ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Δημιουργία συστάσεων με χρήση ενισχυτικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Λεωνίδα Αβδελά

Επιβλέπων: -Εισάγετε το όνομα, αρχικό πατρώνυμο και επίθετο του επιβλέποντα-
-Εισάγετε τον τίτλο του επιβλέποντα-

Αθήνα, Ιανουάριος 2023

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Δημιουργία συστάσεων με χρήση ενισχυτικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Λεωνίδα Αβδελά

Επιβλέπων: -Εισάγετε το όνομα, αρχικό πατρώνυμο και επίθετο του επιβλέποντα-
-Εισάγετε τον τίτλο του επιβλέποντα-

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την -εισάγετε ημερομηνία-.

.....	.....	.....
-Εσαγετε Ονοματεπωνυμο-	-Εσαγετε Ονοματεπωνυμο-	-Εσαγετε Ονοματεπωνυμο-
-Εσαγετε τίτλο-	-Εσαγετε τίτλο-	-Εσαγετε τίτλο-

Αθήνα, Ιανουάριος 2023.

.....

Λεωνίδας Αβδελάς
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

TODO

Λέξεις Κλειδιά

TODO

Abstract

TODO

Keywords

TODO

Ευχαριστίες

Ευχαριστώ την οικογένεια μου και τους καθηγητές μου.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ενισχυτική Μάθηση

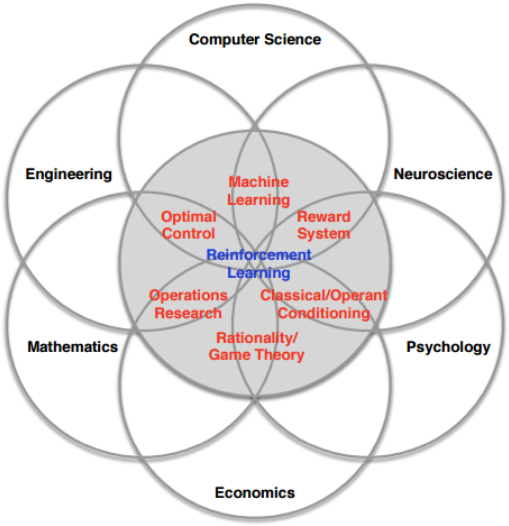
Η ενισχυτική μάθηση (ΕΜ) (Reinforcement Learning (RL)) είναι είναι ένας γενικός όρος που έχει δοθεί σε μια οικογένεια τεχνικών στις οποίες το σύστημα προσπαθεί να μάθει μέσα από την άμεση αλληλεπίδραση με το περιβάλλον.[1] Είναι τομέας της τεχνητής νοημοσύνης και, πιο συγκεκριμένα, της μηχανικής μάθησης.

Πιο συγκεκριμένα, η ΕΜ είναι η διαδικασία κατα την οποία ένας πράκτορας (agent) αλληλεπιδρά με το περιβάλλον του, και μαθαίνει τι να κάνει, παρατηρώντας τις συνέπειες των ενεργειών του (actions). Ο πράκτορας δεν δίνεται πληροφορίες σχετικά με το ποίες ενέργειες να επιλέξει, αλλά πρέπει να ανακαλύψει ποιες δράσεις προσφέρουν την μέγιστη ανταμοιβή (reward), δοκιμάζοντας τες [4]. Επιπλέον, σε πολλές περιπτώσεις οι ενέργειες του πράκτορα δεν επηρεάζουν μόνο την άμεση ανταμοιβή, αλλά και από την ανταμοιβή στην επόμενη κατάσταση και πιθανώς και όλες τις επόμενες ανταμοιβές. Σύμφωνα με τα παραπάνω, η ΕΜ στοχεύει να λύσει προβλήματα μέσω τεχνικών *δοκιμής-και-λάθους (trial-and-error)* σε περιβάλλοντα με *καθυστερημένες ανταμοιβές*.

Η θεωρία της ΕΜ βασίζεται πάνω στην υπόθεση της ανταμοιβής (reward hypothesis), στην ιδέα ότι κάθε στόχος μπορεί να εκφραστεί ως η μεγιστοποίηση την αναμενόμενης αξίας του σωρευτικού (cummulative) αθροίσματος ενός μονοδιάστατου σήματος (ανταμοιβής). Με απλά λόγια, η υπόθεση θέτει την ιδέα ότι κάθε στόχος μπορεί να εκφραστεί σαν την μεγιστοποίηση μιας ανταμοιβής. Η ανταμοιβή αυτή δεν χρειάζεται να είναι θετικός αριθμός, μπορεί να είναι και αρνητικός. Για παράδειγμα, αν ο στόχος είναι η έξοδος από ένα λαβύρινθο, η ανταμοιβή μπορεί να είναι αρνητική σε κάθε βήμα μέχρι την έξοδο, ώστε τελικά η στόχος να είναι η ελαχιστοποίηση της απόλυτης τιμής της ανταμοιβής.

Το πεδίο της ΕΜ έχει τις ρίζες του σε δύο περιοχές. Η πρώτη είναι η συμπεριφορική ψυχολογία, από όπου προέρχεται το παράδειγμα της δοκιμής-και-λάθους, και η δεύτερη είναι

η περιοχή του βέλτιστου ελέγχου, από όπου η ΕΜ δανείζεται τον μαθηματικό φορμαλισμό (κυρίως τον δυναμικό προγραμματισμό) που υποστηρίζει το πεδίο. Η ΕΜ βρίσκονται στην τομή πολλών διαφορετικών επιστημονικών πεδίων, οι οποίοι φαίνται στο Σχήμα 1.1[2].



Σχήμα 1.1: Τα πρόσωπα της ενισχυτικής μάθησης

Η ΕΜ πολλές φορές συγχέεται με τις άλλες τεχνικές μηχανικής μάθησης, παρόλο που έχει αρκετά σημαντικές διαφορές. Αρχικά, η κύρια διαφορά μεταξύ της επιτηρούμενης μάθησης (supervised learning) και της ΕΜ είναι ότι στην επιτηρούμενη μάθηση, το μοντέλο εκπαιδεύεται πάνω σε δείγματα (samples) και ετικέτες (labels), και κάθε πρόβλεψη θεωρείται μοναδικό γεγονός. Από την άλλη, στην ΕΜ, μπορούν να υπάρχουν πολλά βήματα πριν ο πράκτορας μάθει αν η απόφαση που πήρε ήταν σωστή, και μπορεί να μην μάθει ποτέ ποια ήταν η αληθής/βέλτιστη τιμή, αλλά να βλέπει μόνο την επίδραση που είχαν οι πράξεις του στο περιβάλλον.

1.1.1 Γενικά

Πιο φορμαλιστικά, σε ένα περιβάλλον ΕΜ, ένας αυτόνομος πράκτορας, ελεγχόμενος από ένα αλγόριθμο μηχανικής μάθησης, παρατηρεί μια κατάσταση s_t από το περιβάλλον του σε ένα χρονικό βήμα t. Οι καταστάσεις προέρχονται από τον χώρο καταστάσεων $\mathcal{S}$. Ο πράκτορας αλληλεπιδρά με το περιβάλλον επιλέγοντας μια δράση a_t με βάση την κατάσταση s_t, επιλεγμένη από ένα χώρο δράσεων $\mathcal{A}$. Όταν ο πράκτορας επιλέξει την δράση, τότε τόσο το περιβάλλον, οσο και ο πράκτορας μεταβαίνουν σε μια νέα κατάσταση s_{t+1}, με βάση την τρέχουσα κατάσταση και την επιλεγμένη δράση [3].

Ο πράκτορας επίσης λαμβάνει και μια μονοδιάστατη ανταμοιβή R_t η οποία προέρχεται από το ζεύγος κατάστασης-δράσης. Αυτή η ανταμοιβή δρα ως μια μορφή ανατροφοδότησης για

τις δράσεις του πράκτορα. Επιπλέον, ο πράκτορας διατηρεί μια αντιστοίχηση μεταξύ κατάστασης και δράσης, η οποία συμβολίζεται ως $\pi(a_t|s_t)$. Για κάθε ζευγάρι κατάστασης-δράσης, υπάρχει η αντίστοιχη ανταμοιβή R που λαμβάνει ο πράκτορας ως αποτέλεσμα μια συγκεκριμένης δράσης a_t στην κατάσταση s_t. Η βέλτιστη σειρά δράσεων προσδιορίζεται από αυτές τις ανταμοιβές που προμηθεύει το περιβάλλον. Ο τελικός στόχος του πράκτορα είναι να μάθει μια πολιτική π, η οποία μεγιστοποιεί την αναμενόμενη απόδοση (σωρευτική, εκπτωθείσα (discounted) ανταμοιβή). Δοθείσας μιας κατάστασης η πολιτική επιστρέφει την επόμενη δράση την οποία θα κάνει ο πράκτορας. Μια βέλτιστη πολιτική είναι η πολιτική η οποία μεγιστοποιεί την αναμενόμενη απόδοση στο συγκεκριμένο περιβάλλον.

Θέλουμε τα προβλήματα ΕΜ που ασχολούμαστε να ικανοποιούν την Μαρκοβιανή ιδιότητα. Δηλαδή, για κάθε κατάσταση, το μέλλον να εξαρτάται μόνο από την τρέχουσα κατάσταση και όχι τις προηγούμενες. Όταν ισχύει αυτή η ιδιότητα τότε μπορούμε να μοντελοποιήσουμε το πρόβλημα ως μια Μαρκοβιανή Διαδικασία Αποφάσεων (Markov Decision Process), η οποία αποτελείται από:

- ένα σύνολο από καταστάσεις $\mathcal{S}$, και επιπλέον μια κατανομή αρχικών καταστάσεων $p(s_0)$,
- ενα σύνολο από δράσεις $\mathcal{A}$,
- τις πιθανότητες μετάβασης μεταξύ των καταστάσεων
  $\Pr\{S_t = s'|S_{t-1} = s, A_{t-1} = a\}$, οι οποίες αντιστοιχούν ένα ζεύγος κατάστασης-δράσης την στιγμή $t-1$ σε μια κατανομή καταστάσεων την στιγμή t,
- μια συνάρτηση άμεσης ανταμοιβής $R(s_{t-1}, a_{t-1}, s_t)$,
- κατα τον υπολογισμό της σωρευτικής ανταμοιβής, συμπεριλαμβάνεται και ένας παράγοντας ¨έκπτωσης' $\gamma \in [0, 1]$, ο οποίος χρησιμοποιειται για να τοποθετηθεί μικρότερη έμφαση στις πιο παλιές ανταμοιβές.

Σύμφωνα με τα παραπάνω, ορίζουμε ως την απόδοση την εκπτωθείσα, σωρευτική ανταμοιβή μαζί με τον παράγοντα έκπτωσης:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{1.1}$$

Ένα κύριο χαρακτηριστικό κάθε μεθόδου ΕΜ είναι η **συνάρτηση αξίας**, μια πρόβλεψη της αναμενόμενης, σωρευτικής, εκπτωθείσας, μέλλουσας απόδοσης. Ποσοτικοποιεί πόσο καλή είναι μια κατάσταση ή ένα ζεύγος κατάστασης-δράσης. Η αξία μιας κατάστασης

$$v_\pi(s) = \mathbb{E}[R_t|s_t = s] \tag{1.2}$$

είναι η αναμενόμενη απόδοση ακολουθώντας την πολιτική π από την κατάσταση s. Η αξία της δράσης

$$q_\pi(s, a) = \mathbb{E}[R_t|s_t = s, a_t = a] \tag{1.3}$$

είναι η αναμενόμενη απόδοση της επιλογής της δράσης a στην κατάσταση s και ακολουθώντας την πολιτική π μετά.

Επιπλέον, για να περιγράψουμε καλύτερα τις μεθόδους ΕΜ, τις χωρίζουμε σε μεθόδους που λύνουν προβλήματα **βασισμένα σε μοντέλο (model-based)** ή προβλήματα **ανεξάρτητα μοντέλου (model-free)**.

Μορεοερ, το βεττερ δεσςριβε ΡΛ μετηοδς ωε διιδε τηεμ ιν **μοδελ-βασεδ** ανδ **μοδελ-φρεε** προβλεμς. **Μοδελ-βασεδ** μετηοδς αρε υσεδ ωηεν ωε ηαε α ςομπλετε κνοωλεδγε οφ τηε δψναμιςς οφ τηε συρρουνδινγ ενιρονμεντ ανδ ρελψ ον *πλαννινγ* ας τηειρ πριμαρψ ςομπονεντ. Τηε μετηοδς ινςλυδε δψναμις προγραμμινγ ανδ σιμιλαρ μετηοδς. Ωε υσε πολιςψ εαλυατιον το ςαλςυλατε τηε ρετυρν οφ α σπεςιφις πολιςψ ανδ Ον τηε οτηερ ηανδ **μοδελ-φρεε** μετηοδς ρεχυιρε νο πριορ κνοωλεδγε οφ τηε ενιρονμεντ ανδ ρελψ ον *λεαρνινγ*. Βψ α *μοδελ* οφ τηε ενιρονμεντ ωε μεαν ανψτηινγ τηατ αν αγεντ ςαν υσε το πρεδιςτ ηοω τηε ενιρονμεντ ωιλλ ρεσπονδ το ιτς αςτιονς.

Βιβλιογραφία

[1] Ι. Βλαχάβας, Π. Κεφαλάς, Ν. Βασιλειάδης, Φ. Κόκκορας και Η. Σακελλαρίου, *Τεχνητή Νοημοσύνη*. Πανεπιστήμιο Μακεδονίας, 2006.

[2] D. Silver, *Lectures on reinforcement learning*, URL: https://www.davidsilver.uk/teaching/, 2015.

[3] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey", *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, Nov. 2017, ISSN: 1558-0792. DOI: 10.1109/MSP.2017.2743240.

[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018, ISBN: 0262039249.