Στόχος

Στόχος των εξαμηνιαίων εργασιών είναι να σας προετοιμάσουν στην **εφαρμογή αλγορίθμων Μηχανικής Μάθησης σε προβλήματα και δεδομένα του πραγματικού κόσμου.**

Επιλογή προβλήματος

Η βασική πηγή προβλημάτων για το μάθημα της Μηχανικής Μάθησης είναι το αποθετήριο Kaggle, αλλά όχι αποκλειστικά.

Πολλές εξαιρετικές εξαμηνιαίες εργασίες προέρχονται από σπουδαστές οι οποίοι επιλέγουν ένα πεδίο εφαρμογής που τους ενδιαφέρει ή επιλέγουν κάποιο υποπεδίο της Μηχανικής Μάθησης που θέλουν να διερευνήσουν περισσότερο. Συνεπώς, επιλέξτε ένα θέμα με βασικό κριτήριο το πόσο σας ενδιαφέρει το συγκεκριμένο πρόβλημα και όχι αν σας φαίνεται “εύκολο” ή ”βατό”: υπάρχουν ευκολότερα και δυσκολότερα προβλήματα ωστόσο οι απαιτήσεις είναι και αυτές ανάλογες του βαθμού ευκολίας. Μπορείτε πάντα να μας ζητήσετε βοήθεια αν δεν ξέρετε πως να ξεκινήσετε.

Kaggle

Το Kaggle είναι το γνωστότερο αποθετήριο datasets και διαγωνισμών Μηχανικής Μάθησης. Από τους διαγωνισμούς/προβλήματα του Kaggle μπορείτε να διαλέξετε ένα οποιοδήποτε πρόβλημα αρκεί να πληροί τις ακόλουθες προϋποθέσεις:

- Να ανήκει στις κατηγορίες “Featured”, “Research” ή “Playground”
- Να έχει ολοκληρωθεί (να είναι “Completed”)
- Να μην είναι παλαιότερο από 4 χρόνια

Θα βρείτε πολλους διαγωνισμούς σε θεματολογίες όπως computer vision, text mining, recommender systems, time series forecasting, brain-computer interface, medical data, earth sciences…
Διαβάστε εδώ αναλυτικά για τους διαγωνισμούς του Kaggle.

Άλλες πηγές

Μπορείτε να δουλέψετε και με προβλήματα εκτός του Kaggle. Αν για παράδειγμα είστε υποψήφιος διδάκτορας μπορείτε να ασχοληθείτε με ένα θέμα του ερευνητικού σας πεδίου στο διδακτορικό σας. Μπορεί επίσης να σας ενδιαφέρει ένα συγκεκριμένο πρόβλημα που δεν υπάρχει στο Kaggle.

Μπορείτε να δείτε μια λίστα με ιδέες εντός και εκτός Kaggle εδώ.

Τέλος, μπορείτε να βασιστείτε σε datasets που περιγράφονται σε papers ή σε άλλα projects. Δείτε για παράδειγμα τα προηγούμενα projets των τεσσάρων τελευταίων ετών του CS229 (Machine Learning) του Stanford:

2019 (Autumn)	2019 (Spring)	2018	2017	2016	2016 (Spring)

Όπως και να έχει, επιλέξτε το πρόβλημα που επιθυμείτε να μελετήσετε, δεν υπάρχει κάποιο πλεονέκτημα/μειονέκτημα στις εργασίες του Kaggle ως προς τις εργασίες εκτός του Kaggle.

Γλώσσα προγραμματισμού

Μπορείτε να δουλέψετε σε όποια γλώσσα θέλετε (πχ R) αλλά η προτεινόμενη γλώσσα είναι η Python 3.

Ανάπτυξη τοπικά ή σε cloud

Συμβουλευτείτε το Lab 1 Jupyter links για την ανάπτυξη τοπικά ή σε cloud.

In Premises

Η πρότασή μας για τοπική ανάπτυξη είναι να χρησιμοποιήσετε το PyCharm Ultimate IDE που προσφέρεται δωρεάν σε σπουδαστές. Αν δουλέψετε in-prems σε διαγωνισμό Kaggle, πρεπει να κατεβάσετε τοπικά το dataset σας.

Cloud

Στο cloud έχετε πάνω από 4 διαφορετικές επιλογές. Οδηγίες για Colab, Azure, Kernels και Datalore εδώ. Μπορείτε να διαβάσετε μια σύγκριση των επιλογών cloud εδώ.

Kaggle Competitions: ειδικές οδηγίες

Kaggle Kernels and Datasets

Αν δουλεύετε με διαγωνισμούς του Kaggle, μπορεί να σας είναι βολικό να δουλέψετε με Kaggle Kernels (Notebooks). Αφενός μπορείτε να εισάγετε απευθείας το dataset στο notebook workspace σας, αφετέρου οι πυρήνες του Kaggle έχουν καλές προδιαγραφές (CPU, GPU, disk space κλπ). Διαβάστε εδώ (μέχρι το slide

4) για το πως να εισάγετε απευθείας datasets του Kaggle σε πυρήνες και εδώ για τις τεχνικές προδιαγραφές του Kaggle σε σχέση με άλλα cloud Jupyter installations.
Μπορείτε πάντα να κατεβάσετε τα δεδομένα του Kaggle τοπικά ή όπου προτιμάτε. Πρέπει να κάνετε edit ένα νέο notebook στο Kaggle, να εισάγετε στο workspace σας το dataset του διαγωνισμού όπως εδώ και μετά από το δεξί sidebar να κατεβάσετε τα csv που σας χρειάζονται, όπως εδώ.

Documentation - API

Εδώ θα βρείτε το βρείτε τη γενική πληροφόρηση για το Kaggle.
Αν θέλετε να χρησιμοποιήσετε το Kaggle σε άλλο cloud ή τοπικά, διαβάστε εδώ για το Kaggle API.
Σε αυτό το notebook θα δείτε πως με την χρήση του API μπορούμε να εισάγουμε απευθείας δεδομένα του Kaggle στο Google Colab.

Περιορισμοί των cloud notebooks

Ειδικά στην περίπτωση που θα επιλέξετε να εκπαιδεύσετε μεγάλα μοντέλα, η εκπαίδευση μπορεί να πάρει πολύ χρόνο. Τα cloud notebooks ωστόσο έχουν κάποιους περιορισμούς: χρόνος υπολογισμού, μνήμη, χώρος, persistence των δεδομένων και GPU. Διαβάστε σχετικά εδώ, Προσέξτε ειδικά α) τον μέγιστο χρόνο εκτέλεσης κάθε πυρήνα (γενικά 6 με 10 ώρες) και β) τον μέγιστο επιτρεπόμενο χρόνο interactive session (web) που συνήθως είναι 60 λεπτά. Το API δεν έχει τον περιορισμό του interactive session.

Μία πιθανή βοήθεια αν χρειαστεί είναι να μετράτε πόσο χρόνο κάνει ένα epoch και να τρέχετε τόσα epochs ώστε να μην ξεπερνάτε το μέγιστο χρόνο υπολογισμού του cloud. Μπορείτε επίσης να αποθηκεύετε τις παραμέτρους του μοντέλου με joblib dump, pickle ή ειδικές μεθόδους. Στη συνέχεια και εφόσον το cloud έχει data persistence, μπορείτε να ξανανοίγετε το notebook και να φορτώνετε τις αποθηκευμένες παραμέτρους για να συνεχίσετε την εκπαίδευση.

Datasets, μετρικές, πυρήνες Kaggle

Γενικά

- μην διαλέξετε πολύ μικρά προβλήματα / datasets όπως αυτά που είδατε από το UCI
- μην διαλέξετε πάρα πολύ μεγάλα datasets. Αν το dataset είναι πολύ μεγάλο μπορείτε να δουλέψετε με ένα υποσύνολό του. Διαβάστε για τους περιορισμούς του κάθε cloud σε χρόνο και χώρο εδώ.
- προεπεξεργασία: στα πλαίσια του μαθήματος δεν θέλουμε να περάσετε πολύ χρόνο στην προεπεξεργασία πρωτογενών δεδομένων (raw data πχ εικόνες, κείμενα), δηλαδή μπορείτε να χρησιμοποιήσετε έτοιμα χαρακτηριστικά. Ωστόσο η εξαγωγή χαρακτηριστικών είναι σημαντικό κομμάτι της Μηχανικής Μάθησης και σας παροτρύνουμε να την κάνετε εφόσον το πρόβλημα προσφέρεται και έχετε τους απαραίτητους υλικούς και χρονικούς πόρους
- η αξιολόγηση μπορεί να γίνει με πολλούς τρόπους, όχι μόνο με το κριτήριο που χρησιμοποιείται στο leaderboard του Kaggle

Kaggle

- κοιτάξτε τους διάφορους διαθέσιμους πυρήνες για το πρόβλημα. Αν χρησιμοποιήσετε κάποιο μέρος τους πρέπει να μας το αναφέρετε

Deep Learning

Το μάθημα της Μηχανικής Μάθησης ΔΕΝ επικεντρώνεται στη Βαθιά Μάθηση. Στην NTUA ECE υπάρχει ξεχωριστό μάθημα Βαθιά Μάθηση στο εαρινό εξάμηνο όπου ζητάμε εργασίες Deep Learning. Στο μάθημά

μας, στις διαλέξεις και στο εργαστήριο, έχετε δει πολλές μεθοδολογίες εκτός Deep Learning, τις οποίες θέλουμε κυρίως να δοκιμάσετε στην παρούσα εξαμηνιαία εργασία. Αν θέλετε να δοκιμάσετε μεθοδολογίες Deep Learning θα πρέπει υποχρεωτικά να είναι σε σύγκριση με μεθόδους εκτός Deep Learning που είδατε στο μάθημα. Ιδανικά θα παρουσιάστε συγκριτικά αποτελέσματα πολλών μοντέλων (περισσότερα στα παραδοτέα).

Δήλωση θέματος και ομάδων

- Ομάδες των 3 ατόμων (καλύτερα) ή 2
- Δηλώστε το θέμα και την ομάδα σας σε αυτή τη φόρμα
- Θέματα που έχουν ήδη επιλεχθεί εδώ, μην διαλέξετε ίδια θέματα

Παραδοτέα

Τα ζητούμενα θα παραδοθούν εντός zip file στο mycourses του μαθήματος στην ενότητα “Εργασίες”.
Τα παραδοτέα είναι:

1. Αναφορά σε PDF 10 με 15 σελίδες περίπου (εκτός των βιβλιογραφικών αναφορών) στα ελληνικά. Μπορείτε να δείτε μια πρόταση περιεχομένου σε αυτό το pdf για 10 σελίδες (προσαρμόστε ανάλογα με την έκταση του κειμένου σας)
2. Κώδικας/Notebooks. Παρακαλούμε υποβάλετε τα εντός του zip ή ως αναφορά URL στο Github σας. Μέγιστο αποδεκτό μέγεθος upload mycourses 29MB, μην υποβάλετε τα δεδομένα σας

Επικοινωνία

Επικοινωνία με το μάθημα/εργαστήριο: nnlab-grad@islab.ntua.gr