# Μηχανική Μαθηση

## Εισαγωγή

Α. Ν. Γιαννακόπουλος

Ο.Π.Α

Χειμερινό Εξάμηνο 2025-2026

# Εισαγωγή

- Το μάθημα αυτό επικεντρώνεται σε αναλυτικές και αριθμητικές μεθόδους οι οποίες είναι απαραίτητες στην μηχανικής μάθηση
- Τα περισσότερα μοντέλα που χρησιμοποιούνται καθημερινά και που χρησιμοποιείτε και εσείς, όπως π.χ.
  - ο αλγόριθμός του Netflix ή
  - οι προτάσεις για videos στο Youtube, ή
  - ο τρόπος που ψάχνετε στο Google
  - η ιατρική απεικόνιση
  - η συμπίεση δεδομένων π.χ. videos ή ταινιών στις διάφορες ψηφιακές πλατφόρμες,
  - η ιατρική διάγνωση,
  - οι αλγόριθμοι χορήγησης δανείων απο τις τράπεζες, αλλά και
  - ένα σωρό άλλες εφαρμογές

  έχουν στο υπόβαθρο τους καποια μεθοδολογία η οποία καταλήγει σε ένα πρόβλημα μηχανικής μαθησης αριθμητικής ανάλυσης και ένα αλγοριθμικό πλαίσιο η ορθή υλοποίηση του οποίου θα μας οδηγήσει στο σωστό αποτέλεσμα

- Μαθημα επιλογης, 4 ωρες την εβδομαδα
  1. 2 ωρες θεωρια (Τροίας 102, Τεταρτη 11.00-13.00)
  2. 2 ωρες εργαστήριο (Αντωνιαδου, Πεμπτη 11.00-13.00)
- Α. Γιαννακοπουλος, Σ. Βακερούδης
- Εξέταση:
  - Απαλλακτική εργασία

# Τι είναι η μηχανική μαθηση;

Τα ετεροκλήτα παραδείγματα που παρουσιάσαμε εχουν ολα ενα κοινό στοιχείο, το οποίο και θα χρησιμοποιήσουμε ως το βασικό μεθοδολογικό πλαίσιο της μηχανικής μαθησης.

Το βασικο πρόβλημα της μηχανικής μαθησης είναι να συνδέσουμε χαρακτηριστικά $x \in X$ με αποκρίσεις $y \in Y$.

Αυτό που ζητάμε είναι να βρούμε μια συνάρτηση $f : X \to Y$ τετοια ωστε

$$y \simeq f(x), \ \ \forall \ x \in X.$$

Η συναρτηση αυτή πρεπει να κατασκευαστεί απο διαθεσιμα δεδομενα / παρατηρήσεις σχετικά με τα $x$ και $y$,

$$\{(x_1, y_1), \cdots, (x_N, y_N)\} \subset X \times Y.$$

- Το συνολο $X$ ονομαζεται ο χωρος των χαρακτηριστικών (feature space)
- Το σύνολο $Y$ ονομάζεται ο χώρος των αποκρίσεων (response space)

Τα συνολα $X$, $Y$ μπορεί να είναι υποσύνολα καποιου απλού διανυσματικού χώρου π.χ. $X \subset \mathbb{R}^m$, $Y \subset \mathbb{R}^n$ αλλά μπορεί να είναι και υποσύνολα καποιου πιο περίπλοκου συνολου π.χ. ενα σύνολο με μη γραμμική δομή, μέτρα πιθανότητας κλπ.

Ο τρόπος κατασκευής της συνάρτησης $f$ εξαρτάται σε σημαντικό βαθμό απο την φύση των συνόλων $X$ και $Y$.

Προφανώς είναι αδύνατον να περιμένουμε να βρούμε μια συνάρτηση $f : X \to Y$ τέτοια ώστε

$$y_i = f(x_i), \ \ \forall \ \ i = 1, \cdots, N.$$

Το καλύτερο που μπορούμε να κάνουμε είναι να βρούμε μια συνάρτηση $f$ η οποία ικανοποίει την παραπάνω σχέση προσεγγιστικά δηλ.

$$y_i \simeq f(x_i), \ \ \forall \ \ i = 1, \cdots, N,$$

δηλ. ικανοποιεί την σχέση αυτή με καποιο σφάλμα.

Η προσεγγιση αυτή εκφράζεται με βάση την ελαχιστοποιήση καποιας συνάρτησης σφάλματος

$$L(f) := \frac{1}{N} \sum_{i=1}^{N} L_i(y_i, f(x_i)),$$

όπου $L_i$ είναι μια κατάλληλη συνάρτηση η οποία αξιολογεί την απόκλιση της τιμής $f(x_i)$ απο την παρατηρούμενη απόκριση $y_i$.

Η επιλογή της συνάρτησης $L_i$, είναι στην διακριτική ευχέρεια του χρήστη και εξαρτάται επίσης απο το πρόβλημα που αντιμετωπίζουμε.

Πιθανές επιλογές είναι οι

- $L_i(y_i, f(x_i)) = \varphi(y_i, f(y_i))$ όπου $\varphi : Y \times Y \to \mathbb{R}$, είναι μια συνάρτηση συνάφειας μεταξύ δύο οποιονδήποτε στοιχείων του $Y$ – εδω $y_i$ είναι η παρατήρηση της απόκρισης και $f(x_i)$ η πρόβλεψη της βάσει του μοντέλου $f$.
- Αν το $Y$ είναι μετρικός χώρος θα μπορούσε $L_i(y_i, f(x_i)) = d_Y(y_i, f(x_i))$, όπου $d_Y$ είναι η μετρική του χώρου $Y$ (δεν υπάρχει απαραίτητα μια μοναδική επιλογή για την $d_Y$).
- Αν το $Y$ είναι διανυσματικός χώρος με νόρμα θα μπορούσε $L_i(y_i, f(x_i)) = \|y_i - f(x_i)\|_Y$, όπου $\|\cdot\|_Y$ μια νορμα για τον χώρο $Y$.

Οποιοδήποτε προβλημα μηχανικής μάθησης καταλήγει τελικά σε ενα κατάλληλο πρόβλημα βελτιστοποίησης της μορφής

$$\min_{f \in \mathbb{F}} L(f) = \min_{f \in \mathbb{F}} \frac{1}{N} \sum_{i=1}^{N} L_i(y_i, f(x_i)),$$

όπου $\mathbb{F}$ είναι ένα κατάλληλο σύνολο συναρτήσεων.

Το σύνολο συναρτήσεων $\mathbb{F}$ μας καθορίζει την κατηγορία μοντέλου που επιλέγουμε για την μελέτη των δεδομένων μας.

Η συνάρτηση

$$f^* = \arg\min_{f \in \mathbb{F}} L(f),$$

αντιστοιχεί στο καλύτερο μοντέλο για τα δεδομένα μας στην συγκεκριμένη κατηγορία μοντέλων που αντιστοιχεί στο σύνολο $\mathbb{F}$.

# Πιθανές επιλογές για το σύνολο $\mathbb{F}$

- $X = \mathbb{R}^m$, $Y = \mathbb{R}^n$: $\mathbb{F} = \mathcal{L}(X, Y)$ το σύνολο των γραμμικών συναρτήσεων $f : \mathbb{R}^m \to \mathbb{R}^n$ οι οποίες παραμετροποιούνται ως $y = Ax + b$, $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ – Γραμμικά μοντέλα
- $X = \mathbb{R}^m$, $Y = \mathbb{R}^n$: $\mathbb{F} = \mathcal{P}_k(X, Y)$ το σύνολο των πολυωνύμων βαθμού μέχρι $k$, $P_k : \mathbb{R}^m \to \mathbb{R}^n$ οι οποίες παραμετροποιούνται με βάση τους συντελεστές που καθορίζουν το πολυώνυμο – Πολυωνυμικά (μη γραμμικά) μοντέλα
- $X = \mathbb{R}^m$, $Y = \mathbb{R}^n$: $\mathbb{F} = \{f = \sum_{i=1}^{k} a_i \phi_i, \;\; a_i \in \mathbb{R}\}$, όπου $\phi_i : X \to Y$ δεδομένες (προκαθορισμένες) συναρτήσεις βάσης.

Στις παραπάνω περιπτώσεις το προβλημα ελαχιστοποίησης στο οποίο ανάγεται το πρόβλημα μηχανικής μάθησης καταλήγει σε ένα πρόβλημα ελαχιστοποίησης επάνω στο σύνολο των παραμέτρων που καθορίζουν το μοντέλο μας.

# Γραμμική παλινδρόμηση και ομαλοποίηση

Η γραμμική παλινδρόμηση προσπαθεί να συνδέσει χαρακτηριστικά $x = (x_1, \cdots, x_m) \in \mathbb{R}^m$ με μια απόκριση $y \in \mathbb{R}$ μέσω ενός τύπου

$$y \simeq \beta_0 + \sum_{i=1}^{m} \beta_i x_i.$$

Για μία συλλογή δεδομένων $(x_j, y_j) = (x_{1j}, \cdots, x_{mj}, y_j)$, $j = 1, \cdots, N$ θέλουμε να ισχύει

$$y_j \simeq \beta_0 + \sum_{i=1}^{m} \beta_i x_{ij}, \;\; j = 1, \cdots, N$$

Ενισχύοντας τα χαρακτηριστικά με το $x_0 = 1$ μπορούμε να γραψουμε την παραπάνω σχέση με την συμπαγή μορφή

$$y \simeq Xw$$

Το καλύτερο μοντέλο επιλέγεται ως λύση του προβλήματος βελτιστοποίησης

$$L(w) = \frac{1}{N}\|y - Xw\|^2$$

το οποίο μπορεί να λυθεί με διάφορες μεθόδους π.χ. gradient descent , επίλυση γραμμικών συστημάτων (SVD, QR decomposition) κα.

Πολλές φορές μπορούμε να πάρουμε καλύτερα υποδείγματα (π.χ. με μικρότερη διασπορά ή με μικρότερο αριθμός συντελεστών) χρησιμοποιώντας την τεχνική της ομαλοποίησης

$$L(w) = \frac{1}{N}\|y - Xw\|^2 + \lambda\|w\|_2^2, \text{ ridge regression}$$

$$L(w) = \frac{1}{N}\|y - Xw\|^2 + \lambda\|w\|_1, \text{ Lasso regression}$$

$$L(w) = \frac{1}{N}\|y - Xw\|^2 + \lambda_1\|w\|_2^2 + \lambda_2\|w\|_1, \text{ Elastic net}$$

# Πιστωτικός κίνδυνος

Μια τραπεζα η οποία χορηγεί δάνεια επιθυμεί να γνωρίζει την πιθανότητα καποιος πελάτης της να μην αποπληρώσει το δανειο του σαν συναρτηση ορισμενων χαρακτηριστικών του πελάτη

- $x = (x_1, x_2, \cdots, x_m) \in \mathbb{R}^m$ τα χαρακτηριστικά του πελάτη π.χ. εισόδημα, ύψος δανείου κλπ.
- $y \in (0, 1) \subset \mathbb{R}$ η πιθανότητα αποπληρωμής του δανείου.

Θα θελαμε να έχουμε μια συνάρτηση $f : \mathbb{R}^m \to (0, 1)$ τέτοια ωστε

$$f(x_1, \cdots, x_m) = y$$

για καποιον πελάτη με χαρακτηριστικά $x = (x_1, x_2, \cdots, x_m)$.

Η γραμμική συνάρτηση δεν είναι καλό μοντέλο για το πρόβλημα αυτό γιατι μπορεί να δώσει πιθανότητες $y$ οι οποίες είναι αρνητικές!

Λογιστική παλινδρόμηση

$$y = f(x) = \frac{\exp(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)},$$

όπου $\beta_0, \beta_1, \cdots, \beta_m$ παράμετροι του μοντέλου.

Ένας σύνηθισμενος τρόπος ερμηνείας αυτού του μοντέλου είναι να θεωρήσουμε οτι τα δεδομενα $y$ ανήκουν σε 2 κλάσεις την κλαση 1 (επιτυχής αποπληρωμή) και την κλάση 0 (αποτυχία αποπληρωμής) και

$$P(y = 1 \mid x) = f(x), \;\; P(y = 0 \mid x) = 1 - f(x).$$

Το μοντέλο αυτό λοιπόν μπορεί και να χρησιμοποιηθεί και για προβλήματα ταξινόμησης. όπου $w = (w_1, \cdots, w_m) \in \mathbb{R}^m$ παράμετροι οι οποίοι θα πρέπει να καθοριστούν απο τα δεδομένα.

Ας συμβολίσουμε με $\theta = (\beta_0, \beta)$ τις παραμέτρους του μοντέλου και με $h_i(\theta; \mathbf{X}_i)$ όπως παραπάνω την πρόβλεψη του υποδείγματος για την παρατήρηση $\mathbf{X}_i = (x_{1i}, \cdots, x_{mi})$ να ανήκει στην κλάση $Y_i = 1$.

Ορίζουμε την συνάρτηση κόστους

$$\bar{L}_i(\theta; \mathbf{X}_i, Y_i) = \begin{cases} -Y_i \ln(h_i(\theta; \mathbf{X}_i)) & \text{αν} \quad Y_i = 1 \\ -(1 - Y_i) \ln(1 - h_i(\theta; \mathbf{X}_i)) & \text{αν} \quad Y_i = 0 \end{cases} \tag{1}$$

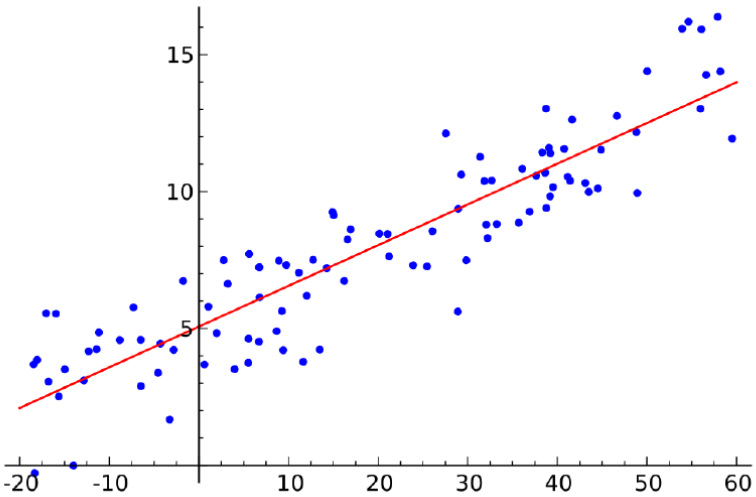
Η συνάρτηση σφάλματος σε όλο το σύνολο των δεδομένων είναι η

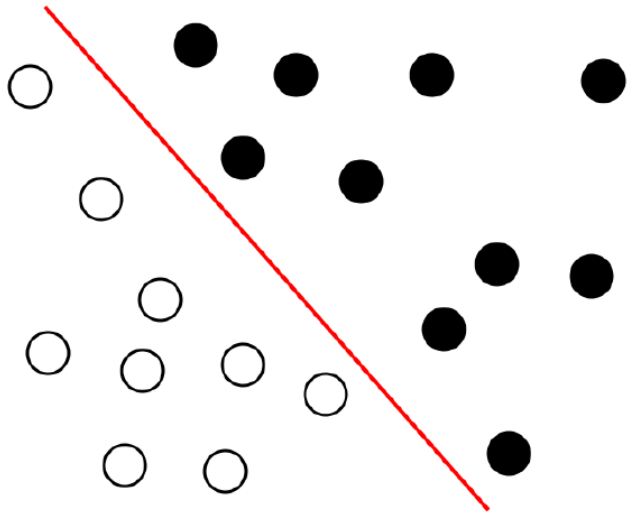
$$L(\theta; \mathbf{X}, Y) = \frac{1}{N} \sum_{i=1}^{N} L_i(\theta; \mathbf{X}_i, Y_i) \tag{2}$$

Το καλύτερο μοντέλο είναι η συνάρτηση $y = f_{\theta^*}(x)$ όπου $\theta^*$ η λύση του προβλήματος βελτιστοποίησης

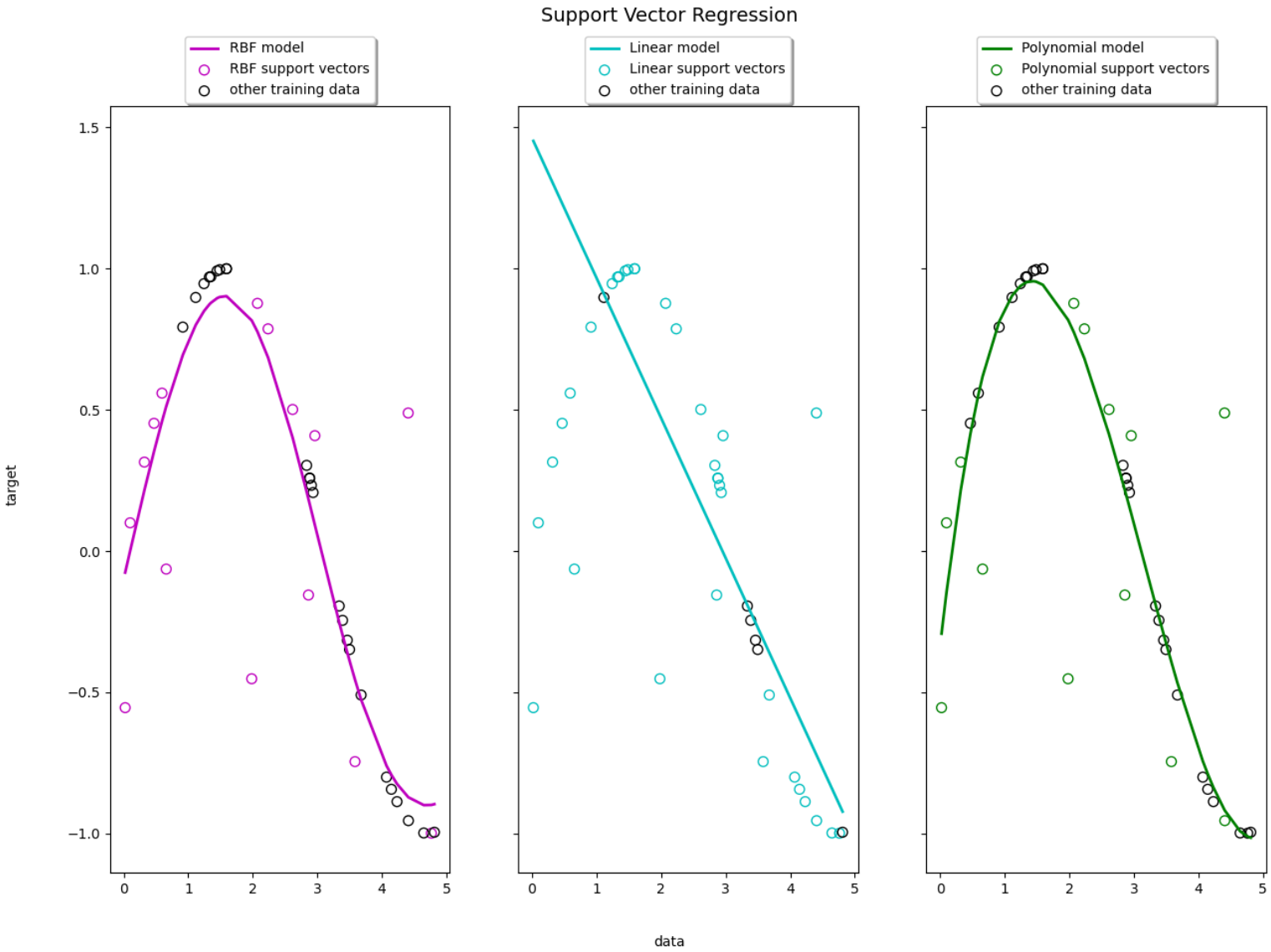
$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^m} L(\theta; \mathbf{X}, Y)$$

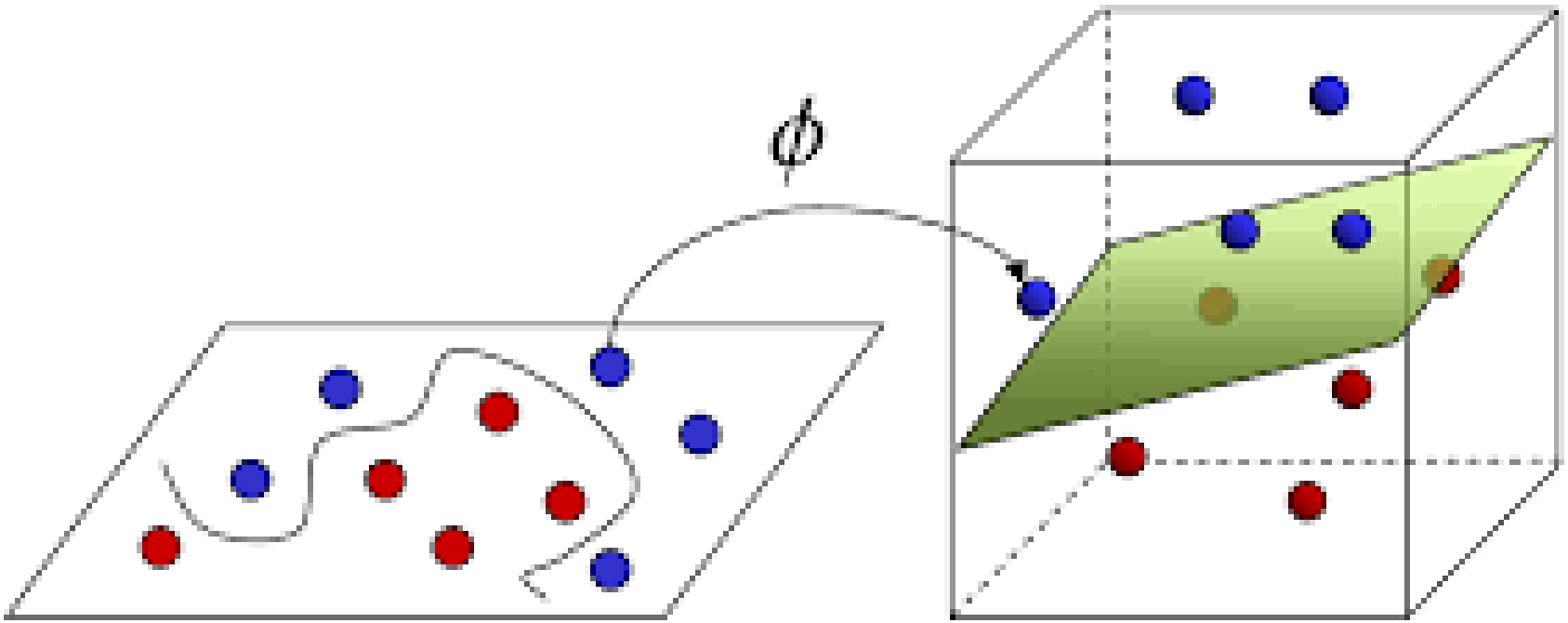
# Γραμμικά μοντέλα – Γραμμικός Διαχωρισμός



# Μη Γραμμικοτητα στα δεδομενα




$\phi$
Input Space
Feature Space

# Ανάγκη αναπτυξης πιο περίπλοκων (μη γραμμικών) μοντέλων

Οι απλές επιλογές για την συνάρτηση $f$ που καθορίζει το μοντέλο δεν είναι πολλές φορές επαρκείς.

Για πιο περίπλοκα δεδομένα χρειάζεται να συμπληρώσουμε το λεξικό των μοντέλων μας με πιο συνθετα μοντέλα που η βαση τους βρίσκεται στην μη γραμμικότητα και την γεωμετρία.

- Μέθοδοι με πυρήνες (Kernel methods)
- Νευρωνικά δικτύα και βαθεία μάθηση
- Εκμάθηση σε πολλαπλότητες (Manifold Learning)

Επίσης λόγω της διαστατικότητας των δεδομένων χρειάζονται τεχνικές απο την

- Πιθανότητες σε υψηλές διαστάσεις
- Στοχαστικές διαδικασίες Gauss

- Μέθοδοι με πυρήνες: Μοντέλα της μορφής

$$f(x) = \sum_{i=1}^{N} a_i \underbrace{K(x_i, x)}_{=\phi_i(x)},$$

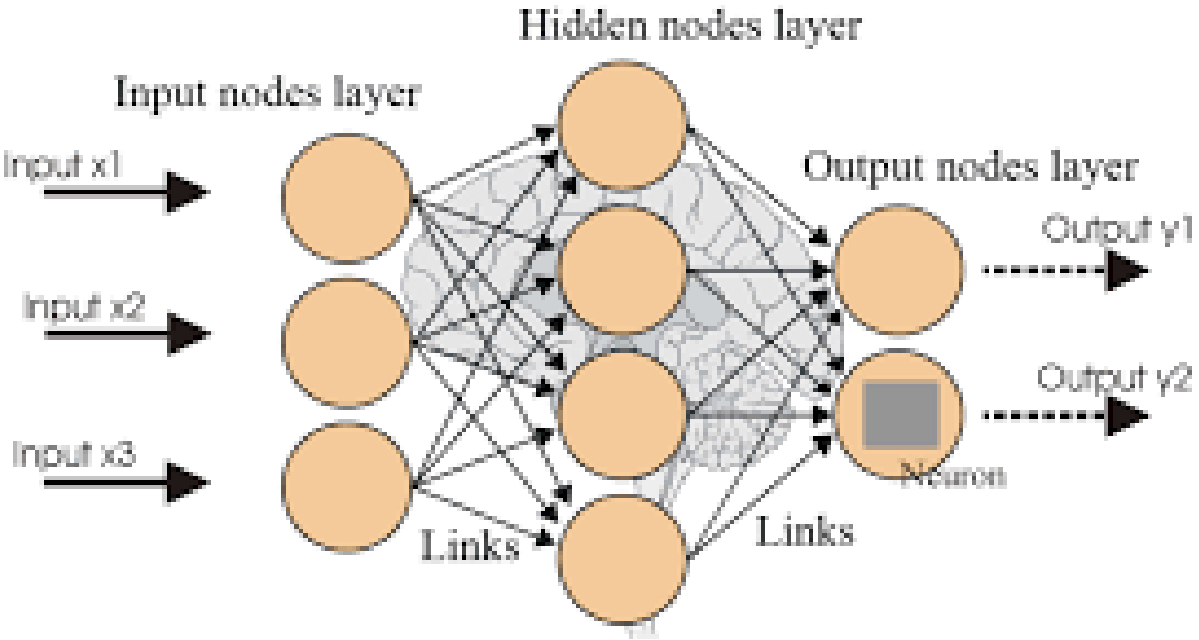
που εκπαιδεύονται σε ενα συγκεκριμένο σύνολο δεδομένων $\{(x_i, y_i),\ i = 1, \cdots, N\}$.

- $K : X \times X \to \mathbb{R}$ προκαθορισμένη συνάρτηση πυρήνα (Kernel function)
- Οι συντελεστές του μοντέλου μπορεί να υπολογιστούν πολύ απλά χρησιμοποιώντας ενα πολύ ισχυρό θεωρητικό πλαίσιο (representer theorem)
- Οι πολύ ισχυρές αυτές μέθοδοι βασίζονται στην ιδέα του μετασχηματισμού του χώρου των χαρακτηριστικών (feature space) $X$ σε ένα νέο χώρο $H$ (υψηλότερης διάστασης) στον οποίο τα δεδομένα είναι πιο εύκολα ή ακόμη και γραμμικά περιγράψιμα.

Η συνέχεια επι της οθόνης ...

# Νευρωνικά δικτυα και βαθιά μάθηση

Προσπαθούμε να αναπαραστήσουμε τα δεδομένα μας χρησιμοποιώντας συναρτήσεις που παράγονται απο σύνθεση μη γραμμικών συναρτήσεων χαμηλής διαστασης (συνήθως 1) και γραμμικών συναρτήσεων $L_k : \mathbb{R}^d \to \mathbb{R}$



Γιατί να δουλεύει μια τέτοια αναπαράσταση και τελικά τι τύπου δεδομένα μπορεί να αναπαραστήσει;

Universal approximation theorems

Πως μπορούμε να το εκπαιδεύσουμε δηλ. να μάθουμε τα βάρη που χρησιμοποιούνται στην αναπαράσταση;

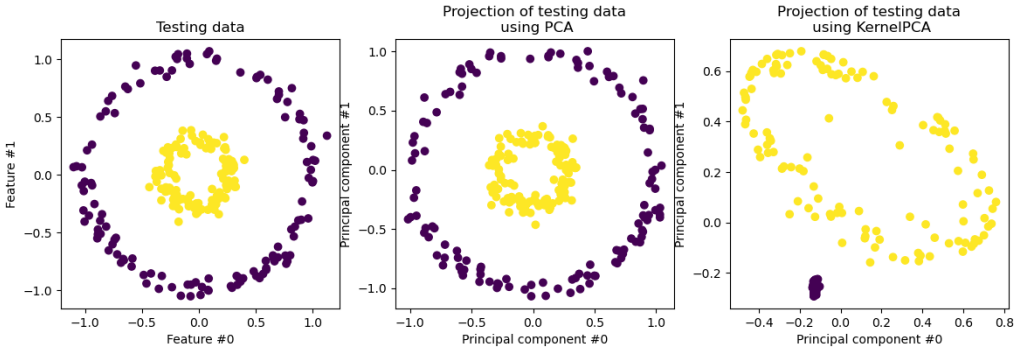
Backpropagation algorithm

Πως μπορούμε να προκαθορίσουμε την δομή του δικτύου ή το πλήθος των στρωμάτων (βάθος) και να μην πηγαίνουμε στα τυφλά;

Η συνέχεια επι της οθόνης ...

# Μείωση διαστατικότητας – Manifold learning

Πολλές φορές τα δεδομένα δίνουν την λανθασμένη εντύπωση υψηλής διάστασης ενώ στην πραγματικότητα η διάσταση τους είναι χαμηλότερη.



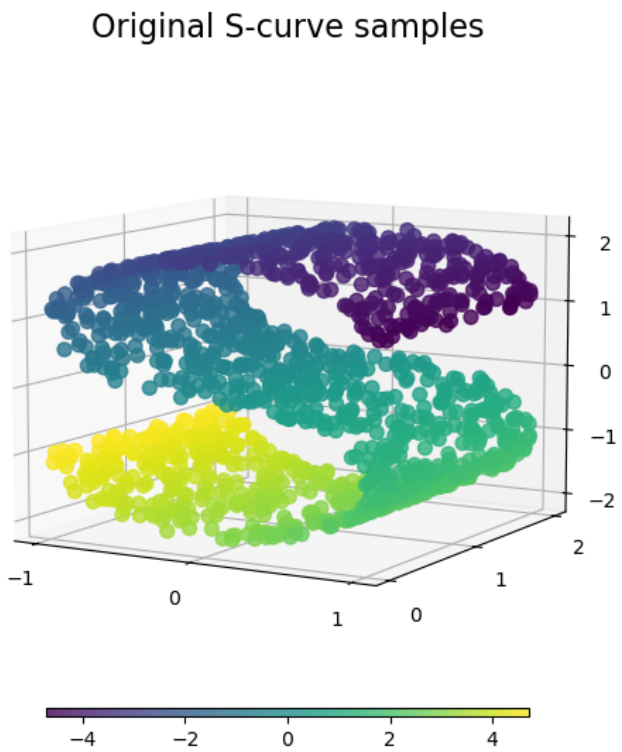

Ο μετασχηματισμός σε πολικές συντεταγμένες μπορεί να μας περιγράψει τα παραπάνω δεδομένα σε 1 διασταση!

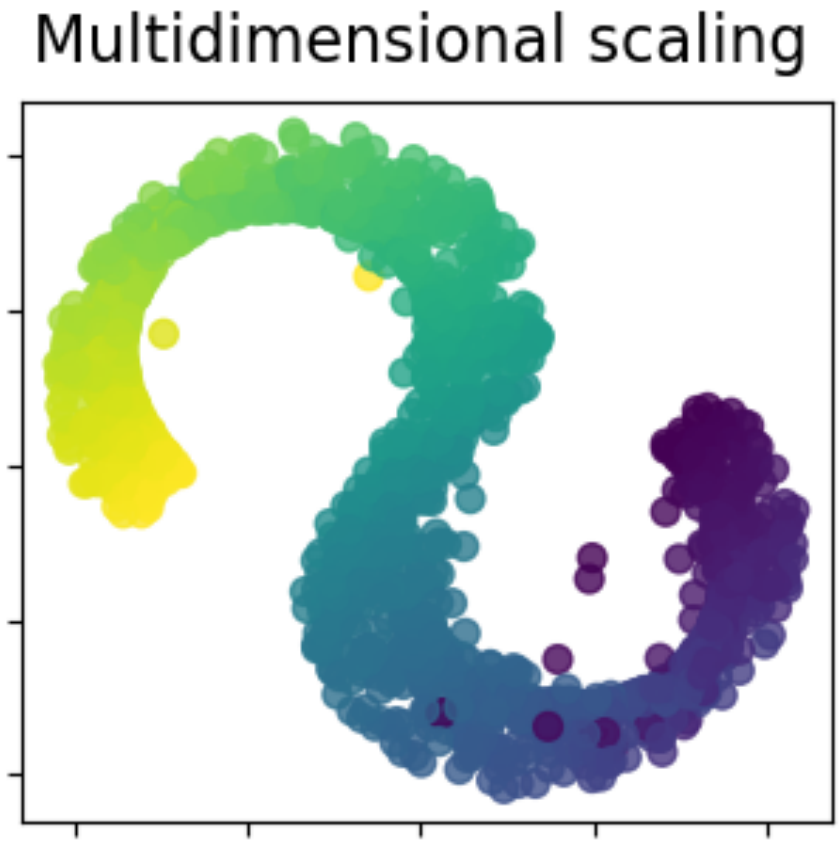
Η μείωση διαστατικότητας βασίζεται σε τεχνικές μετασχηματισμού των δεδομένων $X \in \mathbb{R}^n$ με $n$ μεγάλο σε καινούργιες συντεταγμένες $Y = \Phi(X) \in R^p$, με $p < n$.

Η μελέτη των δεδομένων στην νέα τους μορφή $Y$ που είναι πιο χαμηλής διάστασης μπορεί να είναι πιο απλή και αποτελεσματική, και μάλιστα να μας βοηθήσει στην οπτικοποίηση.

Η γεωμετρία των δεδομένων παίζει πολύ σημαντικό ρόλο στην εύρεση του κατάλληλου μετασχηματισμού $\Phi : \mathbb{R}^n \to \mathbb{R}^p$, $p < n$ που θα μας δώσει την κατάλληλη μείωση διαστατικότητας.

Οι γραμμικές τεχνικές (PCA ) δεν είναι πάντοτε επαρκείς και χρειαζόμαστε νεες μη γραμμικές τεχνικές βασισμένες στην γεωμετρία των δεδομένων.


Original S-curve samples
2
1
0
−1
−2
−1
0
1
0
1
2
−4
−2
0
2
4


Multidimensional scaling

Spectral clustering

Isomap

Laplacian eigenmaps

Stochastic neighbour embedding (SNE)

κ.α. ...