

Νευρωνικά Δίκτυα (Ενδιάμεσο Παραδοτέο)

NEAREST CENTROID & knn IN CYFAR 10

Καλλιμάνης Ιωάννη | 10007 | Ηλεκτρολόγων μηχανικών και μηχανικών ηλεκτρονικών υπολογιστών | [ikallima@ece.auth.gr](mailto:ikallima@ece.auth.gr) | 12-11-2023

Εισαγωγή

Ξεκινώντας την ενασχόληση με την εργασία, αποφάσισα να ασχοληθώ με το database της CYFAR-10 με το πρώτο αυτό ενδιάμεσο παραδοτέο να πραγματεύεται την απόδοση που μπορούν να δώσουν απλή αλγόριθμοι μηχανικής μάθησης σε ένα τόσο μεγάλο πρόβλημα classification.

Περιγραφή κώδικα

Στον κώδικα μου λοιπόν, αξιοποιώντας την συνάρτηση unpickle που μου την παρείχε έτοιμη η ιστοσελίδα της CYFAR-10 για να «φορτώσω» τα δεδομένα μου από τα εκάστοτε mini-batch που είναι αποθηκευμένα. Στην συνέχεια ορίζω της συναρτήσεις που για nearest centroid και k-nearest-neighbors και μια συνάρτηση που επιστρέφει τις μετρικές του κάθε μοντέλου (accuracy, f1\_score, recall). Στην συνέχεια φθάνουμε στην «main» του κώδικα μου όπου σώζω τα δεδομένα μου σε 4 πίνακες (2 για την διαδικασία εκπαίδευσης (δεδομένα και κλάση) και 2 για την διαδικασία εκτίμησης). Εκπαιδεύω στην συνέχεια το καθένα από τα μοντέλα μου με βάση τα δεδομένα εκπαίδευσης και τεστάρω την απόδοση τους εκτυπώνοντας τις μετρικές που αναφέρθηκαν παραπάνω. Τα αποτελέσματα που πείρα είναι:

Centroid Accuracy: 27.74%

Centroid F1 Score: 0.2541

Centroid Recall: 0.2774

1-NN Accuracy: 35.39%

1-NN F1 Score: 0.3495

1-NN Recall: 0.3539

3-NN Accuracy: 33.03%

3-NN F1 Score: 0.3192

3-NN Recall: 0.3303

Να προσθέσω, επίσης, καθώς είχα δοκιμάσει τον κώδικα μου ώστε να δω τα αποτελέ-

σματα του, λόγω των «χαμηλών» μετρικών πρόσθεσα μέσω της pca ένα απλουστευμένο

2D διάγραμμα ώστε να δω πως είναι τα δεδομένα μου στον δισδιάστατο χώρο.(σχ. 1)

Συμπεράσματα

Από το διάγραμμα αυτό, διαπίστωσα ότι τα δεδομένα μου είναι όλα λίγο-πολύ κοντά μεταξύ τους και συγκεντρωμένα σε μια περιοχή, οπότε είναι αρκετά δύσκολο με αλγόριθμους μηχανικής μάθησης σαν τους 2 που χρησιμοποιήθηκαν να πάρουμε ικανοποιητικά υψηλές μετρικές

Σημειώσεις

Ο χρόνος εκτέλεσης του κώδικα μου ανέρχεται στα 5-7 λεπτά. Για προσωπικούς λόγους έτρεξα τον κώδικα μέχρι k = 17 nearest neighbors (ουσιαστικά τους 7 πρώτους πρώτους αριθμούς εξαιρώντας το 2 [1 3 5 7 11 13 17], χρησιμοποιήθηκαν πρώτοι αριθμοί για να έχω πάντα μια επικρατούσα κλάση στους μεγάλους αριθμούς, και 7 καθώς με 1 ακόμα από την centroid να τρέξουν με multi-processing (ή και multithreading) στους 8 πυρήνες του λάπτοπ μου) και διαπίστωσα τουλάχιστον στην μετρική accuracy το καλύτερο μοντέλο να είναι αυτό με k = 1, με μικρές διαφορές της τάξης έως 3-4% από κάθε μοντέλο από τα υπόλοιπα. Καθώς όμως καθυστερούσε αρκετά παραπάνω το πρόγραμμα μου προτίμησα να μην τα συμπεριλάβω στο τελικό παραδοτέο μου

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, διάγραμμα, πολυχρωμία

Περιγραφή που δημιουργήθηκε αυτόματα(σχ. 1)