

Τεχνητή Νοημοσύνη I

Μηχανική μάθηση

Διδάσκων: Τσούρος Δήμος

Εισαγωγή στην (στατιστική) μηχανική μάθηση

Μηχανική Μάθηση

- **Μηχανική Μάθηση** (machine learning) είναι ο τομέας της ΤΝ που ασχολείται με πράκτορες που μαθαίνουν από τις εμπειρίες τους
- **Βασική ιδέα:** Ό,τι προσλαμβάνει ο πράκτορας από το περιβάλλον μπορεί να χρησιμοποιηθεί:
 - για να αποφασίσει τι ενέργειες θα κάνει για να λύσει ένα συγκεκριμένο πρόβλημα
 - και για να μάθει να κάνει καλύτερες ενέργειες μελλοντικά

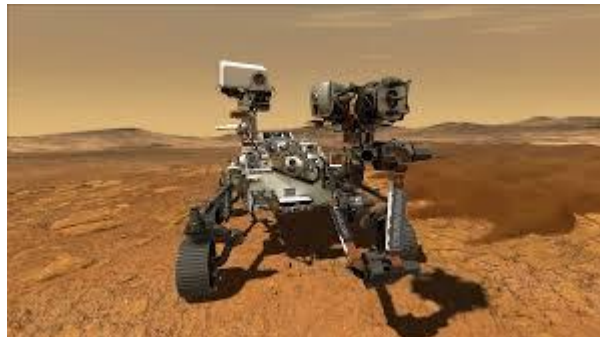
Μηχανική Μάθηση

Γενικός ορισμός: Η Μηχανική Μάθηση (*Machine Learning*) είναι το πεδίο της Τεχνητής Νοημοσύνης που αναπτύσσει αλγορίθμους οι οποίοι **μαθαίνουν αυτόματα μοτίβα από δεδομένα**, χωρίς να έχουν προγραμματιστεί ρητά γι' αυτό.

Τεχνικός ορισμός: Ένα πρόγραμμα μαθαίνει από εμπειρία E σε σχέση με μια εργασία T και ένα μέτρο απόδοσης P , αν η απόδοσή του στην εργασία T , όπως μετριέται από το P , βελτιώνεται με την εμπειρία E .

Παραδείγματα Εφαρμογών

- **Ιατρική διάγνωση:** πρόβλεψη ασθενειών από εξετάσεις
- **Οικονομικά:** εκτίμηση κινδύνου δανείου, πρόβλεψη τιμών
- **Μάρκετινγκ:** πρόβλεψη συμπεριφοράς πελατών
- **Αυτόνομα οχήματα:** λήψη αποφάσεων από αισθητήρες
- **Συστάσεις:** Netflix, Spotify, YouTube



Μηχανική Μάθηση και Στατιστική

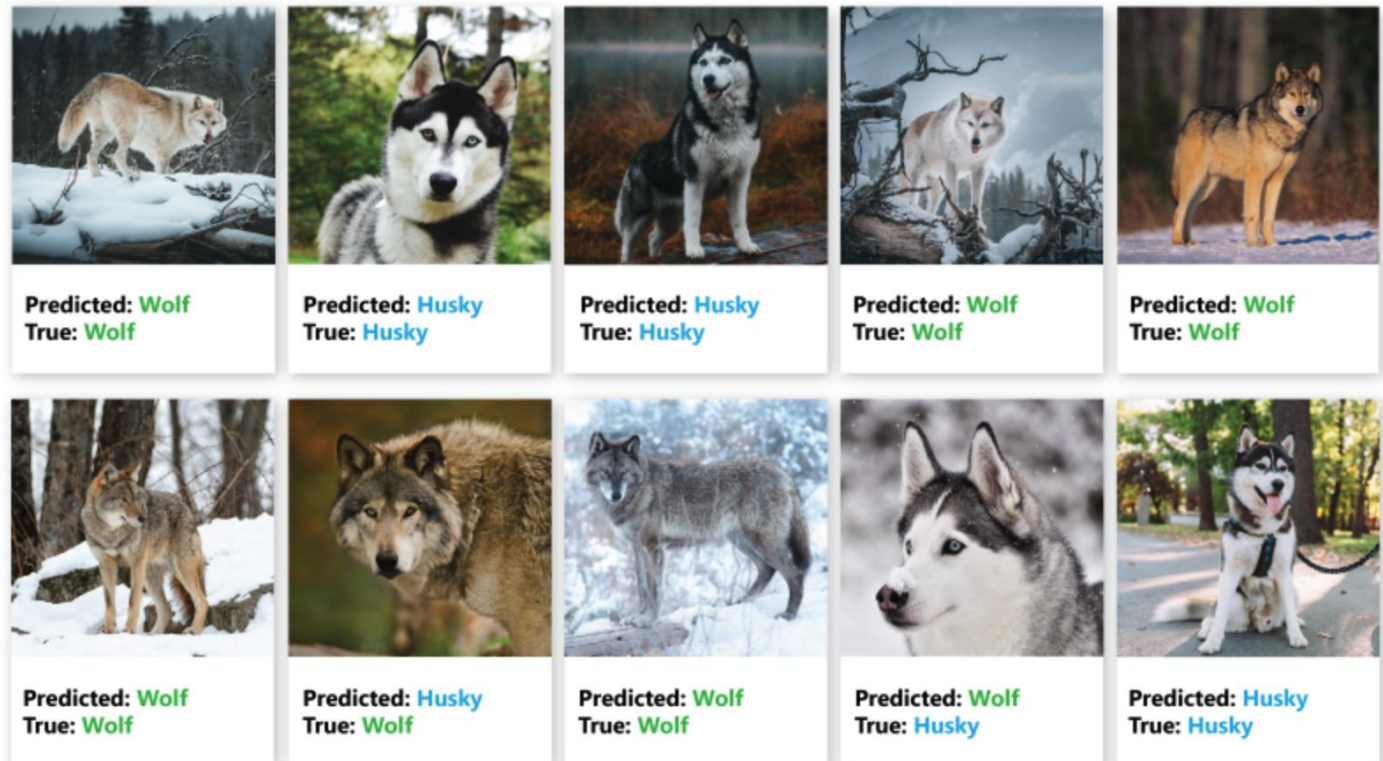
Η Στατιστική και η Μηχανική Μάθηση έχουν κοινό στόχο: **να εξαγάγουν πληροφορία από δεδομένα.**

Στατιστική	Μηχανική Μάθηση
Εστιάζει στην κατανόηση και ερμηνεία των σχέσεων μεταξύ δεδομένων.	Εστιάζει στην πρόβλεψη για το μέλλον και την ακρίβεια των προβλέψεων .
Μοντέλα παραμετρικά, μικρά δείγματα.	Μεγάλα δεδομένα, μη παραμετρικά μοντέλα.
Πιο <i>μαθηματική</i> προσέγγιση.	Πιο <i>αλγοριθμική</i> προσέγγιση.

Η στατιστική εξηγεί το «γιατί», η μηχανική μάθηση προβλέπει το «τι θα γίνει».
Η Μηχανική Μάθηση χρησιμοποιεί στατιστική. Συχνά ονομάζεται «Στατιστική Μηχανική Μάθηση»

Μηχανική Μάθηση

- Χρησιμοποιεί: Δεδομένα για ένα πρόβλημα
 - (ανεξάρτητες μεταβλητές)
- Στόχος: Να μάθει πως σχετίζονται τα δεδομένα
 - (εξαρτημένες μεταβλητές)
- Π.χ. «Μάθε να ξεχωρίζεις λύκους από Husky

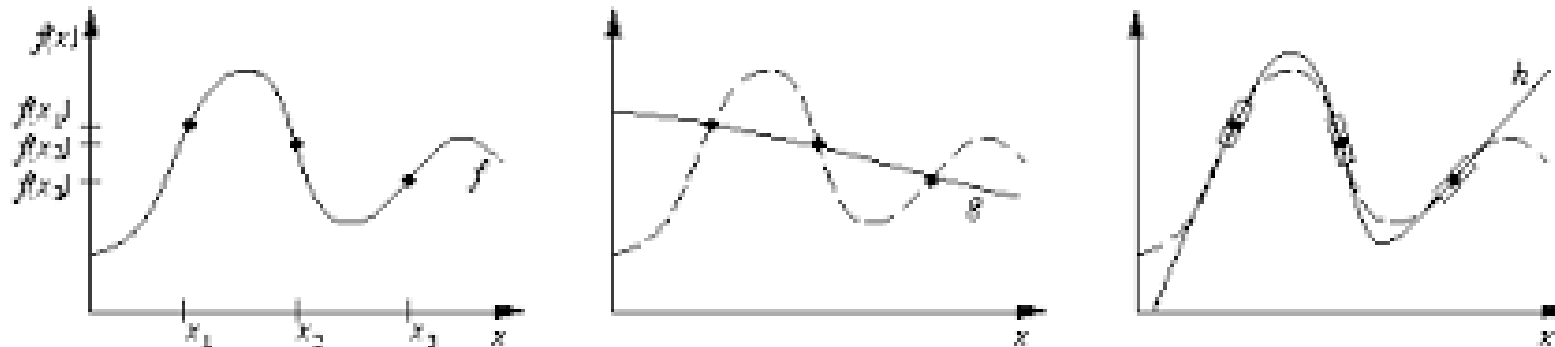


Μηχανική Μάθηση

- Η διαδικασία μάθησης γενικά προσεγγίζεται ως η διαδικασία μάθησης της αναπαράστασης μιας συνάρτησης

Επαγωγική Μάθηση (inductive learning) είναι η μάθηση μιας έννοιας (δηλ. μιας συνάρτησης) μέσω ενός συνόλου παραδειγμάτων:

- Ένα **παράδειγμα** είναι μια δυάδα (x, y) , όπου y είναι το αποτέλεσμα/η σωστή απάντηση, το x είναι η είσοδος, και το θεωρούμε πως $y=f(x)$ είναι η έξοδος μιας άγνωστης συνάρτησης f όταν εφαρμόζεται στο x

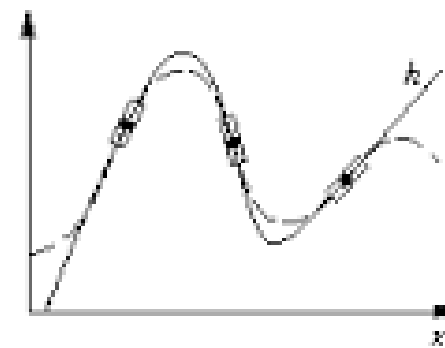
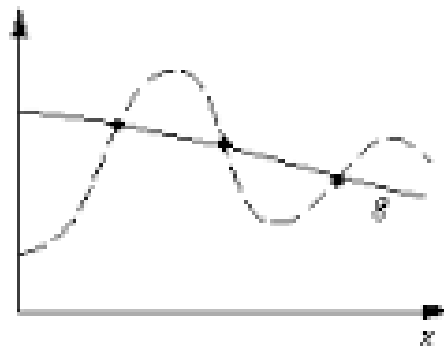
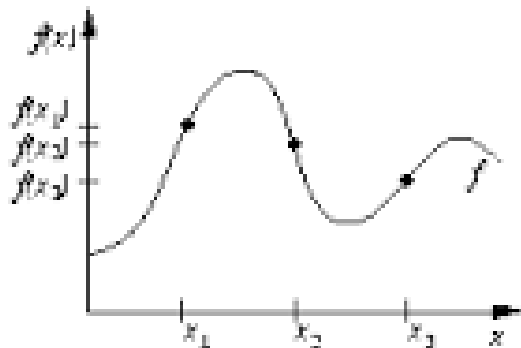


Επαγωγική Μάθηση

Στόχος επαγωγικής μάθησης:

Δοθέντος ενός συνόλου παραδειγμάτων μιας συνάρτησης f , βρες μια συνάρτηση h που προσεγγίζει την f

- Η συνάρτηση h ονομάζεται **υπόθεση**
 - Το πρόβλημα εύρεσης μια καλής υπόθεσης είναι δύσκολο...
 - Όσο περισσότερα δεδομένα έχουμε, τόσο καλύτερα μπορούμε να προσεγγίσουμε την πραγματική συνάρτηση.



Είναι Δυνατόν; (Το Πρόβλημα της Επαγωγής)

- $y = f(x)$
- Δείγματα/δεδομένα: (x, y)
- Βρες μια υπόθεση $h(x)$, τέτοια ώστε να αντικατοπτρίζει την $f(x)$:
 - $h(x) \sim f(x)$
 - Μεγάλος (άπειρος) χώρος υποθέσεων
- Η υπόθεση/συνάρτηση που μαθαίνουμε πρέπει να είναι συνεπείς με τα δεδομένα
- Πολλές υποθέσεις/συναρτήσεις μπορεί να είναι ακριβείς.
 - Πως επιλέγουμε ποια να χρησιμοποιήσουμε;

Είναι Δυνατόν; (Το Πρόβλημα της Επαγωγής)

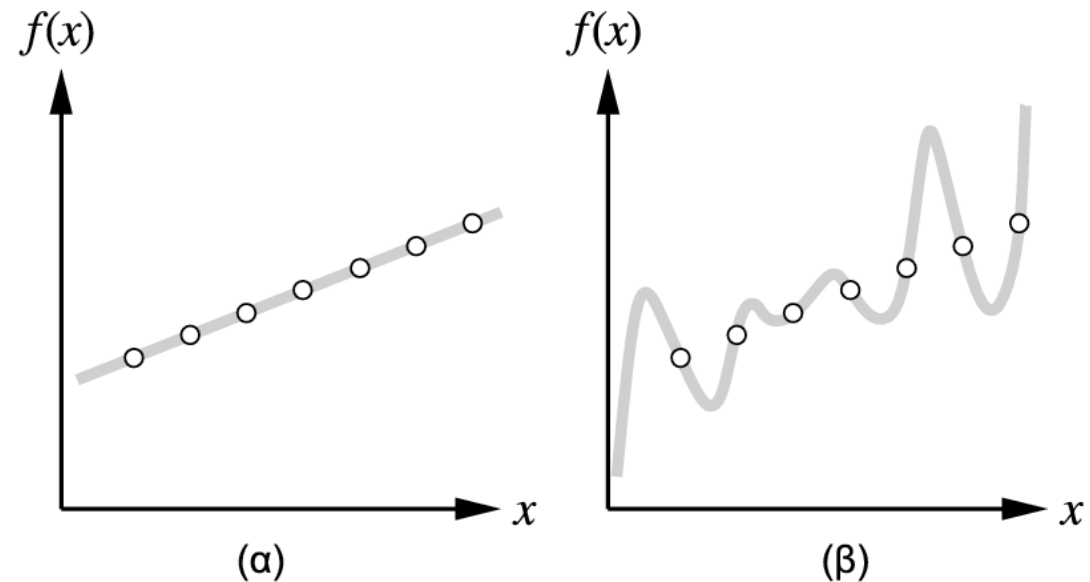
Πολλές υποθέσεις/συναρτήσεις μπορεί να είναι ακριβείς.

Πως επιλέγουμε ποια να χρησιμοποιήσουμε;

Ξυράφι του Ockham:

- «Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από όσες είναι απαραίτητες»
- «από όλες τις θεωρίες που εξηγούν επαρκώς τα δεδομένα, προτιμάται η απλούστερη»

Στόχος η γενίκευση



Είναι Δυνατόν; (Το Πρόβλημα της Επαγωγής)

Συνήθως: Δεδομένα
πολλαπλών μεταβλητών

Viewer

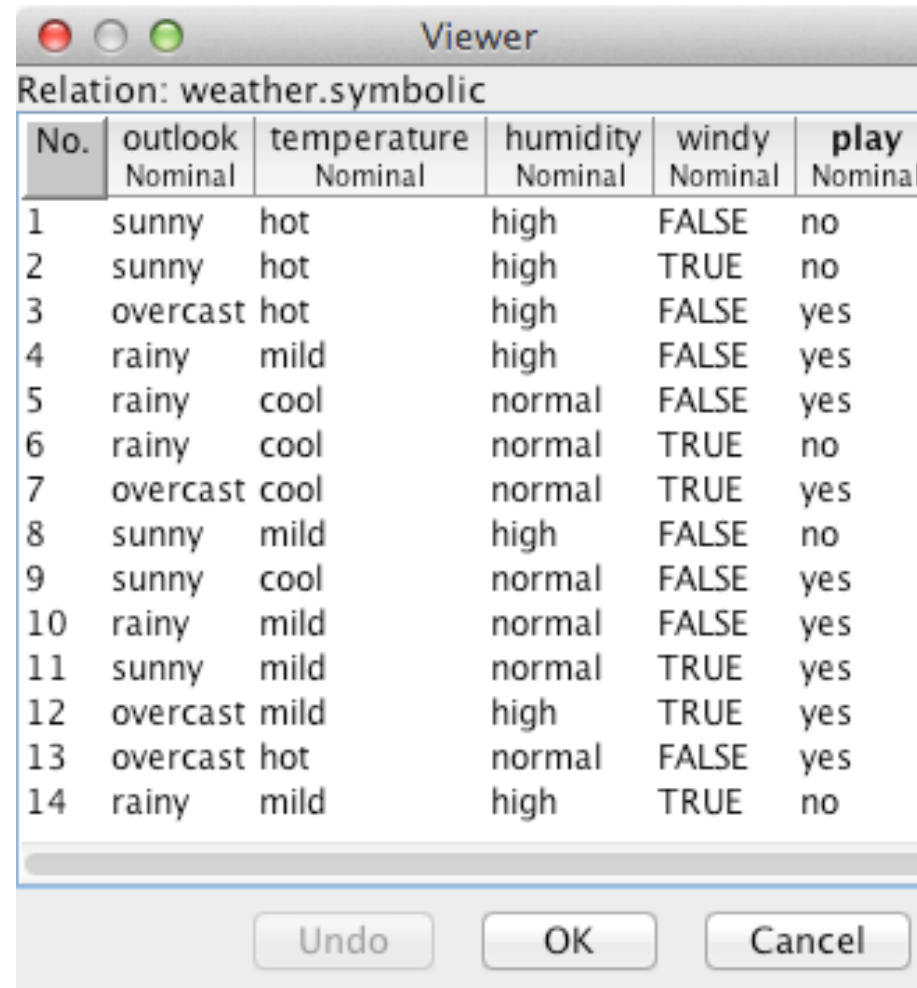
Relation: Glass

No.	RI Numeric	Na Numeric	Mg Numeric	Al Numeric	Si Numeric	K Numeric	Ca Numeric	Ba Numeric	Fe Numeric	Type Nominal
1	1.51793	12.79	3.5	1.12	73.03	0.64	8.77	0.0	0.0	build wind float
2	1.51643	12.16	3.52	1.35	72.89	0.57	8.53	0.0	0.0	vehic wind float
3	1.51793	13.21	3.48	1.41	72.64	0.59	8.43	0.0	0.0	build wind float
4	1.51299	14.4	1.74	1.54	74.55	0.0	7.59	0.0	0.0	tableware
5	1.53393	12.3	0.0	1.0	70.16	0.12	16.19	0.0	0.24	build wind non-float
6	1.51655	12.75	2.85	1.44	73.27	0.57	8.79	0.11	0.22	build wind non-float
7	1.51779	13.64	3.65	0.65	73.0	0.06	8.93	0.0	0.0	vehic wind float
8	1.51837	13.14	2.84	1.28	72.85	0.55	9.07	0.0	0.0	build wind float
9	1.51545	14.14	0.0	2.68	73.39	0.08	9.07	0.61	0.05	headlamps
10	1.51789	13.19	3.9	1.3	72.33	0.55	8.44	0.0	0.28	build wind non-float
11	1.51625	13.36	3.58	1.49	72.72	0.45	8.21	0.0	0.0	build wind non-float
12	1.51743	12.2	3.25	1.16	73.55	0.62	8.9	0.0	0.24	build wind non-float
13	1.52223	13.21	3.77	0.79	71.99	0.13	10.02	0.0	0.0	build wind float
14	1.52121	14.03	3.76	0.58	71.79	0.11	9.65	0.0	0.0	vehic wind float
15	1.51665	13.14	3.45	1.76	72.48	0.6	8.38	0.0	0.17	vehic wind float
16	1.51707	13.48	3.48	1.71	72.52	0.62	7.99	0.0	0.0	build wind non-float
17	1.51719	14.75	0.0	2.0	73.02	0.0	8.53	1.59	0.08	headlamps
18	1.51629	12.71	3.33	1.49	73.28	0.67	8.24	0.0	0.0	build wind non-float
19	1.51994	13.27	0.0	1.76	73.03	0.47	11.32	0.0	0.0	containers
20	1.51811	12.96	2.96	1.43	72.92	0.6	8.79	0.14	0.0	build wind non-float
21	1.52152	13.05	3.65	0.87	72.22	0.19	9.85	0.0	0.17	build wind float
22	1.52475	11.45	0.0	1.88	72.19	0.81	13.24	0.0	0.34	build wind non-float
23	1.51841	12.93	3.74	1.11	72.28	0.64	8.96	0.0	0.22	build wind non-float
24	1.51754	13.39	3.66	1.19	72.79	0.57	8.27	0.0	0.11	build wind float
25	1.52058	12.85	1.61	2.17	72.18	0.76	9.7	0.24	0.51	containers
26	1.51569	13.24	3.49	1.47	73.25	0.38	8.03	0.0	0.0	build wind non-float
27	1.5159	12.82	3.52	1.9	72.86	0.69	7.97	0.0	0.0	build wind non-float
28	1.51683	14.56	0.0	1.98	73.29	0.0	8.52	1.57	0.07	headlamps
29	1.51687	13.23	3.54	1.48	72.84	0.56	8.1	0.0	0.0	build wind non-float
30	1.5161	13.33	3.53	1.34	72.67	0.56	8.33	0.0	0.0	vehic wind float

Undo OK Cancel

Είναι Δυνατόν; (Το Πρόβλημα της Επαγωγής)

Συνήθως: Δεδομένα
πολλαπλών μεταβλητών



Relation: weather.symbolic					
No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

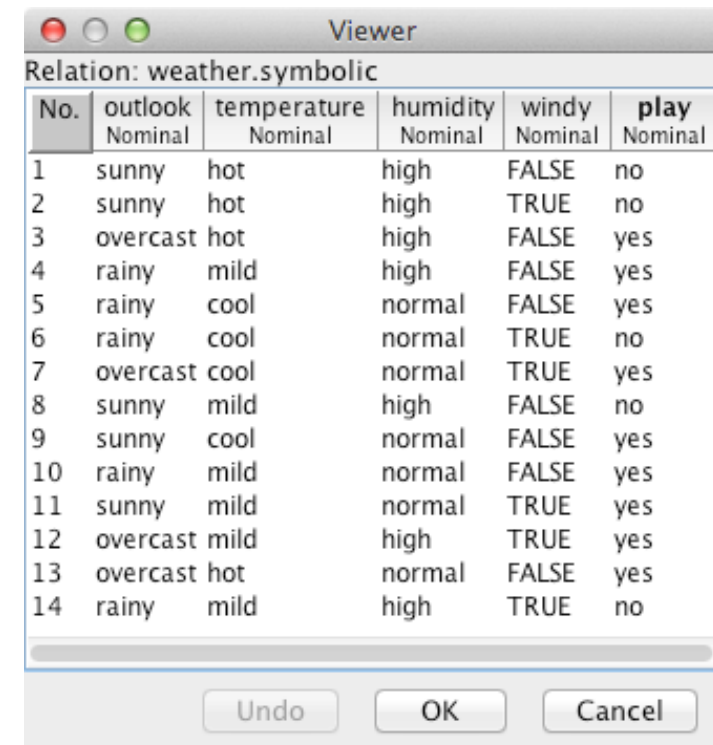
Όχι πάντα με αριθμητικές
τιμές!

Μηχανική Μάθηση

- Έχουν προταθεί πολλές τεχνικές για την επίλυση προβλημάτων μηχανικής μάθησης
 - Δηλ. πολλές τεχνικές για την υλοποίηση του τμήματος μάθησης ενός πράκτορα μάθησης
- Η διαφοροποίηση συνίσταται:
 - Στον τρόπο με τον οποίο γίνεται η **μάθηση**
 - σε ποια στοιχεία του τμήματος απόδοσης επιδρά
 - Στον τρόπο **αναπαράστασης** της γνώσης
 - Στο **feedback** που είναι διαθέσιμο
 - Στην ύπαρξη ή όχι **αρχικής γνώσης**

Βασικές έννοιες

- Μοντέλο (Model, Concept)
- Στιγμιότυπο/παράδειγμα/δείγμα (Instance, Example)
- Χαρακτηριστικό (Feature, Μεταβλητή)
- Κλάση (Class)
- Επίβλεψη (Supervision)
- Θόρυβος (Noise)



Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

Μηχανική Μάθηση

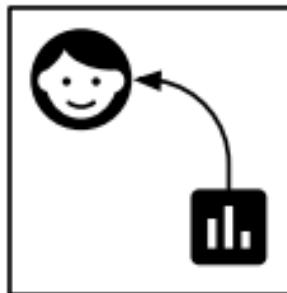
Υπάρχουν τρεις βασικές κατηγορίες μηχανικής μάθησης:

- **Μάθηση με επίβλεψη** (supervised learning)
 - Μαθαίνουμε από δεδομένα, και παραδείγματα τιμών στόχου (ετικέτες-labels) από ειδικούς
- **Μάθηση χωρίς επίβλεψη** (unsupervised learning)
 - μάθηση χωρίς να ξέρουμε το τιμών στόχου, από γενικά δεδομένα
- **Ενισχυτική μάθηση** (reinforcement learning)
 - Ο πράκτορας μαθαίνει μέσω δοκιμής και σφάλματος, λαμβάνοντας ανταμοιβές ή ποινές

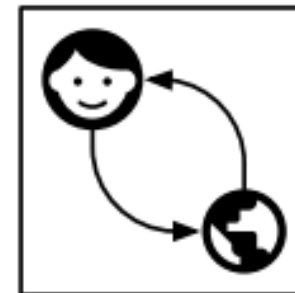
Supervised
Learning



Unsupervised
Learning



Reinforcement
Learning



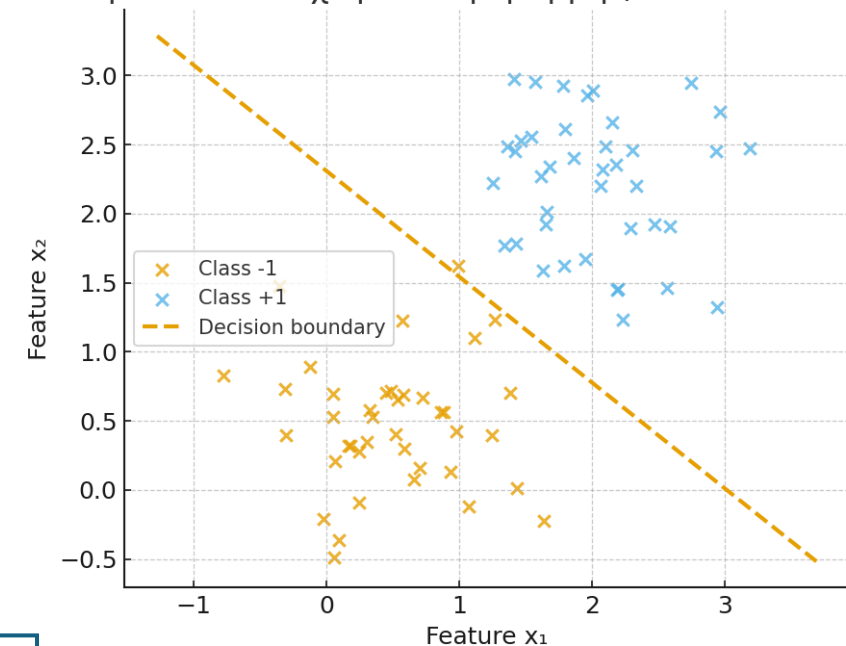
Μηχανική Μάθηση με Επίβλεψη

Στην *εποπτευόμενη μάθηση* δίνεται ένα σύνολο δεδομένων όπου κάθε παράδειγμα συνοδεύεται από τη **σωστή απάντηση** (ετικέτα).

Είδη προβλημάτων:

- **Παλινδρόμηση (Regression):** η έξοδος είναι συνεχής
 - (π.χ. πρόβλεψη τιμής σπιτιού, πρόβλεψη θερμοκρασίας)
- **Ταξινόμηση/κατηγοριοποίηση (Classification):** η έξοδος είναι διακριτή/μια κατηγορία
 - (π.χ. spam / όχι spam, θα βρέξει/δεν θα βρέξει)

Δεδομένα και Διαχωριστική Γραμμή (Decision Boundary)



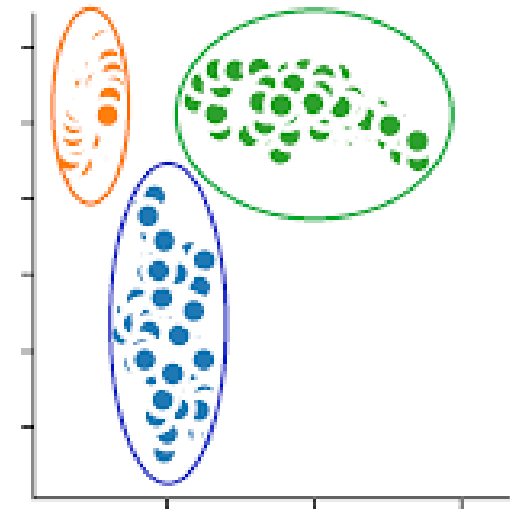
Σε αυτό το μάθημα θα επικεντρώσουμε στη μάθηση με επίβλεψη

Μηχανική Μάθηση χωρίς Επίβλεψη

Στην *μη εποπτευόμενη μάθηση*, τα δεδομένα δεν έχουν γνωστή ετικέτα εξόδου. Στόχος είναι να βρεθούν **κρυφά μοτίβα ή δομές** στα δεδομένα.

Είδη προβλημάτων:

- **Ομαδοποίηση** (Clustering): χωρισμός αντικειμένων σε ομάδες (π.χ. πελάτες με παρόμοιες συνήθειες).
- **Αναγωγή διάστασης** (Dimensionality Reduction): αναπαράσταση δεδομένων σε λιγότερες μεταβλητές, διατηρώντας τη βασική πληροφορία.



Ενισχυτική Μάθηση

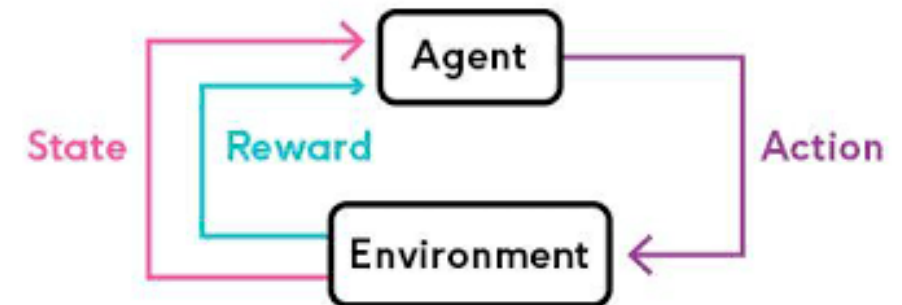
Η ενισχυτική μάθηση (*Reinforcement Learning*) αφορά πράκτορες που μαθαίνουν μέσω αλληλεπίδρασης με το περιβάλλον. Οι πράκτορες επιλέγουν ενέργειες, λαμβάνουν **ανταμοιβές** και προσαρμόζουν τη στρατηγική τους.

Κύρια στοιχεία:

- **Πράκτορας (Agent):** αυτός που μαθαίνει
- **Περιβάλλον (Environment):** με το οποίο αλληλεπιδρά
- **Κατάσταση (State), Ενέργεια (Action), Ανταμοιβή (Reward)**

Παράδειγμα:

Ένα ρομπότ μαθαίνει να περπατάει με συνεχή βελτίωση κινήσεων.



Εκπαίδευση και Αξιολόγηση

Η **εκπαίδευση (*training*)** είναι η διαδικασία εύρεσης των παραμέτρων ενός μοντέλου που ελαχιστοποιούν το σφάλμα στα δεδομένα εκπαίδευσης.

Η **αξιολόγηση (*testing*)** μετρά πόσο καλά το μοντέλο γενικεύει σε νέα δεδομένα.

Διαχωρισμός δεδομένων:

- **Training set:** χρησιμοποιείται για εκμάθηση του μοντέλου
- **Validation/Test set:** χρησιμοποιείται για αξιολόγηση

Παράδειγμα:

Διαχωρίζουμε τα δεδομένα 80% για εκπαίδευση και 20% για έλεγχο.

Βασικό μέτρο απόδοσης: Ακρίβεια

Υπερπροσαρμογή και Υποπροσαρμογή

Ιδανικό μοντέλο:

Πετυχαίνει ισορροπία μεταξύ απλότητας και απόδοσης.

Τα δεδομένα μας μπορεί να περιέχουν **θόρυβο (noise): λάθος δεδομένα που μπορεί να επηρεάσουν το μοντέλο που εκπαιδεύσαμε.**

Υπερπροσαρμογή (Overfitting):

Το μοντέλο μαθαίνει *πολύ καλά* τα δεδομένα εκπαίδευσης, αλλά **αποτυγχάνει σε νέα δεδομένα.**

- Προσαρμόζεται ακόμα και στον “θόρυβο” των δεδομένων.

Υποπροσαρμογή (Underfitting):

Το μοντέλο είναι πολύ απλό και **δεν συλλαμβάνει** τη δομή των δεδομένων.

- Συνήθως όταν δεν έχουμε αρκετά δεδομένα

Μέτρα απόδοσης μοντέλων: Παλινδρόμηση (Regression)

Στόχος: το μοντέλο να προβλέπει όσο το δυνατόν πιο κοντά στις πραγματικές τιμές.

Mean Absolute Error (MAE):

- Μέσο απόλυτο σφάλμα. Ευκολότερα ερμηνεύσιμο
- y_i πραγματική τιμή, $\hat{y}_i$ προβλεπόμενη τιμή

$$\text{MAE} = \frac{1}{n} \sum |y_i - \hat{y}_i|$$

Mean Squared Error (MSE):

- Υπερτονίζει μεγάλα σφάλματα. Χρησιμοποιείται πιο συχνά.
- y_i πραγματική τιμή, $\hat{y}_i$ προβλεπόμενη τιμή

$$\text{MSE} = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

Μέτρα απόδοσης μοντέλων: Κατηγοριοποίηση (Classification)

Στόχος: να προβλέπουμε όσο πιο ακριβές τις κατηγορίες

*Και άλλα που συνδυάζουν
διαφορετικές παραμέτρους
(precision, f1-score)*

Ακρίβεια (Accuracy):

$$\text{Accuracy} = \frac{\text{Σωστές προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

- Δεν φτάνει πάντα!

Recall (ή ευεσθησία):

- Από τα πραγματικά θετικά, πόσα αναγνώρισε σωστά το μοντέλο;
- Σημαντικό για κατηγορίες με πιο μεγάλη σημασία (π.χ. ασθενείς σε ιατρικά δεδομένα)
- TP: True positives (σωστά κατηγοριοποιημένα στην κάθε κατηγορία)
- FN: False negatives (στοιχεία της κατηγορίας που δεν εντοπίστηκαν)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Δεν μας νοιάζουν τα False positives

Βασικές μέθοδοι μηχανικής μάθησης

Βασικές μέθοδοι

1. Επιβλεπόμενη Μάθηση (Supervised Learning)

- **Παλινδρόμηση (Regression):** Linear Regression, Support Vector Regression, Neural Networks
- **Κατηγοριοποίηση (Classification):** Logistic Regression, Decision Trees, Random Forests, k-Nearest Neighbors, Support Vector Machines, Neural Networks

2. Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

- **Ομαδοποίηση (Clustering):** k-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models
- **Μείωση Διαστάσεων (Dimensionality Reduction):** PCA, t-SNE, UMAP, Autoencoders, Neural Networks

3. Ενισχυτική Μάθηση (Reinforcement Learning)

- Q-Learning, SARSA, Deep Q-Networks (DQN), Policy Gradients, Actor-Critic

Παλινδρόμηση (regression)

Γραμμική Παλινδρόμηση (Linear Regression)

Η γραμμική παλινδρόμηση είναι ένα από τα απλούστερα και πιο θεμελιώδη μοντέλα στη μηχανική μάθηση.

- Χρησιμοποιείται για να **προβλέψουμε μια συνεχόμενη μεταβλητή y** με βάση μία ή περισσότερες εισόδους $x_1, x_2, \dots, x_p$.

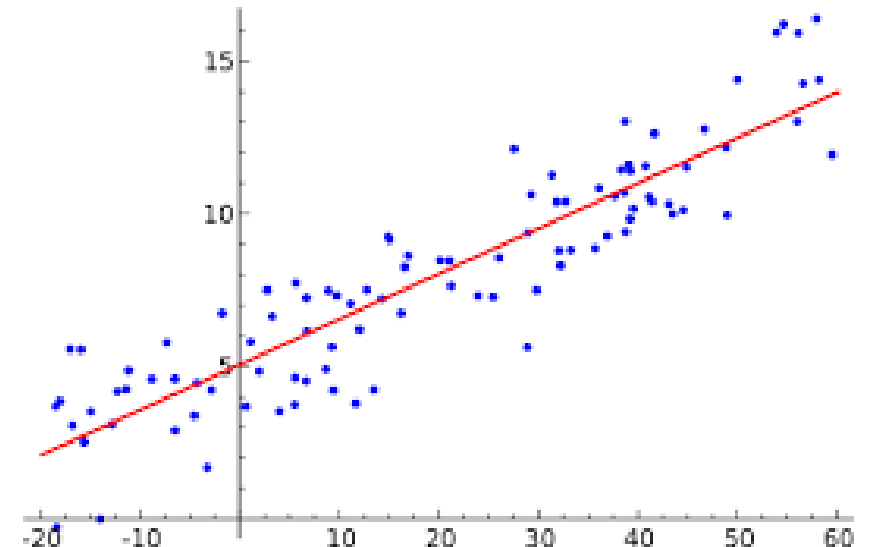
Μαθηματικά:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

όπου:

- $\hat{y}$: η προβλεπόμενη τιμή,
- $\beta_0, \beta_1, \dots, \beta_p$: οι παράμετροι (συντελεστές) που μαθαίνονται από τα δεδομένα.

Το μοντέλο προσπαθεί να “τραβήξει” μια ευθεία (ή υπερεπίπεδο) που προσεγγίζει όσο το δυνατόν καλύτερα τα δεδομένα



Γραμμική Παλινδρόμηση Πώς “μαθαίνει” το μοντέλο

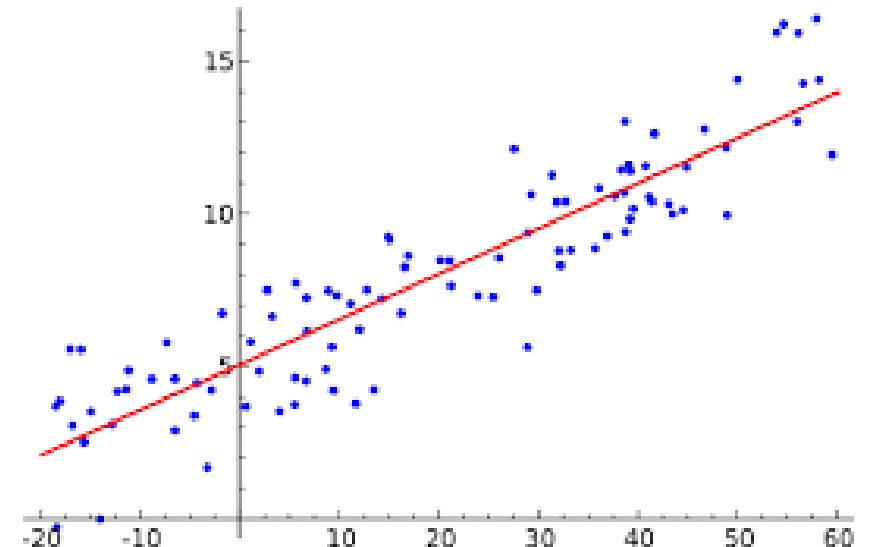
Στη στατιστική, “εκτιμούμε” τους συντελεστές β_i έτσι ώστε η ευθεία να **ταιριάζει καλύτερα** στα δεδομένα.

- Στη μηχανική μάθηση, το ίδιο πράγμα λέγεται “εκπαίδευση” του μοντέλου.
- **Κριτήριο εκμάθησης: Ελαχιστοποίηση σφάλματος**
- Μετράμε το **σφάλμα πρόβλεψης** (διαφορά μεταξύ πραγματικής και προβλεπόμενης τιμής):

Μετράμε το **σφάλμα πρόβλεψης** (διαφορά μεταξύ πραγματικής και προβλεπόμενης τιμής):

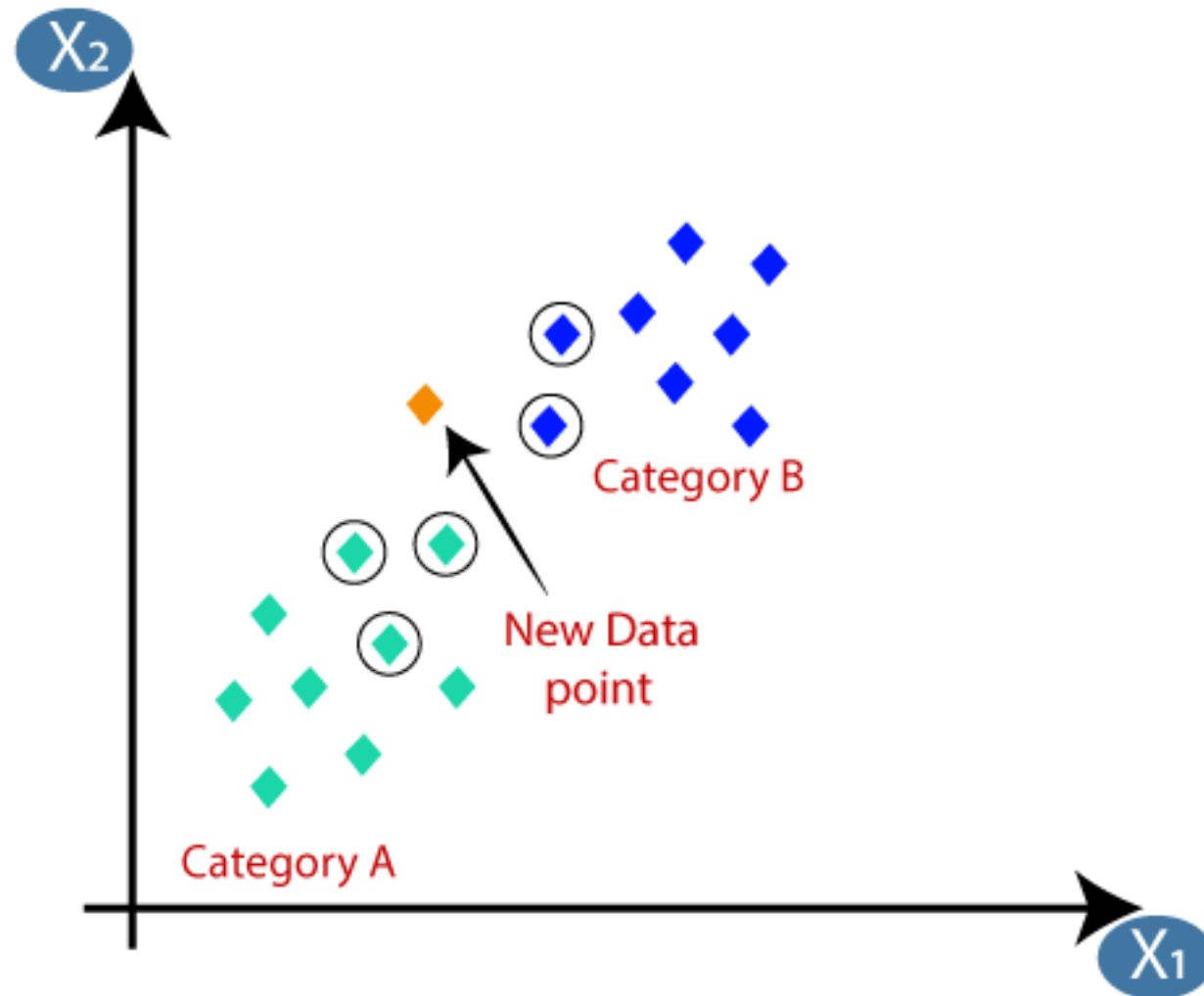
$$E = \sum (y_i - \hat{y}_i)^2$$

Το μοντέλο μαθαίνει επιλέγοντας τις τιμές των β_i που **ελαχιστοποιούν** αυτό το άθροισμα.



Κατηγοριοποίηση (regression)

Κατηγοριοποίηση



*Που θα κατηγοριοποιούσατε το
καινούριο δεδομένο;*

Κατηγοριοποίηση με πλησιέστερους γείτονες

Ο αλγόριθμος k -πλησιέστερων γειτόνων (kNN) κατηγοριοποιεί ένα νέο παράδειγμα με βάση τις **k πιο κοντινές** παρατηρήσεις στο χώρο χαρακτηριστικών.

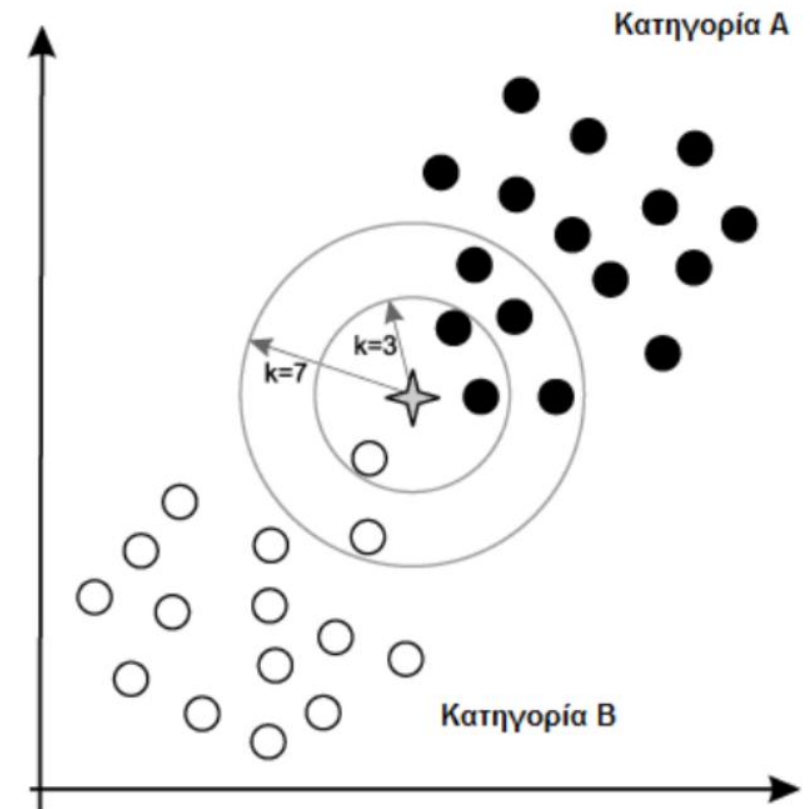
“Πες μου ποιοι είναι οι γείτονές σου και θα σου πω ποιος είσαι.”

Βήματα:

1. Υπολογίζουμε την απόσταση (π.χ. Ευκλείδεια) μεταξύ του νέου σημείου και όλων των υπαρχόντων.
2. Επιλέγουμε τους k πιο κοντινούς.
3. Η κατηγορία που εμφανίζεται πιο συχνά στους γείτονες $\rightarrow$ πρόβλεψη.

Πλεονεκτήματα: απλός, χωρίς εκπαίδευση.

Μειονεκτήματα: αργός σε μεγάλα δεδομένα, ευαίσθητος στην επιλογή k



Naive Bayes

- Ο *Naive Bayes* είναι ένας πιθανοτικός ταξινομητής που βασίζεται στο **Θεώρημα του Bayes** και υποθέτει **ανεξαρτησία** μεταξύ των χαρακτηριστικών.

Ιδέα: Για κάθε κατηγορία, υπολογίζουμε την πιθανότητα τα δεδομένα να ανήκουν σε αυτήν και επιλέγουμε τη μεγαλύτερη.

Υπολογίζουμε την πιθανότητα με βάση το θεώρημα του Bayes:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

όπου C είναι η κατηγορία και $X = (x_1, x_2, \dots, x_n)$ τα χαρακτηριστικά.

Παράδειγμα: Ανάλυση email: πιθανότητα λέξεων “free”, “offer”, “win” όταν είναι spam.

Πλεονεκτήματα:

- Πολύ γρήγορος
- Απλός και αποδοτικός σε προβλήματα κειμένου

Μειονεκτήματα:

- Η υπόθεση ανεξαρτησίας συχνά δεν ισχύει πλήρως

Παράδειγμα

No	outlook	temperatur e	humidit y	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Naive Bayes

- Ο *Naive Bayes* είναι ένας πιθανοτικός ταξινομητής που βασίζεται στο **Θεώρημα του Bayes** και υποθέτει **ανεξαρτησία** μεταξύ των χαρακτηριστικών.

Ιδέα: Για κάθε κατηγορία, υπολογίζουμε την πιθανότητα τα δεδομένα να ανήκουν σε αυτήν και επιλέγουμε τη μεγαλύτερη.

Πιο αναλυτικά σε επόμενη διάλεξη

Frequency Table				Likelihood Table			
Outlook	Play Golf			Outlook	Play Golf		
		Yes	No		Yes	No	
	Sunny	3	2		3/9	2/5	
	Overcast	4	0		4/9	0/5	
Humidity	Play Golf			Humidity	Play Golf		
		Yes	No		Yes	No	
	High	3	4		3/9	4/5	
	Normal	6	1		6/9	1/5	
Temp.	Play Golf			Temp.	Play Golf		
		Yes	No		Yes	No	
	Hot	2	2		2/9	2/5	
	Mild	4	2		4/9	2/5	
Windy	Play Golf			Windy	Play Golf		
		Yes	No		Yes	No	
	False	6	2		6/9	2/5	
	True	3	3		3/9	3/5	

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes | X) = P(Rainy | Yes) \times P(Cool | Yes) \times P(High | Yes) \times P(True | Yes) \times P(Yes)$$

$$P(Yes | X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No | X) = P(Rainy | No) \times P(Cool | No) \times P(High | No) \times P(True | No) \times P(No)$$

$$P(No | X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

Δέντρα Απόφασης (Decision Trees)

Τα δέντρα απόφασης ταξινομούν δεδομένα μέσω **διαδοχικών ερωτήσεων** που χωρίζουν το χώρο χαρακτηριστικών.

- Κάθε εσωτερικός κόμβος είναι ένα τεστ (π.χ. “Εισόδημα $> 30k$;”),
- Κάθε φύλλο αντιστοιχεί σε μια κατηγορία.

Βασική ιδέα: Διαίρει και βασίλευε

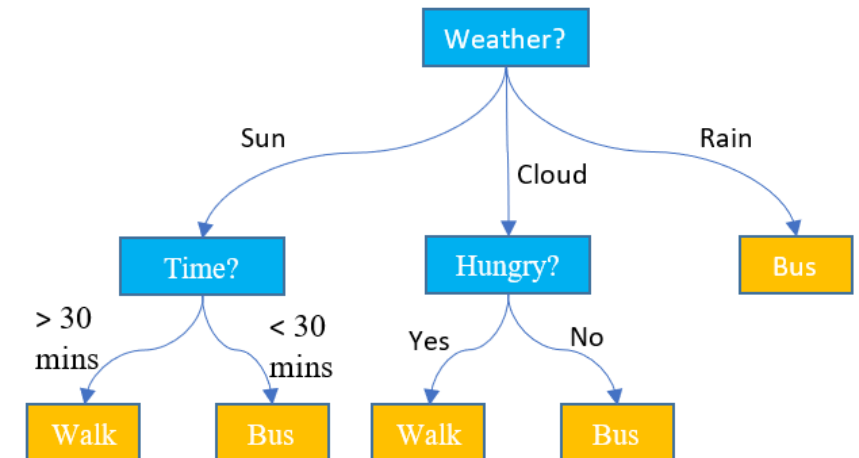
- κάθε ερώτηση διαχωρίζει τα δεδομένα για να πάρει περισσότερη πληροφορία

Πλεονεκτήματα:

- Εύκολη ερμηνεία (κανόνες if–then)
- Αντιμετωπίζει αριθμητικά και κατηγορικά δεδομένα

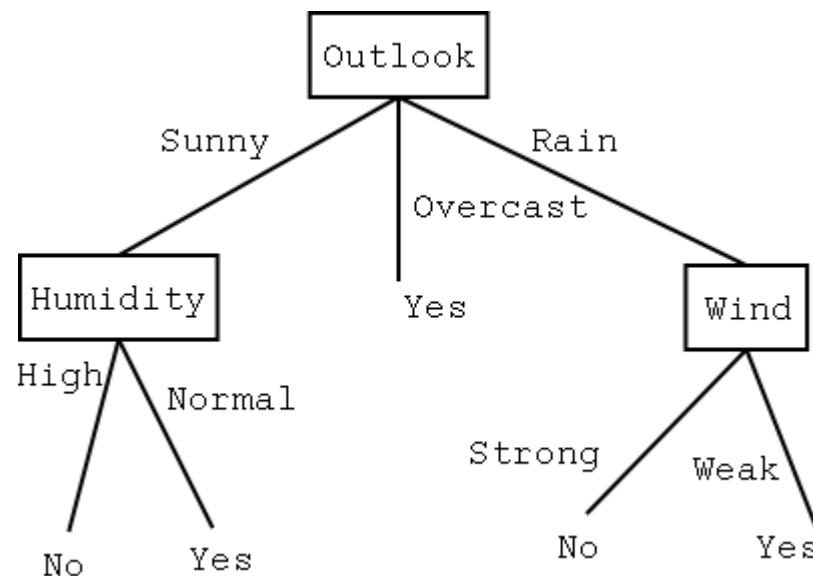
Μειονεκτήματα:

- Τάση για υπερπροσαρμογή
- Ευαισθησία σε μικρές αλλαγές στα δεδομένα



Πώς αναπαρίσταται ένα δέντρο απόφασης;

- ✓ Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού
- ✓ Κάθε κλαδί/σύνδεση ονοματίζεται με ένα κατηγορημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου - πατέρα
- ✓ Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης
- ✓ Χρήση της τεχνικής «διαίρει και βασίλευε» για διαίρεση του χώρου αναζήτησης σε υποσύνολα



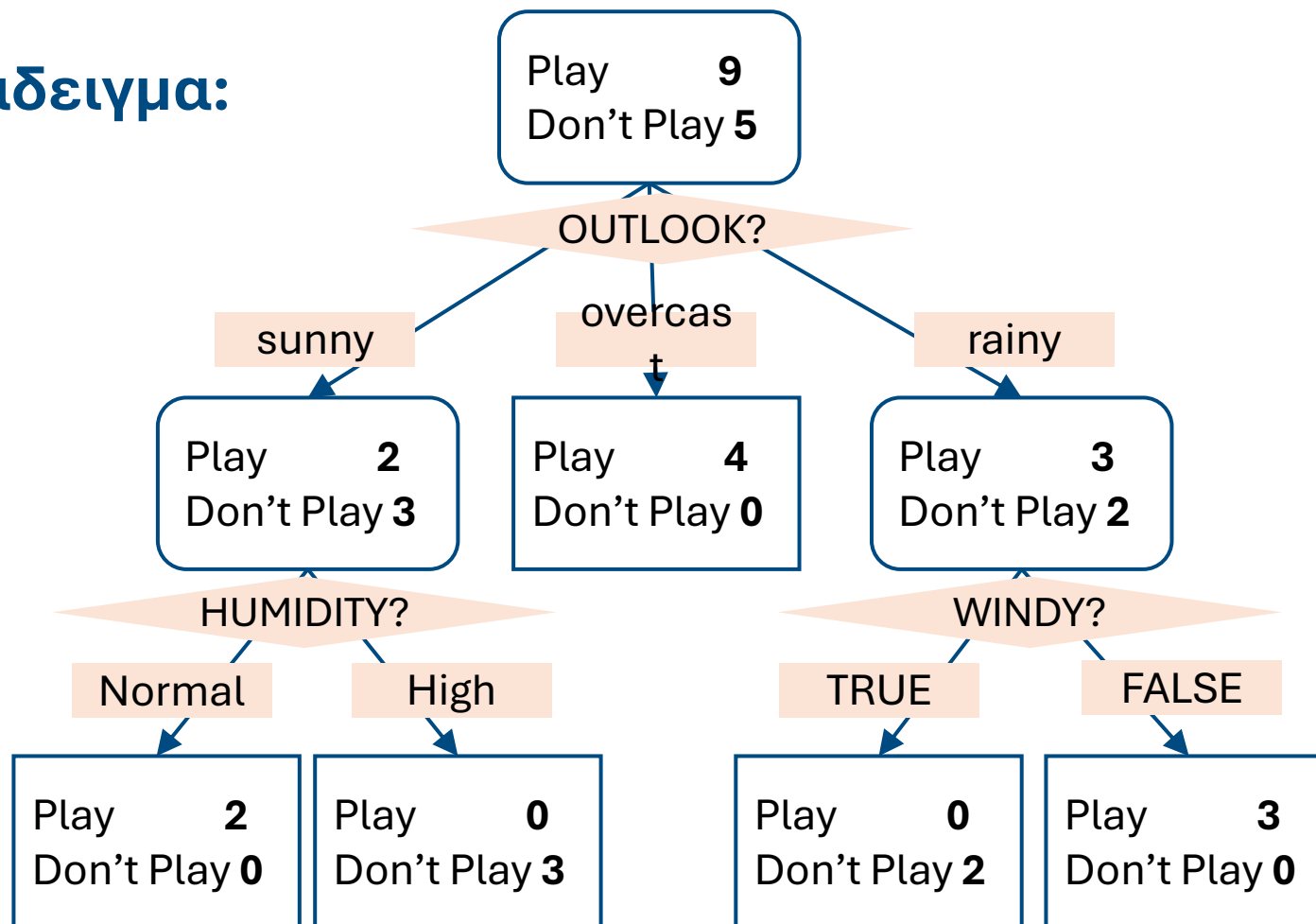
Παράδειγμα για δημιουργία δέντρου απόφασης

No	outlook	temperatur e	humidit y	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Διαχωρισμός δεδομένων

Depended Variable: **PLAY**

Παράδειγμα:



Δέντρα Απόφασης

- Η επαγωγή με **δέντρα απόφασης** (decision trees) είναι μια από τις απλούστερες και πιο πετυχημένες μορφές αλγορίθμων μάθησης

Πως κατασκευάζουμε το δέντρο?

- Αν υπάρχουν παραδείγματα διαφορετικών κατηγοριών, διάλεξε το καλύτερο χαρακτηριστικό για να τα ξεχωρίσεις
- Αν όλα τα εναπομείναντα παραδείγματα ανήκουν σε μία κατηγορία, τέλος
- Αν δεν έχουν μείνει χαρακτηριστικά αλλά έχουν μείνει θετικά και αρνητικά παραδείγματα, έχουμε πρόβλημα!
 - **θόρυβος** (noise) στα δεδομένα ή ανεπαρκή χαρακτηριστικά

Δέντρα Απόφασης

«Αν υπάρχουν θετικά και αρνητικά παραδείγματα, διάλεξε το καλύτερο χαρακτηριστικό για να τα ξεχωρίσεις»

ΠΩΣ;

Άτομο	Έχει Instagram;	Καφέδες/Ημέρα	Χρησιμοποιεί TikTok;
1	Ναι	1	Ναι
2	Όχι	3	Όχι
3	Ναι	2	Ναι
4	Όχι	0	Όχι

Δέντρα Απόφασης

«Αν υπάρχουν θετικά και αρνητικά παραδείγματα, διάλεξε το καλύτερο χαρακτηριστικό για να τα ξεχωρίσεις»

ΠΩΣ;

- Ένας από τους πιο διαδεδομένους μηχανισμούς διαχωρισμού είναι αυτός της **εντροπίας της πληροφορίας** (information entropy)
- Η εντροπία εν γένει εκφράζει το **μέγεθος της ανομοιογένειας** σε ένα σύνολο δεδομένων.
 - Π.χ. αν όλα τα δεδομένα ανήκουν σε μια κατηγορία, τότε δεν υπάρχει ανομοιογένεια: η εντροπία είναι μηδέν.

Δέντρα Απόφασης

Η εντροπία εν γένει εκφράζει το **μέγεθος της ανομοιογένειας** σε ένα σύνολο δεδομένων.

Τύπος (Shannon Entropy):

$$E(D) = - \sum_{i=1}^c p_i \log_2 p_i$$

- Όπου:
 - D: το σύνολο δεδομένων, E(D) η εντροπία του
 - c ο αριθμός κατηγοριών/κλάσεων
 - p_i η πιθανότητα της κλάσης i

Ποια είναι η εντροπία του «Ναι» και του «Όχι» στο προηγούμενο παράδειγμα;

Δέντρα Απόφασης

Το **Κέρδος Πληροφορίας (IG)** μετρά πόση **αβεβαιότητα μειώνεται** όταν χωρίζουμε το σύνολο δεδομένων με ένα χαρακτηριστικό.

- **Τύπος:**

$$IG(D, F) = E(D) - \sum_{v \in Vals(F)} \frac{|D_v|}{|D|} E(D_v)$$

- **Όπου:**

- D: το σύνολο δεδομένων,
- F: το χαρακτηριστικό (feature)
- Vals: Οι διαφορετικές τιμές του χαρακτηριστικού
- D_v το υποσύνολο του D με τιμή v στο χαρακτηριστικό F
- |D|: το μέγεθος του D

Μέση
εντροπία
παραγόμενων
υποσυνόλων

Επιλογή του χαρακτηριστικού F με το μεγαλύτερο κέρδος πληροφορίας!

Δέντρα Απόφασης

Παράδειγμα σε μικρό σύνολο δεδομένων

Πρώτα υπολογίζουμε την εντροπία του συνόλου δεδομένων:

$$\begin{aligned} E(D) &= -(p_{yes} * \log_2 p_{yes} + p_{no} * \log_2 p_{no}) \\ &= -\left(\frac{9}{14} * \log_2 \frac{9}{14} + \frac{5}{14} * \log_2 \frac{5}{14}\right) \\ &= -(0.643 * (-0.637) + 0.357 * (-1.485)) \\ &= -(-0.41 + (-0.53)) = 0.94 \end{aligned}$$

Η εντροπία είναι:

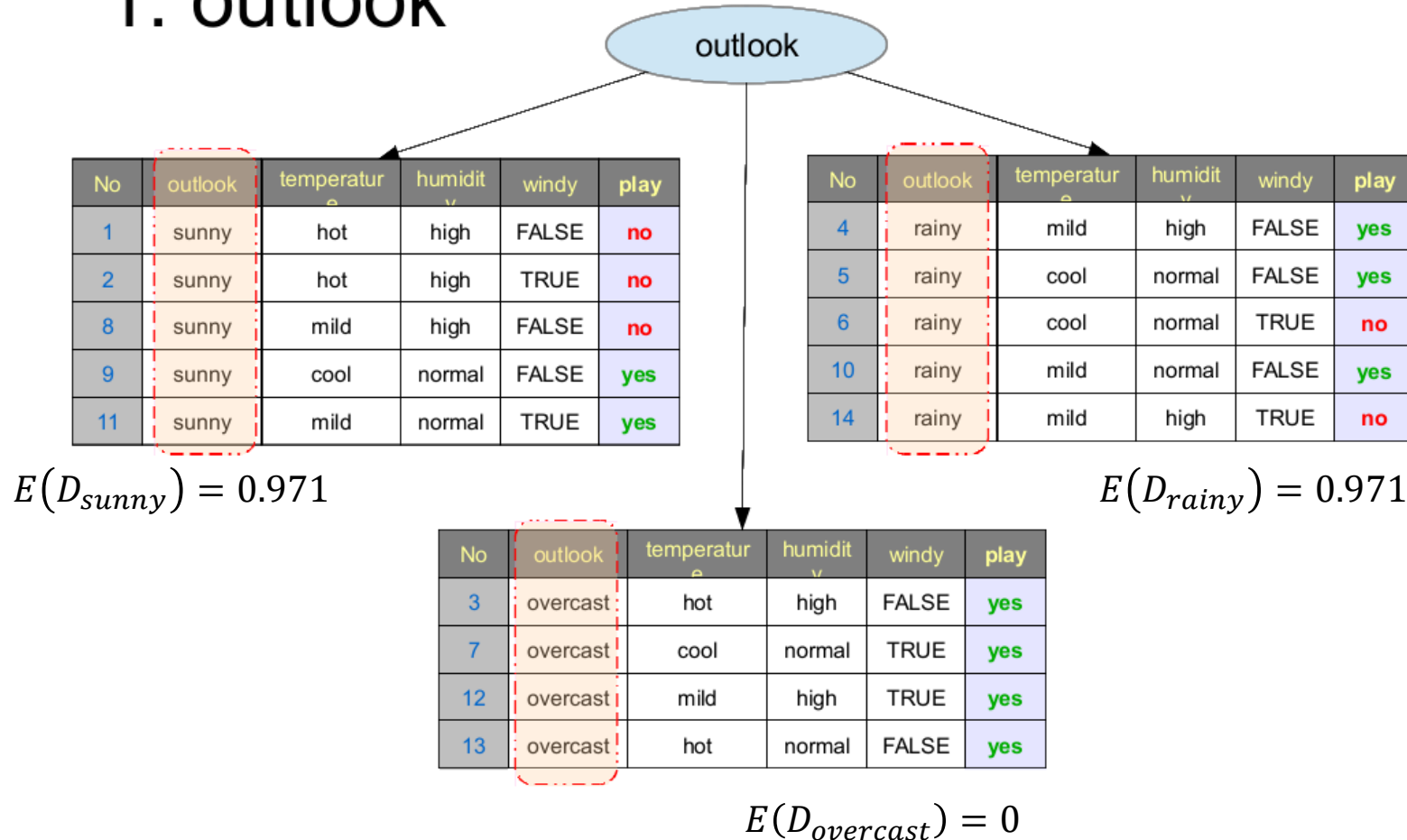
$$E(D) = 0.94$$

Για να επιλέξουμε το καλύτερο χαρακτηριστικό διαχωρισμού, πρέπει να υπολογίσουμε το κέρδος πληροφορίας κάθε χαρακτηριστικού

No	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Δέντρα Απόφασης

1. outlook



Διαδικασία:

- **Διαχωρίζουμε** το σύνολο δεδομένων με βάση τις διαφορετικές τιμές του χαρακτηριστικού,
- **Υπολογίζουμε την εντροπία** του κάθε υποσυνόλου
- **Υπολογίζουμε το κέρδος πληροφορίας**

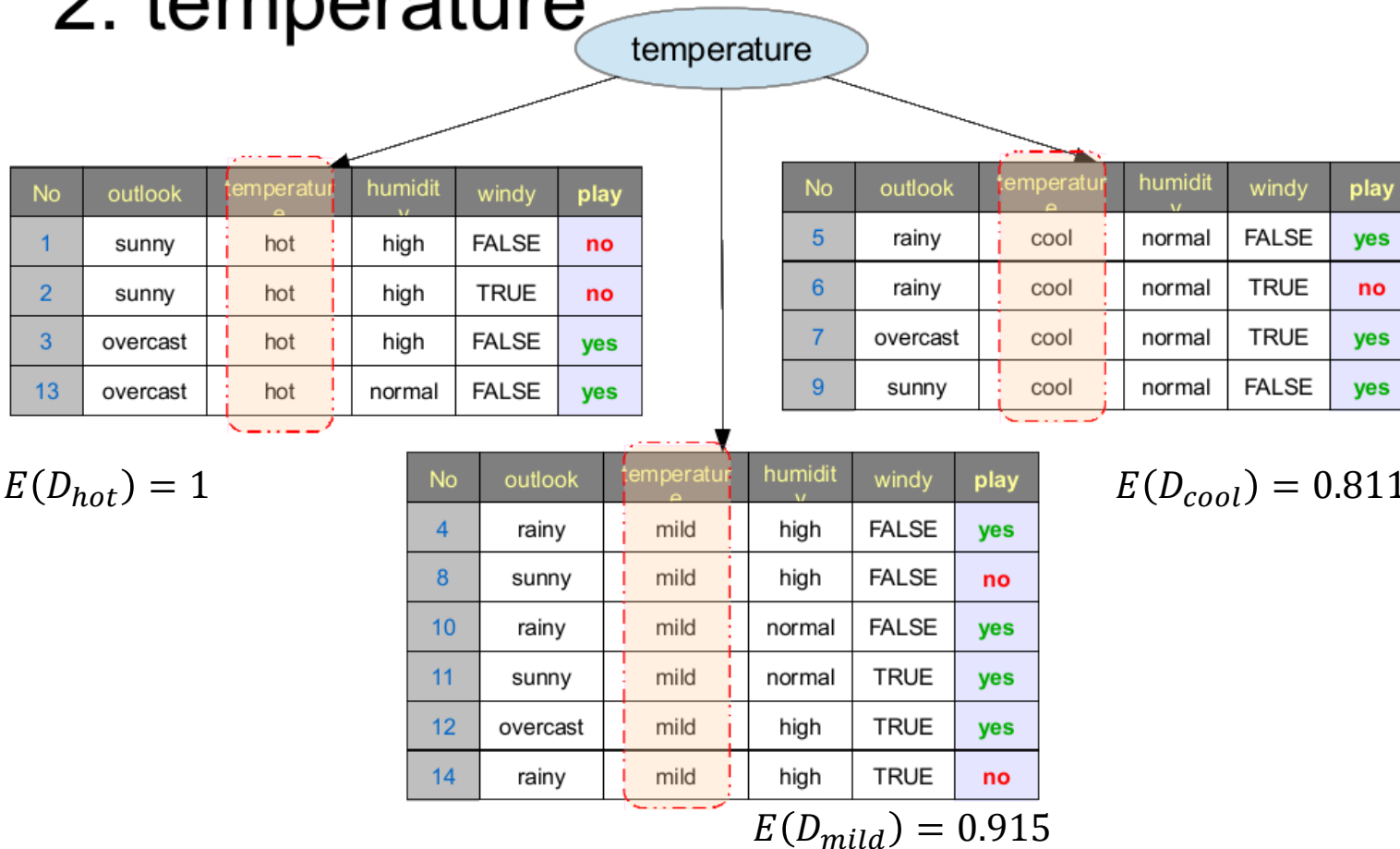
Κέρδος πληροφορίας:

$$\begin{aligned}
 IG(D, outlook) &= E(D) - \sum_{v \in \text{Vals}(\text{outlook})} \frac{|D_v|}{|D|} E(D_v) \\
 &= 0.94 - \left(\frac{5}{14} * 0.971 + \frac{5}{14} * 0.971 + \frac{4}{14} * 0 \right) \\
 &= 0.94 - (0.347 + 0.347 + 0) = 0.246
 \end{aligned}$$

$$IG(D, outlook) = 0.246$$

Δέντρα Απόφασης

2. temperature



Διαδικασία:

- **Διαχωρίζουμε** το σύνολο δεδομένων με βάση τις διαφορετικές τιμές του χαρακτηριστικού,
- **Υπολογίζουμε την εντροπία** του κάθε υποσυνόλου
- **Υπολογίζουμε το κέρδος πληροφορίας**

Κέρδος πληροφορίας:

$$IG(D, temp) = E(D) - \sum_{v \in Vals(temp)} \frac{|D_v|}{|D|} E(D_v)$$

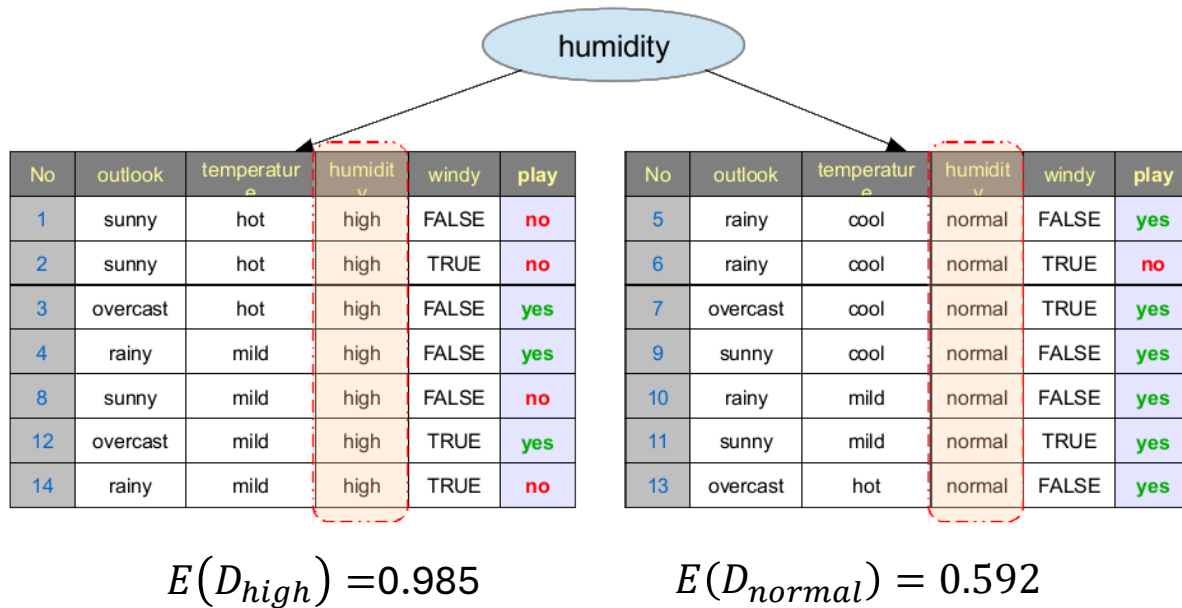
$$= 0.94 - \left(\frac{4}{14} * 1 + \frac{4}{14} * 0.811 + \frac{6}{14} * 0.915 \right)$$

$$= 0.94 - (0.286 + 0.232 + 0.393) = 0.029$$

$$IG(D, temp) = 0.029$$

Δέντρα Απόφασης

3. humidity



Διαδικασία:

- **Διαχωρίζουμε** το σύνολο δεδομένων με βάση τις διαφορετικές τιμές του χαρακτηριστικού,
- **Υπολογίζουμε την εντροπία** του κάθε υποσυνόλου
- **Υπολογίζουμε το κέρδος πληροφορίας**

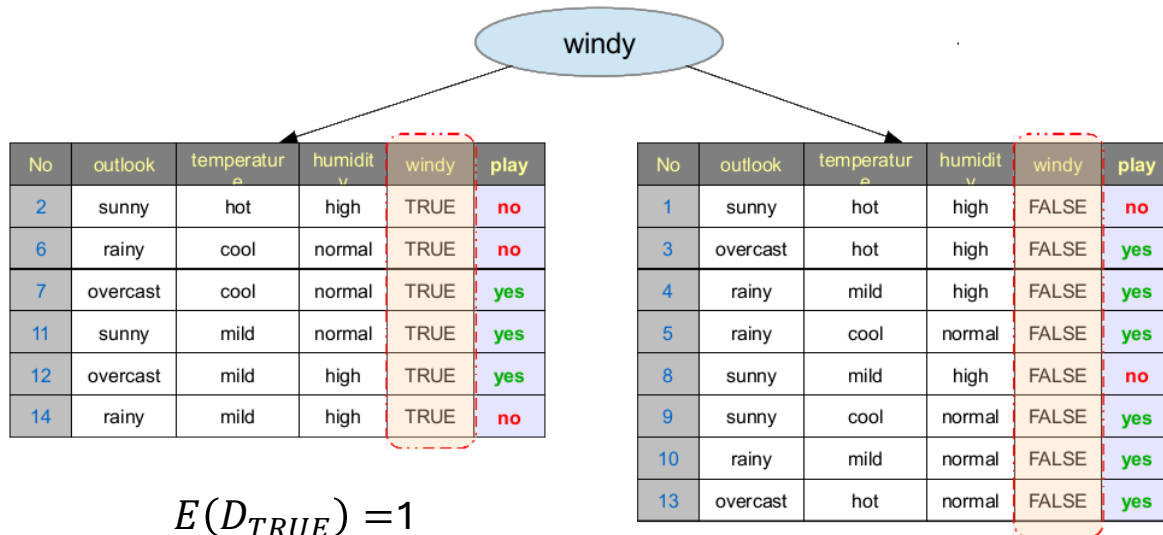
Κέρδος πληροφορίας:

$$\begin{aligned}
 IG(D, humidity) &= E(D) - \sum_{v \in Vals(humidity)} \frac{|D_v|}{|D|} E(D_v) \\
 &= 0.94 - \left(\frac{7}{14} * 0.985 + \frac{7}{14} * 0.592 \right) \\
 &= 0.94 - (0.493 + 0.296) = 0.151
 \end{aligned}$$

$$IG(D, humidity) = 0.151$$

Δέντρα Απόφασης

4. windy



$$E(D_{FALSE}) = 0.811$$

Διαδικασία:

- **Διαχωρίζουμε** το σύνολο δεδομένων με βάση τις διαφορετικές τιμές του χαρακτηριστικού,
- **Υπολογίζουμε την εντροπία** του κάθε υποσυνόλου
- **Υπολογίζουμε το κέρδος πληροφορίας**

Κέρδος πληροφορίας:

$$\begin{aligned}
 IG(D, windy) &= E(D) - \sum_{v \in Vals(windy)} \frac{|D_v|}{|D|} E(D_v) \\
 &= 0.94 - \left(\frac{6}{14} * 1 + \frac{8}{14} * 0.811 \right) \\
 &= 0.94 - (0.429 + 0.463) = 0.048
 \end{aligned}$$

$$IG(D, windy) = 0.048$$

Δέντρα Απόφασης

Παράδειγμα σε μικρό σύνολο δεδομένων

Υπολογίσαμε το κέρδος πληροφορίας για κάθε χαρακτηριστικό:

- $IG(D, outlook) = 0.246$
- $IG(D, temp) = 0.029$
- $IG(D, humidity) = 0.151$
- $IG(D, windy) = 0.048$

Επιλέγουμε ως χαρακτηριστικό διαχωρισμού το χαρακτηριστικό **outlook**, επειδή έχει το μεγαλύτερο κέρδος πληροφορίας!

No	outlook	temperature	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no