



HMMY · ΕΛΜΕΠΑ · 9ο εξάμηνο

Μηχανική Μάθηση και Εξόρυξη Γνώσης

Διάλεξη 1: Εισαγωγή και πλάνο του μαθήματος

Λευτέρης Κουμάκης

Η σημερινή διάλεξη



Μέρος Α: Το μάθημα

Στοιχεία και στόχοι του μαθήματος
Πλάνο 13 διαλέξεων
Συγγράμματα και εργαλεία
Αξιολόγηση και τρόπος δουλειάς



Μέρος Β: Εισαγωγή

Τι είναι MM και Εξόρυξη Γνώσης
KDD και CRISP-DM
Είδη μάθησης
Πώς «μαθαίνει» ένα μοντέλο
Υπερπροσαρμογή, bias–variance
Αξιολόγηση μοντέλων

ΜΕΡΟΣ Α

Το μάθημα

Οργάνωση, πλάνο διαλέξεων, συγγράμματα, αξιολόγηση

Στοιχεία του μαθήματος



Διδάσκων

Λευτέρης Κουμάκης



Email

koumakis@hmu.gr



Ώρες γραφείου

Δευτέρα 11:00–13:00



Τμήμα / εξάμηνο

HMMY, ΕΛΜΕΠΑ · 9ο εξάμηνο



Διαλέξεις

13 διαλέξεις θεωρίας με παραδείγματα σε Python

Τι θα μπορείτε να κάνετε στο τέλος



Να κατανοείτε

Πώς λειτουργούν οι βασικοί αλγόριθμοι: γραμμικά μοντέλα, δένδρα, SVM, ensembles, νευρωνικά δίκτυα.



Να υλοποιείτε

Πλήρη ροή σε Python: από τα ακατέργαστα δεδομένα έως το εκπαιδευμένο μοντέλο.



Να αξιολογείτε

Να μετράτε σωστά την απόδοση και να αποφεύγετε παγίδες όπως η διαρροή δεδομένων.



Να παρακολουθείτε

Τη σύγχρονη βαθιά μάθηση: CNN, LSTM, Transformers και μεγάλα γλωσσικά μοντέλα.

Τι χρειάζεται να θυμάστε



Πιθανότητες & Στατιστική

Κατανομές, μέση τιμή, διασπορά, δεσμευμένη πιθανότητα, Bayes.



Γραμμική Άλγεβρα

Διανύσματα, πίνακες, γινόμενο πινάκων, ανάστροφος.



Λογισμός

Παράγωγοι, μερικές παράγωγοι, κλίση (gradient).



Python

Βασική σύνταξη, NumPy και pandas.

Το πλάνο με μια ματιά



● Θεμέλια

● Κλασική ΜΜ

● Βαθιά μάθηση

Μπλοκ 1: Θεμέλια (Διαλέξεις 1–3)

#	Διάλεξη	Βασικά θέματα	Géron (3η εκδ.)
1	Εισαγωγή	Είδη μάθησης, KDD / CRISP-DM, υπερπροσαρμογή, bias–variance, train/test	Κεφ. 1
2	Δεδομένα και προεπεξεργασία	Καθαρισμός, ελλειπίες τιμές, κλιμάκωση, κωδικοποίηση, pipelines, data leakage	Κεφ. 2
3	Γραμμικά μοντέλα	Γραμμική παλινδρόμηση, gradient descent, Ridge/Lasso, λογιστική παλινδρόμηση	Κεφ. 4

Μπλοκ 2: Κλασική ΜΜ (Διαλέξεις 4–8)

#	Διάλεξη	Βασικά θέματα	Géron (3η εκδ.)
4	Ταξινόμηση I: αξιολόγηση	kNN, Naive Bayes, confusion matrix, precision/recall, ROC–AUC, cross-validation	Κεφ. 3
5	Δένδρα Αποφάσεων	ID3/C4.5/CART, εντροπία, Gini, κλάδεμα, ερμηνευσιμότητα	Κεφ. 6
6	Support Vector Machines	Μέγιστο περιθώριο, soft margin, kernel trick, SVR	Κεφ. 5
7	Ensemble μέθοδοι	Bagging, Random Forests, Boosting (XGBoost), stacking	Κεφ. 7
8	Μη επιβλεπόμενη μάθηση	PCA, k-means, DBSCAN, ανίχνευση ανωμαλιών, κανόνες συσχέτισης	Κεφ. 8–9

Μπλοκ 3: Βαθιά μάθηση (Διαλέξεις 9–13)

#	Διάλεξη	Βασικά θέματα	Géron (3η εκδ.)
9	Νευρωνικά Δίκτυα	Perceptron, MLP, backpropagation, optimizers, dropout, batch norm	Κεφ. 10–11
10	CNN	Συνέλιξη, pooling, αρχιτεκτονικές, transfer learning	Κεφ. 14
11	RNN και LSTM	Ακολουθίες, BPTT, LSTM/GRU, χρονοσειρές, encoder–decoder	Κεφ. 15–16
12	Transformers και LLMs	Attention, BERT/GPT, pretraining, instruction tuning, RLHF, RAG	Κεφ. 16 + J&M
13	Γενετικά μοντέλα & Ενισχυτική	Autoencoders, VAE, GANs, Q-learning, policy gradients	Κεφ. 17–18

Συγγράμματα



Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 3η έκδ. (O'Reilly)

Βασικό σύγγραμμα · πρακτική με κώδικα



Κεφ. 4 «Μηχανική Μάθηση» (Κάλλιπος)

Δένδρα απόφασης, ΤΝΔ, γενετικοί αλγόριθμοι



Λυκοθανάσης & Κουτσομητρόπουλος, Υπολογιστική νοημοσύνη και βαθιά μάθηση (Κάλλιπος)

Νευρωνικά δίκτυα και βαθιά μάθηση στα ελληνικά



Jurafsky & Martin, Speech and Language Processing, 3rd ed. draft

Συμπληρωματικό · Transformers και LLMs

Εργαλεία που θα χρησιμοποιήσουμε



Python

Η κύρια γλώσσα του μαθήματος



Jupyter / Colab

Notebooks στον browser, με δωρεάν GPU



NumPy & pandas

Διαχείριση και ανάλυση δεδομένων



scikit-learn

Κλασική ΜΜ: Διαλέξεις 2–8



Keras / TensorFlow

Νευρωνικά δίκτυα: Διαλέξεις 9–13



Hugging Face

Έτοιμα Transformers και LLMs

Αξιολόγηση



Τελική εξέταση

50 %



Project

50 %



Προϋποθέσεις

>= 5 ελάχιστος βαθμός ανά σκέλος

Οι ερωτήσεις Slido δεν βαθμολογούνται: χρησιμεύουν για να ελέγχουμε την κατανόηση την ώρα του μαθήματος.

Πώς δουλεύουμε στην τάξη



Slido

Ερωτήσεις σε κάθε διάλεξη, απαντάτε από το κινητό.



Notebooks

Κάθε ενότητα συνοδεύεται από notebook για πειραματισμό.



Διάβασμα

Προτεινόμενα κεφάλαια πριν από κάθε διάλεξη.



Ερωτήσεις

Ανά πάσα στιγμή, στην τάξη ή στις ώρες γραφείου.

ΜΕΡΟΣ Β

Τι είναι η Μηχανική Μάθηση;

Ορισμοί, παραδείγματα και η σχέση της με την Εξόρυξη Γνώσης

Η Μηχανική Μάθηση γύρω μας



Φίλτρα spam

Χαρακτηρίζουν κάθε νέο email



Συστάσεις

Ταινίες, μουσική, προϊόντα



Φωνητικοί βοηθοί

Αναγνώριση ομιλίας



Ανίχνευση απάτης

Ύποπτες συναλλαγές σε κάρτες



Αυτόματη μετάφραση

Κείμενο από γλώσσα σε γλώσσα



Υγεία

Υποστήριξη διάγνωσης από ιατρικά δεδομένα

Ορισμός 1: Arthur Samuel (1959)

*«Το πεδίο μελέτης που δίνει στους υπολογιστές την
ικανότητα να μαθαίνουν χωρίς να έχουν
προγραμματιστεί ρητά.»*

Arthur Samuel, πρόγραμμα ντάμας της IBM, 1959

Ορισμός 2: Tom Mitchell (1997)

«Ένα πρόγραμμα μαθαίνει από την εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο απόδοσης P , αν η απόδοσή του στις T , όπως μετριέται με το P , βελτιώνεται με την E .»



Εργασία (Task)

Τι πρέπει να κάνει το σύστημα



Εμπειρία (Experience)

Τα δεδομένα από τα οποία
μαθαίνει



Απόδοση (Performance)

Πώς μετράμε αν τα πάει καλά

Παράδειγμα: το φίλτρο spam ως T, E, P



Να χαρακτηρίζει κάθε νέο email ως spam ή κανονικό



Χιλιάδες email που έχουν ήδη σημειώσει οι χρήστες ως spam ή μη



Ποσοστό σωστά χαρακτηρισμένων email (ακρίβεια, accuracy)

Προσοχή: αν απλώς κατεβάσουμε όλη τη Wikipedia, έχουμε περισσότερα δεδομένα, αλλά όχι μάθηση.
Ποια εργασία βελτιώθηκε;

Παραδείγματα T, E, P

Εργασία T	Εμπειρία E	Μέτρο απόδοσης P
Να παίζει σκάκι	Παρτίδες εναντίον του εαυτού του	Ποσοστό νικών απέναντι σε αντιπάλους
Να προβλέπει τη ζήτηση ρεύματος για την επόμενη ημέρα	Ιστορικά δεδομένα κατανάλωσης και καιρού	Μέσο σφάλμα πρόβλεψης (π.χ. σε MWh)
Να αναγνωρίζει χειρόγραφα ψηφία	Εικόνες ψηφίων με τη σωστή ετικέτα	Ποσοστό σωστά αναγνωρισμένων ψηφίων
Να προτείνει ταινίες σε χρήστες	Ιστορικό προβολών και βαθμολογιών	Ποσοστό προτάσεων που τελικά προβλήθηκαν

Παραδοσιακός προγραμματισμός έναντι MM

Παραδοσιακός προγραμματισμός



Μηχανική Μάθηση



Το μοντέλο εφαρμόζεται μετά σε νέα δεδομένα, όπως ένα κανονικό πρόγραμμα.

Πότε αξίζει να χρησιμοποιήσουμε MM;



Πάρα πολλοί κανόνες

Ένα φίλτρο spam με χιλιάδες χειροποίητους κανόνες είναι αδύνατο να συντηρηθεί.



Δεν υπάρχει γνωστός αλγόριθμος

Αναγνώριση ομιλίας: κανείς δεν μπορεί να γράψει τους κανόνες για όλες τις φωνές.



Περιβάλλον που αλλάζει

Οι απατεώνες αλλάζουν τακτική· το μοντέλο επανεκπαιδεύεται σε νέα δεδομένα.



Γνώση από τα δεδομένα

Ποιοι παράγοντες προβλέπουν την αποχώρηση πελάτη; Εδώ ξεκινά η Εξόρυξη Γνώσης.

Παράδειγμα: γιατί οι κανόνες «χάνουν» από το spam

Κανόνας: μπλόκαρε ό,τι περιέχει
«4U»

›

Οι spammers γράφουν «For U»

›

Νέος κανόνας... και ξανά από την
αρχή



Με MM

Οι χρήστες σημειώνουν νέα email ως spam → το μοντέλο επανεκπαιδεύεται → μαθαίνει μόνο του ότι το «For U» είναι πλέον ένδειξη spam.

ΤΝ, Μηχανική Μάθηση, Βαθιά Μάθηση

Τεχνητή Νοημοσύνη

Μηχανική Μάθηση

Βαθιά Μάθηση

Νευρωνικά δίκτυα με πολλά επίπεδα: CNN, LSTM, Transformers, LLMs

ΤΝ χωρίς ΜΜ

Έμπειρα συστήματα, αναζήτηση, χειροποίητοι κανόνες

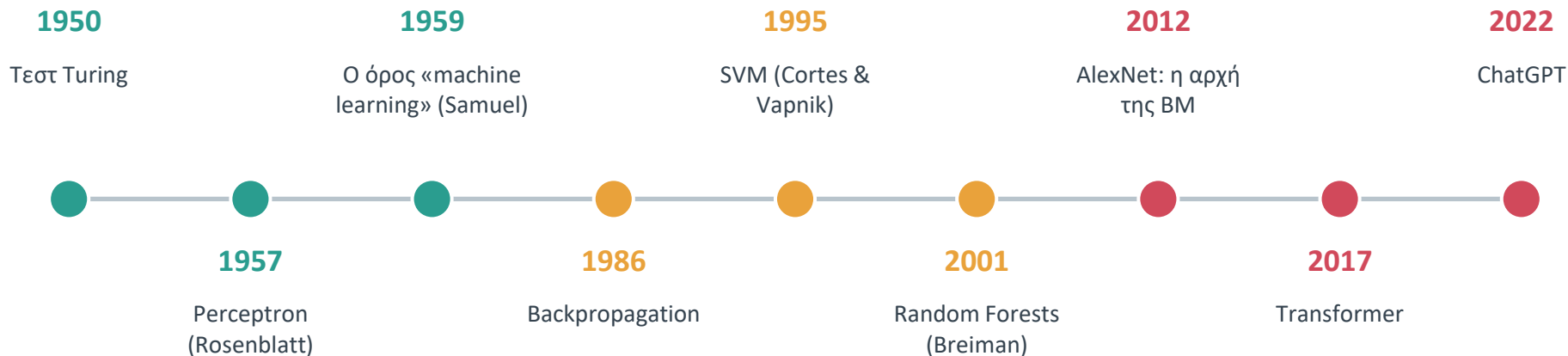
ΜΜ χωρίς ΒΜ

Γραμμικά μοντέλα, δένδρα, SVM, k-means

ΒΜ

Αναγνώριση εικόνας, μετάφραση, chatbots

Σύντομη ιστορία της Μηχανικής Μάθησης



Τι είναι η Εξόρυξη Γνώσης (Data Mining);

Η μη τετριμμένη διαδικασία εντοπισμού προτύπων σε δεδομένα που είναι:



Έγκυρα

Ισχύουν και σε νέα
δεδομένα



Καινοφανή

Δεν τα γνωρίζαμε
ήδη



Χρήσιμα

Οδηγούν σε
αποφάσεις ή
ενέργειες



Κατανοητά

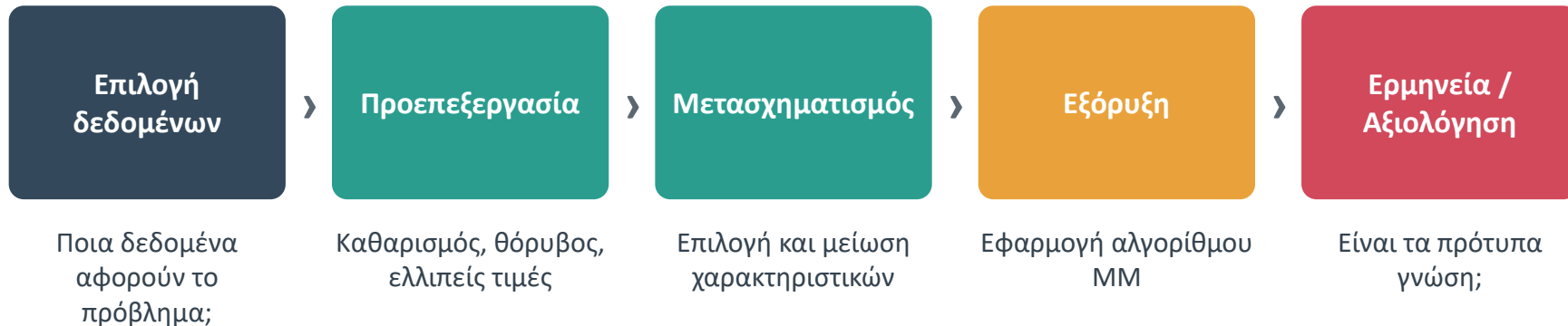
Ερμηνεύονται από
ανθρώπους

Fayyad, Piatetsky-Shapiro & Smyth, 1996

MM, Εξόρυξη Γνώσης και Στατιστική

	Στατιστική	Μηχανική Μάθηση	Εξόρυξη Γνώσης
Κύριος στόχος	Συμπερασμα για τον πληθυσμό	Ακριβής πρόβλεψη σε νέα δεδομένα	Ανακάλυψη προτύπων σε μεγάλα δεδομένα
Τυπικό ερώτημα	«Είναι η διαφορά στατιστικά σημαντική;»	«Τι θα συμβεί στο επόμενο δείγμα;»	«Τι ενδιαφέρον κρύβουν τα δεδομένα;»
Έμφαση	Υποθέσεις, διαστήματα εμπιστοσύνης	Γενίκευση, απόδοση σε test set	Κλιμάκωση, ερμηνεία, αξιοποίηση
Παράδειγμα	Επηρεάζει το φάρμακο την πίεση;	Θα αποχωρήσει αυτός ο πελάτης;	Ποιες ομάδες πελατών υπάρχουν;

Η διαδικασία KDD (Knowledge Discovery in Databases)



Η διαδικασία είναι επαναληπτική: τα αποτελέσματα της ερμηνείας συχνά μας στέλνουν πίσω σε προηγούμενα βήματα.

CRISP-DM: το πρότυπο της βιομηχανίας



Παράδειγμα CRISP-DM: αποχώρηση πελατών κινητής

Φάση	Τι κάνουμε στο παράδειγμα
1. Πρόβλημα	Στόχος: να εντοπίσουμε πελάτες που πιθανότατα θα φύγουν τον επόμενο μήνα, για να τους κάνουμε προσφορά
2. Δεδομένα	Τι έχουμε; Χρήση λεπτών/δεδομένων, παράπονα, διάρκεια συμβολαίου, ιστορικό πληρωμών
3. Προετοιμασία	Συνένωση πινάκων, ελλιπείς τιμές, νέα χαρακτηριστικά (π.χ. μεταβολή χρήσης τελευταίου τριμήνου)
4. Μοντελοποίηση	Δοκιμή λογιστικής παλινδρόμησης, δένδρων απόφασης, random forest
5. Αξιολόγηση	Εντοπίζει το μοντέλο τους πελάτες που πράγματι φεύγουν; Συμφέρει οικονομικά η καμπάνια;
6. Ανάπτυξη	Μηνιαία λίστα πελατών υψηλού κινδύνου στο τμήμα πωλήσεων, παρακολούθηση απόδοσης

ΕΝΟΤΗΤΑ

Δεδομένα και είδη μάθησης

Ορολογία, τύποι χαρακτηριστικών, επιβλεπόμενη, μη επιβλεπόμενη, ενισχυτική

Ορολογία: δείγματα, χαρακτηριστικά, ετικέτα

Εμβαδόν (m ²)	Δωμάτια	Περιοχή	Έτος κατασκευής	Τιμή (χιλ. €)
60	2	Κέντρο	1985	120
75	2	Αλικαρνασός	2005	150
80	3	Κέντρο	1998	155
95	3	Αμμουδάρα	2010	185
110	4	Κέντρο	2015	210



Γραμμή = δείγμα

Ένα σπίτι (instance, sample)



Στήλη = χαρακτηριστικό

Εμβαδόν, δωμάτια... (feature)



Στόχος / ετικέτα

Αυτό που θέλουμε να προβλέψουμε (label, target)

Συμβολισμός: m δείγματα, n χαρακτηριστικά · πίνακας X ($m \times n$) · διάνυσμα στόχων y

Τύποι χαρακτηριστικών



Αριθμητικά

Συνεχή: θερμοκρασία, τιμή

Διακριτά: αριθμός δωματίων



Κατηγορικά

Ονομαστικά: περιοχή, χρώμα

Διατακτικά: μέγεθος $S < M < L$



Μη δομημένα

Κείμενο, εικόνα, ήχος: χρειάζονται μετατροπή σε αριθμούς



Χρονοσειρές

Μετρήσεις με χρονική σειρά: κατανάλωση ανά ώρα, ΗΚΓ

Είδη μάθησης ανάλογα με την «επίβλεψη»



Επιβλεπόμενη

Όλα τα δεδομένα έχουν
ετικέτα



Μη επιβλεπόμενη

Καμία ετικέτα



Ημι-επιβλεπόμενη

Λίγες ετικέτες, πολλά
δεδομένα χωρίς



Αυτο-επιβλεπόμενη

Οι ετικέτες προκύπτουν
από τα ίδια τα δεδομένα



Ενισχυτική

Ανταμοιβές από το
περιβάλλον

Επιβλεπόμενη μάθηση: ταξινόμηση και παλινδρόμηση



Ταξινόμηση (classification)

Ο στόχος είναι κατηγορία.

- Spam ή όχι;
- Καλοήθης ή κακοήθης όγκος;
- Ποιο ψηφίο (0–9) είναι στην εικόνα;
- Θα αποπληρώσει το δάνειο;

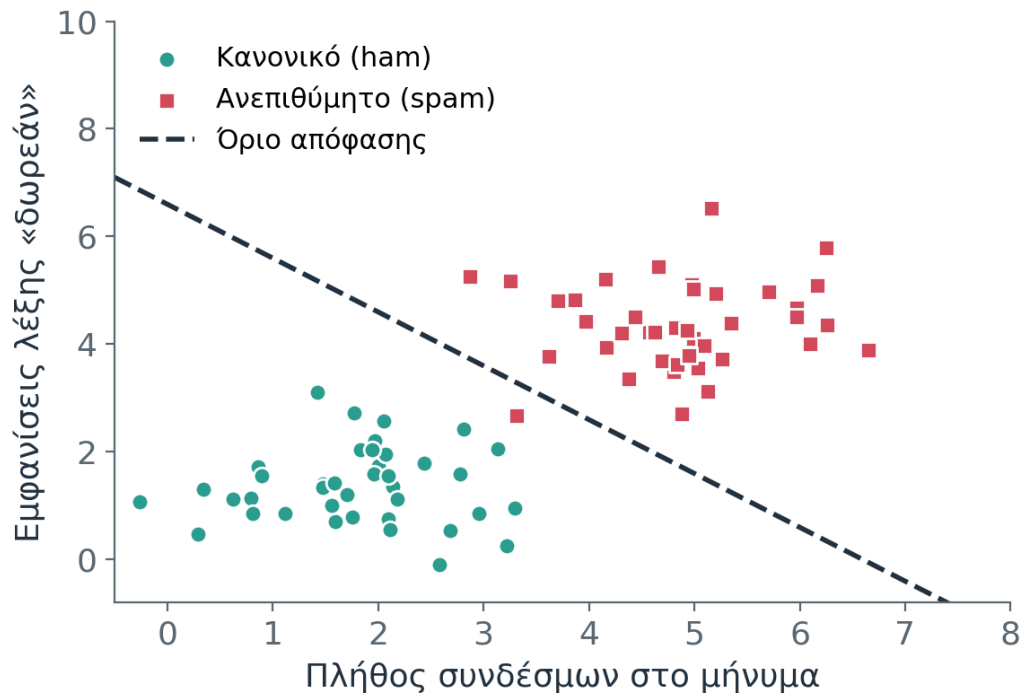


Παλινδρόμηση (regression)

Ο στόχος είναι αριθμός.

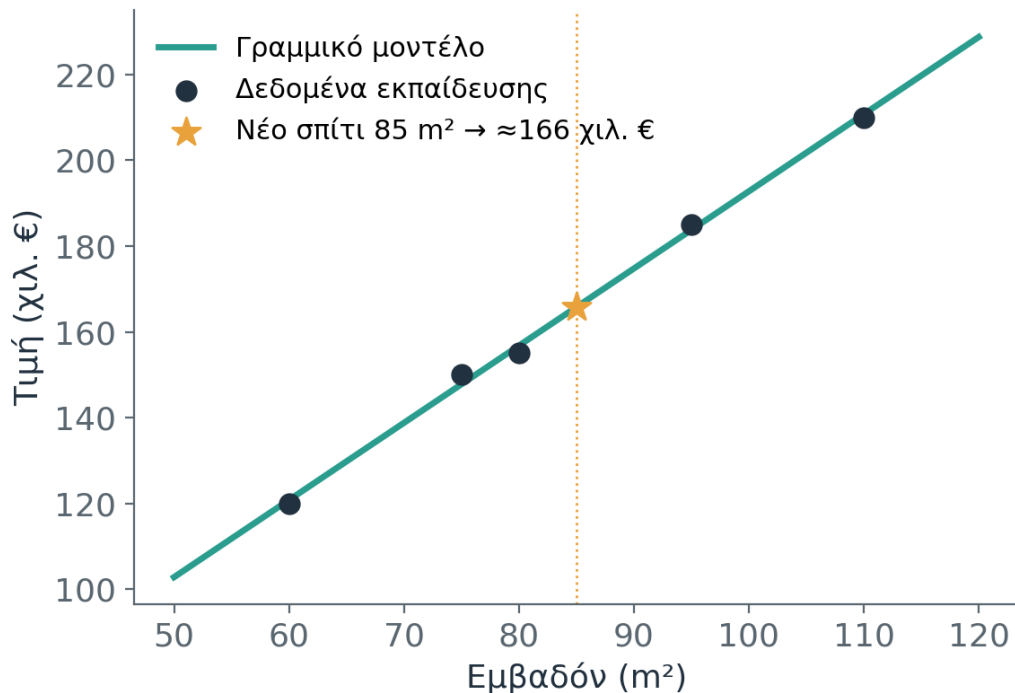
- Τιμή ενός σπιτιού
- Ζήτηση ρεύματος αύριο
- Χρόνος παράδοσης μιας παραγγελίας
- Θερμοκρασία σε 6 ώρες

Παράδειγμα ταξινόμησης: spam με δύο χαρακτηριστικά



- Κάθε email = ένα σημείο στον χώρο χαρακτηριστικών
- Ο αλγόριθμος μαθαίνει ένα όριο απόφασης από τα παραδείγματα
- Νέο email: βλέπουμε σε ποια πλευρά του ορίου πέφτει
- Στην πράξη: χιλιάδες χαρακτηριστικά, όχι δύο

Παράδειγμα παλινδρόμησης: τιμή σπιτιού



Μοντέλο

τιμή $\approx 12,9 + 1,80 \times \text{εμβαδόν}$

Πρόβλεψη για 85 m²

$12,9 + 1,80 \times 85 \approx 165,8$ χιλ. €

Ερμηνεία: κάθε επιπλέον m² ανεβάζει την τιμή κατά περίπου 1.800 €.

Μη επιβλεπόμενη μάθηση: χωρίς ετικέτες



Ομαδοποίηση

Ομάδες πελατών με παρόμοια συμπεριφορά (k-means, DBSCAN)



Μείωση διάστασης

Από 100 χαρακτηριστικά σε 2 για οπτικοποίηση (PCA, t-SNE)



Ανίχνευση ανωμαλιών

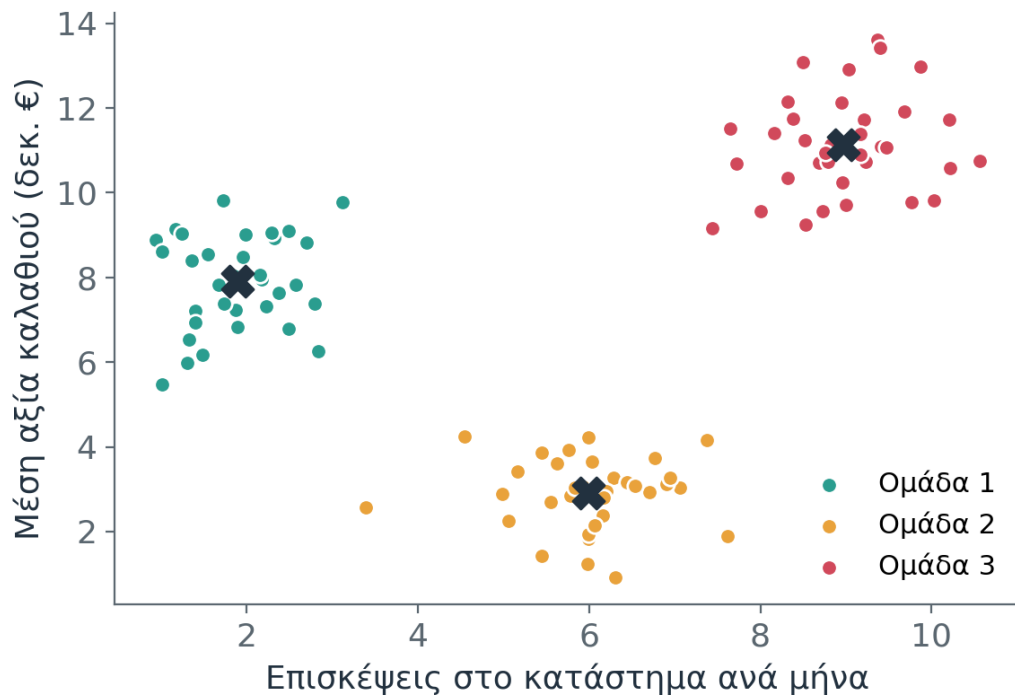
Ασυνήθιστες συναλλαγές, ελαττωματικά προϊόντα



Κανόνες συσχέτισης

Ποια προϊόντα αγοράζονται μαζί (Apriori)

Παράδειγμα ομαδοποίησης: τμηματοποίηση πελατών



Ομάδα 1

Λίγες επισκέψεις, μεγάλα καλάθια:
«εβδομαδιαίοι αγοραστές»



Ομάδα 2

Μέτριες επισκέψεις, μικρά καλάθια:
«περαστικοί»



Ομάδα 3

Συχνές επισκέψεις, μεγάλα καλάθια:
«πιστοί πελάτες»

Παράδειγμα: κανόνες συσχέτισης στο καλάθι αγορών

#	Προϊόντα στο καλάθι
Σ1	ψωμί, γάλα
Σ2	ψωμί, πάνες, μπίρα, αυγά
Σ3	γάλα, πάνες, μπίρα, αναψυκτικό
Σ4	ψωμί, γάλα, πάνες, μπίρα
Σ5	ψωμί, γάλα, πάνες, αναψυκτικό

Κανόνας: {πάνες} → {μπίρα}

Υποστήριξη (support)

Πόσο συχνά εμφανίζονται μαζί: $3/5 = 0,60$

Εμπιστοσύνη (confidence)

Όταν υπάρχουν πάνες, πόσο συχνά υπάρχει και μπίρα: $3/4 = 0,75$

Άρα: το 75% όσων αγοράζουν πάνες αγοράζει και μπίρα.

Ημι-επιβλεπόμενη και αυτο-επιβλεπόμενη μάθηση



Ημι-επιβλεπόμενη

Λίγα δεδομένα με ετικέτα, πολλά χωρίς.

Παράδειγμα: εφαρμογή φωτογραφιών.

Ομαδοποιεί τα πρόσωπα που εμφανίζονται στις φωτογραφίες (μη επιβλεπόμενο βήμα). Δίνετε ένα όνομα ανά πρόσωπο και το όνομα διαδίδεται σε όλες τις φωτογραφίες.

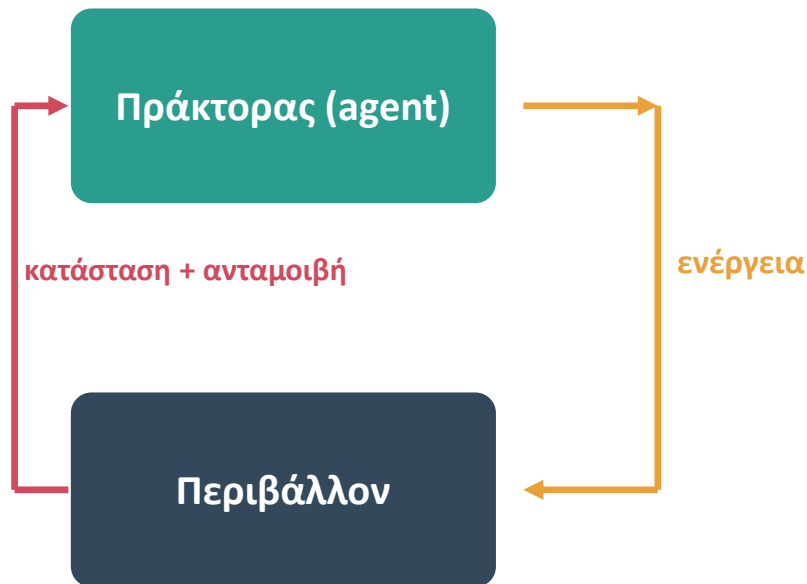


Αυτο-επιβλεπόμενη

Οι ετικέτες κατασκευάζονται από τα ίδια τα δεδομένα.

Παράδειγμα: «Η Κρήτη είναι το μεγαλύτερο ____ της Ελλάδας». Το μοντέλο μαθαίνει να προβλέπει τη λέξη που λείπει. Έτσι προεκπαιδεύονται τα μεγάλα γλωσσικά μοντέλα.

Ενισχυτική μάθηση: μάθηση από ανταμοιβές



Παραδείγματα

- AlphaGo: έμαθε Go παίζοντας εκατομμύρια παρτίδες με τον εαυτό του
- Ρομπότ που μαθαίνουν να περπατούν
- Διαχείριση ενέργειας σε κτίρια
- RLHF: ευθυγράμμιση chatbots με τις ανθρώπινες προτιμήσεις



Ερώτηση Slido 1 από 3



Μια αλυσίδα σούπερ μάρκετ θέλει να χωρίσει τους πελάτες της σε ομάδες με παρόμοια αγοραστική συμπεριφορά, χωρίς να έχει προκαθορισμένες κατηγορίες. Τι πρόβλημα είναι;

A

Ταξινόμηση

B

Παλινδρόμηση

Γ

Ομαδοποίηση

Δ

Ενισχυτική μάθηση

Batch έναντι online μάθησης



Batch (εκτός σύνδεσης)

Εκπαίδευση σε όλα τα δεδομένα μαζί, μετά το μοντέλο «παγώνει».

- Νέα δεδομένα → επανεκπαίδευση από την αρχή
- Απλό, αλλά αργό και ακριβό για μεγάλους όγκους

Παράδειγμα: μοντέλο πιστοληπτικής αξιολόγησης που ανανεώνεται κάθε μήνα.



Online (σταδιακή)

Το μοντέλο ενημερώνεται συνεχώς με μικρές παρτίδες (mini-batches).

- Γρήγορη προσαρμογή σε αλλαγές
- Κίνδυνος: κακά δεδομένα χαλούν γρήγορα το μοντέλο

Παράδειγμα: συστάσεις ειδήσεων που αλλάζουν μέσα στη μέρα.

Μάθηση βάσει στιγμιοτύπων έναντι βάσει μοντέλου



Βάσει στιγμιοτύπων (kNN)

Απομνημονεύει τα παραδείγματα και συγκρίνει με ένα μέτρο ομοιότητας.

Νέο σπίτι 85 m^2 , $k = 3$ πλησιέστεροι:

$80 \text{ m}^2 \rightarrow 155 \cdot 75 \text{ m}^2 \rightarrow 150 \cdot 95 \text{ m}^2 \rightarrow 185$

Πρόβλεψη = $(155 + 150 + 185) / 3 \approx 163,3$ χιλ. €



Βάσει μοντέλου (γραμμικό)

Χτίζει ένα μοντέλο από τα δεδομένα και μετά «ξεχνά» τα δεδομένα.

Μοντέλο: τιμή $\approx 12,9 + 1,80 \times \text{εμβαδόν}$

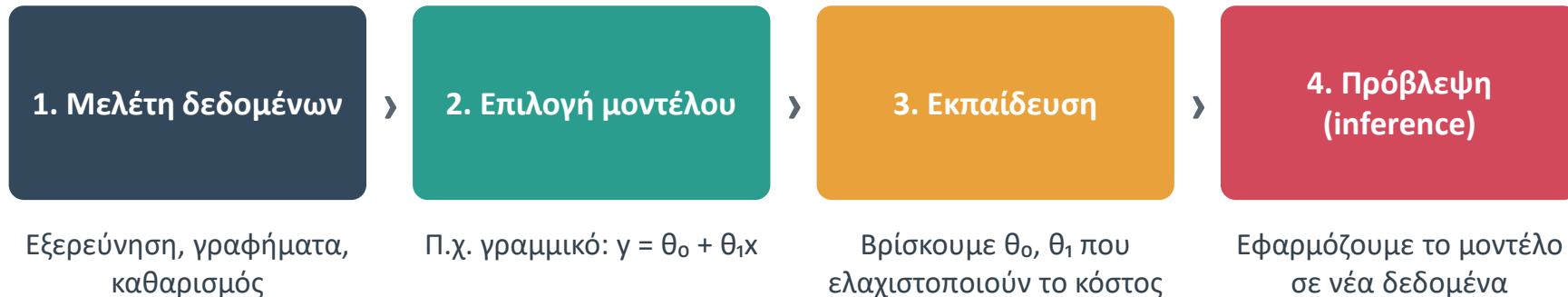
Πρόβλεψη = $12,9 + 1,80 \times 85 \approx 165,8$ χιλ. €

ΕΝΟΤΗΤΑ

Πώς «μαθαίνει» ένα μοντέλο

Μοντέλο, παράμετροι, συνάρτηση κόστους

Η τυπική ροή ενός έργου MM



Επιλογή μοντέλου = επιλογή τύπου και αρχιτεκτονικής. Εκπαίδευση = εύρεση των τιμών των παραμέτρων.

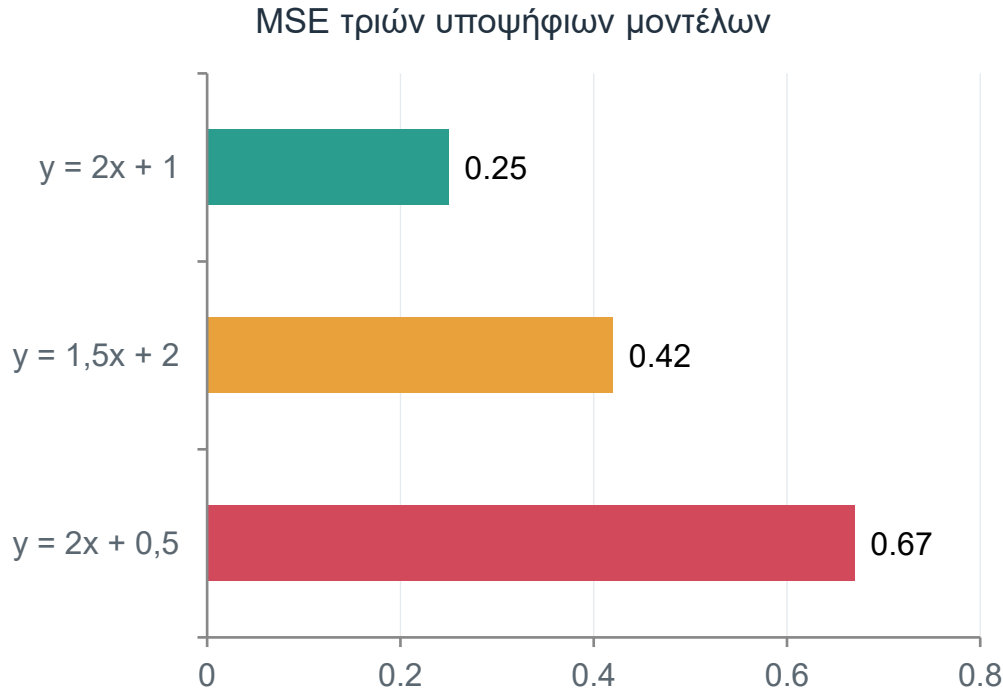
Συνάρτηση κόστους: μέσο τετραγωνικό σφάλμα (MSE)

$$\text{MSE} = (1/m) \cdot \sum (\hat{y}_i - y_i)^2$$

Μοντέλο: $\hat{y} = 2x + 1$

x	Πραγματικό y	Πρόβλεψη $\hat{y} = 2x + 1$	Σφάλμα $\hat{y} - y$	Σφάλμα ²
1	3,5	3	-0,5	0,25
2	4,5	5	+0,5	0,25
3	7,5	7	-0,5	0,25
			MSE =	0,25

Εκπαίδευση = αναζήτηση των καλύτερων παραμέτρων



- Κάθε συνδυασμός θ_0, θ_1 δίνει διαφορετικό κόστος
- Εκπαίδευση: βρίσκουμε τον συνδυασμό με το ελάχιστο κόστος
- Εδώ νικά το $y = 2x + 1$ (MSE = 0,25)
- Στη Διάλεξη 3: πώς το κάνει αυτόματα το gradient descent

Παράμετροι έναντι υπερπαραμέτρων

Μοντέλο	Παράμετροι (τις μαθαίνει)	Υπερπαραμέτροι (τις ορίζουμε εμείς)
Γραμμική παλινδρόμηση	Συντελεστές $\theta_0, \theta_1, \dots$	Βαθμός πολυωνύμου, ισχύς ομαλοποίησης λ
k-πλησιέστεροι γείτονες	— (αποθηκεύει τα δεδομένα)	Πλήθος γειτόνων k, μέτρο απόστασης
Δένδρο απόφασης	Χαρακτηριστικά και κατώφλια διάσπασης	Μέγιστο βάθος, ελάχιστα δείγματα ανά φύλλο
Νευρωνικό δίκτυο	Βάρη και πολώσεις	Πλήθος επιπέδων, ρυθμός μάθησης, εποχές

ΕΝΟΤΗΤΑ

Οι κύριες προκλήσεις

Κακά δεδομένα, κακά μοντέλα, υπερπροσαρμογή

Όταν φτάνει τα δεδομένα



Ανεπαρκή δεδομένα

Για σύνθετα προβλήματα χρειάζονται χιλιάδες ή εκατομμύρια παραδείγματα.



Μη αντιπροσωπευτικά

Μοντέλο που εκπαιδεύτηκε μόνο σε ασθενείς ενός νοσοκομείου μπορεί να αποτύχει σε άλλο.



Κακής ποιότητας

Λάθη μέτρησης, ακραίες τιμές, ελλιπείς τιμές.



Άσχετα χαρακτηριστικά

Π.χ. το όνομα της χώρας ως χαρακτηριστικό πρόβλεψης ικανοποίησης ζωής.

Παράδειγμα μεροληψίας δείγματος: Literary Digest, 1936

10 εκατ.

ερωτηματολόγια στάλθηκαν

2,4 εκατ.

απαντήσεις

57%

πρόβλεψη για Landon

62%

πήρε τελικά ο Roosevelt

- Διευθύνσεις από τηλεφωνικούς καταλόγους και συνδρομητές περιοδικών → υπερεκπροσώπηση εύπορων ψηφοφόρων
- Απάντησε κάτω από το 25% → μεροληψία μη απόκρισης
- Δίδαγμα: ένα μεγάλο αλλά μεροληπτικό δείγμα δεν είναι αντιπροσωπευτικό

Μηχανική χαρακτηριστικών (feature engineering)



Επιλογή

Κρατάμε τα πιο χρήσιμα χαρακτηριστικά από όσα έχουμε.

Παράδειγμα:
Αφαιρούμε το «ID πελάτη»



Εξαγωγή

Συνδυάζουμε χαρακτηριστικά σε ένα πιο χρήσιμο.

Παράδειγμα:
Χιλιόμετρα + ηλικία
αυτοκινήτου → «φθορά»



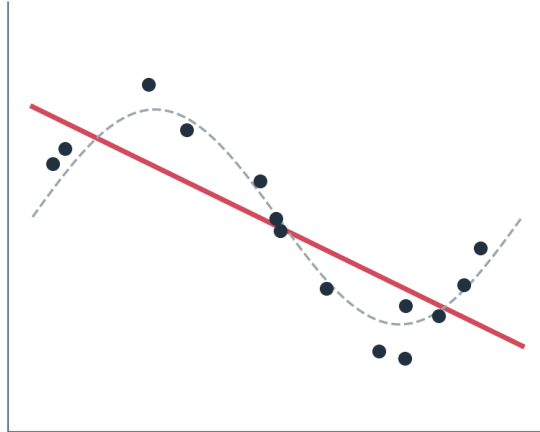
Δημιουργία

Φτιάχνουμε νέα χαρακτηριστικά από τα ακατέργαστα.

Παράδειγμα:
Ημερομηνία γέννησης →
ηλικία

Υποπροσαρμογή, καλή γενίκευση, υπερπροσαρμογή

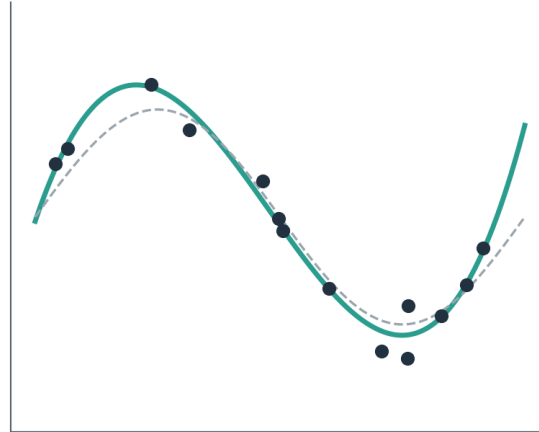
Βαθμός 1: υποπροσαρμογή



Βαθμός 1

Πολύ απλό: δεν «πιάνει» ούτε τα δεδομένα εκπαίδευσης

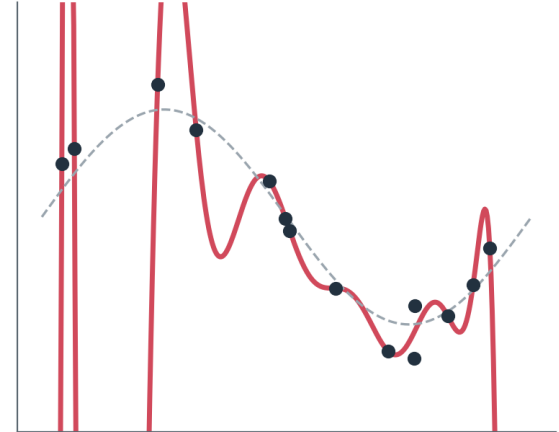
Βαθμός 3: καλή γενίκευση



Βαθμός 3

Ακολουθεί την πραγματική τάση

Βαθμός 12: υπερπροσαρμογή



Βαθμός 12

Αποστηθίζει τον θόρυβο

Υπερπροσαρμογή και υποπροσαρμογή: τι κάνουμε



Υπερπροσαρμογή (overfitting)

Σύμπτωμα: άριστο στην εκπαίδευση, κακό σε νέα δεδομένα.

- Απλούστερο μοντέλο ή λιγότερα χαρακτηριστικά
- Ομαλοποίηση (regularization)
- Περισσότερα δεδομένα
- Καθαρισμός θορύβου



Υποπροσαρμογή (underfitting)

Σύμπτωμα: κακό ακόμη και στα δεδομένα εκπαίδευσης.

- Πιο ισχυρό μοντέλο (περισσότερες παράμετροι)
- Καλύτερα χαρακτηριστικά
- Λιγότερη ομαλοποίηση



Ερώτηση Slido 2 από 3



Εκπαιδεύσατε τρία μοντέλα. Σφάλμα εκπαίδευσης / σφάλμα ελέγχου: A = 12% / 13%, B = 1% / 25%, Γ = 30% / 31%. Ποιο παρουσιάζει υπερπροσαρμογή;

A

Το μοντέλο A

B

Το μοντέλο B

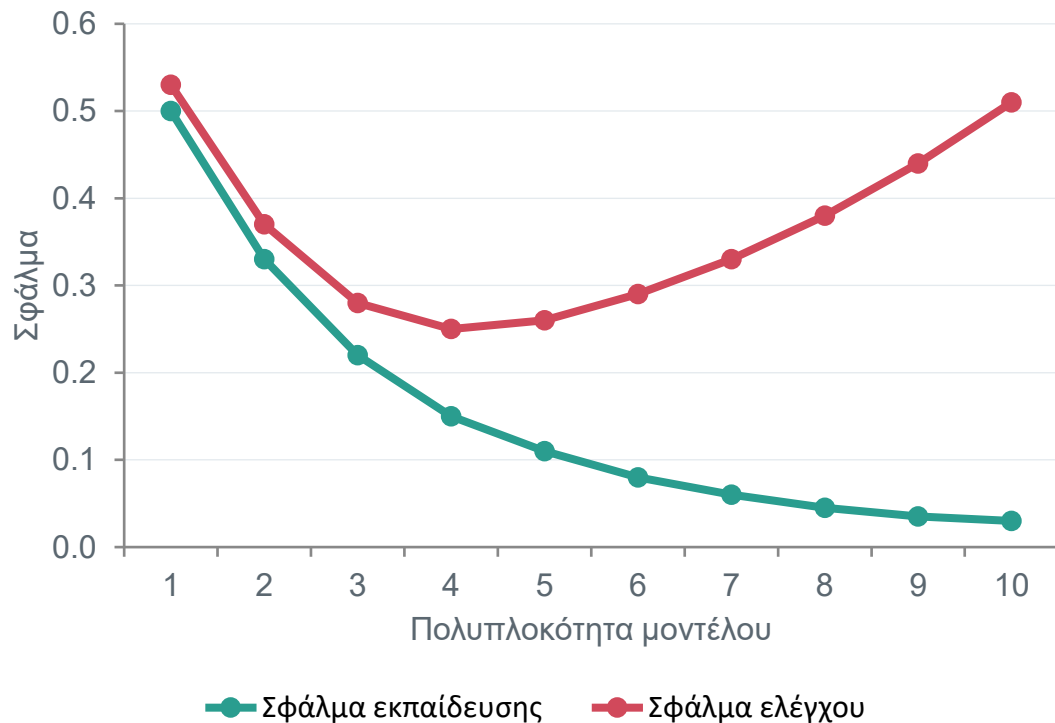
Γ

Το μοντέλο Γ

Δ

Κανένα

Το δίλημμα μεροληψίας-διακύμανσης



Μεροληψία (bias)

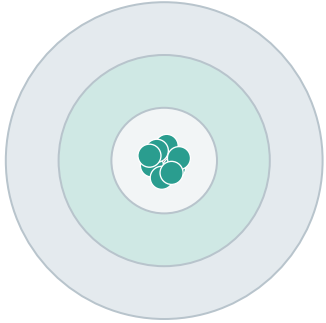
Λάθος απλουστευτικές υποθέσεις → υποπροσαρμογή

Διακύμανση (variance)

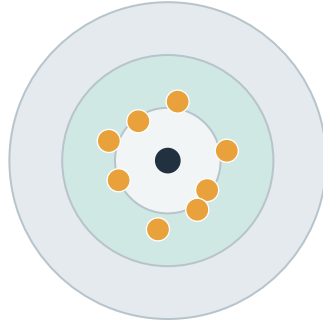
Υπερβολική ευαισθησία στα δεδομένα εκπαίδευσης → υπερπροσαρμογή

Στόχος: το ελάχιστο της κόκκινης καμπύλης

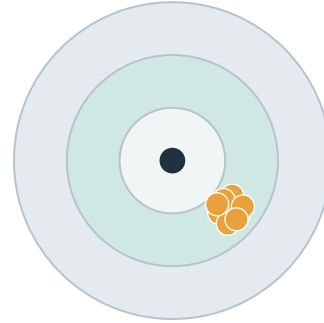
Μεροληψία και διακύμανση ως στόχος σκοποβολής



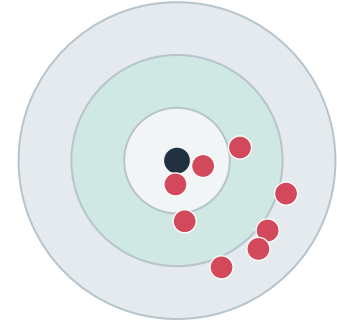
Χαμηλή μεροληψία, χαμηλή
διακύμανση



Χαμηλή μεροληψία, υψηλή
διακύμανση



Υψηλή μεροληψία, χαμηλή
διακύμανση



Υψηλή μεροληψία, υψηλή
διακύμανση

ΕΝΟΤΗΤΑ

Αξιολόγηση και επιλογή μοντέλου

Train / validation / test, cross-validation, διαρροή δεδομένων

Διαχωρισμός σε σύνολο εκπαίδευσης και ελέγχου

Εκπαίδευση (train) 80% · 8.000 δείγματα

Έλεγχος 20% · 2.000

Παράδειγμα: σύνολο 10.000 δειγμάτων

- Εκπαιδεύουμε MONO στο train set
- Το σφάλμα στο test set εκτιμά το σφάλμα γενίκευσης (out-of-sample error)
- Χαμηλό σφάλμα εκπαίδευσης + υψηλό σφάλμα ελέγχου = υπερπροσαρμογή
- Το test set το «ανοίγουμε» μία φορά, στο τέλος

Σύνολο επικύρωσης (validation) για τις υπερπαραμέτρους



To validation set λέγεται και dev set (development set).

Διασταυρωμένη επικύρωση k-fold ($k = 5$)

Επανάληψη 1	valid.					Ακρίβεια 0,82
Επανάληψη 2		valid.				Ακρίβεια 0,85
Επανάληψη 3			valid.			Ακρίβεια 0,80
Επανάληψη 4				valid.		Ακρίβεια 0,84
Επανάληψη 5					valid.	Ακρίβεια 0,83

Εκτίμηση απόδοσης = μέσος όρος = 0,828

Διαρροή δεδομένων (data leakage)



**Προεπεξεργασία σε όλα
τα δεδομένα**

Κανονικοποίηση με μέση τιμή
που υπολογίστηκε και από το
test set, πριν τον διαχωρισμό.



**Ο ίδιος ασθενής και στα
δύο σύνολα**

Πολλές μετρήσεις του ίδιου
ασθενή μοιράζονται σε train
και test: το μοντέλο
«αναγνωρίζει» τον ασθενή.



**Χαρακτηριστικό που
«προδίδει» τον στόχο**

Στην πρόβλεψη αποχώρησης:
στήλη «ημερομηνία
ακύρωσης συμβολαίου».



Ερώτηση Slido 3 από 3



Δοκιμάσατε 100 τιμές μιας υπερπαραμέτρου αξιολογώντας κάθε φορά στο test set. Η καλύτερη δίνει 95% ακρίβεια. Τι ισχύει;

A

Το 95% είναι αξιόπιστη εκτίμηση για νέα δεδομένα

B

Το 95% είναι πιθανότατα υπερεκτίμηση

Γ

Πρέπει να μεγαλώσουμε το test set στο 50%

Δ

Το μοντέλο υποπροσαρμόζεται

Δεν υπάρχει «δωρεάν γεύμα» (No Free Lunch)

Δεν υπάρχει μοντέλο που να είναι εγγυημένα το καλύτερο για κάθε πρόβλημα.

- Κάθε μοντέλο κάνει υποθέσεις (π.χ. το γραμμικό υποθέτει γραμμική σχέση)
- Σε κάποια δεδομένα κερδίζει το γραμμικό μοντέλο, σε άλλα ένα νευρωνικό δίκτυο
- Στην πράξη: λογικές υποθέσεις + σύγκριση λίγων υποψήφίων μοντέλων

Wolpert, 1996

Σύνοψη

1

MM: βελτίωση σε μια εργασία T μέσω εμπειρίας E , μετρημένη με P

2

Εξόρυξη Γνώσης: ανακάλυψη έγκυρων, νέων, χρήσιμων προτύπων (KDD, CRISP-DM)

3

Είδη: επιβλεπόμενη, μη επιβλεπόμενη, ημι-, αυτο-επιβλεπόμενη, ενισχυτική

4

Εκπαίδευση = εύρεση παραμέτρων που ελαχιστοποιούν μια συνάρτηση κόστους

5

Υπερπροσαρμογή έναντι υποπροσαρμογής: το δίλημμα μεροληψίας–διακύμανσης

6

Αξιολόγηση: train / validation / test, cross-validation, προσοχή στη διαρροή

Για την επόμενη διάλεξη



Διάβασμα

- Géron, Κεφ. 1 (και ερωτήσεις στο τέλος)
- Géron, Κεφ. 2 (για τη Διάλεξη 2)
- ISL, Κεφ. 2
- Κάλλιπος «Μηχανική Μάθηση», 4.1–4.2



Προετοιμασία

- Λογαριασμός Google για το Colab
- Επανάληψη βασικών Python, NumPy, pandas

Διάλεξη 2: Δεδομένα και προεπεξεργασία,
ένα έργο MM από την αρχή ως το τέλος



Ερωτήσεις;

koumakis@hmu.gr · Ώρες γραφείου: Δευτέρα 11:00–13:00