ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΒΑΘΙΑ ΜΑΘΗΣΗ
3Η ΕΡΓΑΣΙΑ

Ιωάννης Δεϊρμεντζόγλου 10015

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΕΜ 10015 Email deirmentz@ece.auth.gr

ΕΙΣΑΓΩΓΗ

Στην πρώτη εργασία του μαθήματος παρουσιάστηκαν τα αποτελέσματα εφαρμογής κατηγοριοποιητών Nearest Centroid, k-Nearest Neighbor, MultiLayer Perceptron και Convolutional Neural Network στη βάση δεδομένων CIFAR-10, όπου η σύγκριση έδειξε ότι το καλύτερο μοντέλο ήταν αυτό ενός CNN.

Στην 2η εργασία ζητούμενο είναι κατηγοριοποίηση με χρήση ενός SVM για διαχωρισμό 2 ή περισσοτέρων κλάσεων. Το dataset που χρησιμοποιήθηκε ήταν το UCI – Breast Cancer Wisconsin **.** Πρόκειται για ένα σύνολο δεδομένων που χρησιμοποιείται για την ταξινόμηση καρκινικών κυττάρων του μαστού σε κακοήθη- Malignant(M) και καλοήθη- Benign(B).Τα αποτελέσματα που προέκυψαν ήταν ικανοποιητικά και φάνηκε ότι το SVM ήταν αποτελεσματικό στο συγκεκριμένο binary classification πρόβλημα

Σε αυτήν την εργασία του μαθήματος επιλέχθηκε να υλοποιηθεί ένα RBF NN (Radial Basis Function Neural Network) για ένα multiclass classification πρόβλημα. Σε αυτήν την εργασία χρησιμοποιήθηκε η βάση δεδομένων cifar 10 . Η CIFAR-10 περιλαμβάνει 60.000 εικόνες σε 10 διαφορετικές κατηγορίες. Κάθε εικόνα ανήκει σε μία από τις κατηγορίες: αεροπλάνα, αυτοκίνητα, πουλιά, γάτες, ελάφια, σκύλοι, βατράχια, ίπποι, πλοία και φορτηγά. Οι εικόνες έχουν διαστάσεις 32x32x3 (3 χρωματικά κανάλια – RGB εικόνες) .

Προεπεξεργασία των δεδομένων

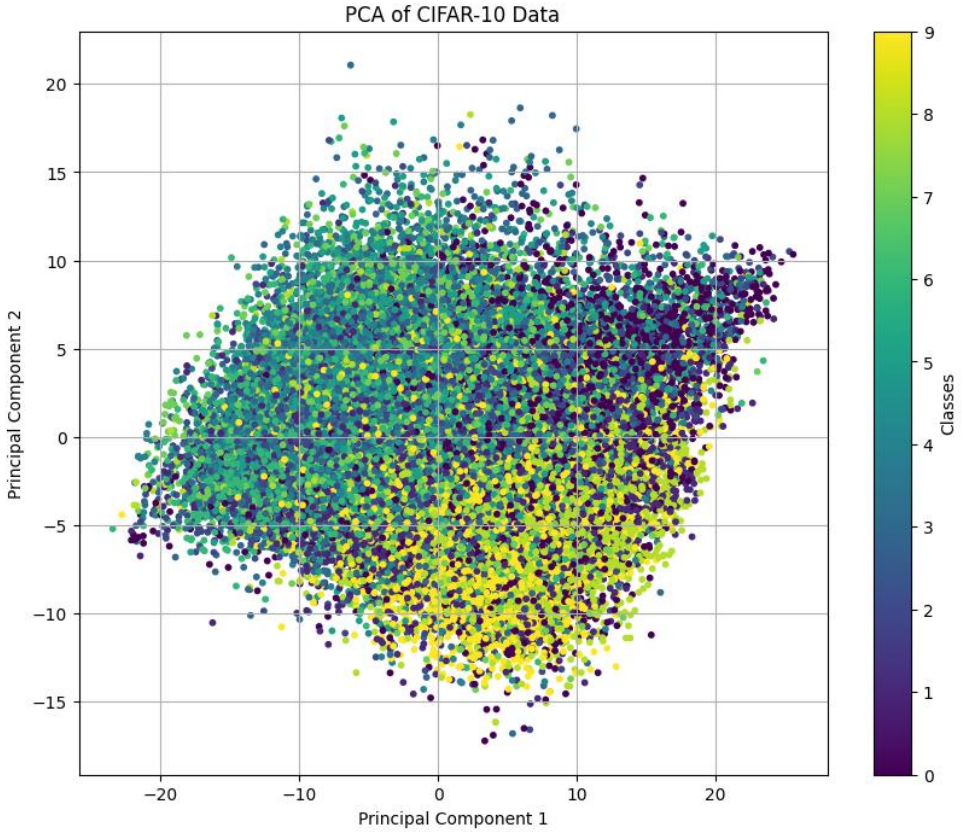
Αρχικά , τα δεδομένα που φορτώνονται από την cifar 10 χωρίζονται σε train και test set με διαστάσεις 50.000x32x32x3 (50k εικόνες) και 10.000x32x32x3 αντίστοιχα . Οι εικόνες προκειμένου να κατηγοριοποιηθούν , κανονικοποιουνται μεταξύ 0 και 1 και μετατρέπονται σε 1d πίνακες. Επίσης , στα labels του train και test εφαρμόζεται one hot encoding , ώστε το μοντέλο να διαχωρίσει καλυτέρα τα labels μεταξύ τους και να έχει καλύτερη απόδοση .

Ανάλυση κυρίων συνιστωσών (PCA)

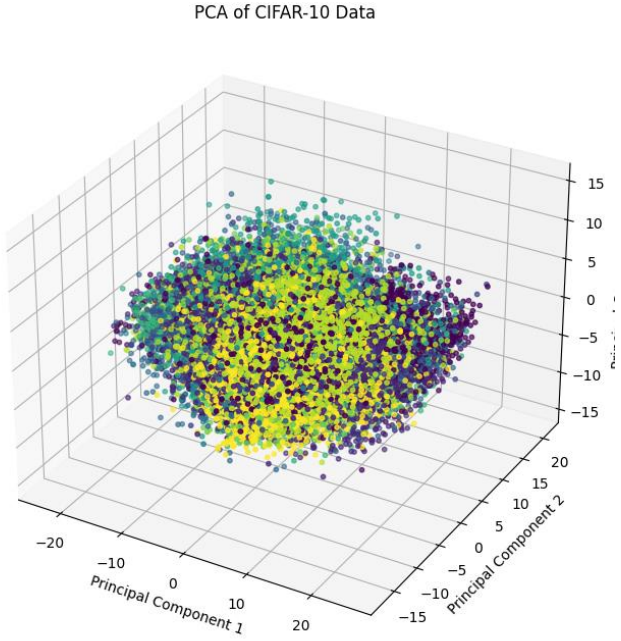
Με ανάλυση κύριων συνιστωσών μειώνεται ο αριθμός των features τέτοιος ώστε να διατηρείται το 90% ή αντίστοιχα το 95 % της διακύμανσης των αρχικών δεδομένων . Γενικά , στόχος είναι να δημιουργήσουμε νέα features που προκύπτουν από το αρχικό σύνολο(3072 για την cifar 10) με σκοπό να διατηρηθεί όσο περισσότερη πληροφορία γίνεται σε πολύ μικρότερο αριθμό features (έχουμε 3072 features χωρίς PCA).

2D and 3D visualization using PCA

Τα παρακάτω σχήματα αναπαριστούν την εφαρμογή του PCA για μείωση διαστάσεων στο σύνολο δεδομένων CIFAR-10. Δημιουργούνται τόσο 2D όσο και 3D scatter plots για να προβάλει το σύνολο δεδομένων σε μειωμένες διαστάσεις (2 και 3 principal components). Η γραφική παράσταση 2D βοηθά στην οπτικοποίηση των δεδομένων με δύο principal components, ενώ η τρισδιάστατη γραφική παράσταση παρέχει πρόσθετη εικόνα ενσωματώνοντας ένα 3ο principal component.



Σχήμα 1 : 2D visualization



Σχήμα 2 : 3D visualization

Παρατηρείται ότι οι κλασεις δεν φαίνονται να είναι διαχωρίσιμες με μόλις 2 ή 3 components και φαίνεται να χάνεται αρκετή πληροφορία κάτι που είναι λογικό αφού για διατήρηση ενός ικανοποιητικού ποσοστού πχ. 95% απαιτούνται 217 features .

1. KNN και Nearest Centroid Κατηγοριοποιητές

Παρακάτω παρουσιάζονται τα αποτελέσματα εφαρμογής των κατηγοριοποιητών Nearest Centroid και k-Nearest Neighbor στην **CIFAR-10.** Τα αποτελέσματα δεν ήταν καθόλου ικανοποιητικά.

Παρακάτω παρουσιάζονται τα αποτελέσματα των κατηγοριοποιητών :

- **KNN-- Nearest Neighbor with k =1 Neighbor**

```
Training time: 0.181 seconds

Testing time: 17.228 seconds

For the 1-NN classifier accuracy is: 0.1660
For the 1-NN classifier, F1 score is: 0.1661
```

- **KNN-- Nearest Neighbor with k=3 Neighbors**

```
Training time: 0.081 seconds

Testing time: 18.975 seconds

For the 3-NN classifier accuracy is: 0.0956
For the 3-NN classifier, F1 score is: 0.1332
```

- **Nearest Centroid with Euclidean Distance metric**

```
Training time: 0.014 seconds

Testing time: 0.079 seconds

For the Nearest Centroid classifier accuracy is: 0.0956
For the Nearest Centroid classifier, F1 score is: 0.2004
```

- **Nearest Centroid with Manhattan Distance metric**

```
Training time: 0.019 seconds

Testing time: 0.015 seconds

For the Nearest Centroid classifier accuracy is: 0.0956
For the Nearest Centroid classifier, F1 score is: 0.2054
```

Παρατηρείται ότι τα αποτελέσματα δεν είναι ικανοποιητικά και οι παραπάνω κατηγοριοποιητες δεν διαχωρίζουν σωστά τις εικόνες του cifar 10 . Επίσης , φαίνεται ότι ο κατηγοριοποιητες Nearest Centroid απαιτεί λιγότερο χρόνο testing και έχει καλύτερο accuracy score .

2. RBF Neural Network

Ένα **RBF Neural Network** για ταξινόμηση πολλαπλών κλάσεων αποτελείται συνήθως από ένα στρώμα εισόδου (**Input Layer**) που λαμβάνει χαρακτηριστικά δεδομένων εισόδου, κρυφό στρώμα (**Hidden Layer**) το οποίο μετασχηματίζει κάθε είσοδο με μια Radial Basis Function (μετατροπή την εισόδου σε χώρο υψηλότερης διάστασης) και ένα στρώμα εξόδου **(Output Layer)** το οποίο εκτελεί τον κλασσικό γραμμικό ΜΣ στην έξοδο του RBF στρώματος. Η βασική διαφορά ενός RBF δικτύου με ένα MLP είναι ότι το RBF μοντέλο έχει δύο στάδια εκπαίδευσης. Πρώτα υπολογίζονται τα κέντρα μέσω του kmeans και η διασπορά σ κάθε νευρώνα, και ύστερα εκπαιδεύεται το υπόλοιπο δίκτυο με Back Propagation. Συγκεκριμένα, τα κέντρα θα υπολογιστούν με χρήση του αλγορίθμου **KMeans** της sklearn, και ως σ επιλέγεται η εξής τιμή για όλους τους νευρώνες:

Χρήσιμες συναρτήσεις για την υλοποίηση

- **rbfKernel ():** Η συνάρτηση αυτή υπολογίζει τον πίνακα πυρήνα Radial Basis Function (RBF) μεταξύ δύο συνόλων X και Y. Ο πυρήνας RBF υπολογίζεται μετρώντας την τετραγωνική Ευκλείδεια απόσταση μεταξύ κάθε δυνατού ζεύγους διανυσμάτων και εφαρμόζοντας τον τύπο του πυρήνα RBF. Επιστρέφει τον πίνακα K . Στην υλοποίηση της εργασίας οπού X είναι το train set και οπού Y το σύνολο των κέντρων που υπολογίζονται από τον Kmeans.
- **rbfKernelDerivative ():** Η συνάρτηση υπολογίζει την παράγωγο του πυρήνα Radial Basis Function (RBF) ως προς X μεταξύ δύο συνόλων μεταξύ δύο συνόλων X και Y.
- **f1Score ():** Η συνάρτηση υπολογίζει την μετρική f1 score.
- **MSE ():** Η συνάρτηση υπολογίζει το Mean Squared Error.
- **MSEDerivative ():** Η συνάρτηση υπολογίζει την παράγωγο του Mean Squared Error.

RBF Implementation

Η κλάση RBF υλοποιεί ένα νευρωνικό δίκτυο Radial Basis Function (RBF) για ταξινόμηση. Ξεκινά χρησιμοποιώντας τον αλγόριθμο K-Means για τον υπολογισμό των κέντρων για τους πυρήνες RBF με βάση τον καθορισμένο αριθμό κέντρων (numCenters). Αυτά τα κέντρα προέρχονται από το σύνολο δεδομένων εκπαίδευσης (x_train). Ο κώδικας προχωρά στον υπολογισμό της διακύμανσης , η οποία υπολογίζεται ως σ **= maxDistance/√2**P **,** όπου **P**: πλήθος Centroid, και **maxDistance**: μέγιστη απόσταση μεταξύ οποιονδήποτε 2 Centroid. Η επιλογή αυτή για τις διασπορές είναι η πιο συνηθισμένη. Στην συνέχεια , υπολογίζονται τα w (βάρη) ως εξής : υπολογίζεται αρχικά κάθε ευκλείδεια απόσταση μεταξύ δειγμάτων του XTrain και των κέντρων (που προέκυψαν από τον kmeans) και στην συνέχεια τα βάρη προκύπτουν από το γινόμενο του yTrain με τον ψευδο-αντίστροφο του kernelXTrain. Ο πίνακας b περιέχει τα bias για κάθε νευρώνα εξόδου . Στην συνέχεια , εάν η μεταβλητή trainEpochs είναι true τότε το δίκτυο εκπαιδεύεται για έναν συγκεκριμένο

αριθμό εποχών οπού τα βάρη ανανεώνονται με βάση τα errors στην λογική του backpropagation. Τέλος , υπολογίζεται αρχικά κάθε ευκλείδεια απόσταση μεταξύ δειγμάτων του XTest και των κέντρων (που προέκυψαν από τον kmeans) και με βάση τα τελικά βάρη w προκύπτουν οι προβλέψεις για το ytest υπολογίζεται το accuracy score . Εάν η μεταβλητή metrics είναι true εκτυπώνονται classification report , confusion matrix κλπ. .

Αποτελέσματα

Για το κάθε ένα δίκτυο ο αριθμός των εποχών τέθηκε ίσος με 20 ενώ το learning rate στο 0.01 . Ο αριθμός των κέντρων στην πρώτη δοκιμή έγινε για κέντρα ίσα με τον αριθμό των features του dataset μετα το PCA δηλαδή 217. Τα αποτελέσματα που προέκυψαν παρατίθενται παρακάτω :

```
Number of centers: 217

KMeans elapsed time: 255.874 seconds
Transforming training data: 115.675 seconds
Calculating weights: 3.166 seconds
```

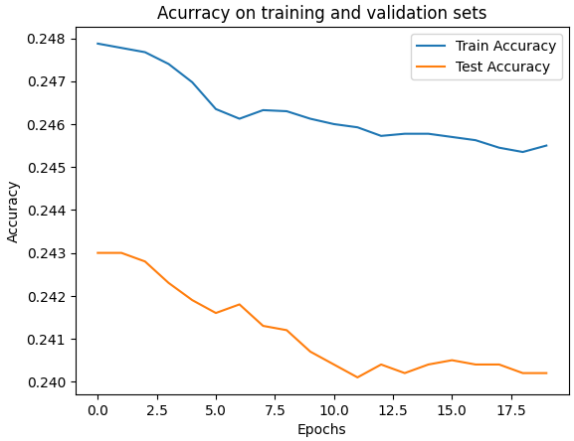
Βλέπουμε πως η διαδικασία είναι αρκετά χρονοβόρα , ειδικά ο υπολογισμός των κέντρων από τον Kmeans . Η συνολική διαδικασία διήρκησε 7 λεπτά .

```
Training set accuracy:  0.24438
Test set accuracy:  0.2477
```

Το accuracy δεν είναι υψηλό , με αποτέλεσμα εκτός τον μεγάλο χρόνο που απαιτείται το δίκτυο να μην φαίνεται να παρέχει και ικανοποιητικά αποτελέσματα. Παρακατω παρατίθεται και το classification report οπού τα αποτελέσματα κάθε άλλο πάρα ικανοποιητικά είναι :

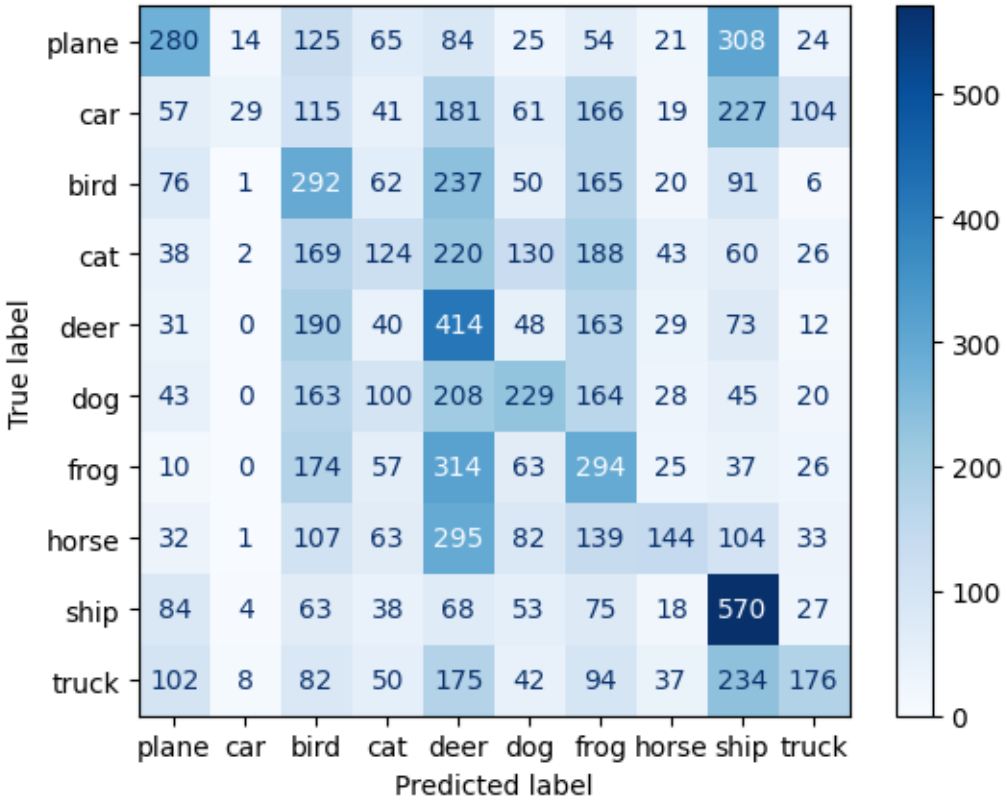
	precision	recall	f1-score	support
0	0.36	0.31	0.33	1000
1	0.35	0.04	0.08	1000
2	0.18	0.34	0.24	1000
3	0.19	0.12	0.15	1000
4	0.21	0.39	0.27	1000
5	0.28	0.18	0.22	1000
6	0.18	0.28	0.22	1000
7	0.43	0.11	0.18	1000
8	0.30	0.56	0.40	1000
9	0.39	0.13	0.20	1000

Εικόνα : Classification Report for 217 centers



Εικόνα: Accuracy on train and validation sets

Όπως φαίνεται και από το σχήμα , παρατηρήθηκε ότι η εκπαίδευση με epochs δεν βελτίωνε ιδιαίτερα το accuracy του validation (κάποιες δοκιμές το μείωναν κιόλας-πιθανό σφάλμα στην υλοποίηση) , οπότε οι επόμενες δοκιμές έγιναν χωρίς εκπαίδευση και ανανέωση των βαρών, ώστε να μειωθεί και ο χρόνος . Τέλος , παρατίθεται και ο confusion matrix :

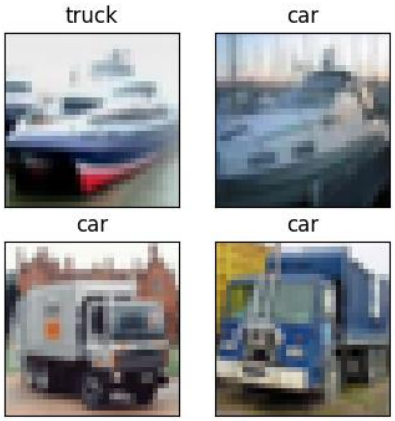


Εικόνα: Confusion Matrix

Και από τον πίνακα σύγχυσης εξάγεται το συμπέρασμα ότι το μοντέλο δεν έχει καλή επίδοση. Παρατηρείται ότι μόνο το πλοίο και λίγο το ελάφι κατηγοριοποιούνται σε σχετικά ικανοποιητικό βαθμό. Το αμάξι φαίνεται να έχει την χειρότερη κατηγοριοποίηση οπού δεν αναγνωρίζεται σχεδόν καθόλου.

Παρακατω παρουσιάζονται παραδείγματα σωστής και λανθασμένης κατηγοριοποίησης :

Εικόνα : Σωστή κατηγοριοποίηση



Εικόνα : Λανθασμένη κατηγοριοποίηση

Στην συνέχεια , έγιναν δοκιμές με διαφορετικό αριθμό κέντρων για σύγκριση .

- **numCenters = 10**

```
Number of centers: 10

KMeans elapsed time: 43.729 seconds
Transforming training data: 5.692 seconds
Calculating weights: 0.045 seconds
Training set accuracy:  0.20038
Test set accuracy:  0.2092
```

- **numCenters = 20**

```
Number of centers: 20

KMeans elapsed time: 86.443 seconds
Transforming training data: 25.820 seconds
Calculating weights: 0.216 seconds
Training set accuracy:  0.21194
Test set accuracy:  0.2194
```

- **numCenters = 100**

```
Number of centers: 100

KMeans elapsed time: 163.853 seconds
Transforming training data: 53.104 seconds
Calculating weights: 0.874 seconds
Testing elapsed time: 11.146 seconds

Training set accuracy:  0.24816
Test set accuracy:  0.2505
```

- **numCenters = 1000**

```
Number of centers: 1000

KMeans elapsed time: 1068.091 seconds
Transforming training data: 578.899 seconds
Calculating weights: 28.269 seconds

Training set accuracy:  0.2662
Test set accuracy:  0.2509
```

Παρατηρείται ότι με αύξηση του αριθμού των κέντρων αυξάνεται και το accuracy score χωρίς όμως κανένα να πλησιάσει σε ικανοποιητικό επίπεδο πάρα τον μεγάλο αριθμό κέντρων (για 1000 κέντρα μόνο 25% accuracy στο τεστ σετ). Είναι όμως εμφανές ότι βελτιώνει την επίδοση η αύξηση των σε κάποιο βαθμό. Παρόλα αυτά , αυξάνεται ραγδαία και ο χρόνος υπολογισμών , ειδικά αυτός του υπολογισμού των κέντρων.

Γενικά , φαίνεται τα δίκτυα RBF να λειτουργούν ελαφρώς πιο αποτελεσματικά από τους κατηγοριοποιητες KNN και Nearest Centroid αλλά είναι αρκετά πιο χρονοβόρα. Το γενικό συμπέρασμα είναι ότι και ακόμα για αρκετά μεγάλο αριθμό κέντρων δεν φαίνεται να λειτουργούν ικανοποιητικά στο dataset της cifar 10 .