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

Δημιουργία συστάσεων με χρήση contextual bandits

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 contextual bandits

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.

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

Ενισχυτική Μάθηση

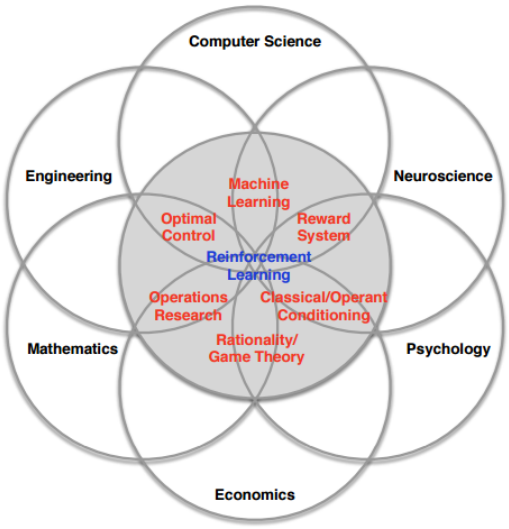
1.1 Γενικά

Η ενισχυτική μάθηση (ΕΜ) (Reinforcement Learning (RL)) είναι είναι ένας γενικός όρος που έχει δοθεί σε μια οικογένεια τεχνικών στις οποίες ένα 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 πράξεων του. Ο πράκτορας δεν δίνεται πληροφορίες σχετικά με το ποίες ενέργειες (actions) να επιλέξει, αλλά πρέπει να ανακαλύψει ποιες ενέργειες προσφέρουν την μέγιστη ανταμοιβή (reward), δοκιμάζοντας τες [4]. Επιπλέον, σε πολλές περιπτώσεις οι ενέργειες του πράκτορα δεν επηρεάζουν μόνο την άμεση ανταμοιβή που θα πάρει, αλλά και από την ανταμοιβή στην επόμενη κατάσταση, και πιθανώς και όλες τις επόμενες ανταμοιβές. Ετσι,μπορεί να υπάρξουν καταστάσεις που ο πρακτορας θα πρέπει να θυσιάσει την άμεση ανταμοιβή για να αποκτήσει καλύτερες ανταμοιβές μακροπρόθεσμα. Σύμφωνα με τα παραπάνω, η ΕΜ στοχεύει να λύσει προβλήματα μέσω τεχνικών *δοκιμής-και-λάθους (trial-and-error)* σε περιβάλλοντα με *καθυστερημένες ανταμοιβές*.

Μια κομβική ιδέα, πάνω στην οποία στηρίζεται η ΕΜ, είναι η υπόθεση της ανταμοιβής (reward hypothesis), η ιδέα ότι κάθε στόχος μπορεί να εκφραστεί ως η μεγιστοποίηση της αναμενόμενης αξίας του σωρευτικού (cummulative) αθροίσματος ενός μονοδιάστατου σήματος. Με απλά λόγια, η υπόθεση θέτει την ιδέα ότι κάθε στόχος μπορεί να εκφραστεί σαν την μεγιστοποίηση μιας ανταμοιβής. Η ανταμοιβή αυτή δεν χρειάζεται απαραίτητα να είναι θετικός αριθμός, αλλά ακόμα και για αρνητικές τιμές της, συνεχίζουμε να καλούμε τον όρο ανταμοιβή. Για παράδειγμα, αν ο στόχος είναι η έξοδος από ένα λαβύρινθο, η ανταμοιβή μπορεί να είναι αρνητική σε κάθε βήμα μέχρι την έξοδο, οπού και γίνεται 0. Τότε ο στόχος τελικά είναι η ελαχιστοποίηση της απόλυτης τιμής της ανταμοιβής, δηλαδή η έξοδος στα λιγότερα βήματα.

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. Είναι σημαντικό να γνωρίζουμε ότι η ΕΜ βρίσκεται στην τομή πολλών διαφορετικών επιστημονικών πεδίων, οι οποίοι φαίνονται στο Σχήμα 1.1[2]. Όλα αυτά τα πεδία προσεγγίζουν ένα παρόμοιο πρόβλημα, αλλά απο διαφορετική σκοπιά και με διαφορετικές παραμέτρους.



Σχήμα 1.1: Τα πρόσωπα της ενισχυτικής μάθησης

Η ΕΜ πολλές φορές συγχέεται με τις άλλες τεχνικές μηχανικής μάθησης, την επιβλεπόμενη και την μη επιβλεπόμενη μάθηση, παρόλο που έχει αρκετά σημαντικές διαφορές.

Αρχικά, η κύρια διαφορά μεταξύ της επιβλεπόμενης μάθησης (supervised learning) και της ΕΜ είναι ότι στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται πάνω σε δείγματα (samples) και ετικέτες (labels), και κάθε πρόβλεψη θεωρείται μοναδικό γεγονός. Στόχος είναι η προσέγγιση μιας άγνωστης συνάρτησης με βάση τα δεδομένα. Αντίθετα, στην ΕΜ, μπορούν να υπάρχουν πολλά βήματα πριν ο πράκτορας μάθει αν η απόφαση που πήρε ήταν σωστή, και είναι πιθανό να μην μάθει ποτέ ποια ήταν η αληθής/βέλτιστη τιμή. Το μόνο που παρατηρεί είναι η επίδραση που είχαν οι πράξεις του στο περιβάλλον.

Όσον αφορά την μη επιβλεπόμενη μάθηση, μπορεί αρχικά να φαίνεται παρόμοια με την ΕΜ. Όμως, στόχος της μη επιβλεπόμενης μάθησης είναι η εύρεση της κρυμμένης δομής δεδομένων τα οποία δεν έχουν ετικέτες. Η ΕΜ έχει διαφορετικό στόχο, ο οποίος είναι η μεγιστοποίηση του σήματος ανταμοιβής. Παρόλο που η εύρεση δομής είναι πολύ σημαντική και στην ΕΜ ώστε να μπορέσει ο πράκτορας να επιλέξει τις κατάλληλες κινήσεις, αυτό από μόνο του δεν

επιτυγχάνει τον στόχο της ΕΜ.

Ενα από τα κύρια προβλήματα της ΕΜ, που δεν συναντάται στις άλλες μορφές μηχανικής μάθησης είναι ο συμβιβασμός μεταξύ εξερεύνησης και εκμετάλλευσης. Για να αποκτήσει ο πράκτορας μεγάλη ανταμοιβή θα προτιμήσει τις ενέργειες που δοκίμασε στο παρελθόν και του προσέφεραν μεγαλύτερη ανταμοιβή. Αλλά για να ανακαλύψει τέτοιες πράξεις, πρέπει να δοκιμάσει πράξεις που δεν έχει δοκιμάσει στο παρελθόν. Έτσι ο πράκτορας πρέπει να *εκμεταλλευτεί* το τι έχει ήδη βιώσει ώστε να αποκτήσει ανταμοιβές, αλλά πρέπει και να *εξερευνήσει*, ώστε να πάρει καλύτερες αποφάσεις στο μέλλον. Έτσι το δίλημμα είναι ποια από τις δύο στρατηγικές να επιλέξει κάθε φορά, καθώς καμία δεν μπορεί επιδιωχθεί αποκλειστικά, έτσι ώστε να επιτευχθεί ο στόχος του πράκτορα. Ως αποτέλεσμα, ο πράκτορας πρέπει να δοκιμάσει διάφορες κινήσεις και προοδευτικά να προτιμήσει αυτές που θεωρεί καλύτερες.Καθώς πολλές από τα προβλήματα που θέλουμε να λύσουμε με ΕΜ είναι στοχαστικά, ο πράκτορας πρέπει να δοκιμάσει κάθε πράξη πολλές φορές για να πάρει μια αξιόπιστη εκτίμηση της προσδοκώμενης ανταμοιβής.

1.2 Στοιχεία της ΕΜ

Πέρα από την πράκτορα και το περιβάλλον, υπάρχουν ακόμα τέσσερα κύρια στοιχεία σε ένα σύστημα ΕΜ. Αυτά είναι:

- Η πολιτική (policy), η οποία ορίζει την συμπεριφορά του πράκτορα για μια δοθείσα χρονική στιγμη. Με απλά λόγια, η πολιτική είναι μια χαρτογράφηση από τις καταστάσεις που αντιλαμβάνεται ο πράκτορας στις ενέργειες που παίρνει σε αυτές τις καταστάσεις. Οι πολιτικές μπορεί να είναι και στοχαστικές, προσδιορίζοντας μια πιθανότητα για κάθε ενέργεια.

- Το σήμα ανταμοιβής, το οποίο αναφέρθηκε και νωρίτερα. Το σήμα αυτό προσδιορίζει στόχος ενός προβλήματος ΕΜ. Σε κάθε χρονικό βήμα, το περιβάλλον στέλνει στον πράκτορα έναν αριθμό, την ανταμοιβή. Ο μόνος στόχος του πράκτορα είναι να μεγιστοποιήσει την συνολική ανταμοιβή που λαμβάνει μακροπρόθεσμα. Έτσι το σήμα της ανταμοιβής προσδιορίζει ποιά είναι τα καλά και τα κακά γεγονότα για τον πράκτορα. Το σήμα είναι η κύρια βάση λόγω της οποίας αλλάζει η πολιτική. Αν ο πράκτορας επιλέξει μια ενέργεια με μικρή ανταμοιβή, τότε η πολιτική του ίσως να αλλάξει για να επιλέξει κάποια άλλη ενέργεια, όταν υπάρξει η ίδια κατάσταση στο μέλλον. Γενικά, οι ανταμοιβές μπορεί να είναι στοχαστικές συναρτήσεις της κατάστασης του περιβάλλοντος και των ενεργειών που επιλέχθηκαν.

- Η συνάρτηση αξίας κάθε κατάστασης (value function). Αντίθετα από το σήμα ανταμοιβής που μας επιστρέφει το τί είναι καλό άμεσα, η συνάρτηση αξίας προσδιορίζει τι είναι καλό μακροπρόθεσμα. Σε γενικές γραμμές, η αξία μιας κατάστασης είναι η συνολική ανταμοιβή που μπορεί να περιμένει να αποκτήσει ένας πράκτορας στο μέλλον, ξεκινώντας από την συγκεκριμένη κατάσταση. Έτσι, ενώ οι ανταμοιβές προσδιορίζουν

την άμεση και εσωτερική επιθυμητότητα των καταστάσεων του περιβάλλοντος, η αξία υποδεικνύει την μακροπρόθεσμη επιθυμητότητα των καταστάσεων, παίρνοντας υπόψιν τις καταστάσεις που θα ακολουθήσουν και τις διαθέσιμες ανταμοιβές σε αυτές τις καταστάσεις. Έτσι, μια κατάσταση με χαμηλή ανταμοιβή, μπορεί να έχει μεγάλη αξία γιατί οδηγεί σε καταστάσεις με μεγαλύτερες ανταμοιβές.

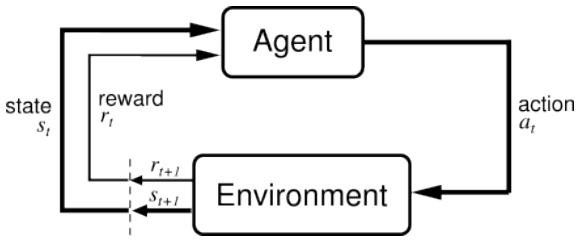
- Προαιρετικά,ένα μοντέλο του περιβάλλοντος (model). Το μοντέλο ενός συστήματος ΕΜ μιμήται την συμπεριφορά του συστήματος, ή πιο γενικά, επιτρέπει την δημιουργία συμπερασμάτων για το πώς θα συμπεριφερθεί το περιβάλλον. Τα μοντέλα χρησιμοποιούνται για σχεδιασμό (planning), δηλαδή την επίλογή της σειράς των δράσεων παίρνοντας υπόψιν πιθανές μελλοντικες καταστάσεις, πριν τις βιώσει ο πράκτορας. Μέθοδοι επίλυσης προβλημάτων ΕΜ που χρησιμοποιούν μοντέλα και σχεδιασμό λέγονται μέθοδοι βασισμένοι σε μοντέλα (model-based). Αν οι μέθοδοι δεν έχουν μοντέλο, δηλαδή μέθοδοι που μαθαίνουν ρητά μέσω δοκιμής-και-λάθους, λέγονται μέθοδοι χωρίς μοντέλο (model-free).

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 ενέργεια a_t με βάση την κατάσταση s_t, επιλεγμένη από ένα χώρο ενεργειών $\mathcal{A}$. Όταν ο πράκτορας εκτελέσει την ενέργεια, τότε τόσο το περιβάλλον, μεταβαίνει σε μια νέα κατάσταση s_{t+1}, με βάση την τρέχουσα κατάσταση και την επιλεγμένη ενέργεια [3]. Σε κάθε τριπλέτα (κατάστασης, ενέργειας, νέας κατάστασης), αντιστοιχεί μια πιθανότητα μετάβασης $\Pr\{S_{t+1} = s' | S_t = s, A_t = a\}$. Σε κάθε κατάσταση,ο πράκτορας μπορεί είναι να παρατηρήσει την πλήρη δυναμική του περιβάλλοντος ή μέρος της. Ο πράκτορας επίσης λαμβάνει και μια μονοδιάστατη ανταμοιβή R_t η οποία προέρχεται από την τριπλέτα (κατάστασης, ενέργειας, νέας κατάστασης), και συμβολίζεται ως $R(s_{t-1}, a_{t-1}, s_t)$. Αυτή η ανταμοιβή δεν είναι γνωστή στον πράκτορα στην αρχή και δρα ως μια μορφή ανατροφοδότησης για τις δράσεις του πράκτορα. Αυτή η διαδικασία αναπαριστάται οπτικά στο Σχήμα 1.2.

Συνήθως ο πράκτορας διατηρεί μια εσωτερική κατάσταση, η οποία περιλαμβάνει κομμάτια της κατάστασης του περιβάλλοντος τα οποία θεωρούνται σημαντικά, καθώς και άλλες πληροφορίες. Σε αυτή την εσωτερική κατάσταση, ο πράκτορας διατηρεί μια αντιστοίχηση μεταξύ κατάστασης και ενέργειας, η οποία συμβολίζεται ως $\Pr(a_t|s_t)$.

Η βέλτιστη σειρά ενεργειών προσδιορίζεται από 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 αποφασίζει την επόμενη ενέργεια την οποία θα κάνει ο πράκτορας. Μια βέλτιστη πολιτική είναι η πολιτική η οποία μεγιστοποιεί την αναμενόμενη απόδοση στο συγκεκριμένο περιβάλλον.

Πέρα απο την περιγραφή του συστήματος ΕΜ με βάση την ύπαρξη ή όχι μοντέλου του



Σχήμα 1.2: Αλληλεπίδραση πράκτορα και περιβάλλοντος

περιβάλλοντος, ένας άλλος τρόπος να περιγράψουμε τα συστήματα ΕΜ είναι με βάση το περιβάλλον στο οποίο βρίσκονται οι πράκτορες. Η μία περίπτωση είναι να είναι αυτό το περιβάλλον πλήρως παρατηρήσιμο, δηλαδή ο πράκτορας μπορεί να παρατηρήσει κάθε πληροφορία για την δυναμική του περιβάλλοντος. Αυτό φυσικά δεν σημαίνει ότι κάθε παρατήρηση θα είναι χρήσιμη. Έτσι η κατάσταση του πράκτορα θα είναι το υποσύνολο των παρατηρήσεων που είναι χρήσιμες. Αυτά τα περιβάλλοντα ικανοποιούν την Μαρκοβιανή ιδιότητα. Δηλαδή, για κάθε κατάσταση, το μέλλον 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). Αντίθετα, υπάρχουν περιβάλλοντα που δεν είναι πλήρως παρατηρήσιμα, όπως για παράδειγμα ένα δωμάτιο μέσα στο οποίο κινειται ένα ρομπότ. Σε αυτή την περίπτωση, το ρομπότ δεν γίνεται σε κάθε κίνηση του να γνωρίζει τα πάντα για το περιβάλλον γιατί υπάρχουν πάρα πολλές παράμετροι.

Η Μαρκοβιανή ιδιότητα ορίζεται ως:

$$p(r, s|S_t, A_t) = p(r, s|\mathcal{H}_t, A_t) \tag{1.1}$$

το οποίο σημαίνει ότι η πιθανότητα να βρεθούμε στην κατάσταση s με ανταμοιβή r, αν γνωρίζουμε ολόκληρη την ιστορία της αλληλεπίδρασης του πράκτορα με το περιβάλλον και παίρνοντας και κάνοντας την ενέργεια A_t (αριστερό μέρος) είναι ίδια με την πιθανότητα να βρεθούμε στην κατάσταση s με ανταμοιβή r γνωρίζοντας μόνο την τελευταία κατάσταση S_t στην οποία βρισκόταν ο πράκτορας και την ενέργεια A_t που έκανε (δεξί μέρος).

Σε μια Μαρκοβιανή Διαδικασία Αποφάσεων δημιουργούμε μια εκτίμηση της βέλτιστης $q_*(s, \alpha)$ κάθε ενέργειας α σε κάθε κατάσταση s ή μια εκτίμηση της αξίας $u_*(s)$ κάθε κατάστασης δεδομένης μιας βέλτιστης επιλογής ενεργειών.

Κεφάλαιο 2

Contextual Bandits

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.