ΗΜΜΥ ΕΜΠ

Αναγνώριση Προτύπων (2022-2023)

Προπαρασκευή 1ης Εργαστηριακής Άσκησης

Ηλιόπουλος Γεώργιος:	03118815	giliopoulos301@gmail.com
Σερλής Εμμανουήλ Αναστάσιος:	03118125	manosserlis@gmail.com

Θέμα: Οπτική Αναγνώριση Ψηφίων

Σκοπός είναι η υλοποίηση ενός συστήματος οπτικής αναγνώρισης ψηφίων. Τα δεδομένα προέρχονται από την US Postal Service (γραμμένα στο χέρι σε ταχυδρομικούς φακέλους και σκαναρισμένα) και περιέχουν τα ψηφία από το 0 έως το 9 και διακρίνονται σε train και test.

Βήμα 1: Εισαγωγή των δεδομένων

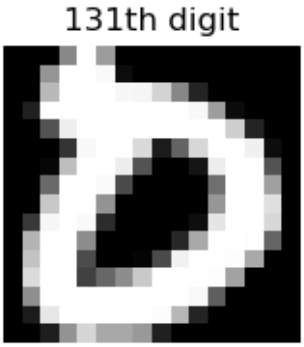
Στο βήμα έγινε η μεταβίβαση των δεδομένων από `txt` αρχεία σε πίνακες. Αρχικά τα δεδομένα περάστηκαν ως ένα data frame κάνοντας χρήση της βιβλιοθήκης pandas και στη συνέχεια μετατράπηκαν σε `NumPy arrays` διάστασης $N_{samples} \times 256$, όπου 256 είναι ο αριθμός των χαρακτηριστικών κάθε ψηφίου που ταυτίζεται με τον αριθμό των pixel (εικόνες 16×16). Από τους πίνακες απομονώθηκε η πρώτη στήλη που περιείχε το label κάθε δείγματος σε ένα `NumPy array y` διάστασης $N_{samples} \times 1$, και τα υπόλοιπα στοιχεία σε ένα `NumPy array` διάστασης $N_{samples} \times 256$, όπου 256 είναι ο αριθμός των χαρακτηριστικών κάθε ψηφίου που ταυτίζεται με τον αριθμό των pixel (εικόνες 16×16).

Έτσι μετά το πρώτο βήμα έχουμε 4 NumPy arrays:

Όνομα πίνακα	Περιγραφή	Διάσταση
`X_train`	Features των δεδομένων εκπαίδευσης	7291×256
`y_train`	Labels των δεδομένων εκπαίδευσης	7291×1
`X_test`	Features των δεδομένων εκπαίδευσης	2007×256
`y_test`	Labels των δεδομένων εκπαίδευσης	2007×1

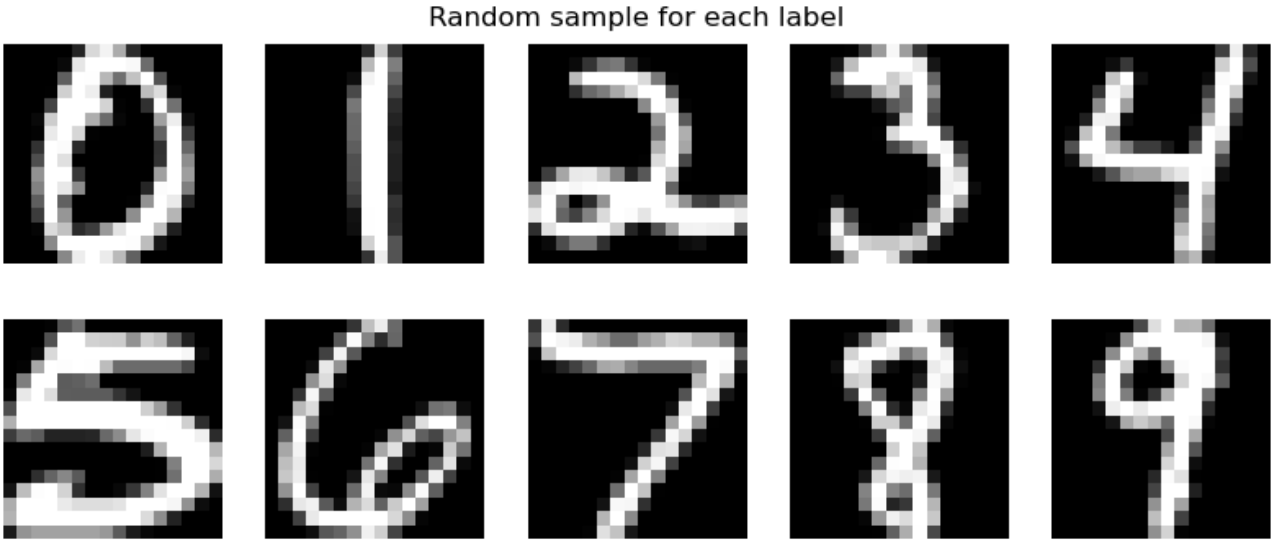
Βήμα 2: Απεικόνιση 131ου ψηφίου

Για να δούμε πως μοιάζουν οπτικά τα δεδομένα μας θα απεικονίσουμε το 131ο δεδομένο του train set. Επιλέγουμε την σειρά 130 (λαμβάνοντας υπόψιν πως η μέτρηση ξεκινάει από το 0) του πίνακα `X_train` και το κάνουμε `reshape(16, 16)` για να οργανώσουμε τα pixel σε ένα grid 16×16.



Βήμα 3: Απεικόνιση ενός τυχαίου δείγματος από κάθε κλάση

Στη συνέχεια παρουσιάζεται ένα τυχαίο δείγμα από κάθε label. Για να επιλεχθεί τυχαία ένα δείγμα από κάθε κλάση δημιουργούμε μία λίστα `samples` μήκους 10 (μία θέση για την κάθε κλάση) που κάθε στοιχείο της είναι ένας NumPy array διατάσεων $N_{each\ class\ \#samples} \times 256$ που έχει τα feature όλων των δειγμάτων που ανήκουν σε αυτή τη κλάση. Έτσι με `random.randint(0, len(samples[i]) - 1)` μπορούμε να επιλέξουμε τυχαία ένα δείγμα για κάθε ψηφίο.



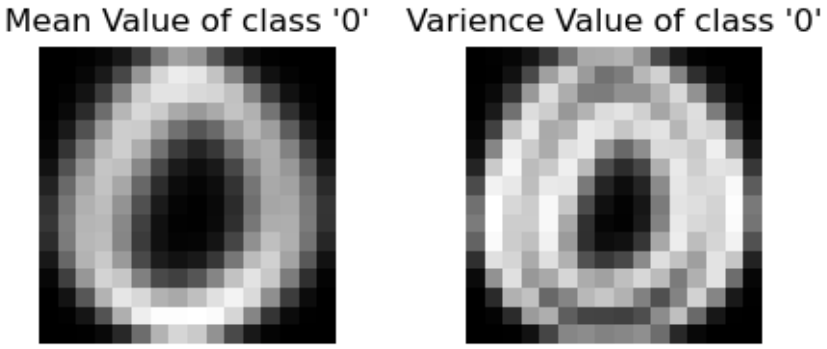
Βήμα 4 – 5: Υπολογισμός μέση τιμής και διασποράς του pixel (10, 10) των μηδενικών

Αφού έχουμε αποθηκευμένα όλα τα δείγματα κάθε κλάσης σε πίνακες ο υπολογισμός της μέσης τιμής και της διασποράς του pixel (10, 10) όλων των δειγμάτων "0" με χρήση των `np.mean` και `np.var`. Λαμβάνοντας υπόψι πως η αρίθμηση ξεκινάει από το 0 το pixel βρίσκεται στη θέση $10 \cdot 16 + 10 - 1 = 169$ και όχι 170.

	μέση τιμή	**διασπορά**
pixel (10, 10) όλων των μηδενικών	`-0.8568`	`0.1699`

Βήμα 6 - 8: Υπολογισμός και απεικόνιση μέσης τιμής και διασποράς όλων των pixel όλων των μηδενικών

Για τον υπολογισμό της μέσης τιμής και της διασποράς όλων pixel των δειγμάτων από την κλάση των μηδενικών κινηθήκαμε ανάλογα με τα βήματα 4 και 5 αλλά οι np.mean και np.var εφαρμόζονται σε όλα τα pixel (`np.mean(samples[0], axis=0)`).

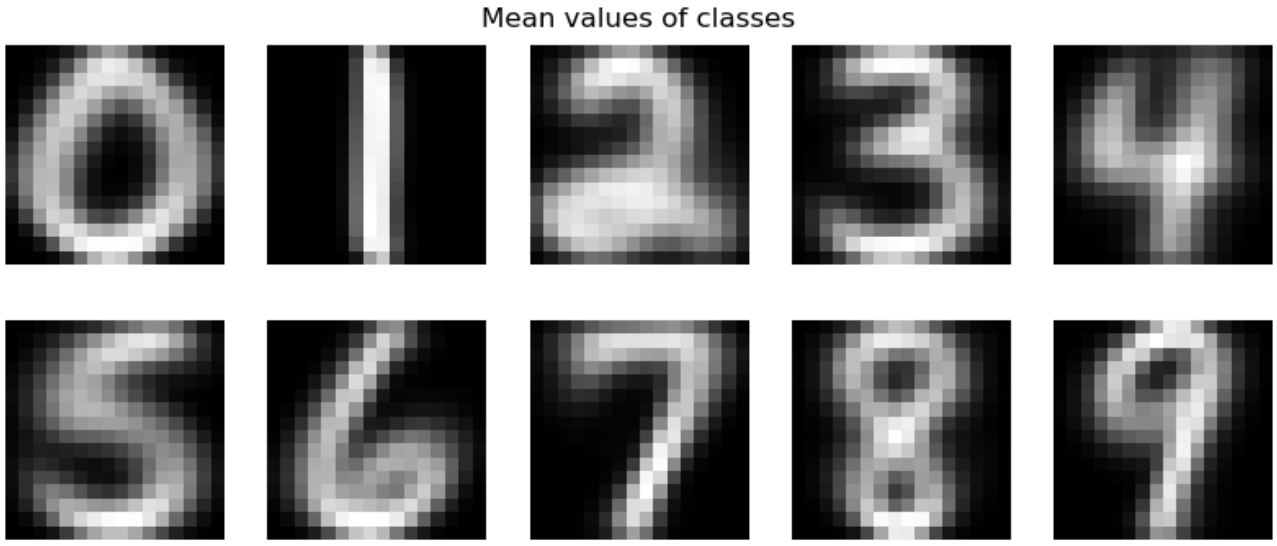


Στην απεικόνιση της μέσης τιμής παρατηρούμε πως παίρνουμε μία πιο θολή εκδοχή του ψηφίου "0", το οποίο είναι λογικό καθώς είναι ο μέσος όρων όλων των χειρόγραφων μηδενικών. Για την διασπορά παρατηρούμε πως τα pixel του περιγράμματος έχουν μεγαλύτερη τιμή, αφού σε αυτά τα pixel η διασπορά είναι υψηλότερη.

Και στις δύο περιπτώσεις το μηδενικό είναι ευδιάκριτο.

Βήμα 9: Υπολογισμός και απεικόνιση μέσης τιμής και διασποράς όλων κλάσεων

Για τον υπολογισμό της μέσης τιμής και της διασποράς όλων pixel των δειγμάτων από όλες τις κλάσεις κινηθήκαμε με ανάλογο τρόπο. Παρακάτω φαίνονται οι μέσες τιμές από όλες τις κλάσεις.



Παρατηρούμε πως σε όλες τις κλάσεις τα ψηφία είναι ευδιάκριτα. Το ψηφίο “1” είναι πιο ευδιάκριτο από τα υπόλοιπα ψηφία, γεγονός που υποδεικνύει περισσότερη ομοιογένεια στην γραφή του “1”.

Βήμα 10: Ταξινόμηση του 101ου ψηφίου του test set βάσει του ευκλείδειου ταξινομητή

Για τον υπολογισμό της Ευκλείδειας απόστασης των χαρακτηριστικών του 101ου ψηφίου χρησιμοποιήθηκε η ακόλουθη εντολή:

```
dist = np.linalg.norm(samples_mean - X_test[digit], ord=None, axis=1)
```

Το `np.linalg.norm` όρισμα `ord=None` υπολογίζει την Frobenius νόρμα της διαφοράς (2η κανονική νόρμα) η οποία ταυτίζεται με την Ευκλείδεια απόσταση.

Ευκλείδεια απόσταση	$d(p,q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_m - q_m)^2}$, όπου p η πραγματική τιμή ενός pixel και q η μέση τιμή του (συνολικά m pixel)
Frobenius νόρμα	$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left\|a_{ij}\right\|^2}$, όπου a_{ij} η διαφορά της μέσης τιμής από την πραγματική τιμή ενός pixel

Έτσι με τα τον υπολογισμό της ευκλείδειας απόστασης μεταξύ των pixel του 101ου ψηφίου και της μέσης τιμής κάθε κλάσης παίρνουμε με την συνάρτηση np.argmin την κλάσης όπου ελαχιστοποιείται η απόσταση.

```
The digit we want to classify is the 101th didgit of the test
set and as we can see below it is a '0'.
The euclidean classifier classifies it as a '0' which is True.
```

Η ταξινόμηση του 101ου ψηφίου έγινε σωστά ως ‘0’.

Βήμα 11: Ταξινόμηση όλων των ψηφίων βάσει του ευκλείδειου ταξινομητή και υπολογισμός ποσοστού επιτυχίας

Για την ταξινόμηση όλων των ψηφίων σε μία από τις 10 κατηγορίες αρχικά υπολογίζουμε την ευκλείδια απόσταση κάθε ψηφίου από τις μέσες τιμές κάθε κατηγορίας και στη συνέχεια επιλέγουμε την ελάχιστη. Οι προβλέψεις για κάθε ψηφίο αποθηκεύονται σε μία λίστα `predicts`.

Για το υπολογισμό του ποσοστού επιτυχίας του ταξινομητή που φτιάξαμε μετράμε πόσες φορές η λίστα predicts ισούται με τη λίστα των labels και στη συνέχεια το διαιρούμαι με το σύνολο των ψηφίων προς ταξινόμηση.

Ποσοστό επιτυχίας Ευκλειδιου ταξινομητή	81.42%

Βήμα 12: Υλοποίηση του ευκλείδειου ταξινομητή ως μία κλάση σαν ένα scikit-learn estimator

Τώρα καλούμαστε να υλοποιήσουμε τον παραπάνω ταξινομητή σε μία κλάση. Χρησιμοποιούμε τις ίδιες εντολές απλά πλέον τις βάζουμε μέσα σε συναρτήσεις της κλάσης `EuclideanDistanceClassifier`. Η κλάση μας έχει 3 συναρτήσεις ώστε να λειτουργεί ως ένας scikit-learn estimator και να μπορούμε να χρησιμοποιήσουμε τις έτοιμες συναρτήσεις από την `scikit-learn`.

Συνάρτηση	return	Περιγραφή
`fit(self, X, y)`	`self`	Δέχεται ως είσοδο δύο πίνακες `X` και `y` με δείγματα και τα labels τους αντίστοιχα και επιστρέφει τον ταξινομητή εκπαιδευμένο
`predict(self, X)`	`predicts`	Δέχεται ως όρισμα έναν πίνακα `X` με δείγματα προς ταξινόμηση και επιστρέφει τις προβλέψεις
`score(self, X, y)`	`acc`	Δέχεται ως όρισμα δύο πίνακες `X` και `y` με δείγματα προς ταξινόμηση και τα label τους και επιστρέφει το ποσοστό επιτυχίας

Βήμα 13α: Υπολογισμός score του ευκλείδειου ταξινομητή με χρήση 5-fold cross-validation

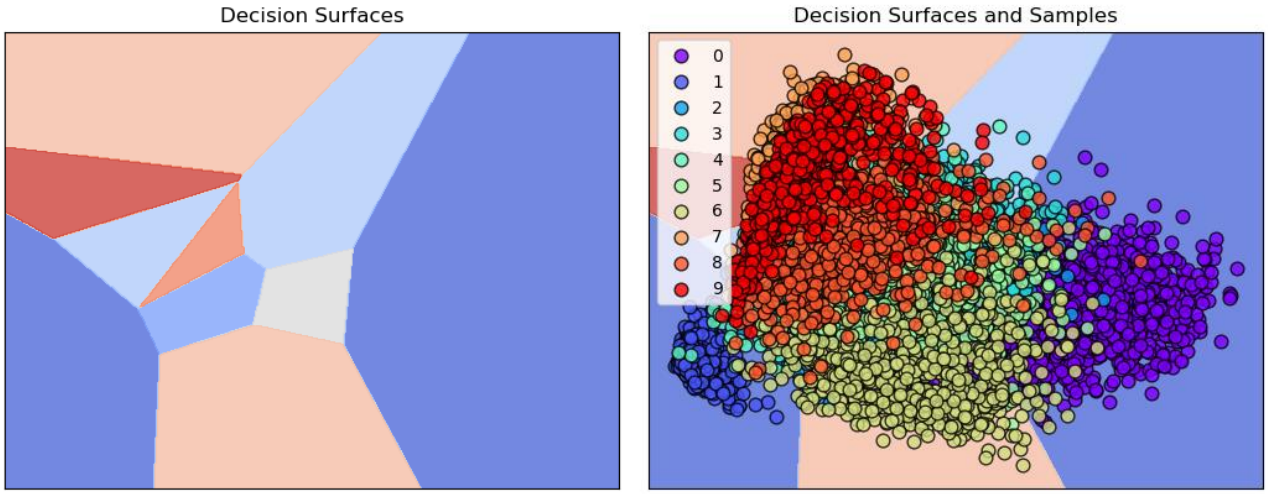
Τώρα θα εφαρμόσουμε την *5-fold-cross-validation* μέθοδο για να αξιολογήσουμε καλύτερα τον ταξινομητή. Βάσει αυτής της μεθόδου ανακατεύουμε το train set και το χωρίζουμε σε k (για εμάς $k = 5$) μέρη και γίνονται k εκπαιδεύσεις. Σε κάθε εκπαίδευση λαμβάνουμε το 1 μέρος ως trin set και τα υπόλοιπα $k - 1$ μέρη ως test set. Εν τέλει λαμβάνουμε ως ποσοστό επιτυχίας το μέσο όρος των επιμέρους ποσοστών. Για να υλοποιηθεί αυτή η μέθοδος χρησιμοποιήθηκε η συνάρτηση `cross_validate`.

```
The score of the Euclidean Classifie with 5-fold cross-validation is 85.14%.
Before 5-fold cross-validation it was 81.42%.
So it is more by 3.73%.
```

Η μικρή αύξηση του ποσοστού επιτυχίας οφείλεται στο γεγονός πως τώρα το test set μας έχει πολύ περισσότερα δείγματα από το προηγούμενο test set.

Βήμα 13β: Περιοχές απόφασης του ευκλείδειου ταξινομητή

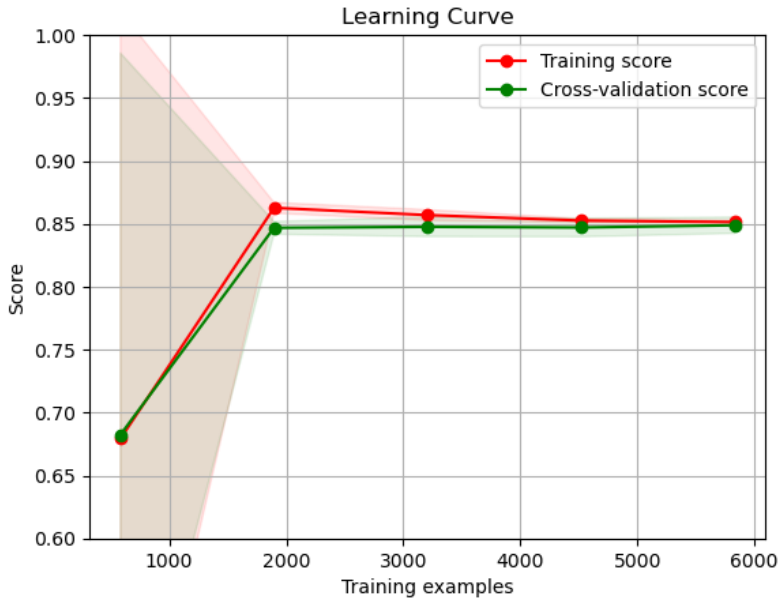
Η σχεδίαση των περιοχών απόφασης με 256 χαρακτηριστικά δεν είναι δυνατή και για αυτό χρησιμοποιούμε *Principal Component Analysis* (*PCA*) για να μειώσουμε σε 2 χαρακτηριστικά. Αφού μειώσουμε σε 2 χαρακτηριστικά κάνουμε fit τον ταξινομητή μας και απεικονίζουμε τις περιοχές απόφασης με την συνάρτηση `plot_clf` που μας δίνεται.



Παρατηρούμε πως τα δείγματα έχουν ταξινομηθεί κατά κύριο λόγο σωστά στις κατηγορίες τους.

Βήμα 13γ: Καμπύλη εκμάθησης του ευκλείδειου ταξινομητή (learning curve)

Χρησιμοποιώντας την συνάρτηση `plot_learning_curve` που μας δίνεται μπορούμε να αναπαραστήσουμε τις καμπύλες εκμάθησης για διαφορετικό αριθμό train samples.



Όσο αυξάνεται ο αριθμός των δειγμάτων το training score πέφτει γιατί το πρόβλημα δυσκολεύει, αφού έχει να ταξινομήσει περισσότερα χειρόγραφα ψηφία και εν τέλει

σταθεροποιείται και δεν μπορεί να βελτιωθεί περαιτέρω. Αντίθετα το cross-validation score αυξάνεται γιατί έρχεται αντιμέτωπο με μεγαλύτερη γκάμα περιπτώσεων. Παρατηρούμε πως με την αύξηση των δειγμάτων τα δύο ποσοστά συγκλίνουν. Η αύξηση δηλαδή των δειγμάτων εκπαίδευσης άνω των 6000 φαίνεται πως δεν έχει νόημα.