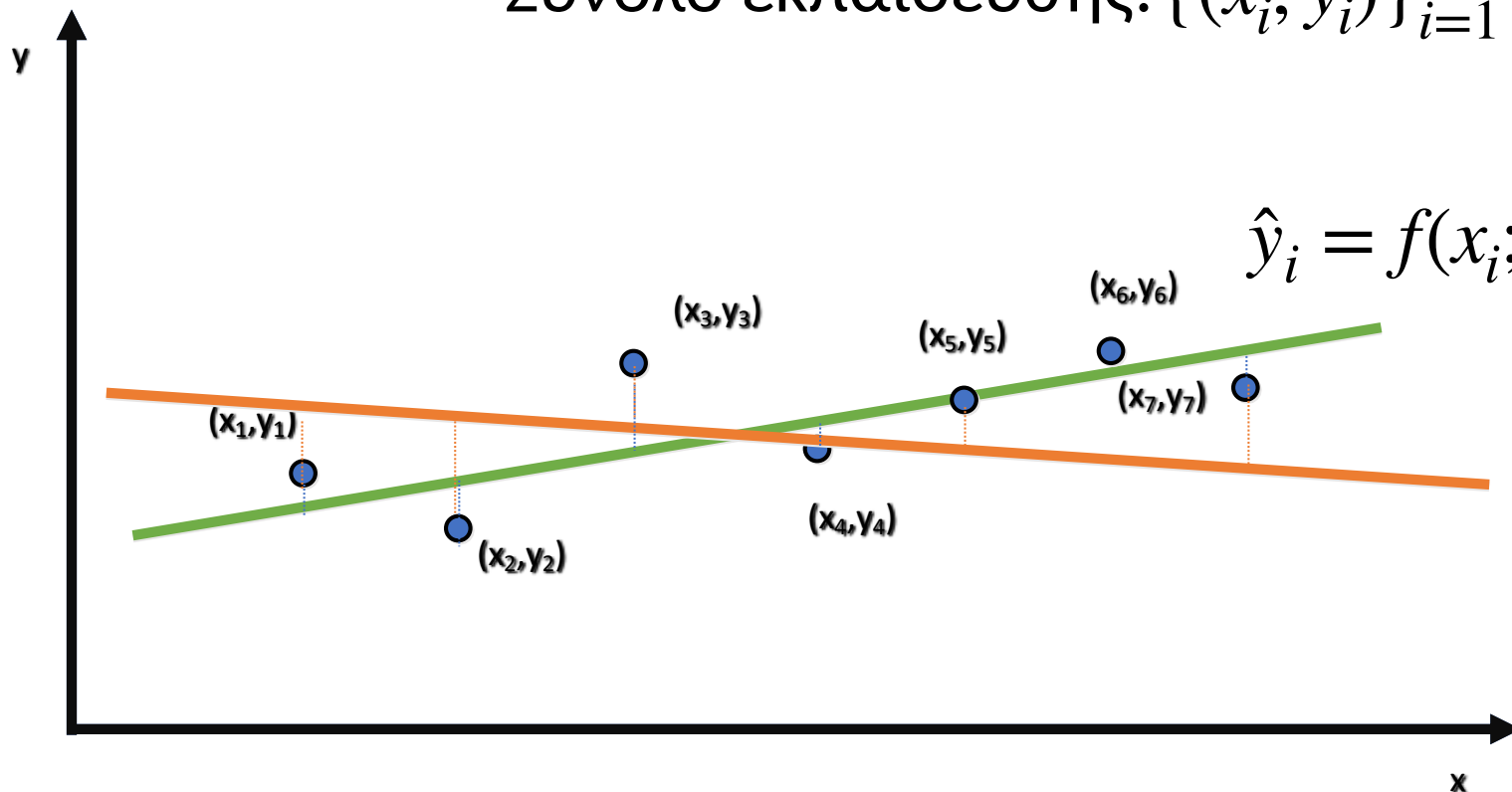
— **Υπόθεση για τα δεδομένα**: γραμμική σχέση ανάμεσα στις μεταβλητές x και y . Για να μοντελοποιήσουμε αυτή τη σχέση αρκεί να παραμετροποιήσουμε την άγνωστη συνάρτηση $f(\cdot)$ ως εξής:

$$y_i = f(x_i; \mathbf{w} = [w_0, w_1]^T) = w_0 + w_1 x_i$$

— Η **μάθηση** της άγνωστης συνάρτησης $f(\cdot)$ ανάγεται στο προσδιορισμό των άγνωστων παραμέτρων, δηλαδή του διανύσματος \mathbf{w} , μέσω του συνόλου εκπαίδευσης.

Linear regression

Σύνολο εκπαίδευσης: $\{(x_i, y_i)\}_{i=1}^7$

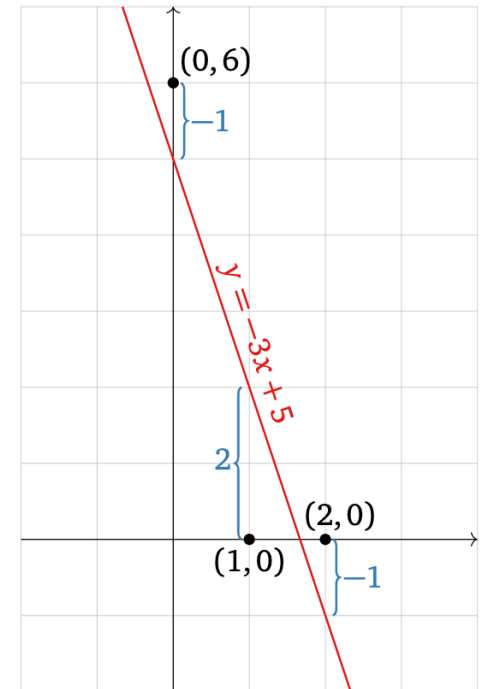


Εκπαίδευση: Προσδιορισμός των παραμέτρων του μοντέλου, έτσι ώστε η εκτίμηση της συνάρτησης $f(\cdot)$ να είναι “**κοντά**” στα δεδομένα εκπαίδευσης.

Linear regression

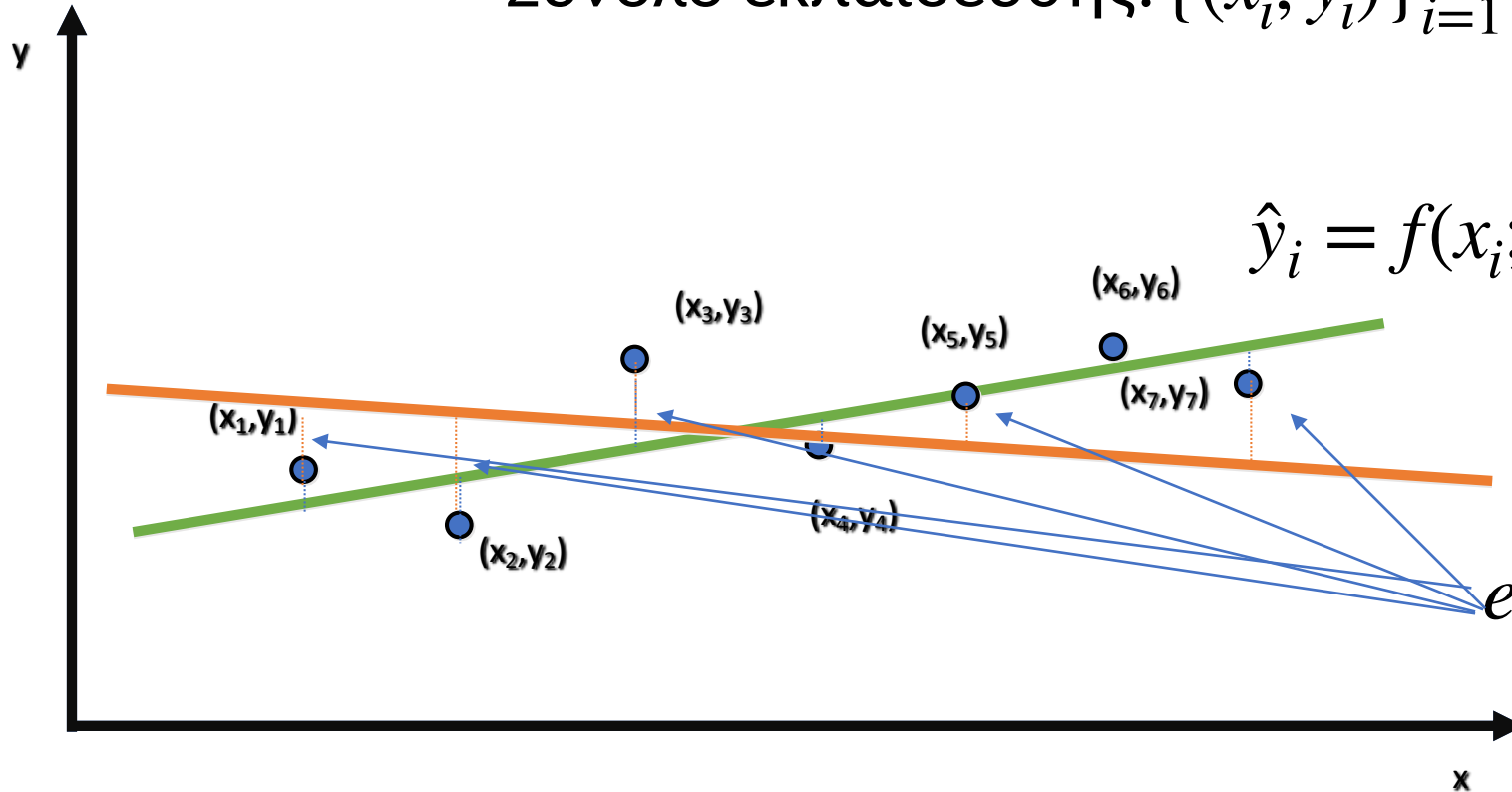
- Υπάρχουν πολλά μέτρα εγγύτητας, π.χ. νόρμες που θα μπορούσαμε να χρησιμοποιήσουμε. Θα περιορίσουμε την προσοχή μας στην μέθοδο **ελαχίστων τετραγώνων (least squares)**.
- Ας ορίσουμε ως **υπόλοιπο (residual)** ορίζεται ως η διαφορά της i -στης μεταβλητής στόχου από την εκτίμηση της:

$$e_i = y_i - \hat{y}_i = y_i - f(x_i; \hat{\mathbf{w}}) = y_i - (\hat{w}_0 + \hat{w}_1 x_i)$$



Linear regression

Σύνολο εκπαίδευσης: $\{(x_i, y_i)\}_{i=1}^7$



$$\hat{y}_i = f(x_i; \hat{\mathbf{w}} = [\hat{w}_0, \hat{w}_1]^T) = \hat{w}_0 + \hat{w}_1 x_i$$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{w}_0 + \hat{w}_1 x_i)$$

Linear regression

— Ορίζουμε ως κριτήριο εγγύτητας το υπόλοιπο άθροισμα τετραγώνων (residual sum of squares)

$$RSS = e_1^2 + e_2^2 + \dots + e_N^2$$

$$= \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^N (y_i - (\hat{w}_0 + \hat{w}_1 x_i))^2$$

— Το RSS αποτελεί τη **συνάρτηση απώλειας (loss function)** της μεθόδου εκτιμησης παραμέτρων που είναι γνωστή ως **ελάχιστα τετράγωνα (least squares)**. Ουσιαστικά η loss function ποσοτικοποιεί το κόστος της πρόβλεψης $f(x;w)$ στη θέση του y .

Linear regression

— Το RSS σε διανυσματική μορφή γράφεται ως

$$RSS = \sum_{i=1}^N (y_i - (\hat{w}_1 x_i + \hat{w}_0))^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

όπου

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

Linear regression

— Το βέλτιστο διάνυσμα παραμέτρων αποτελεί λύση του προβλήματος βελτιστοποίησης

$$\min_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

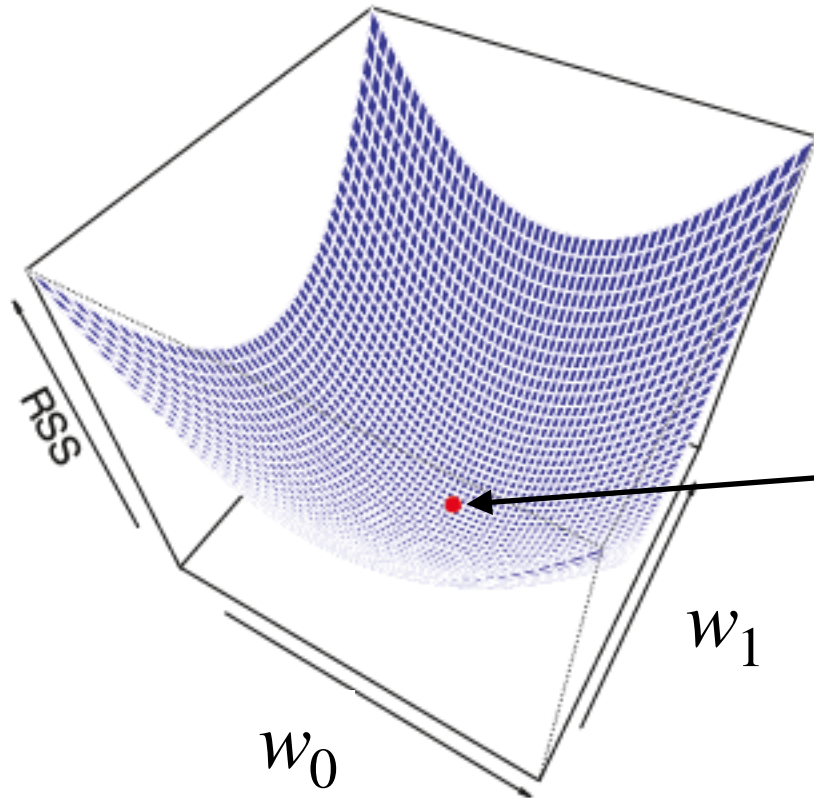
Q: Πώς λύνουμε αυτό το πρόβλημα βελτιστοποίησης (ελαχιστοποίησης);

Linear regression

$$\min_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

— Η Ευκλείδεια νόρμα είναι κυρτή συνάρτηση, συνεχής και παραγωγίσιμη.

$$l(\mathbf{w}) = RSS$$



(Ολικό) ελάχιστο

Linear regression

— Επομένως, το βέλτιστο διάνυσμα παραμέτρων μπορεί να βρεθεί θέτοντας την πρώτη παράγωγο της συνάρτησης ίση με το μηδέν και λύνοντας ως προς \mathbf{w} .

$$l(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) \rightarrow$$

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} = -\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \mathbf{w} = 0$$

$$\mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Σύνοψη: Γραμμική παλινδρόμηση (I/II)

- **Υπόθεση:** Γραμμική σχέση ανάμεσα στα δεδομένα X και Y .
- **Στατιστικό μοντέλο:** γραμμική παραμετρική συνάρτηση

$$y_i = f(x_i; \mathbf{w} = [w_0, w_1]^T) = w_0 + w_1 x_i$$

Ο **χώρος υποθέσεων** \mathcal{H} περιλαμβάνει όλες τις συναρτήσεις που παράγονται από ένα στατιστικό μοντέλο. Για παράδειγμα, ο χώρος υποθέσεων του γραμμικού μοντέλου παλινδρόμησης είναι:

$$\mathcal{H} = \{y = w_0 + w_1 x, \forall \mathbf{w} \in \mathbb{R}^2\}$$

Σύνοψη: Γραμμική παλινδρόμηση (II/II)

— **Υπόθεση:** Γραμμική σχέση ανάμεσα στα δεδομένα X και Y .

— **Στατιστικό μοντέλο:** γραμμική παραμετρική συνάρτηση

$$y_i = f(x_i; \mathbf{w} = [w_0, w_1]^T) = w_0 + w_1 x_i$$

— **Συνάρτηση απώλειας (loss function):** Το κριτήριο των ελαχίστων τετραγώνων αποτελεί μια από τις πολλές επιλογές $l(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$

— **Εκπαίδευση:** Προσδιορισμός των βέλτιστων παραμέτρων του στατιστικού μοντέλου, χρησιμοποιώντας το σύνολο εκπαίδευσης.

$$\min l(\mathbf{w}) = \min \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

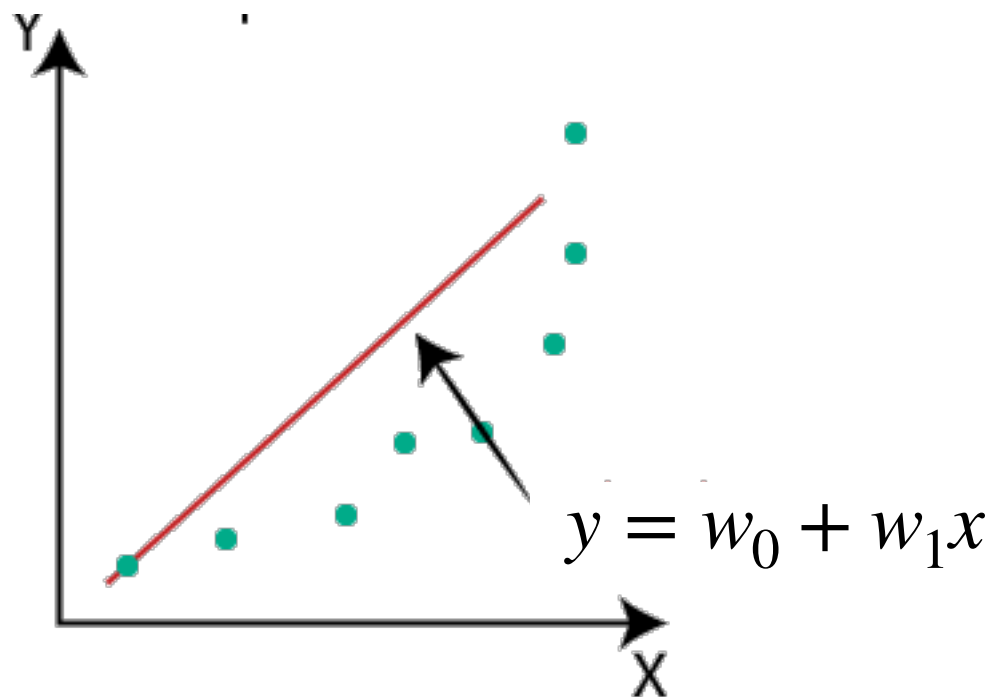
$$\nabla_{\mathbf{w}} l(\mathbf{w}) = \nabla_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = 0 \Rightarrow \hat{\mathbf{w}} = [\hat{w}_0, \hat{w}_1]^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

— **Πρόβλεψη:** Πρόβλεψη της τιμής της μεταβλητής στόχου όταν είναι διαθέσιμο ένα νέο (δηλ. εκτός συνόλου εκπαίδευσης) δεδομένο εισόδου x .

$$\hat{y}_i = \hat{w}_0 + \hat{w}_1 x$$

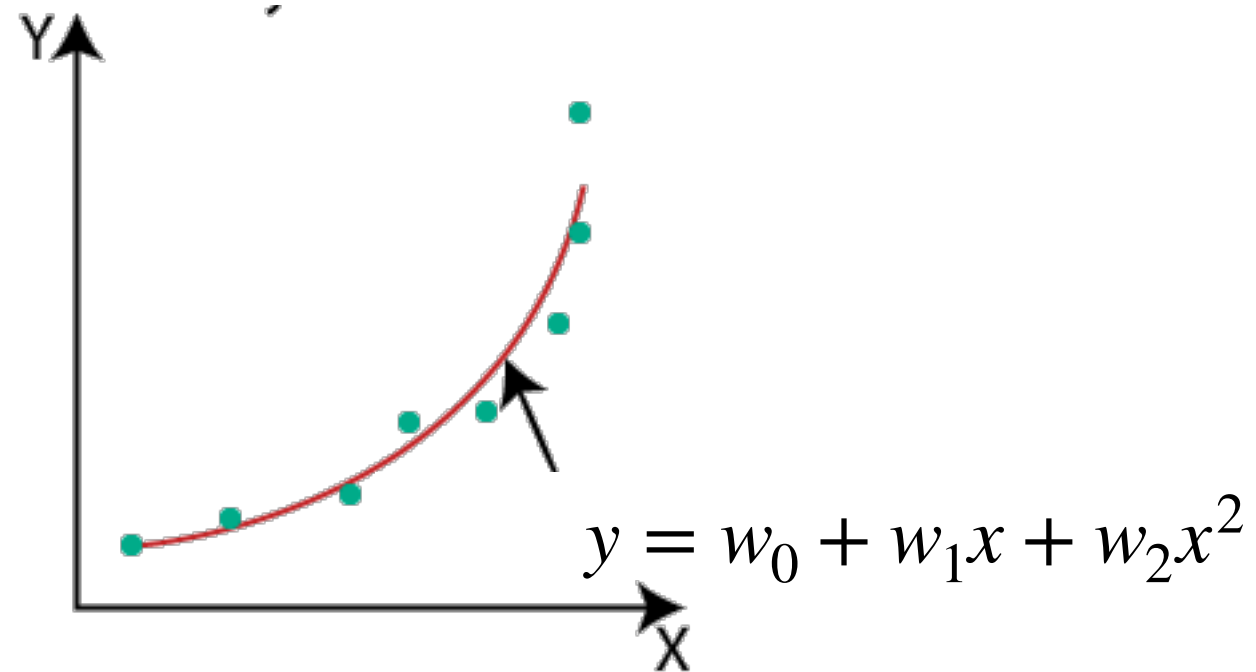
Πολυωνυμική παλινδρόμηση

Στα περισσότερα προβλήματα οι μεταβλητές X και Y παρουσιάζουν **μη-γραμμική** σχέση.



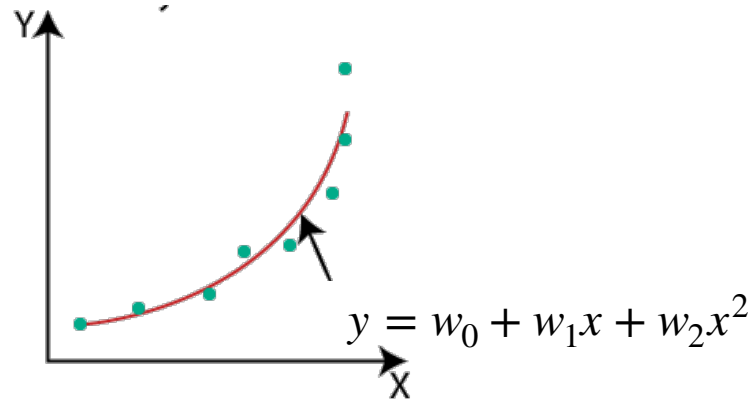
Q: Πώς θα επεκτείνουμε το χώρο υποθέσεων πέρα των γραμμικών συναρτήσεων;

— Υπόθεση: Πολυωνυμική σχέση ανάμεσα στα δεδομένα X και Y .



Πολυωνυμική παλινδρόμηση (polynomial regression)

— **Υπόθεση:** Πολυωνυμική σχέση ανάμεσα στα δεδομένα X και Y .



— **Μοντέλο:** Πολυωνυμική παραμετρική συνάρτηση βαθμού M

$$y_i = f(x_i; \mathbf{w} = [w_0, w_1, \dots, w_M]^T) = w_0 + w_1x_i + \dots + w_Mx_i^M = \sum_{j=0}^M w_jx_i^j$$

* Ο βαθμός του πολυώνυμου (M) αποτελεί **υπερ-παράμετρο (hyper-parameter)** του μοντέλου και προσδιορίζεται από το χρήστη αξιοποιώντας το validation set.

Πολυωνυμική παλινδρόμηση (polynomial regression)

— Συνάρτηση απώλειας (loss function): RSS για πολυωνυμική παλινδρόμηση

$$RSS = \sum_{i=1}^N \left(y_i - \sum_{j=0}^M w_j x_i^j \right)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

όπου

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^M \\ 1 & x_2 & x_2^2 & \dots & x_2^M \\ 1 & x_3 & x_3^2 & \dots & x_3^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & x_N^M \end{bmatrix} \in \mathbb{R}^{N \times M+1} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix} \in \mathbb{R}^{M+1}$$

Πολυωνυμική παλινδρόμηση (polynomial regression)

— Συνάρτηση απώλειας (loss function): RSS για πολυωνυμική παλινδρόμηση

$$RSS = \sum_{i=1}^N \left(y_i - \sum_{j=0}^M w_j x_i^j \right)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$

— Εκπαίδευση:

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$$
$$\hat{\mathbf{w}} = [\hat{w}_0, \hat{w}_1, \dots, \hat{w}_M]^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

— Πρόβλεψη:

$$\hat{y} = f(x_i; \hat{\mathbf{w}} = [w_0, w_1, \dots, w_M]^T) = \hat{w}_0 + \hat{w}_1 x + \dots + \hat{w}_M x^M = \sum_{j=0}^M \hat{w}_j x^j$$

Ψευδοαντίστροφος πίνακας (pseudoinverse)

$$\hat{\mathbf{w}} = \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}}{\mathbf{X}^\dagger}$$

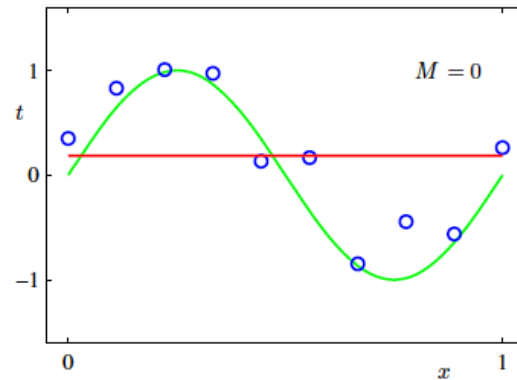
Ψευδοαντίστροφος ή Moore-Penrose inverse: $\mathbf{X}^\dagger = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$

np.linalg.pinv στη numpy

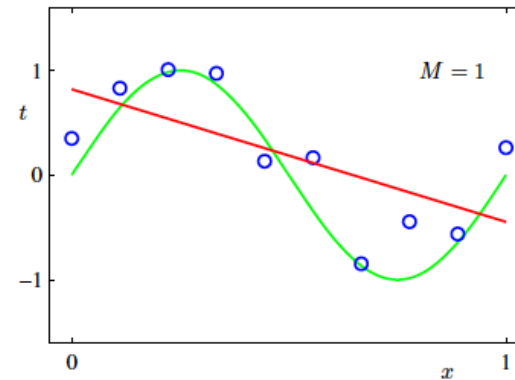
Επιλογή υπερ-παραμέτρου M

— Διαφορετικές επιλογές του βαθμού του πολυώνυμου, M , έχουν επίπτωση στη συμπεριφορά του μοντέλου.

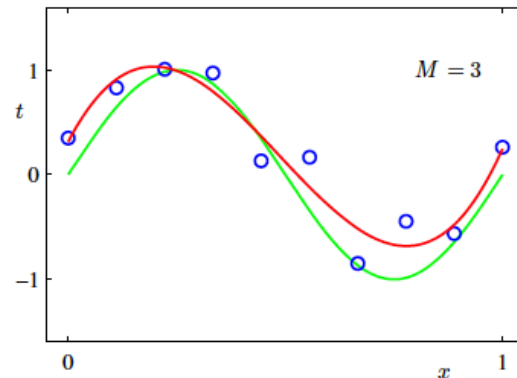
(α)



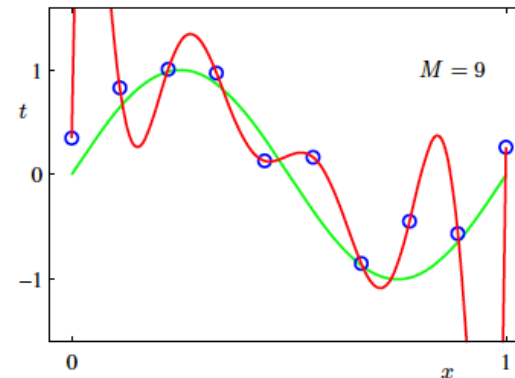
(β)



(γ)

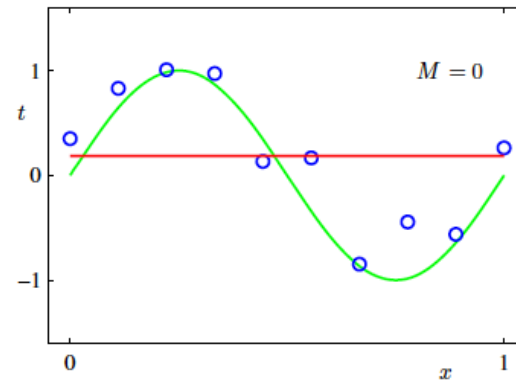


(δ)

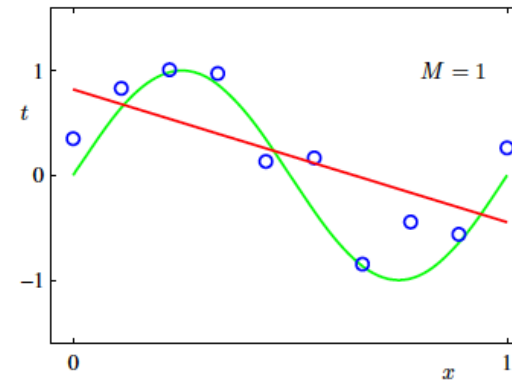


Επιλογή υπερ-παραμέτρου M

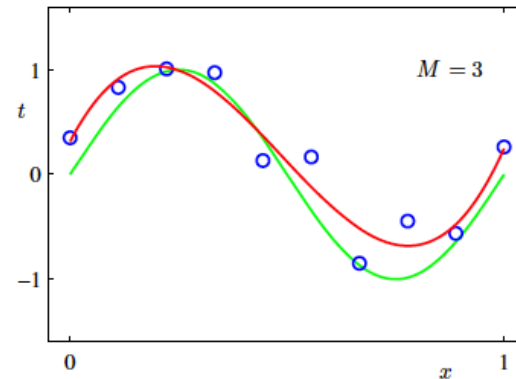
(α)



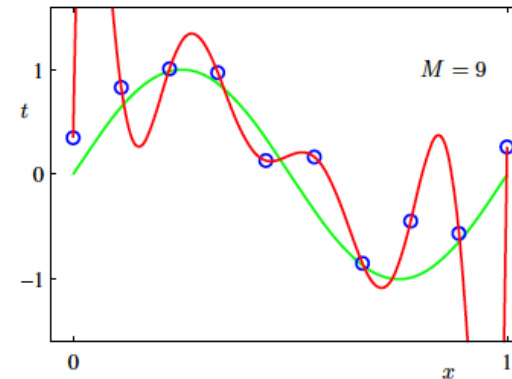
(β)



(γ)



(δ)



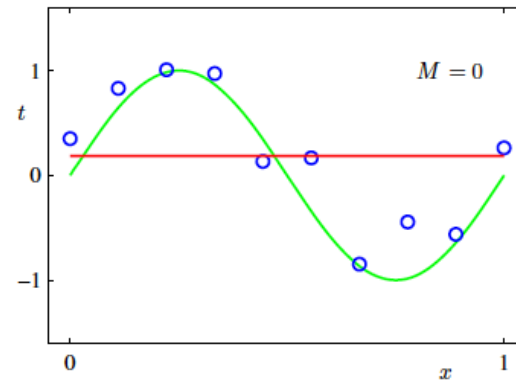
(α),(β) Περιορισμένη ικανότητα γενίκευσης (under-fitting), $M=0$, $M=1$

(γ) Ικανοποιητική ικανότητα γενίκευσης, $M=3$

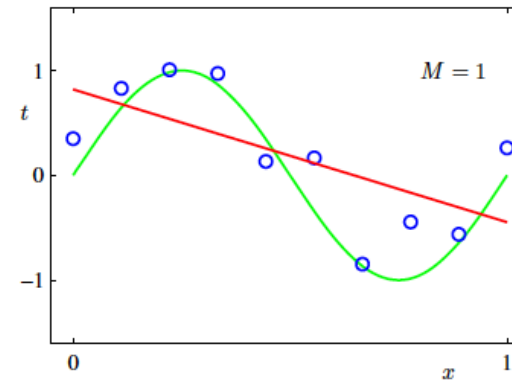
(δ) Περιορισμένη ικανότητα γενίκευσης (over-fitting), $M=9$

Επιλογή υπερ-παραμέτρου M

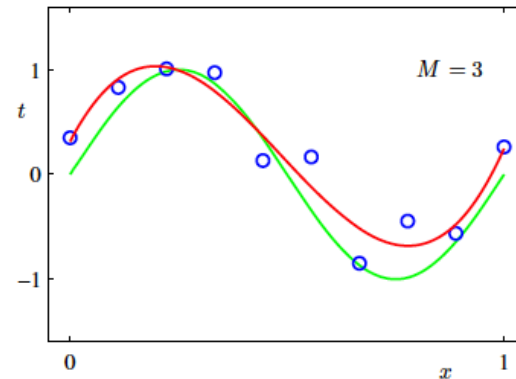
(α)



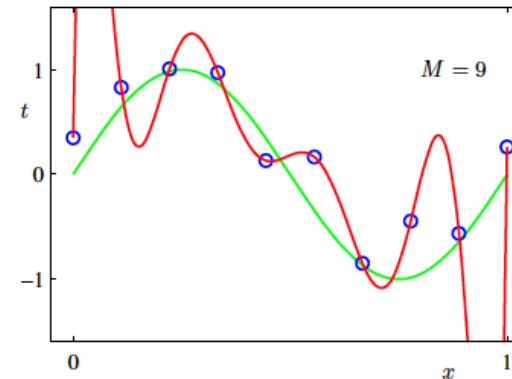
(β)



(γ)



(δ)

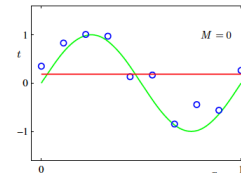


— Ο προσδιορισμός του βαθμού του πολυωνύμου (και γενικά των υπερ-παραμέτρων τα στατιστικά μοντέλα μάθησης) απαιτεί μεθοδολογίες επιλογής μοντέλου (*model selection*) ώστε να μεγιστοποιήσουμε τη γενίκευση.

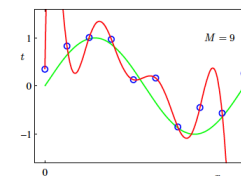
Η χωρητικότητα μοντέλου μάθησης και οι συνέπειες της

- Η χωρητικότητα ενός μοντέλου (model capacity) μηχανικής μάθησης αφορά στην ικανότητα του να μάθει ένα εύρη σύνολο συναρτήσεων μέσω των δεδομένων εκπαίδευσης.
- Ένας τρόπος για να ελέγξουμε τη χωρητικότητα ενός μοντέλου μάθησης είναι με το να τροποποιήσουμε των χώρο υποθέσεων.

- Μοντέλα μικρής χωρητικότητας δεν μπορούν να περιγράψουν σύνθετα δεδομένα ικανοποιητικά (under-fitting)



- Μοντέλα μεγάλης χωρητικότητας μπορεί να απομνημονεύσουν ιδιότητες και χαρακτηριστικά του συνόλου εκπαίδευσης τα οποία δεν παρουσιάζονται στο σύνολο ελέγχου (over-fitting)

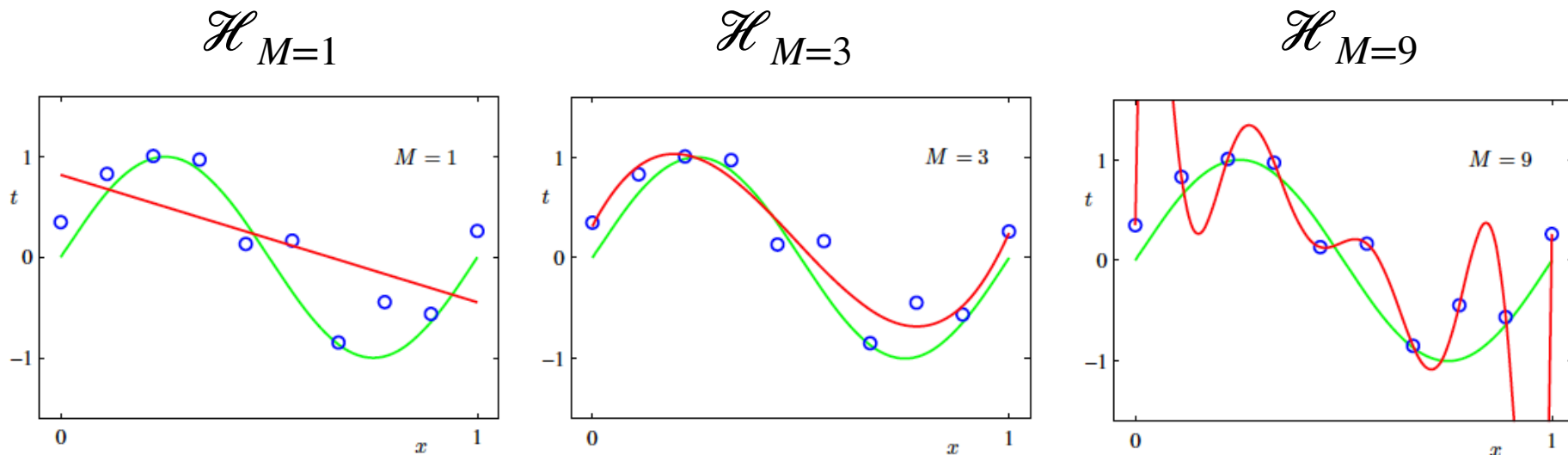


Η χωρητικότητα μοντέλου μάθησης και οι συνέπειες της

— Ο χώρος υποθέσεων μοντέλου πολυωνυμικής παλινδρόμησης

$$\mathcal{H}_M = \{y = \sum_{j=0}^M w_j x^j, \forall \mathbf{w} \in \mathbb{R}^{M+1}\}$$

Παράδειγμα: Πολυωνυμική παλινδρόμηση με μεταβλητό βαθμό πολυώνυμου (M). Μεγαλύτερος βαθμός πολυώνυμου συνεπάγεται μεγαλύτερη χωρητικότητα μοντέλου, δηλαδή $\mathcal{H}_{M=1} < \mathcal{H}_{M=2} < \mathcal{H}_{M=3} \dots$



Γραμμικά μοντέλα συναρτήσεων βάσης

Γραμμικά μοντέλα συναρτήσεων βάσης

—Σύνολο εκπαίδευσης: N ζεύγη διανύσματος-βαθμωτού $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \sim p$

Δεδομένα εισόδου

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$$

Μεταβλητές στόχου

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$

— Γραμμική παλινδρόμηση με διανυσματικά δεδομένα εισόδου:
γραμμικός συνδυασμός των στοιχείων του διανύσματος εισόδου

$$y = f(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + \dots + w_D x_D$$

Σημειώστε: Το παραπάνω μοντέλο αποτελεί γραμμική συνάρτηση των παραμέτρων \mathbf{w} και των μεταβλητών εισόδου.

Γραμμικά μοντέλα συναρτήσεων βάσης

Q: Πως θα επεκτείνουμε την κλάση των μοντέλων (ή ισοδύναμα το χώρο υποθέσεων);

A: Εξετάζοντας γραμμικούς συνδυασμούς **μη-γραμμικών συναρτήσεων** των μεταβλητών εισόδου.

Γραμμικά μοντέλα συναρτήσεων βάσης

—Σύνολο εκπαίδευσης: N ζεύγη διανύσματος-βαθμωτού $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \sim p$

Δεδομένα εισόδου

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$$

Μεταβλητές στόχου

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$

— Γραμμικό μοντέλο συναρτήσεων βάσης: γραμμικός συνδυασμός **μη-γραμμικών συναρτήσεων** των μεταβλητών εισόδου.

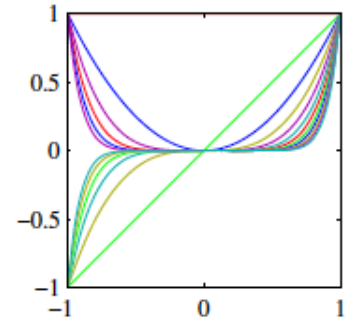
$$y = f(\mathbf{x}; \mathbf{w}) = w_0 + \sum_{j=1}^{J-1} w_j \phi_j(\mathbf{x})$$

Σημειώστε: Το μοντέλο αποτελεί γραμμική συνάρτηση των παραμέτρων \mathbf{w} αλλά **όχι** των μεταβλητών εισόδου, εφόσον οι συναρτήσεις βάσης είναι μη-γραμμικές.

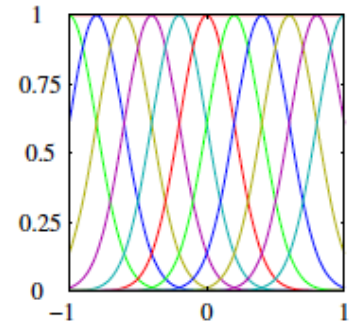
Γραμμικά μοντέλα συναρτήσεων βάσης

— Παραδείγματα συναρτήσεων βάσης:

- Πολυωνυμικές συναρτήσεις βάσης: $\phi_j(x) = x^j$



- Gaussian συναρτήσεις βάσης: $\phi_j(x) = \exp\left(-\frac{(x - \mu_j)^2}{2s^2}\right)$



— Οι συναρτήσεις βάσης $\phi(\cdot)$ αποτελούν κάποιας μορφής σταθερής προεπεξεργασίας ή εξαγωγής χαρακτηριστικών από τα αρχικά δεδομένα εισόδου.

Γραμμικά μοντέλα συναρτήσεων βάσης

—Σύνολο εκπαίδευσης: N ζεύγη διανύσματος-βαθμωτού $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \sim p$

Δεδομένα εισόδου

$$\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$$

Μεταβλητές στόχου

$$\{y_1, y_2, \dots, y_N\}, y_i \in \mathbb{R}$$

$$y = f(\mathbf{x}; \mathbf{w}) = w_0 + \sum_{j=1}^{J-1} w_j \phi_j(\mathbf{x})$$

—Ορίζουμε:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N], \in \mathbb{R}^{D \times N}$$

$$\mathbf{y} = [y_1, y_2, \dots, y_N]^T \in \mathbb{R}^N$$

$$\phi_i \doteq \phi(\mathbf{x}_i) = [\phi_0(\mathbf{x}_i), \phi_1(\mathbf{x}_i), \dots, \phi_{J-1}(\mathbf{x}_i)] \in \mathbb{R}^{1 \times J}$$

$$\Phi \doteq \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix} \in \mathbb{R}^{N \times J}$$

Γραμμικά μοντέλα συναρτήσεων βάσης

— Το γραμμικό μοντέλο συναρτήσεων βάσης σε μορφή πινάκων:

$$\mathbf{y} = \mathbf{\Phi} \mathbf{w}$$

— Εκπαίδευση: $\hat{\mathbf{w}} = \min_{\mathbf{w}} l(\mathbf{w}) = \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{\Phi} \mathbf{w}\|_2^2$

Normal equation: $\hat{\mathbf{w}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{y} = \mathbf{\Phi}^\dagger \mathbf{y}$

Σημειώστε: Η λύση ενδέχεται να μην είναι μοναδική

— Πρόβλεψη: $\hat{y} = \phi(\mathbf{x}) \hat{\mathbf{w}}$

Αξιολόγηση μοντέλων παλινδρόμησης

Αξιολόγηση μοντέλων παλινδρόμησης

Η αξιολόγηση των μοντέλων παλινδρόμησης εστιάζει κυρίως σε δύο πτυχές:

- 1) Πόσο καλά το μοντέλο που έχει εκτιμηθεί μπορεί να εξηγήσει τη διακύμανση της εξαρτημένης μεταβλητής (Y) στο σύνολο δεδομένων εκπαίδευσης. Με άλλα λόγια, πόσο καλά περιγράφει το μοντέλο τα δεδομένα.
- 2) Πόσο κοντά στην πραγματική τιμή βρίσκεται η τιμή που προβλέπει το μοντέλο παλινδρόμησης.

Συντελεστής προσδιορισμού R^2

— Η αξιολόγηση του πόσο καλά ένα μοντέλο γραμμικής παλινδρόμησης εξηγεί τη διακύμανση της τιμής της μεταβλητής στόχου γίνεται με υπολογισμό του **συντελεστή προσδιορισμού** (coefficient of determination)

R^2 (R-squared):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2 \qquad TSS = \sum_{i=1}^m (y_i - \bar{y})^2$$

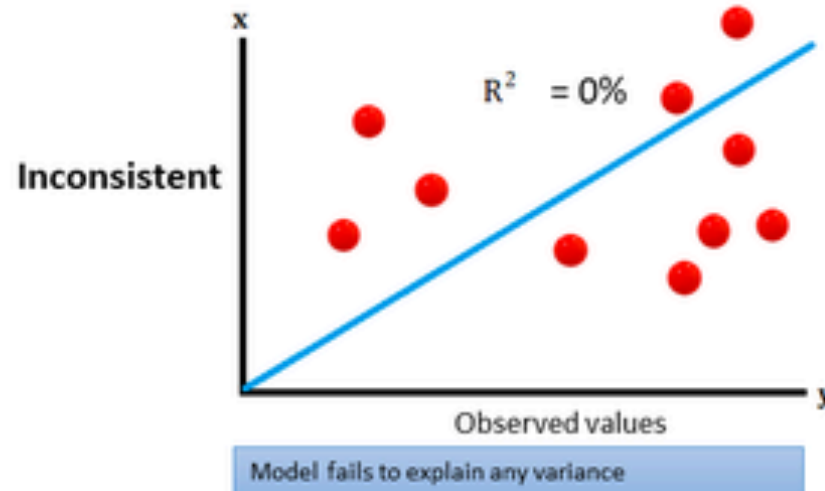
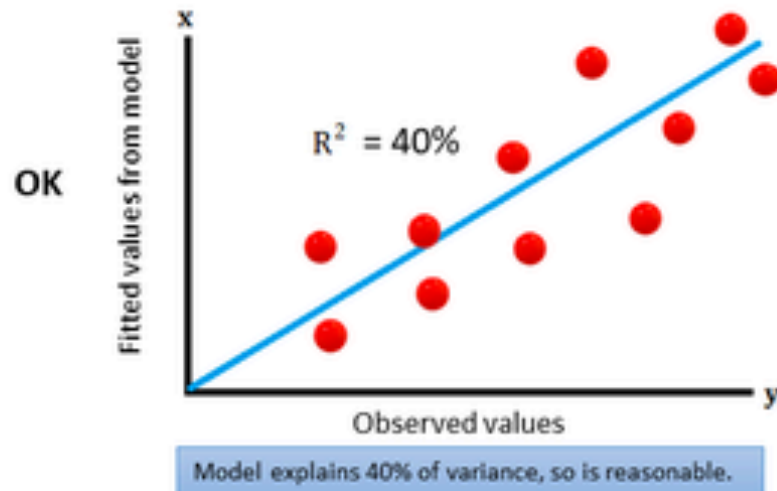
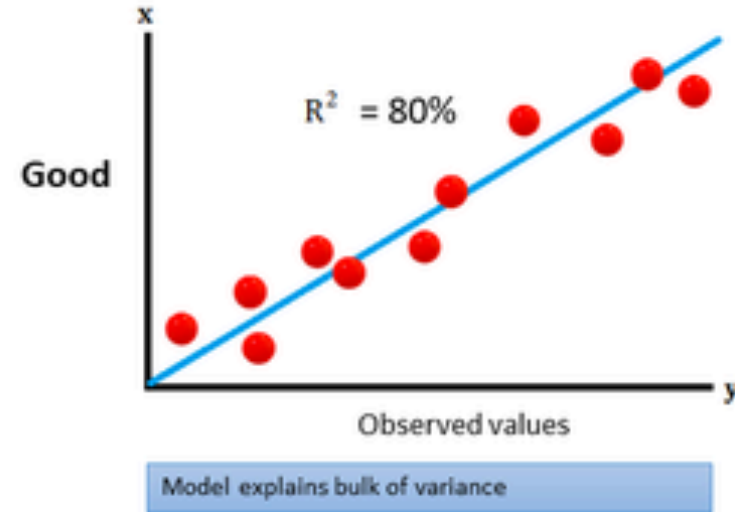
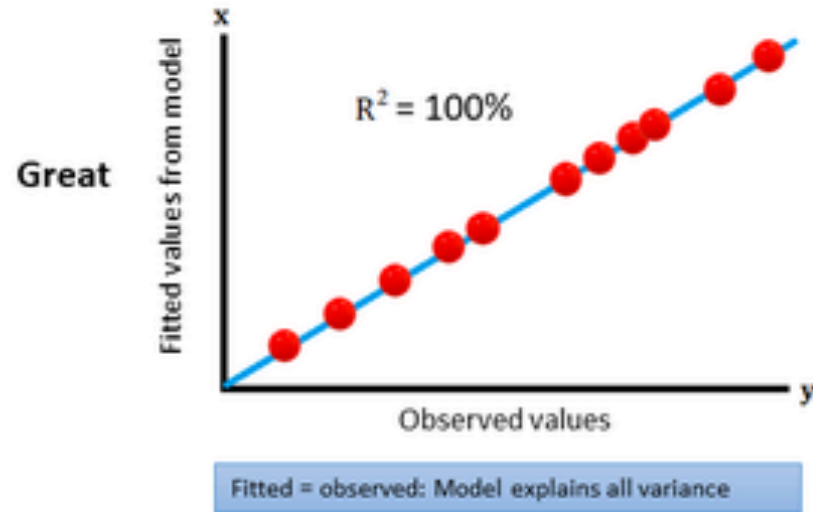
— Λαμβάνει τιμές από το 0 έως και το 1 και εκφράζει το ποσοστό της διακύμανσης που κατορθώνει να εξηγήσει το μοντέλο παλινδρόμησης.

Συντελεστής προσδιορισμού R^2

- Ο συντελεστής R squared δεν έχει μονάδες μέτρησης καθότι εκφράζει ποσοστό.
- Γενικά, όσο πιο υψηλή η τιμή του συντελεστή R^2 τόσο καλύτερα επιτυγχάνει το μοντέλο να εξηγήσει τη διακύμανση της ανεξάρτητης μεταβλητής και επομένως να εξηγήσει τα δεδομένα.
- Μπορεί να χρησιμοποιηθεί για έλεγχο υποθέσεων.

Συντελεστής προσδιορισμού R^2

Comparison of R-Squared for Different Linear Models (Same Data Set)



Αξιολόγηση μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη

— Η αξιολόγηση ενός μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη γίνεται με τη χρήση ορισμένων μετρικών οι οποίες εκτιμούν πόσο κοντά στην πραγματική τιμή βρίσκεται η τιμή που προβλέπει το μοντέλο παλινδρόμησης.

— Οι μετρικές αυτές, μετρούν το σφάλμα (που προσδιορίζεται από τη διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής της μεταβλητής Y) που κάνει το μοντέλο παλινδρόμησης στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής και συνεπώς οι μετρικές αυτές καλούνται και **μετρικές σφάλματος**.

Μετρικές εκτίμησης σφάλματος πρόβλεψης

— Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE):

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

— Μέσο Τετραγωνισμένο Σφάλμα (Mean Squared Error - MSE):

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Μετρικές εκτίμησης σφάλματος πρόβλεψης

—Μέσο Τετραγωνικό Σφάλμα (Root Mean Squared Error – RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

—Μέσο Απόλυτο Εκατοστιαίο Σφάλμα (Mean Absolute Percentage Error – MAPE):

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Μετρικές εκτίμησης σφάλματος πρόβλεψης

- Όλες οι παραπάνω μετρικές δίνουν τιμές σφάλματος μεταξύ 0 και άπειρο με εξαίρεση τη μετρική του Μέσου Απόλυτου Εκατοστιαίου Σφάλματος που κυμαίνεται μεταξύ 0 και 1.
- Προφανώς, όσο πιο μικρή η τιμή του σφάλματος τόσο πιο καλή η πρόβλεψη που κάνει το μοντέλο.
- Η επιλογή της κατάλληλης μετρικής εξαρτάται από το πρόβλημα που μελετάται καθώς οι μετρικές αυτές χειρίζονται τα σφάλματα πρόβλεψης με διαφορετικό τρόπο.

Μετρικές εκτίμησης σφάλματος πρόβλεψης

- Το Μέσο Απόλυτο Σφάλμα δίνει την ίδια βαρύτητα σε όλα τα υπόλοιπα (residuals) που προκύπτουν και δεν διακρίνει εάν η προβλεπόμενη τιμή υπερτιμά ή όχι την πραγματική τιμή.
- Το Μέσο Τετραγωνισμένο Σφάλμα δίνει διαφορετική βαρύτητα στο μέγεθος των υπολοίπων (residual) με τα μεγαλύτερα υπόλοιπα να συνεισφέρουν στο συνολικό σφάλμα παραπάνω εξαιτίας της ύψωσης στο τετράγωνο και προτιμάται εάν είναι επιθυμητό να “τιμωρηθούν” μεγάλες τιμές υπόλοιπου και ακραίες προβλεπόμενες τιμές.
- Το Μέσο Τετραγωνικό Σφάλμα από την άλλη έχει καλύτερη ερμηνευτική δύναμη καθώς το σφάλμα εκφράζεται σε μονάδες της ανεξάρτητης μεταβλητής – σε αντίθεση με το Μέσο Τετραγωνισμένο Σφάλμα.

Μετρικές εκτίμησης σφάλματος πρόβλεψης

- Το Μέσο Απόλυτο Εκατοστιαίο Σφάλμα έχει εύκολη ερμηνεία. Όμως δεν μπορεί να χρησιμοποιηθεί εάν η εξαρτημένη μεταβλητή μπορεί να λάβει την τιμή μηδέν (0) εξαιτίας της πράξης της διαίρεσης που εμφανίζεται.
- Επιπλέον, η μετρική αυτή μεροληπτεί υπέρ προβλεπόμενων τιμών που είναι συστηματικά μικρότερες από την πραγματική τιμή. Ειδικότερα, η μετρική αυτή θα είναι μικρότερη όπου οι προβλεπόμενες τιμές είναι μικρότερες από την πραγματική εάν συγκριθεί με άλλο μοντέλο το οποίο παρουσιάζει τιμές που προβλέπει τιμές μεγαλύτερες από τις πραγματικές κατά το ίδιο μέγεθος.
