Ενδιάμεση εργασία Χρυσολόγου Γεώργιος (ΑΕΜ: 10782)

Για την εκπόνηση της ενδιάμεσης εργασίας, επέλεξα το data set CIFAR-10 και σχεδίασα τους αλγορίθμους 1-NN, 3-NN και Nearest Centroid ο ίδιος ώστε να καλύπτουν τις ανάγκες του συγκεκριμένου data set. Το CIFAR-10 set αποτελείται από 50000 δείγματα εκπαίδευσης και 10000 δείγματα ελέγχου. Κάθε δείγμα αποτελεί ένα διάνυσμα 3072 στοιχείων . Το πρόγραμμα ,το οποίο υλοποίησα σε Python, αποθηκεύει τα 50000 δεδομένα εκπαίδευσης του σε έναν πίνακα 50000 x 3072 με όνομα data_tr, όπου κάθε γραμμή αποτελεί ένα δείγμα εκπαίδευσης. Αντίστοιχα, αποθηκεύει τα 10000 δείγματα ελέγχου σε έναν πίνακα 10000 x 3072 με όνομα data_test, όπου κάθε γραμμή αποτελεί ένα δείγμα ελέγχου. Κατόπιν, εκτελεί τους 3 αλγορίθμους ,με την σειρά, σε ολόκληρο το data set και εκτυπώνει το ποσοστό αποτυχίας και τον συνολικό χρόνο εκτέλεσης του κάθε αλγορίθμου. Παρακάτω αναφέρονται τα αποτελέσματα αυτά:

	Ποσοστό αποτυχίας(%)	Χρόνος εκτέλεσης
1-NN	74.65	115 λεπτά
3-NN	75.41	118 λεπτά
Nearest Centroid	72.26	1.45 δευτερόλεπτα
1-NN (τροποποίηση τιμών στοιχείων σε float64)	64.61	87 λεπτά
3-NN (τροποποίηση τιμών στοιχείων σε float64)	64.39	92 λεπτά
Nearest Centroid (τροποποίηση τιμών στοιχείων σε float64)	72.26	2.13 δευτερόλεπτα
1-NN (με κανονικοποίηση)	64.61	85 λεπτά
3-NN (με κανονικοποίηση	64.39	88 λεπτά
Nearest Centroid (με κανονικοποίση)	72.26	2.04 δευτερόλεπτα

Τα στοιχεία κάθε δείγματος (διάνυσμα 1 x 3072) είναι ακέραιοι αριθμοί από 1 έως 255. Λόγω της χρήσης της βιβλιοθήκης numpy, προκαλούνται σφάλματα εξαιτίας των ακεραίων αριθμών. Με την τροποποίηση τους σε float, παρατηρείται μεγάλη μείωση του ποσοστού αποτυχίας και των τριών αλγορίθμων. Τα ποσοστά αυτά είναι ίδια με την περίπτωση της κανονικοποίησης, κατά την οποία κάθε στοιχείο (ακέραιος αριθμός) διαιρείται διά 255. Αυτό είναι λογικό επειδή το εύρος των τιμών κάθε στοιχείου είναι αρκετά περιορισμένο (1- 255) οπότε η κανονικοποίηση δεν προκαλεί βελτίωση της απόδοσης (η σχετική απόσταση μεταξύ των δειγμάτων παραμένει ίδια).

Όσον αφορά τους χρόνους εκτέλεσης, συμπεραίνουμε ότι ο γρηγορότερος, με μεγάλη διαφορά, αλγόριθμος Centroid. Μεταξύ των K-NN αλγορίθμων, ο 1-NN είναι ελαφρώς ταχύτερος. Ωστόσο, τα συμπεράσματα αυτά δεν είναι απολύτως ασφαλή, καθώς σε κάθε νέο πείραμα ο χρόνος εκτέλεσης των αλγορίθμων δεν παρέμενε σταθερός (η κατάταξη τους με βάση τον χρόνο , όμως, έμενε ίδια).

ΣΗΜΑΝΤΙΚΟ : Για να τρέξει σωστά το πρόγραμμα, θα πρέπει να βρίσκεται στον ίδιο φάκελο με τα 5 data batches και το 1 test batch του CIFAR-10.