2023

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

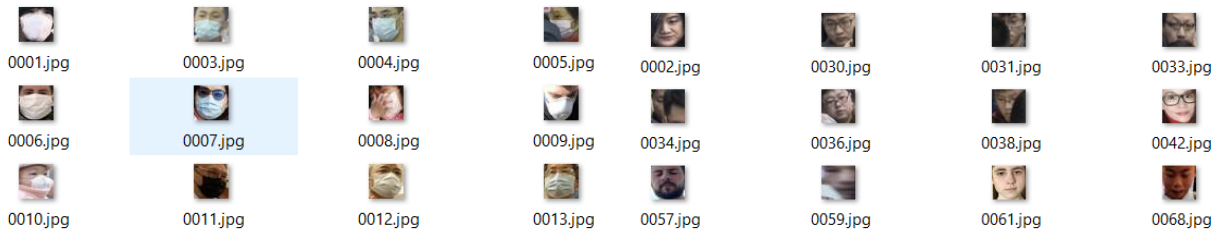
ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Υπεύθυνος Καθηγητής: Η. Θεοδωρακόπουλος

ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ

1. Μέρος 1ο – Ταξινόμηση Εικόνων (50%)

Στο πρώτο μέρος καλείστε να επιλύσετε ένα πρόβλημα ταξινόμησης εικόνων που προκύπτουν από αλγορίθμους ανίχνευσης προσώπων, με στόχο την επιβεβαίωση χρήσης μάσκας. Τα δεδομένα που θα χρησιμοποιήσετε βρίσκονται στο αρχείο Mask_DB.zip, και αποτελούνται από 1044 εικόνες προσώπων χωρίς τη χρήση μάσκας (without_mask), ισάριθμες εικόνες προσώπων που χρησιμοποιούν μάσκα (with_mask), και 56 εικόνες προσώπων που χρησιμοποιούν μάσκα εσφαλμένα (mask_incorrect_use).



a. Ως δεδομένα εκπαίδευσης, επικύρωσης και δοκιμής θα χρησιμοποιηθούν **μόνο** εικόνες από τις κλάσεις with_mask και without_mask με τις αντίστοιχες ετικέτες. Χωρίστε τα δεδομένα τυχαία σε 60% δεδομένα εκπαίδευσης, 20% δεδομένα επικύρωσης και 20% δεδομένα δοκιμής.
b. Υλοποιείστε και εκπαιδεύστε ένα σύστημα αναγνώρισης προτύπων χρησιμοποιώντας κατάλληλους αλγορίθμους της επιλογής σας συμβατικούς[1] ή/και νευρωνικούς. Δικαιολογήστε το σκεπτικό σας.
c. Καθορίστε τις βέλτιστες παραμέτρους του μοντέλου σας (πχ. αριθμό ελάττωσης διαστάσεων, παραμέτρους ταξινομητών, μοντέλα δικτύων κλπ) χρησιμοποιώντας το σύνολο επικύρωσης. Δώστε μερικά παραδείγματα παραλλαγών που δοκιμάσατε και την επίδοσή τους στο validation set. Για το καλύτερο μοντέλο σας υπολογίστε το σφάλμα ταξινόμησης στο σύνολο δοκιμής, καθώς και τις καμπύλες precision-recall και το AUC.
d. Εφαρμόστε το εκπαιδευμένο μοντέλο σας στις εικόνες της κλάσης «mask_incorrect_use» και υπολογίστε το ποσοστό αυτών που ταξινομούνται εσφαλμένα ως αποδεκτή χρήση μάσκας. Προτείνετε

[1] Στην περίπτωση χρήσης μόνο συμβατικών τεχνικών να συμπεριλάβετε και στάδιο ελάττωσης διαστάσεων.

τρόπους που μπορεί να αυξηθεί ο αριθμός των εικόνων της κλάσης mask_incorrect_use που ταξινομούνται ως μη χρήση μάσκας, χωρίς να επηρεαστούν σημαντικά τα αποτελέσματα του test set, χρησιμοποιώντας μόνο τα υπάρχοντα δεδομένα. Δοκιμάστε κάτι από αυτά που προτείνατε και δείξτε εάν έχει το επιθυμητό αποτέλεσμα.

2. Μέρος 2ο – Πρόβλεψη πρόσδεσης χημικών μορίων σε βιολογικούς υποδοχείς (50%)

Στο δεύτερο μέρος καλείστε να επιλύσετε ένα πρόβλημα πρόβλεψης του βαθμού πρόσδεσης (binding) χημικών μορίων σε βιολογικούς υποδοχείς. Η διαδικασία αυτή χρησιμοποιείται συχνά στη φαρμακολογία και τη βιοχημεία για την ανακάλυψη χημικών μορίων που δυνητικά μπορούν να αποτελέσουν ενεργές ουσίες που προσδένονται σε κάποιο υποδοχέα-στόχο και αναστέλλουν τη δράση του. Στο αρχείο Data_Receptors.zip θα βρείτε δεδομένα εκπαίδευσης και δοκιμής για 1115 και 124 χημικά μόρια αντίστοιχα, για καθένα από τα οποία έχουν υπολογισθεί 3473 χαρακτηριστικά. Τα πρώτα 1425 χαρακτηριστικά (κατά στήλες) μπορούν να λάβουν συνεχείς τιμές, ενώ τα 2048 επόμενα αποτελούν binary περιγραφείς των μορίων. Για τα δεδομένα εκπαίδευσης σας δίνονται και οι ετικέτες που καθορίζουν εάν το κάθε μόριο μπορεί να προσδεθεί αποτελεσματικά στον υποδοχέα στόχο (label 1) ή όχι (label 0).

a. Σχεδιάστε και υλοποιήστε ένα μοντέλο μηχανικής μάθησης χρησιμοποιώντας κατάλληλους αλγορίθμους της επιλογής σας, συμβατικούς ή/και νευρωνικούς, που δέχεται σαν είσοδο τα χαρακτηριστικά νέων χημικών μορίων και προβλέπει αν το κάθε μόριο μπορεί να συνδεθεί στον υποδοχέα στόχο. Δικαιολογήστε το σκεπτικό σας.
b. Τα διαθέσιμα δεδομένα για βελτιστοποίηση και εκπαίδευση του μοντέλου σας είναι μονάχα αυτά που περιλαμβάνονται στο αρχείο Train_features.csv για τα οποία έχετε διαθέσιμες ετικέτες (Train_labels.csv). Ακολουθήστε το κατάλληλο πρωτόκολλο ώστε να βελτιστοποιήσετε τις παραμέτρους του

μοντέλου σας, και να εκτιμήσετε με τον καλύτερο δυνατό τρόπο το πιθανό σφάλμα στο test set.

c. Δώστε την εκτίμησή σας για το σφάλμα που θα έχει το εκπαιδευμένο μοντέλο σας στο test set, δεδομένου ότι η κατανομή κλάσεων στο test set είναι παρόμοια με αυτή των δεδομένων εκπαίδευσης. Να εκτιμήσετε επίσης το AUC για την καμπύλη TP-FP. Τέλος, να παράξετε ένα αρχείο με όνομα *test_predictions.csv* στο οποίο θα καταχωρήσετε την πρόβλεψη του μοντέλου σας για κάθε δείγμα από το test set στη μορφή *predicted_label,prediction_score* σε κάθε γραμμή, με τη σειρά που περιέχονται τα δεδομένα στο test_features.csv. Το *prediction_score* αποτελεί την πρωτογενή έξοδο του ταξινομητή σας και πρέπει να είναι είτε η πιθανότητα πρόσδεσης είτε να έχει αύξουσα τιμή για αυξανόμενη πιθανότητα πρόσδεσης. Το αρχείο να συμπεριληφθεί μαζί με τα υπόλοιπα παραδοτέα της εργασίας στην υποβολή σας.

Σημείωση:

Σε περίπτωση που χρησιμοποιήσετε νευρωνικά δίκτυα προτείνεται η υλοποίηση με PyTorch, αλλά γίνονται δεκτές υλοποιήσεις και σε Matlab. Για τους συμβατικούς αλγορίθμους μπορείτε να επιλέξετε να υλοποιήσετε τα ζητούμενα σε Matlab ή Python.

ΠΑΡΑΔΟΤΕΑ

1. Αναλυτική αναφορά (σε pdf) που θα περιγράφει τη διαδικασία εκτέλεσης κάθε βήματος με σχολιασμό των αποτελεσμάτων, συμπεριλαμβανομένης περιγραφής των επιλογών σας για κάθε ταξινομητή. Αντί αναφοράς μπορείτε να υποβάλετε και Matlab live scripts κατάλληλα δομημένα ή python notebooks σε περίπτωση που υλοποιήσετε τα ζητούμενα σε γλώσσα Python, προσέχοντας να συμπεριλάβετε αναλυτικό σχολιασμό όπως ζητείται παραπάνω.
2. Πηγαίος κώδικας των υλοποιήσεών σας.
3. Αρχείο με τις προβλέψεις (test_predictions.csv) από το μέρος 2.
4. Αρχείο PowerPoint παρουσίασης (1-2 διαφάνειες) της εργασίας σας. Η παρουσίαση θα γίνει σε ώρα και τόπο που θα καθορισθούν.