Machine Learning & Knowledge Extraction

DR KONSTANTINOS KARAMPIDIS

Πληροφορίες Μαθήματος

- Ωράριο:
  - Θεωρία: **Κάθε Τρίτη 09:00-13:00 – Αίθουσα 207**
  - Εργαστήριο: **13:00-14:00 ΕΡΓ6 (σύμφωνα με το πρόγραμμα που είναι αναρτημένο στο eclass)**
- Εργασίες
  - **1 project – Ομάδες έως 2 ατόμων – 80%**
  - **Εργαστηριακές ασκήσεις – 20%**
- Προαπαιτούμενα: Κανένα

Περιεχόμενο Μαθήματος

- **Εισαγωγή στη Μηχανική Μάθηση - τι είναι, γιατί μας ενδιαφέρει, παραδείγματα προβλημάτων, η μηχανική μάθηση ως αναζήτηση, υπόθεση επαγωγικής μάθησης**
- Επεξεργασία εισόδου – Μείωση διαστατικότητας- Αξιόλογηση
- Μέθοδοι επιβλεπόμενης μάθησης
- Νευρωνικά Δίκτυα
- Εξελικτική Μάθηση – Γενετικοί Αλγόριθμοι
- Μέθοδοι μη επιβλεπόμενης μάθησης
- Βαθιά Μάθηση

Εισαγωγή στη Μηχανική Μάθηση

Μάθηση

Η διαδικασία βελτίωσης της επίδοσης ενός συστήματος σε μια συγκεκριμένη εργασία μετά την παρατήρηση πολλών παραδειγμάτων

Εισαγωγή στη Μηχανική Μάθηση

Για να υπάρξει μάθηση απαιτούνται τρία βασικά συστατικά:

- Ένα περιβάλλον το οποίο να προσφέρει δεδομένα υπό μορφή παραδειγμάτων στο σύστημα
- Ένα κριτήριο αξιολόγησης της επίδοσης του συστήματος
- Μια συγκεκριμένη εργασία την οποία το σύστημα καλείται να εκτελέσει.

Εισαγωγή στη Μηχανική Μάθηση

- Η μάθηση σε ένα γνωστικό σύστημα, όπως γίνεται αντιληπτή στην καθημερινή ζωή, μπορεί να συνδεθεί με δύο βασικές ιδιότητες:
  - την ικανότητα στην πρόσκτηση γνώσης κατά την αλληλεπίδρασή του με το περιβάλλον,
  - την ικανότητα να βελτιώνει με την επανάληψη τον τρόπο εκτέλεσης μία ενέργειας.
- Ευκαιρίες: Η κοινωνία παρέχει μεγάλα πλήθη από δεδομένα από πηγές όπως:
  - επιχειρήσεις, επιστήμη, ιατρική, οικονομία, γεωγραφία, περιβάλλον, αθλήματα, κλπ.

Τα δεδομένα αυτά είναι πολύτιμα όμως είναι πολλές φορές χαμηλού επιπέδου και άρα μη εύκολα εκμεταλλεύσιμα.

Ορισμός Μηχανικής Μάθησης

- Ο άνθρωπος προσπαθεί να κατανοήσει το περιβάλλον του παρατηρώντας το και δημιουργώντας μια απλοποιημένη (αφαιρετική) εκδοχή του που ονομάζεται μοντέλο (model).
  - Η δημιουργία ενός τέτοιου μοντέλου, ονομάζεται **επαγωγική μάθηση** (inductive learning)
  - ενώ η διαδικασία γενικότερα ονομάζεται **επαγωγή** (induction).
- Επιπλέον ο άνθρωπος έχει τη δυνατότητα να οργανώνει και να συσχετίζει τις εμπειρίες και τις παραστάσεις του δημιουργώντας νέες δομές που ονομάζονται πρότυπα (patterns).
- Η δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων, από ένα υπολογιστικό σύστημα, ονομάζεται **μηχανική μάθηση** (machine learning).

Ορισμός Μηχανικής Μάθησης

Η Μηχανική Μάθηση είναι ένας τομέας της Τεχνητής Νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων μάθησης, δηλαδή αλγορίθμων που βελτιώνουν την επίδοση ενός συστήματος σε διάφορα προβλήματα.

Μηχανική Μάθηση - Άλλοι ορισμοί

- Η μελέτη υπολογιστικών μεθόδων για την απόκτηση νέας γνώσης , νέων δεξιοτήτων και νέων τρόπων οργάνωσης της υπάρχουσας γνώσης (Carbonell, 1987).
- Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία (E) ως προς μια κλάση εργασιών (T) και ένα μέτρο επίδοσης (P), αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E (Mitchell, 1997).
- Κάτι μαθαίνει όταν αλλάζει τη συμπεριφορά του κατά τέτοιο τρόπο ώστε να αποδίδει καλύτερα στο μέλλον (Witten & Frank, 2000).

Μηχανική Μάθηση - Παραδείγματα

Αναγνώριση αντικειμένων πχ. είδη ρουχισμού

Παρ' όλο που τα αντικείμενα αυτά έχουν μεγάλη ποικιλία σχημάτων, χρωμάτων και μεγεθών, ένα σύστημα μπορεί να τα μάθει και να τα αναγνωρίζει μέσω της παρατήρησης και της ανάλυσης μεγάλου πλήθους δειγμάτων από κάθε κατηγορία.

Μηχανική Μάθηση - Παραδείγματα

Πρόβλεψη μια τιμής ή αξίας

Πρόβλεψη θερμοκρασίας όταν είναι γνωστές κάποιες συνθήκες (πίεση, υγρασία κλπ.)

Συσταδοποίηση ομοειδών αντικειμένων

Ομαδοποίηση γραπτών κειμένων με βάση κάποιο κριτήριο ομοιότητας πχ. το ποσοστό λέξεων που είναι κοινές.

Μηχανική Μάθηση - Παραδείγματα

Ανάλυση δεδομένων

Εύρεση κρυφών παραγόντων που επηρεάζουν τις τιμές των μετοχών στο χρηματιστήριο, παρατηρώντας και αναλύοντας μεγάλο πλήθος διακυμάνσεων των τιμών για μεγάλο χρονικό διάστημα.

Ανάπτυξη στρατηγικών

Παιχνίδια ή προβλήματα ρομποτικής.

Μηχανική Μάθηση - Παραδείγματα

Ανάλυση δεδομένων

Εύρεση κρυφών παραγόντων που επηρεάζουν τις τιμές των μετοχών στο χρηματιστήριο, παρατηρώντας και αναλύοντας μεγάλο πλήθος διακυμάνσεων των τιμών για μεγάλο χρονικό διάστημα.

Ανάπτυξη στρατηγικών

Παιχνίδια ή προβλήματα ρομποτικής.

Μηχανική Μάθηση - Στόχος

Στόχος της μηχανικής μάθησης είναι η δυνατότητα παραγωγής σωστών εκτιμήσεων σχετικά με δεδομένα τα οποία αντιμετωπίζονται πρώτη φορά από το σύστημα.

Μηχανική Μάθηση – Τύποι

- Μάθηση με επίβλεψη (supervised learning) ή μάθηση με παραδείγματα (learning from examples),
- Μάθηση χωρίς επίβλεψη (unsupervised learning) ή μάθηση από παρατήρηση (learning from observation).
- Ενισχυτική μάθηση (reinforcement learning)

  Μάθηση μέσω ενίσχυσης (επιβράβευσης)

Μηχανική Μάθηση – Τύποι

- Στη μάθηση με επίβλεψη το σύστημα καλείται να "μάθει" μια έννοια ή συνάρτηση από ένα σύνολο δεδομένων, η οποία αποτελεί περιγραφή ενός μοντέλου.

- Στη μάθηση χωρίς επίβλεψη το σύστημα πρέπει μόνο του να ανακαλύψει συσχετίσεις ή ομάδες σε ένα σύνολο δεδομένων, δημιουργώντας πρότυπα, χωρίς να είναι γνωστό αν υπάρχουν, πόσα και ποια είναι.

Μάθηση με Επίβλεψη

- Στη μάθηση με επίβλεψη το σύστημα πρέπει να "μάθει" επαγωγικά μια συνάρτηση που ονομάζεται συνάρτηση στόχος (target function) και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα.
- Η συνάρτηση στόχος χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής, που ονομάζεται εξαρτημένη μεταβλητή ή μεταβλητή εξόδου, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται ανεξάρτητες μεταβλητές ή μεταβλητές εισόδου ή χαρακτηριστικά.

Μπορούμε να δούμε τη διαδικασία μάθησης γενικά ως τη διαδικασία μάθησης της αναπαράστασης μιας συνάρτησης.

Μάθηση με Επίβλεψη

- Η επαγωγική μάθηση (inductive learning) στηρίζεται στην "υπόθεση επαγωγικής μάθησης" (inductive learning hypothesis), δηλ. είναι η μάθηση μιας έννοιας μέσω ενός συνόλου παραδειγμάτων.

Σύμφωνα με την οποία:

- Κάθε υπόθεση h που προσεγγίζει καλά τη συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο παραδειγμάτων, θα προσεγγίζει το ίδιο καλά τη συνάρτηση στόχο και για περιπτώσεις που δεν έχει εξετάσει.

Προβλήματα στην επαγωγική μάθηση

- Πώς προσδιορίζουμε πότε μια υπόθεση είναι καλή;

- Μια καλή υπόθεση πρέπει να **γενικεύεται** σωστά, δηλαδή να προβλέπει ορθά παραδείγματα που δεν έχουν εξεταστεί.

- Πώς επιλέγουμε ανάμεσα σε πολλές, συνεπείς με τα παραδείγματα, υποθέσεις;

– Προτιμήστε την απλούστερη υπόθεση που συμφωνεί με τα παραδείγματα.

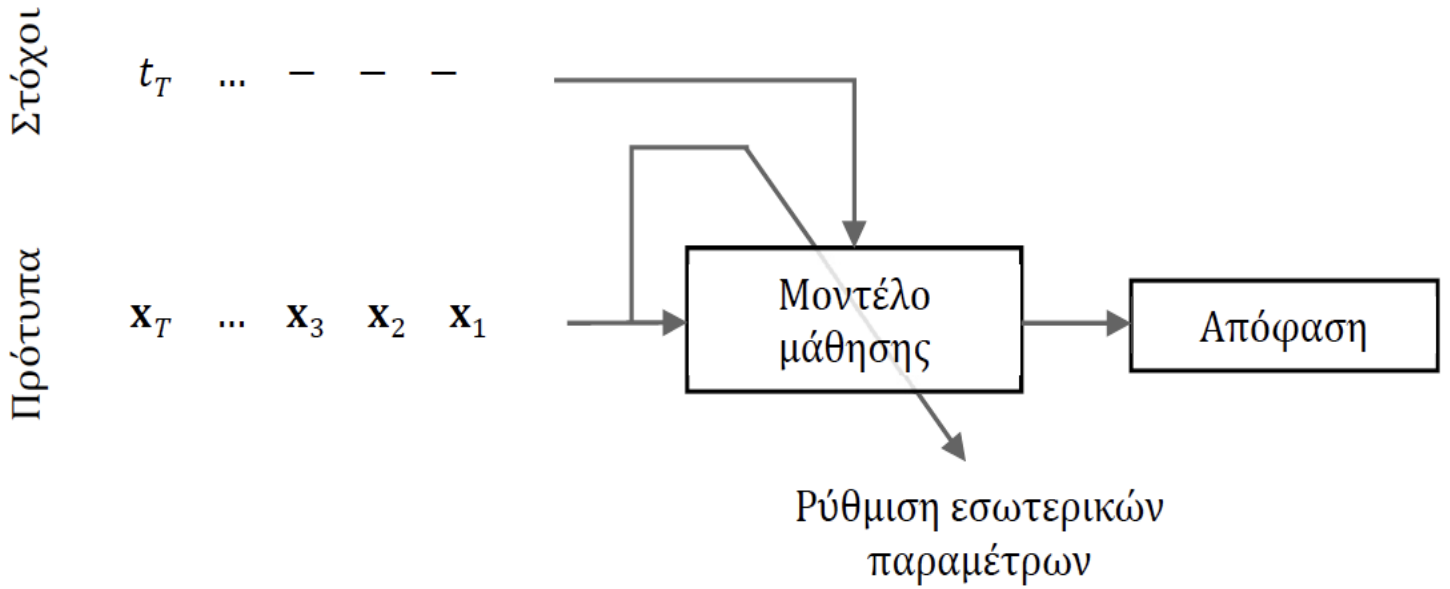
- Πώς προσδιορίζουμε αν ένα πρόβλημα μάθησης είναι εφικτό;

– Πώς είμαστε σίγουροι ότι ο χώρος των υποθέσεων περιλαμβάνει την πραγματική συνάρτηση, αφού δεν γνωρίζουμε ποια είναι η πραγματική συνάρτηση;

Μάθηση με Επίβλεψη

- Στην μάθηση με επίβλεψη διακρίνονται δυο είδη προβλημάτων (learning tasks).
  - Η ταξινόμηση (classification) αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων/κατηγοριών) (π.χ. ομάδα αίματος).
  - Η παρεμβολή (regression) αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών (π.χ. πρόβλεψη ισοτιμίας νομισμάτων ή τιμής μετοχής).

Μάθηση με Επίβλεψη



Μάθηση με Επίβλεψη

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none

Μάθηση με Επίβλεψη

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true    then play = no
If outlook = overcast                  then play = yes
If humidity = normal                   then play = yes
If none of the above                   then play = yes
```

Είδη Μάθησης με Επίβλεψη

- Μάθηση εννοιών:
  - Η μάθηση εννοιών είναι τυπικό παράδειγμα επαγωγικής μάθησης κατά την οποία:
    - Το σύστημα τροφοδοτείται με παραδείγματα που ανήκουν (θετικά παραδείγματα) ή δεν ανήκουν (αρνητικά παραδείγματα) στη συγκεκριμένη έννοια.
    - Ακολούθως πρέπει να παραχθεί κάποια γενικευμένη περιγραφή της έννοιας, δηλαδή να δημιουργηθεί ένα μοντέλο, ώστε να είναι δυνατό στη συνέχεια να αποφασιστεί αν μια άγνωστη περίπτωση ανήκει σε αυτήν την έννοια.

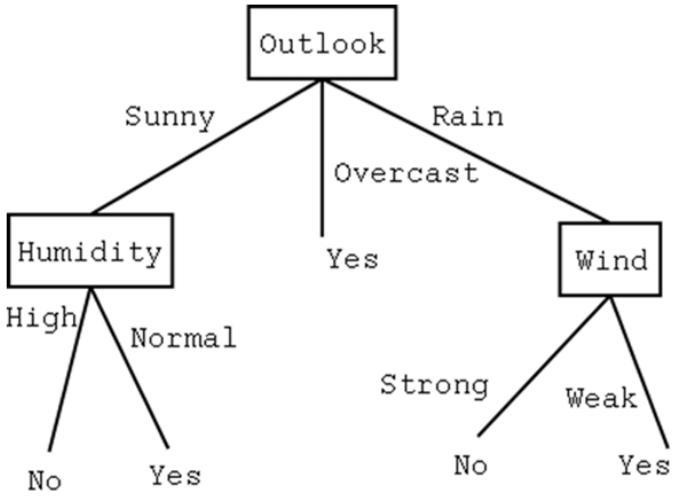
Είδη Μάθησης με Επίβλεψη

- Δένδρα Ταξινόμησης / Απόφασης / Κανόνες Ταξινόμησης:
  - Δενδροειδής δομή που με γραφικό τρόπο περιγράφει τα δεδομένα.
  - Εναλλακτικά, το δένδρο μπορεί να αναπαρασταθεί και ως σύνολο κανόνων if-then, που ονομάζονται κανόνες ταξινόμησης (classification rules).
  - Τα δένδρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών).

Είδη Μάθησης με Επίβλεψη

Δέντρο Απόφασης ή Δέντρο Κατηγοριοποίησης είναι ένα δέντρο με τις ακόλουθες ιδιότητες:

- Κάθε εσωτερικός κόμβος και η ρίζα ονοματίζεται με το όνομα ενός χαρακτηριστικού.
- Κάθε κλάδος ονοματίζεται με ένα κατηγόρημα διάσπασης του χαρακτηριστικού που αποτελεί το όνομα του κόμβου-πατέρα.
- Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης



Είδη Μάθησης με Επίβλεψη

- Μάθηση Κανόνων Ταξινόμησης
  - Κυριότερες κατηγορίες κανόνων είναι οι προτασιακοί και οι κατηγορηματικοί
  - Αφορά σε προβλήματα όπου δεν απαιτείται η αναπαράσταση σχέσεων ανάμεσα στις τιμές των διαφόρων χαρακτηριστικών (όπως και στα δένδρα ταξινόμησης/απόφασης).

Είδη Μάθησης με Επίβλεψη

- Ταξινόμηση των εγγραφών με βάση ένα σύνολο από κανόνες της μορφής "if…then…"

- Κανόνας: (*Συνθήκη*) $\rightarrow y$

  όπου

  - *Συνθήκη* (*Condition*) είναι σύζευξη συνθηκών στα γνωρίσματα
  - y η ετικέτα της κλάσης

Είδη Μάθησης με Επίβλεψη

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes
R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals
R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles
R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Είδη Μάθησης με Επίβλεψη

Ένας κανόνας r καλύπτει (covers) ένα στιγμιότυπο (εγγραφή) αν τα γνωρίσματα του στιγμιότυπου ικανοποιούν τη συνθήκη του κανόνα

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

- Ο κανόνας R1 καλύπτει το hawk (ή αλλιώς το hawk ενεργοποιεί (trigger) τον κανόνα) => Bird
- Ο κανόνας R3 καλύπτει το grizzly bear => Mammal

Είδη Μάθησης με Επίβλεψη

- Μάθηση κατά Περίπτωση
  - Τα δεδομένα εκπαίδευσης διατηρούνται αυτούσια σε αντίθεση με τις άλλες μεθόδους μηχανικής μάθησης οι οποίες κωδικοποιούν τα παραδείγματα εκπαίδευσης σε μια συμπαγή περιγραφή.
  - Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την κατηγορία μιας νέας περίπτωσης, εξετάζει εκείνη τη στιγμή τη σχέση της με τα ήδη αποθηκευμένα παραδείγματα.

Είδη Μάθησης με Επίβλεψη

- Μάθηση κατά Bayes
  - Στη μάθηση κατά Bayes (Bayesian learning) κάθε παράδειγμα εκπαίδευσης μπορεί σταδιακά να μειώσει ή να αυξήσει την πιθανότητα να είναι σωστή μια υπόθεση.
  - Μια πρακτική δυσκολία στην εφαρμογή της μάθησης κατά Bayes είναι η απαίτηση για τη γνώση πολλών τιμών πιθανοτήτων.
  - Όταν αυτές οι τιμές δεν είναι δυνατό να υπολογιστούν επακριβώς, υπολογίζονται κατ' εκτίμηση από παλαιότερες υποθέσεις, εμπειρική γνώση, κτλ.

Είδη Μάθησης με Επίβλεψη

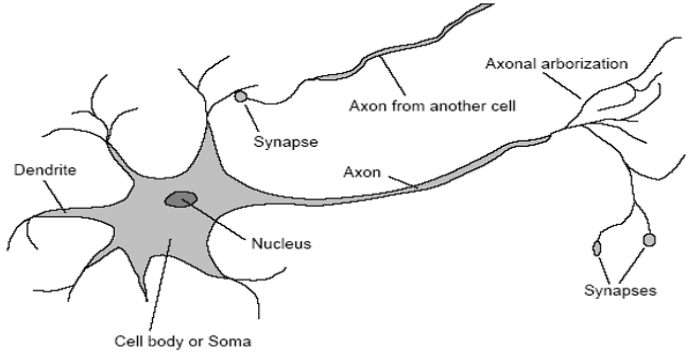
- Παρεμβολή ή Παλινδρόμηση
  - Η διαδικασία προσδιορισμού της σχέσης μιας μεταβλητής y (εξαρτημένη μεταβλητή ή έξοδος) με μια ή περισσότερες άλλες μεταβλητές x1, x2, …, xn (ανεξάρτητες μεταβλητές ή είσοδοι).
  - Σκοπός της είναι η πρόβλεψη της τιμής της εξόδου όταν είναι γνωστές οι είσοδοι.

Είδη Μάθησης με Επίβλεψη

- Νευρωνικά Δίκτυα
  - Παρέχουν ένα πρακτικό (εύκολο) τρόπο για την εκμάθηση αριθμητικών και διανυσματικών συναρτήσεων ορισμένων σε συνεχή ή διακριτά μεγέθη.
  - Χρησιμοποιούνται τόσο για παρεμβολή (γραμμική και μη γραμμική) όσο και για ταξινόμηση.
  - Έχουν το μεγάλο πλεονέκτημα της ανοχής που παρουσιάζουν σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή δεδομένα που περιστασιακά έχουν λανθασμένες τιμές (π.χ. λάθη καταχώρησης).
  - Αδυνατούν όμως να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν.

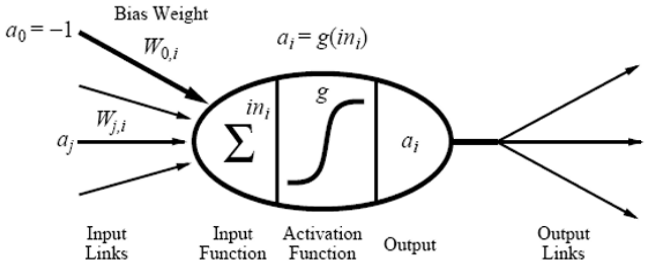
Είδη Μάθησης με Επίβλεψη

- Ανθρώπινος νευρώνας ανθρώπινου εγκεφάλου

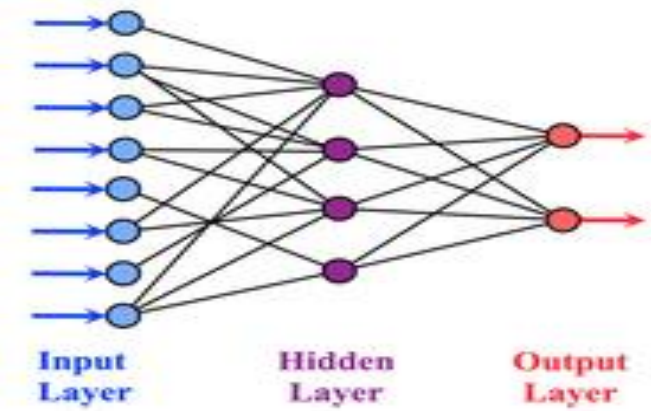


- Τεχνητός νευρώνας

$$a_i \leftarrow g(in_i) = g\left(\Sigma_j W_{j,i} a_j\right)$$

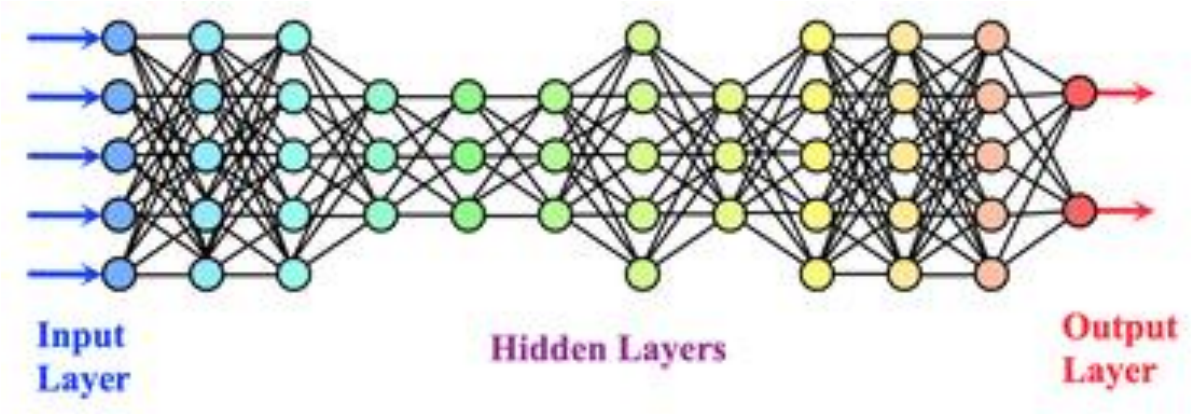


Είδη Μάθησης με Επίβλεψη



Shallow Neural Network

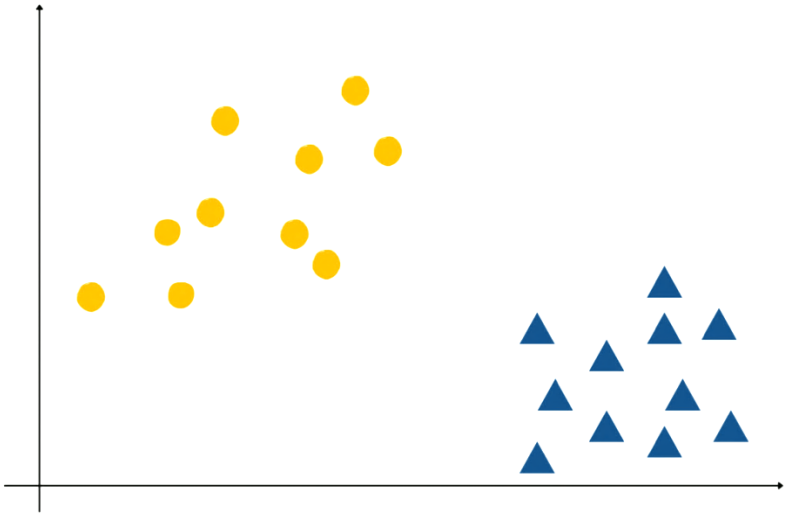
Deep Neural Network



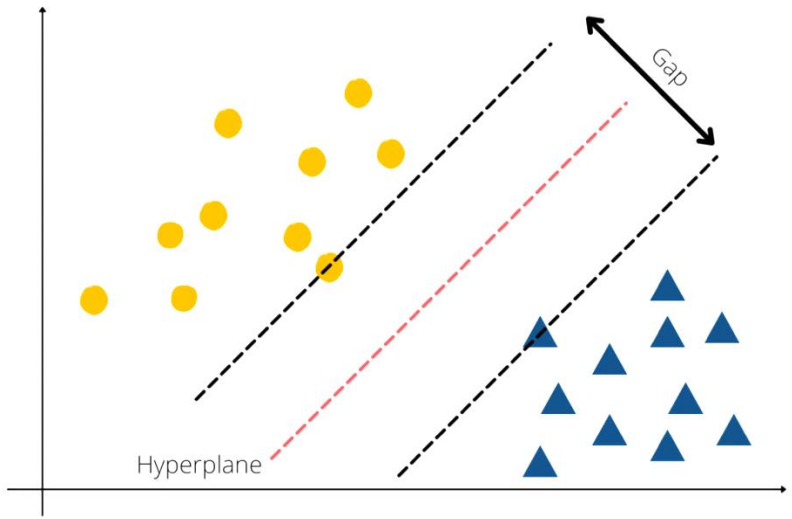
Είδη Μάθησης με Επίβλεψη

- Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ) – Support Vector Machines (SVMs)
  - Στηρίζονται στη Θεωρία Στατιστικής Μάθησης (Statistical Learning Theory) και στα νευρωνικά δίκτυα τύπου Perceptron
  - Στην περίπτωση της ταξινόμησης, οι ΜΔΥ προσπαθούν να βρουν μια υπερ-επιφάνεια (hypersurface) που να διαχωρίζει στο χώρο των παραδειγμάτων τα αρνητικά από τα θετικά παραδείγματα.
  - Η υπερεπιφάνεια αυτή επιλέγεται έτσι, ώστε να απέχει όσο το δυνατόν περισσότερο από τα κοντινότερα θετικά και αρνητικά παραδείγματα (maximum margin hypersurface).

Είδη Μάθησης με Επίβλεψη



Training Data



Best Hyperplane

Μάθηση χωρίς Επίβλεψη

- Στη μάθηση χωρίς επίβλεψη το σύστημα έχει στόχο να ανακαλύψει συσχετίσεις και ομάδες από τα δεδομένα, βασιζόμενο μόνο στις ιδιότητές τους.
- Σαν αποτέλεσμα προκύπτουν πρότυπα (περιγραφές), κάθε ένα από τα οποία περιγράφει ένα μέρος από τα δεδομένα.
- Παραδείγματα προτύπων πληροφόρησης είναι οι κανόνες συσχέτισης (association rules) και οι ομάδες (clusters), οι οποίες προκύπτουν από τη διαδικασία της ομαδοποίησης (clustering).

Κανόνες Συσχέτισης

- Η ανακάλυψη ή εξόρυξη κανόνων συσχέτισης (association rule mining) εμφανίστηκε αρκετά αργότερα από τη μηχανική μάθηση και έχει περισσότερες επιρροές από την ερευνητική περιοχή των βάσεων δεδομένων.
  - Προτάθηκε στις αρχές της δεκαετίας του '90 από τον Rakesh Agrawal ως τεχνική ανάλυσης καλαθιού αγορών (market basket analysis) όπου το ζητούμενο είναι η ανακάλυψη συσχετίσεων ανάμεσα στα αντικείμενα μιας βάσης δεδομένων.
  - Στο συγκεκριμένο πρόβλημα υπάρχει ένας μεγάλος αριθμός αντικειμένων (items), για παράδειγμα ψωμί, γάλα, κτλ. Οι πελάτες γεμίζουν τα καλάθια τους με κάποιο υποσύνολο αυτών των αντικειμένων και το ζητούμενο είναι να βρεθεί ποια από αυτά τα αντικείμενα αγοράζονται μαζί, χωρίς να ενδιαφέρει ποιος είναι ο αγοραστής.

Κανόνες Συσχέτισης

- Οι κανόνες συσχέτισης είναι προτάσεις της μορφής {X1,...,Xn}→Y, που σημαίνει ότι αν βρεθούν όλα τα X1, ..., Xn στο καλάθι (στην ανάλυση καλαθιού αγορών) τότε είναι πιθανό να βρεθεί και το Y.
  - Για παράδειγμα, ένας τέτοιος κανόνας θα μπορούσε να λέει: "όποιος αγοράζει καφέ (X1) και ζάχαρη (X2) αγοράζει και αναψυκτικά (Y)"

Ομάδες (Clusters)

- Είναι πρότυπα πληροφόρησης που προκύπτουν με ομαδοποίηση (clustering) δηλαδή διαχωρισμό ενός συνόλου (συνήθως πολυδιάστατων) δεδομένων σε ομάδες, ώστε:
  - σημεία που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο και
  - σημεία που ανήκουν σε διαφορετικές ομάδες να διαφέρουν όσο το δυνατόν περισσότερο.

Μάθηση χωρίς Επίβλεψη

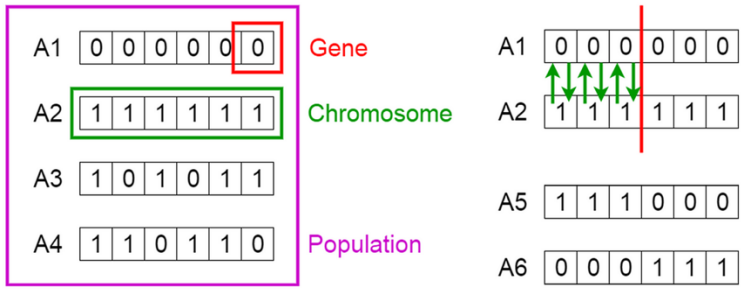
- Αλγόριθμοι ομαδοποίησης:
  - Υπάρχουν τρεις γενικές κατηγορίες αλγορίθμων ομαδοποίησης:
    - Οι αλγόριθμοι βασισμένοι σε διαχωρισμούς (partition based), που προσπαθούν να βρουν τον καλύτερο διαχωρισμό ενός συνόλου δεδομένων σε ένα συγκεκριμένο αριθμό ομάδων.
    - Οι ιεραρχικοί (hierarchical) αλγόριθμοι, που προσπαθούν με ιεραρχικό τρόπο να ανακαλύψουν τον αριθμό και τη δομή των ομάδων.
    - Οι πιθανοκρατικοί (probabilistic) αλγόριθμοι που βασίζονται σε μοντέλα πιθανοτήτων.

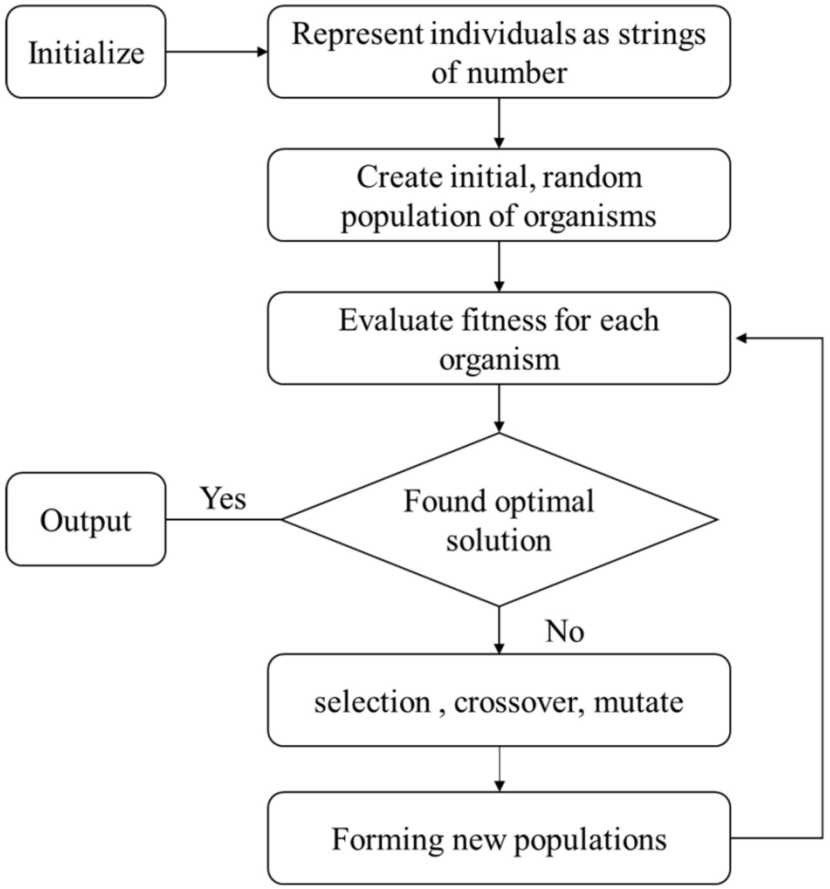
Μάθηση χωρίς Επίβλεψη

- Γενετικοί Αλγόριθμοι:
  - Μέθοδος μάθησης που βασίζεται στην προσομοίωση του φυσικού φαινομένου της εξέλιξης (evolution). Η μάθηση αντιμετωπίζεται σαν μία ειδική περίπτωση βελτιστοποίησης.
  - Οι υποθέσεις συνήθως αναπαριστώνται από ακολουθίες bit (bit-strings).
  - Η αναζήτηση της κατάλληλης υπόθεσης ξεκινάει τυχαία με έναν πληθυσμό (μια συλλογή) αρχικών υποθέσεων, τα μέλη του οποίου παράγουν τη νέα "γενιά" μέσω διαδικασιών αναπαραγωγής αντίστοιχων των βιολογικών, όπως:
    - διασταύρωση (crossover)
    - τυχαία μετάλλαξη (random mutation)
  - Σε κάθε βήμα, οι υποθέσεις του τρέχοντος πληθυσμού αξιολογούνται βάσει μιας προκαθορισμένης συνάρτησης καταλληλότητας (fitness function).
  - Βάσει αυτής επιλέγονται για το αν θα υφίστανται ή όχι στην επόμενη γενιά.

Μάθηση χωρίς Επίβλεψη

Genetic Algorithms





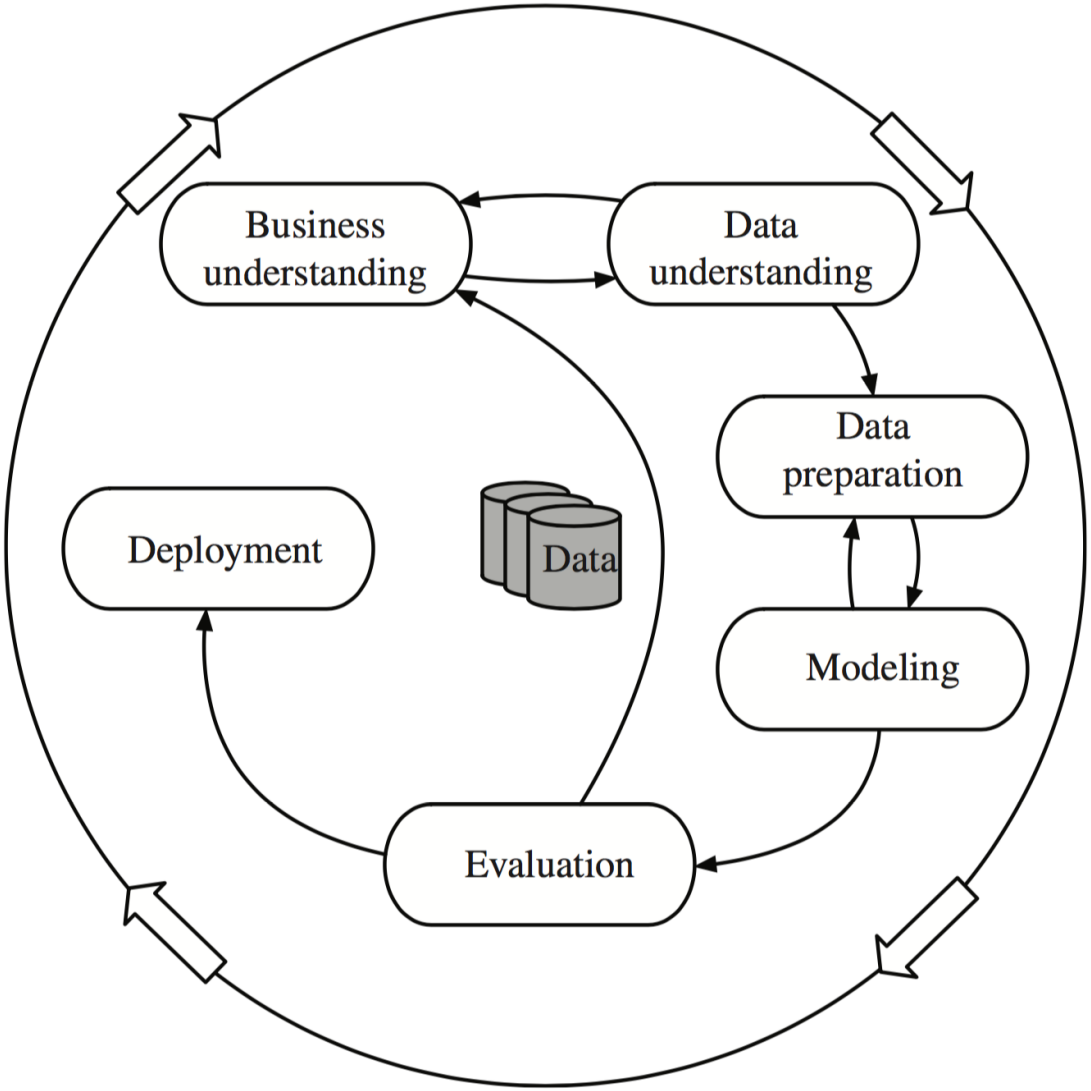
Μάθηση χωρίς Επίβλεψη

- **Ενισχυτική Μάθηση:**
  - Γενική περιγραφή οικογένειας τεχνικών στις οποίες το σύστημα μάθησης προσπαθεί να μάθει μέσω άμεσης αλληλεπίδρασης με το περιβάλλον.
  - Είναι εμπνευσμένη από τα αντίστοιχα ανάλογα της μάθησης με επιβράβευση και τιμωρία που συναντώνται στα έμβια όντα.
  - Σκοπός του συστήματος μάθησης: να μεγιστοποιήσει μια συνάρτηση του αριθμητικού σήματος ενίσχυσης (ανταμοιβή), για παράδειγμα την αναμενόμενη τιμή του σήματος ενίσχυσης στο επόμενο βήμα.
  - Το σύστημα δεν καθοδηγείται από κάποιον εξωτερικό επιβλέποντα για το ποια ενέργεια θα πρέπει να ακολουθήσει αλλά πρέπει να ανακαλύψει μόνο του ποιες ενέργειες είναι αυτές που θα του αποφέρουν το μεγαλύτερο κέρδος.

Εξόρυξη Δεδομένων

- Αναζήτηση προτύπων στα δεδομένα για παροχή γνώσης που επιτρέπει τη γρήγορη και ακριβή λήψη αποφάσεων
- Κρίσιμο θέμα: εύρεση προτύπων που είναι τόσο ακριβή και δυνατά
  - Πρόβλημα 1: τα περισσότερα πρότυπα δεν είναι ενδιαφέροντα
  - Πρόβλημα 2: τα πρότυπα μπορεί να είναι μη ακριβή ή πλαστά
  - Πρόβλημα 3: τα δεδομένα μπορεί να διαστραφούν ή να έχουν υπολειπόμενες τιμές
- Οι τεχνικές μηχανικής μάθησης ανακαλύπτουν πρότυπα στα δεδομένα και παρέχουν πολλαπλά εργαλεία για εξόρυξη δεδομένων.
- Οι περισσότερο ενδιαφέρουσες τεχνικές είναι αυτές που παρέχουν δομημένες περιγραφές.

Διαδικασία Εξόρυξης Δεδομένων



Μηχανική Μάθηση και Στατιστική

- Ιστορική διαφοροποίηση
  - Στατιστική: έλεγχος υποθέσεων
  - Μηχανική Μάθηση: εύρεση των σωστών υποθέσεων
- Αλλά έχουν και μεγάλη επικάλυψη
  - Δένδρα αποφάσεων (C4.5 and CART)
  - Μέθοδοι εύρεσης κοντινότερων γειτόνων (nearest neighbour methods)
- Σήμερα: ύπαρξη σύγκλισης
  - Οι περισσότεροι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν στατιστικές μεθόδους

Εφαρμογές Μηχανικής Μάθησης

- Χιλιάδες
- Ενδεικτικά:
  - Επεξεργασία αιτήσεων δανείων
  - Πρόβλεψη Παροχής Ηλεκτρικής Ενέργειας
  - Διάγνωση Λαθών Μηχανής
  - Πωλήσεις και Μάρκετινγκ

Επεξεργασία Αιτήσεων Δανείων

- Δεδομένα: συμπληρωμένο ερωτηματολόγιο με οικονομικές και ιδιωτικές πληροφορίες + αίτηση
- Ερώτηση: να εγκριθεί η αίτηση;
- Μια απλή στατιστική μέθοδος καλύπτει το 90% των περιπτώσεων -> αυτοματισμός
- Οι οριακές περιπτώσεις ανατίθενται στους υπαλλήλους δανείων.
- Όμως 50% των αποδεχόμενων περιπτώσεων αυτού του είδους είχαν αθετηθεί στο παρελθόν.
- Λύση: μη αποδοχή των οριακών περιπτώσεων;
  - Όχι διότι αυτές οι περιπτώσεις αντιστοιχούν στους πιο ενεργούς πελάτες

Επεξεργασία Αιτήσεων Δανείων

- Εφαρμογή τεχνικών μηχανικής μάθησης:
  - Χρήση 1000 παραδειγμάτων εκπαίδευσης οριακών περιπτώσεων
  - Θεώρηση 20 ιδιοτήτων:
    - ηλικία
    - αριθμός ετών με τον τρέχοντα εργοδότη
    - αριθμός ετών με την τρέχουσα διεύθυνση
    - αριθμός ετών με την τράπεζα
    - κατοχή άλλων καρτών πίστωσης κα.
  - Μαθημένοι Κανόνες: σωστοί σε 70% των περιπτώσεων
    - ενώ οι ειδικοί μόνο σε 50%
  - Οι κανόνες μπορούν να χρησιμοποιηθούν για την επεξήγηση αποφάσεων στους πελάτες ενώ βοηθούν στην παροχή υποστήριξης στην διαδικασία έγκρισης των δανείων.

Πρόβλεψη Παροχής Ηλεκτρικής Ενέργειας

- Οι εταιρείες παροχής ηλεκτρικής ενέργειας χρειάζονται την πρόβλεψη της μελλοντικής ανάγκης για ενέργεια
- Οι προβλέψεις για μέγιστο/ελάχιστο φόρτο για κάθε ώρα οδηγούν σε σημαντικές αποταμιεύσεις σχετιζόμενες με τη ρύθμιση της λειτουργικής ρεζέρβας, το προγραμματισμό διατήρησης και τη διαχείριση της καταγραφής των καυσίμων
- Δεδομένο: χειρωνακτικά κατασκευασμένο μοντέλο φόρτου από δεδομένα 15 ετών που υποθέτει κανονικές κλιματικές συνθήκες
  - Το στατικό μοντέλο αποτελείται από:
    - Το βασικό φόρτο για όλο το έτος
    - Τη περιοδικότητα του φόρτου κατά μήκος του έτους
    - Την επίδραση των διακοπών / αργιών
- Πρόβλημα: προσαρμογή στις κλιματικές συνθήκες

Πρόβλεψη Παροχής Ηλεκτρικής Ενέργειας

- Λύση:
  - Διόρθωση πρόβλεψης με τη χρήση των πιο παρόμοιων ημερών.
  - Η μέση διαφορά μεταξύ των πιο παρόμοιων ημερών προστέθηκε στο στατικό μοντέλο
  - Χρήση γραμμικής παρεμβολής για την αποτίμηση της επίδρασης στο φόρτο των ακόλουθων ιδιοτήτων:
    - θερμοκρασία
    - υγρασία
    - ταχύτητα ανέμου
    - μετρήσεις κάλυψης νέφους
    - διαφορά μεταξύ του τρέχοντος και του προβλεπόμενου φόρτου
  - Οι συντελεστές γραμμικής παρεμβολής σχηματίζουν τα βάρη των ιδιοτήτων στη συνάρτηση ομοιότητας.

Διάγνωση Λαθών Μηχανής

- Κλασσικός τομές των εμπειρικών συστημάτων
- Δεδομένα: Η ανάλυση Fourier των δονήσεων μετρημένων σε διάφορα σημεία που αφορούν την τοποθέτηση (mounting) της συσκευής
- Κάλυψη: Αποτρεπτική διαχείριση των ηλεκτρομαγνητικών κινητήρων και γεννητόρων
- Η πληροφορία που λαμβάνεται έχει πολύ θόρυβο
- Τρέχουσα κατάσταση: διάγνωση με τη χρήση εμπειρικών / χειρονακτικών κανόνων
  - Οι εμπειρογνώμονες δεν ήταν ικανοποιημένοι με το αρχικό σύνολο κανόνων λόγω της μη συναφής με την γνώση τομέα.

Διάγνωση Λαθών Μηχανής

- Διαθέσιμα: 600 λάθη με διάγνωση εμπειρογνώμονα
- Μόνο 300 ικανοποιητικά -> χρήση στην εκπαίδευση
- Ζητούμενο: ποιο λάθος παρουσιάζεται εφόσον ξέρουμε ότι υπάρχει;
- Οι ιδιότητες ήταν σε χαμηλό επίπεδο και επαυξήθηκαν με ενδιάμεσες έννοιες που εμπεριείχαν την αιτιολογική γνώση τομέα (causal domain knowledge)
- Η επιπλέον γνώση είχε ως αποτέλεσμα την δημιουργία πιο ικανοποιητικών και πολύπλοκων κανόνων μέσω της χρήσης επαγωγικού αλγορίθμου
- Οι νέοι κανόνες ξεπέρασαν σε απόδοση τους παλιούς

Πωλήσεις και Μάρκετινγκ

- Οι εταιρείες καταγράφουν με ακριβή τρόπο χιλιάδες εγγραφές από δεδομένα μάρκετινγκ και πωλήσεων
- Εφαρμογές:
  - Πιστότητα (loyalty) Πελατών:
    προσδιορισμός πελατών που μπορεί να αθετήσουν ή να σπάσουν τρέχοντα συμβόλαια μέσω της ανίχνευσης αλλαγής στη συμπεριφορά τους
    (πχ. Τραπεζικές και τηλεφωνικές εταιρείες)
  - Ειδικές προσφορές:
    προσδιορισμός πελατών με υψηλή πιθανότητα αποδοχής των προσφορών (πχ. αξιόπιστοι κάτοχοι πιστωτικών καρτών που χρειάζονται επιπλέον χρήματα σε περίοδο διακοπών)

Πωλήσεις και Μάρκετινγκ

- Ανάλυση καλαθιού αγοράς
  - Χρήση τεχνικών εύρεσης κανόνων συσχέτισης για την ανακάλυψη αντικειμένων που τείνουν να εμφανίζονται μαζί σε μια δοσοληψία (πχ. σε υπεραγορές)
- Ιστορική ανάλυση προτύπων αγοράς
- Προσδιορισμός δυνητικών πελατών
  - Εκμετάλλευση στην εκτέλεση στοχευμένων εκστρατειών σε αντίθεση με δαπανηρές μαζικές εκστρατείες

Μηχανική Μάθηση & Εξόρυξη Γνώσης

Ερωτήσεις

?

Βιβλιογραφία

- Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας, Η. Σακελλαρίου, Τεχνητή Νοημοσύνη - Γ' Έκδοση, ISBN: 978-960-8396-64-7, Έκδοση/Διάθεση: Εκδόσεις Πανεπιστημίου Μακεδονίας, 2011
- Ian H. Witten and Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Κ. Διαμαντάρας, Ι. Μπότσης, Μηχανική Μάθηση – Α' Έκδοση, ISBN: 978-960-461-955-5, Εκδόσεις Κλειδάριθμος, 2019
- P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining», Addison Wesley, 2006