Νευρωνικά Δίκτυα - Βαθιά Μάθηση
Ενδιάμεση Εργασία

8 Νοεμβρίου 2023

Δημήτριος Αλεξόπουλος
aadimitri@ece.auth.gr
ΑΕΜ 10091

Περιεχόμενα

1 Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η σύγκριση της απόδοσης του κατηγοριοποιητή πλησιέστερου γείτονα (Nearest Neighbor Classifier) με 1 και 3 πλησιέστερους γείτονες με τον κατηγοριοποιητή πλησιέστερου κέντρου (Nearest Centroid Classifier) στην βάση δεδομένων της επιλογής μας.
Για τους σκοπούς της εργασίας επιλέγεται η βάση δεδομένων $CIFAR-10$, η οποία αποτελείται από 60000 έγχρωμες εικόνες 32×32 σε 10 κλάσεις, με 6000 εικόνες ανά κλάση. Υπάρχουν 50000 εικόνες εκπαίδευσης και 10000 εικόνες δοκιμής.
Το σύνολο δεδομένων χωρίζεται σε πέντε παρτίδες εκπαίδευσης και μία παρτίδα δοκιμής, κάθε μία με 10000 εικόνες. Η παρτίδα δοκιμής περιέχει ακριβώς 1000 τυχαία επιλεγμένες εικόνες από κάθε κλάση. Οι παρτίδες εκπαίδευσης περιέχουν τις υπόλοιπες εικόνες με τυχαία σειρά, αλλά ορισμένες παρτίδες εκπαίδευσης μπορεί να περιέχουν περισσότερες εικόνες από μια κλάση από ό,τι από μια άλλη. Μεταξύ τους, οι παρτίδες εκπαίδευσης περιέχουν ακριβώς 5000 εικόνες από κάθε κλάση.
Παρακάτω θα διαμορφωθεί κατάλληλα το σύνολο των δεδομένων, ώστε να διαβάζουμε τα δεδομένα εκπαίδευσης (*training*) και τα δεδομένα ελέγχου (*test*) και να μετράμε την απόδοση των δύο κατηγοριοποιητών.

2 Nearest Neighbor Classifier

Ο ταξινομητής πλησιέστερου γείτονα είναι ένας απλός αλλά αποτελεσματικός αλγόριθμος για την επιβλεπόμενη μηχανική μάθηση, καθώς ταξινομεί ένα στοιχείο από τα δεδομένα εισόδου εντοπίζοντας το/τα πλησιέστερο/-α στοιχείο/-α δεδομένων στο σύνολο δεδομένων εκπαίδευσης και αναθέτοντας την ετικέτα κλάσης του/των πλησιέστερου/-ων γείτονα/-ων. Η βασική υπόθεση είναι ότι παρόμοια στοιχεία δεδομένων στο χώρο χαρακτηριστικών (feature space) πρέπει να ανήκουν στην ίδια κλάση.

2.1 Μαθηματική Ανάλυση

Ας συμβολίσουμε το σύνολο δεδομένων εκπαίδευσης ως ένα σύνολο στοιχείων με τις αντίστοιχες ετικέτες:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\},$$

όπου το x_i αντιπροσωπεύει ένα διάνυσμα χαρακτηριστικών και το y_i είναι η ετικέτα κλάσης που σχετίζεται με το x_i. Ο στόχος είναι η ταξινόμηση ενός νέου στοιχείου δεδομένων εισόδου, το οποίο συμβολίζεται ως x_{new}.
Η βασική ιδέα πίσω από τον ταξινομητή πλησιέστερου γείτονα είναι να βρεθεί το στοιχείο δεδομένων εκπαίδευσης, x_i, το οποίο είναι πλησιέστερο στο x_{new} από την άποψη μιας μετρικής απόστασης, συνήθως της ευκλείδειας απόστασης (L_2 *norm*):

$$d(x_i, x_{new}) = \sqrt{\sum_{j=1}^{d} (x_i^{(j)} - x_{new}^{(j)})^2},$$

όπου d είναι ο αριθμός των χαρακτηριστικών (features) και $x_i^{(j)}$ αντιπροσωπεύει την j-οστή συνιστώσα του διανύσματος χαρακτηριστικών x_i.

Μόλις βρεθεί ο πλησιέστερος γείτονας, ο ταξινομητής αναθέτει την ετικέτα κλάσης του x_i στο x_{new}.

2.2 Υλοποίηση

Για την εκτέλεση του αλγορίθμου του κατηγοριοποιητή πλησιεστέρου γείτονα κατασκευάσαμε την συνάρτηση *accuracy = train_and_evaluate_knn*(*X_train*, *y_train*, *X_test*, *y_test*, *k* = 1), με τα εξής ορίσματα κι εξόδους:

- X_{train}: το σύνολο εκπαίδευσης που περιλαμβάνει τα χαρακτηριστικά των δεδομένων εκπαίδευσης
- y_{train}: οι ετικέτες των κατηγοριών των δεδομένων εκπαίδευσης
- X_{test}: το σύνολο δοκιμής που περιλαμβάνει τα χαρακτηριστικά των δεδομένων δοκιμής
- y_{test}: οι αναμενόμενες ετικέτες των κατηγοριών των δεδομένων δοκιμής
- k: το πλήθος των κοντινότερων γειτόνων που θα χρησιμοποιηθεί για την αξιολόγηση (Προεπιλεγμένη τιμή: 1)
- *accuracy*: το ποσοστό των σωστών προβλέψεων του ταξινομητή k-κοντινότερων γειτόνων ($K-NN$) στα δεδομένα δοκιμής

3 Nearest Centroid Classifier

Ο ταξινομητής πλησιέστερου κέντρου είναι επίσης ένας απλός αλλά αποτελεσματικός αλγόριθμος επιβλεπόμενης μηχανικής μάθησης. Η πρωταρχική του ιδέα περιστρέφεται γύρω από τη χρήση των κέντρων κάθε κλάσης στο χώρο των χαρακτηριστικών (feature space) για την ταξινόμηση νέων στοιχείων δεδομένων.
Σε αντίθεση με τον αλγόριθμο k-Nearest Neighbors (K-NN) που βασίζεται στις αποστάσεις από τους πλησιέστερους γείτονες, ο ταξινομητής Nearest Centroid χρησιμοποιεί τις θέσεις των κέντρων των κλάσεων για την ταξινόμηση. Το μετρικό της απόστασης υπολογίζεται με ανάλογο τρόπο με τον ταξινομητή πλησιεστέρων γειτόνων, όπως είδαμε παραπάνω.

3.1 Υλοποίηση

Για την εκτέλεση του αλγορίθμου του κατηγοριοποιητή πλησιεστέρου κέντρου κατασκευάσαμε την συνάρτηση *accuracy = train_and_evaluate_nearest_centroid*(*X_train*, *y_train*, *X_test*, *y_test*), με τα εξής ορίσματα κι εξόδους:

- X_{train}: το σύνολο εκπαίδευσης που περιλαμβάνει τα χαρακτηριστικά των δεδομένων εκπαίδευσης
- y_{train}: οι ετικέτες των κατηγοριών των δεδομένων εκπαίδευσης
- X_{test}: το σύνολο δοκιμής που περιλαμβάνει τα χαρακτηριστικά των δεδομένων δοκιμής
- y_{test}: οι αναμενόμενες ετικέτες των κατηγοριών των δεδομένων δοκιμής

- ***accuracy***: το ποσοστό των σωστών προβλέψεων του ταξινομητή κοντινότερου κέντρου στα δεδομένα δοκιμής

4 Αποτελέσματα - Σχολιασμοί

Εκτελούμε τους δύο παραπάνω αλγορίθμους ταξινόμησης στο σύνολο δεδομένων CIFAR-10, τον αλγόριθμο πλησιέστερων γειτόνας με τιμές $k = 1$ ή $k = 3$ και τον αλγόριθμο πλησιέστερου κέντρου. Τα ποσοστά των σωστών προβλέψεων σε κάθε περίπτωση φαίνονται στον παρακάτω πίνακα:

Classifier	*Accuracy*
K-Nearest Neighbors (k=1)	0.3539
K-Nearest Neighbors (k=3)	0.3303
Nearest Centroid Classifier	0.2774

Σχολιασμοί:

- Παρατηρούμε ότι ο αλγόριθμος πλησιεστέρων γειτόνων έχει καλύτερη απόδοση σε σύγκριση με τον αλγόριθμο πλησιέστερου κέντρου. Μάλιστα, την καλύτερη απόδοση όλων εμφανίζει ο αλγόριθμος πλησιέστερων γειτόνων με μόλις $k = 1$, δηλαδή ο αλγόριθμος που ταξινομεί ένα νέο στοιχείο των δεδομένων μας στην κλάση που περιέχει των πλησιέστερό του (σε χαρακτηριστικά) γείτονα.

- Το γεγονός ότι ο συμψηφισμός περισσότερων του ενός γειτόνων οδηγεί σε λιγότερο αποδοτικό αποτέλεσμα ενδεχομένως οφείλεται στο ότι τα δεδομένα έχουν πολλές παρεμφερείς λεπτομέρειες ή σημαντική διακύμανση εντός της ίδιας κλάσης. Πιθανόν οι τιμές των πίξελς των φωτογραφιών δεν αποτελούν ικανοποιητικό μετρικό για την ταξινόμηση των στοιχείων. Θα μπορούσαν να προταθούν τεχνικές ψηφιακής επεξεργασίας εικόνας, όπως εξαγωγή περιγράμματος αντικειμένου, που θα αύξαναν την αποδοτικότητα του αλγορίθμου.

- Όσον αφορά την υπολογιστική πολυπλοκότητα, ο αλγόριθμος πλησιέστερου κέντρου είναι σαφώς ταχύτερος από τον αλγόριθμο πλησιέστερων γειτόνων.

Αναφορές

[1] CIFAR-10 and CIFAR-100 datasets. *Canadian Institute for Advanced Research (CIFAR)*. https://www.cs.toronto.edu/~kriz/cifar.html

[2] K-nearest neighbors algorithm. *Wikipedia*. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[3] Nearest centroid classifier. *Wikipedia*. https://en.wikipedia.org/wiki/Nearest_centroid_classifier