ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΜΑΘΗΣΗΣ

Εργασία μηχανικής μάθησης 2020 - 2021

Ομάδες

Μπορείτε να δουλέψετε σε ομάδες των 1 ή 2 ατόμων. Για να εγγραφείτε σε ομάδα:
1. Το πρώτο μέλος της ομάδας μεταβαίνει στην περιοχή "Ομάδες χρηστών" του μαθήματος.
2. Διαλέγει μια ομάδα με 0 μέλη, μπαίνει μέσα (κλικ στο όνομα της ομάδας και κάνει εγγραφή.
3. Στέλνει τον αριθμό της ομάδας στο άλλο μέλος να κάνει το ίδιο,

Datasets

Σε κάθε ομάδα έχουν αντιστοιχηθεί με μοναδικό τρόπο τρία datasets, τα S, B, και C. Μπορείτε να βρείτε τα datasets που σας αντιστοιχούν στον πίνακα "Ομάδες - datasets". Σημειώστε τους κωδικούς "S", "B", και "C" που σας αντιστοιχούν.

Μέρη της εργασίας και βαθμολόγηση

Η εργασία αποτελείται από δύο ανεξάρτητα μέρη:

- Μέρος Α. Επιβλεπόμενη μάθηση: παλινδρόμηση. Θα χρησιμοποιήσετε τα datasets που αντιστοιχούν στα S και B. (1 μονάδα)
- Μέρος Β. Βαθιά μάθηση: συνελικτικά δίκτυα. Θα χρησιμοποιήσετε το dataset που αντιστοιχεί στο C. (1 μονάδα)

Α. Επιβλεπόμενη μάθηση: παλινδρόμηση

Δεδομένα

Έχουν προεπιλεχθεί για κάθε ομάδα δύο datasets από το αποθετήριο του UCI Machine Learning με σκοπό την μελέτη του προβλήματος της παλινδρόμησης (regression). Τα datasets έχουν χαρακτηριστεί ως μικρά (Small) και μεγάλα (Big) χρησιμοποιώντας ένα ad-hoc κριτήριο πολυπλοκότητας log(# Instances x # Attributes). Θα βρείτε τα (S,B) που σας αντιστοιχούν στον πίνακα Ομάδες - Datasets. Για να δείτε σε ποια datasets του UCI αντιστοιχούν συμβουλευτείτε τον πίνακα UCI Regression Datasets.

Τα δεδομένα του κάθε dataset προέρχονται από συλλογή δεδομένων από προβλήματα του αληθινού κόσμου. Ως τέτοια, μπορεί να έχουν απουσιάζουσες τιμές, διαφορετικούς τύπους χαρακτηριστικών, χαρακτηριστικά με υψηλή συσχέτιση μεταξύ τους ή χαμηλή διακύμανση, μεγάλο αριθμό δειγμάτων, λίγα χαρακτηριστικά κ.ο.κ. Σε κάποια dataset πρέπει να προβλέψετε την τιμή μιας μεταβλητής, σε κάποια περισσότερων (multi-output regression)

Αλγόριθμοι

Μπορείτε να χρησιμοποιήσετε αλγόριθμους παλινδρόμησης από το scikit-learn. Τους περισσότερους τους έχουμε δει στο μάθημα ή είναι μικρές παραλλαγές των βασικών αλγορίθμων.

Linear models

Ordinary Least Squares
Ridge regression
Lasso
Multi-task Lasso (πρόβλεψη περισσότερων μεταβλητών από κοινού - multiple regression)
Elastic-Net (συνδυάζει ομαλοποίηση L1 και L2)
Multi-task Elastic-Net (πρόβλεψη περισσότερων μεταβλητών από κοινού - multiple regression)
Polynomial regression

Kernel-based models

Kernel ridge regression
Support Vector regression

Stochastic Gradient Descent

Stochastic Gradient Descent regression

Instance-based models

Nearest Neighbors regression (αντί της κλάσης, υπολογίζουμε από τους γείτονες την τιμή του δείγματος)

Decision Trees[1]

Decision Trees regression (χρησιμοποιούμε MAE ή παρόμοια μετρική αντί της εντροπίας για το splitting)

Boosting - Ensemble methods

Forests of randomized trees: RandomForestRegressor, ExtraTreesRegressor
Gradient Tree Boosting: GradientBoostingRegressor
Histogram-Based Gradient Boosting: HistGradientBoostingRegressor
VotingRegressor

Multi-output algorithms

Πρόβλεψη πολλών τιμών για κάθε δείγμα
Multi-output regression: MultiOutputRegressor, RegressorChain

[1] Οι σπουδαστές του ΕΔΕΜΜ έχουν δει τα δέντρα αποφάσεων στο υποχρεωτικό μάθημα της Εξόρυξης γνώσης από δεδομένα. Αν είστε εκτός ΕΔΕΜΜ μπορείτε είτε να μην τα χρησιμοποιήσετε καθόλου είτε, εφόσον σας ενδιαφέρει, να δείτε τα βίντεο των διαλέξεων 3 και 4 στο MS Teams του μαθήματος (vj66vs1), κανάλι “Διαλέξεις - Εργαστήριο” και tab “Βίντεο”. Σας προτείνουμε το δεύτερο καθώς οι μέθοδοι Boosting βασίζονται κυρίως σε δέντρα αποφάσεων.

Neural Networks

Multi-layer Perceptron regression

Στόχος και διαδικασία

Στόχος

Για καθένα από τα δύο dataset στόχος σας είναι να βρείτε α) τη βέλτιστη αρχιτεκτονική μετασχηματιστών (στάδια προ-επεξεργασίας) και β) τις βέλτιστες υπερ-παραμέτρους (τόσο των μετασχηματιστών όσο και του αλγόριθμου παλινδρόμησης) μέσω grid search και cross validation.

Διαδικασία

Η διαδικασία της προεπεξεργασίας και αξιολόγησης είναι αντίστοιχη με αυτή που ακολουθούμε στο άλλο πρόβλημα της επιβλεπόμενης μάθησης, την ταξινόμηση, προφανώς με διαφοροποιήσεις όπως οι μετρικές αξιολόγησης. Η διαδικασία παρουσιάστηκε στην διάλεξη 10 του μαθήματος και ήταν το θέμα της προαιρετικής εργασίας του εξαμήνου.
Τα σχετικά notebooks θα τα βρείτε όλα στο φάκελο: εισαγωγικά notebooks για ταξινόμηση (υπάρχει και στο eclass στο εργαστήριο ως “00. Διαδικασία εκπαίδευσης ταξινομητών”). Αν τα ανοίξετε απευθείας στο Colab παρακαλούμε κάντε ένα δικό σας αντίγραφο προτού τρέξετε οτιδήποτε (File -> Save a copy in drive)

Ανάλυση dataset

Για κάθε dataset αναφέρετε τις ακόλουθες βασικές πληροφορίες:

1. Σύντομη παρουσίαση του dataset (τι περιγράφει).
2. Αριθμός δειγμάτων και χαρακτηριστικών, είδος χαρακτηριστικών. Υπάρχουν μη διατεταγμένα χαρακτηριστικά και ποια είναι αυτά;
3. Υπάρχουν επικεφαλίδες; Αρίθμηση γραμμών;
4. Ποια / ποιες είναι οι κολόνες με τις μεταβλητές - στόχους;
5. Χρειάστηκε να κάνετε μετατροπές στα αρχεία text και ποιες?
6. Υπάρχουν απουσιάζουσες τιμές; Πόσα είναι τα δείγματα με απουσιάζουσες τιμές και ποιο το ποσοστό τους επί του συνόλου;
7. Διαχωρίστε σε train και test set. Εάν υπάρχουν απουσιάζουσες τιμές και μη διατεταγμένα χαρακτηριστικά διαχειριστείτε τα και αιτιολογήστε τις επιλογές σας.

Ελάχιστος αριθμός μοντέλων - πρόβλεψη πολλών μεταβλητών

Θα πρέπει να εκπαιδεύσετε και συγκρίνετε αλγόριθμους κατ’ ελάχιστον:

- 3 γραμμικά μοντέλα
- 2 μοντέλα από διαφορετικά είδη μοντέλων μεταξύ των Kernel, SGD, Decision Trees, Neural Networks
- 2 μοντέλα Boosting - Ensemble

Επιπρόσθετα:
- αν εκπαιδεύσετε τα κατ’ ελάχιστον σε πλήθος μοντέλα, επιλέξτε διαφορετικά για το dataset S και το dataset B.

- Αν το πρόβλημά σας έχει να κάνει με πρόβλεψη πολλών μεταβλητών, θα χρησιμοποιήσετε αλγόριθμους multi-task και μεθοδολογίες multi-output.

Βελτιστοποίηση και σύγκριση αλγορίθμων παλινδρόμησης

1. Για κάθε αλγόριθμο βελτιστοποιήστε την απόδοσή του στο training set μέσω της διαδικασίας προεπεξεργασίας και εύρεσης βέλτιστων υπερπαραμέτρων. Κάντε εκτίμηση στο test set και τυπώστε για κάθε estimator τις μετρικές αξιολόγησής σας. Ποιες είναι οι σημαντικότερες υπερπαράμετροι του κάθε αλγορίθμου;
2. Για το τελικό fit του κάθε αλγορίθμου στο σύνολο του training set και για το predict στο test set εκτυπώστε πίνακες με τους χρόνους εκτέλεσης.
3. Για κάθε averaged metric, εκτυπώστε bar plot σύγκρισης μεταξύ όλων των αλγορίθμων.
4. Τυπώστε πίνακα με τη μεταβολή της επίδοσης των αλγορίθμων πριν και μετά τη βελτιστοποίησή τους.
5. Σχολιάστε τα αποτελέσματα των plots, των τιμών των μετρικών, τη μεταβολή της απόδοσης και τους χρόνους εκτέλεσης.

Παράδοση

Παράδοση zip στο eclass. Εκτός του ή των ipynb, υποχρεωτικά θα συμπεριλάβετε τον κώδικά python (.py) των notebooks ο οποίος αναμένεται να είναι μοναδικός εφόσον κάθε ομάδα δουλεύει σε διαφορετικό πρόβλημα. Αν σας βολεύει μπορείτε να συμπεριλάβατε σύντομη αναφορά. Όχι δεδομένα προφανώς, και τα κελιά πάντοτε εκτελεσμένα με ορατό output.

Β. Βαθιά μάθηση: συνελικτικά δίκτυα

Στόχος

Στόχος σας είναι να βελτιστοποιήσετε την απόδοση μοντέλων βαθιάς μάθησης στο σύνολο δεδομένων CIFAR-100 χρησιμοποιώντας την βιβλιοθήκη TensorFlow 2. το βασικό παραδοτέο μπορεί να είναι και μια γραπτή αναφορά (PDF, Word κλπ) που να συνοδεύεται από τα αρχεία .ipynb.

Το TF2 είναι το μόνο δεκτό ML framework για την άσκηση (δεν πρέπει να γίνει ούτε χρήση του tensorflow.compat.v1).

Δεδομένα

Για όλες τις υλοποιήσεις που θα κάνετε θα δουλέψετε με κάποιες υποκατηγορίες του CIFAR-100 που σας αντιστοιχούν στην ομάδα σας με μοναδικό τρόπο, θέτοντας στο notebook ως team_seed τον αριθμό “C” που σας αντιστοιχεί στον πίνακα “Ομάδες - datasets”.

Στα προβλήματα βαθιάς μάθησης είναι απαραίτητη η επιτάχυνση με GPU. Από τις λύσεις cloud μπορείτε να δουλέψετε είτε στο Colab είτε στο Kaggle:

- Ανοίξτε το στο Colab και κάντε ένα αντίγραφο στο drive σας
- Κάντε fork τον πυρήνα στο Kaggle

Εάν διαθέτετε κάρτα γραφικών μπορείτε να δουλέψετε τοπικά κατεβάζοντας το αρχείο ipynb.

Στο notebook θα βρείτε στην ενότητα “Εισαγωγή και επισκόπηση του συνόλου δεδομένων” τον κώδικα που σας δίνει τα ονόματα των κλάσεων σας καθώς και τα σχετικά index τα οποία πρέπει να χρησιμοποιήσετε σε όποια υλοποίηση και αν κάνετε.

Διαδικασία

Ξεκινήστε από τα μοντέλα του notebook και προχωρήστε στον ορισμό νέων μοντέλων, είτε εκ του μηδενός ("from scratch") είτε με μεταφορά μάθησης (transfer learning). Για κάθε μοντέλο συμπεριλάβετε μια σύντομη περιγραφή της αρχιτεκτονικής του και των βασικών ιδιοτήτων του. Βελτιστοποιήστε τα ακολουθώντας όσα αναφέρονται στην ενότητα “Βελτίωση της απόδοσης με πειράματα” ή/και πρόσθετες βελτιστοποιήσεις.

Περιγράψτε τη βελτιστοποίησή των μοντέλων σας, τις επιλογές και τα συμπεράσματά σας.

Παρατηρήσεις ως προς τη βελτιστοποίηση

- Ένα μοντέλο βελτιστοποιείται ως προς τον εαυτό του, όχι ως προς τα άλλα μοντέλα. Αν για παράδειγμα φέρετε με μεταφορά μάθησης ένα state-of-the-art μοντέλο, θα εξετάσετε πόσο μπορείτε να βελτιστοποιήσετε το ίδιο και δεν θα το συγκρίνετε με ένα πολύ απλό δίκτυο “from scratch”. Αυτό δεν σημαίνει ότι δεν μπορείτε να παρουσιάσετε συγκρίσεις μεταξύ μοντέλων σε πίνακες και γραφήματα (βλ. και επόμενη παρατήρηση).
- Η βελτιστοποίηση μπορεί να αφορά στην απόδοση ως προς τη μετρική που χρησιμοποιούμε (ορθότητα) στο σύνολο ελέγχου αλλά και σε άλλες ιδιότητες που έχουν να κάνουν με την εκπαίδευση: απαιτούμενη μνήμη, αριθμός παραμέτρων, χρόνος εκπαίδευσης, συμπεριφορά ως προς την υπερεκπαιδευση κ.ο.κ. Μπορείτε να παρουσιάσετε συγκριτικά αποτελέσματα “πριν και μετά” της βελτίωσης αυτών ιδιοτήτων ποσοτήτων ακόμα και για το ίδιο μοντέλο, ασχέτως της απόλυτης απόδοσής του ως προς την ορθότητα σε σχέση με άλλα μοντέλα.
- Προσπαθήστε να δοκιμάσετε όλες τις δυνατότητες βελτιστοποίησης που αναφέρονται στο notebook.

Ο ρόλος του διαφορετικού αριθμού κατηγοριών (20 μέχρι 80)

Μικρός αριθμός κατηγοριών γενικά σημαίνει ευκολότερο πρόβλημα αλλά και λιγότερα συνολικά δεδομένα και αντιθετοαντίστροφα για μεγάλο αριθμό κατηγοριών. Αυτό προφανώς έχει επίδραση στους χρόνους εκτέλεσης, στην επίδοση και στην υπερεκπαίδευση ή μη μικρών και μεγάλων δικτύων.

Μας ενδιαφέρει να δούμε τι πετυχαίνετε με 80 κατηγορίες και να μας αναφέρετε χρόνους εκπαίδευσης.

Σημειώσεις για την επίδοση των GPUs.

Σας παραθέτουμε τις μετρήσεις που λάβαμε για τους χρόνους εκτέλεσης ανά βήμα (step) -lower better- και το συνολικό AI-Score -higher better- για το απλό συνελικτικό και το VGG16 του notebook στο Colab, στο Kaggle και σε μια φυσική GeForce RTX 2080 Ti καθώς και το AI-Score τους -higher is better- συμφωνα με τη βιβλιοθήκη ai-benchmark. Το απλό CNN έχει 128.420 εκπαιδεύσιμους παράγοντες ενώ το VGG16 14.765.988.

	Colab (Tesla T4)	Kaggle (Tesla P100)	RTX 2080 Ti
simple CNN (ms/step)	15	15	7
VGG16 (ms/step)	75	53	28
AI-Score	14248	20983	28368

Παρακαλούμε αναφέρετε μας σε ποιο cloud δουλεύετε. Αν χρησιμοποιείτε φυσική GPU αναφέρετε μας το AI-Score της χρησιμοποιώντας το ai-benchmark.

Συμπληρωματικά ερωτήματα

Πέραν της προηγούμενης αναφοράς σας ως προς τα μοντέλα, τη βελτιστοποίησή και την απόδοσή τους, αναλύστε με πίνακες και γραφήματα τα ακόλουθα θέματα, πάντα με βάση τα πειράματα που εκτελέσατε:

- Μεταφορά μάθησης vs εκπαίδευση εκ του μηδενός (“from scratch”)
- Επίδραση της επαύξησης δεδομένων (data augmentation)
- Επίδραση του πλήθους των επιπέδων που θα εκπαιδευτούν (fine-tuning) κατά τη μεταφορά μάθησης
- Επίδραση του πλήθους των δεδομένων/κλάσεων στην απόδοση του μοντέλου
- Επίδραση του ρυθμού μάθησης (learning rate)
- Επίδραση του αλγόριθμου βελτιστοποίησης (optimizer)
- Επίδραση του μεγέθους δέσμης (batch size)
- Επίδραση του μεγέθους των εικόνων (resize input)

Παράδοση

Μπορείτε να δουλέψετε ελεύθερα σε δικά σας notebooks, δεν χρειάζεται απαραίτητα να επεκτείνετε απαραίτητα αυτό της εκφώνησης, αρκεί να χρησιμοποιήσετε τις προκαθορισμένες κατηγορίες εικόνων που σας έχουν αντιστοιχηθεί..

Παράδοση zip στο eclass. Εκτός του ή των ipynb, υποχρεωτικά θα συμπεριλάβετε τον κώδικά python (.py) των notebooks ο οποίος αναμένεται να είναι μοναδικός εφόσον κάθε ομάδα δουλεύει σε διαφορετικό πρόβλημα. Αν σας βολεύει μπορείτε να συμπεριλάβατε σύντομη αναφορά. Όχι δεδομένα προφανώς, και τα κελιά πάντοτε εκτελεσμένα με ορατό output.

Ερωτήσεις & απαντήσεις

Στο forum του μαθήματος στο eclass, στην περιοχή συζητήσεων της άσκησης.