Αναγνώριση Προτύπων

Εργασία 1η

Μιχάλης Στεφανής
Βασίλης Μηναδάκης

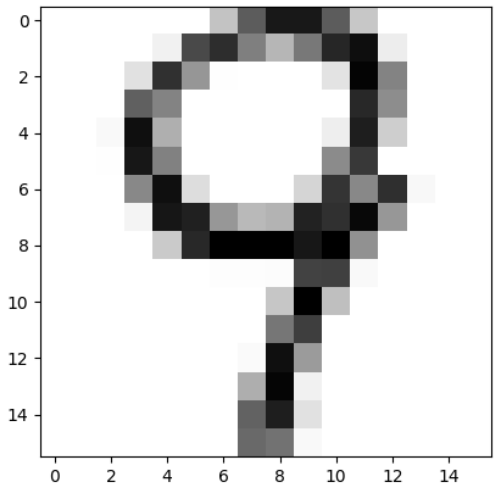
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβιο Πολυτεχνείο
Αθήνα
24/10/2021

Για την επίλυση όλων των ερωτημάτων έγινε η χρήση των βιβλιοθηών numpy, scikit-learn και matplotlib της python.
Το πρώτο στάδιο αφορά την εισαγωγή των train και test δεδομένων από μορφή αρχείου κειμένου (.txt) στον κώδικα της python. Αυτό επιτυγχάνεται με τη χρήση της συνάρτησης $np.genfromtxt()$. Έπειτα, ακολουθώντας τις οδηγίες της εκφώνησης δημιουργήθηκαν ρα εξής arrays δεδομένων:

Ονομασία	Διαστάσεις
train_features	7291x256
train_labels	7291
test_features	2007x256
test_labels	2007

Πίνακας 1: Διαστάσεις των training και test dataset.

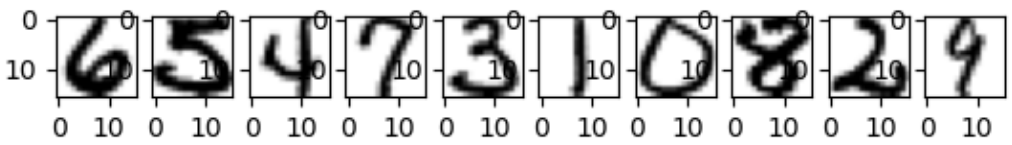
Στο 2ο βήμα σχεδιάστηκε το υπ'αριθμόν 131 ψηφίο των train δεδομένων χρησιμοποιώντας συναρτήσεις της βιβλιοθήκης matplotlib και συγκεκριμένα της $plt.imshow()$ και $plt.show()$. Η πρώτη εφαρμόζεται ουσιαστικά πάνω στο 131ο στοιχείο των train_features και η εικόνα που παράγεται είναι η κάτωθι.



Εικόνα 1: Απεικόνιση του 131ου ψηφίου των train_features.

Για το επόμενο βήμα (Βήμα 3) δομήθηκε η συνάρτηση $plot\_digits\_samples(X, y)$, της οποίας τα ορίσματα είναι τα training_features και τα training_labels αντίστοιχα. Συνοπτικά, η συγκεκριμένη συνάρτηση λειτουργεί με μία επαναληπτική διαδικασία από το 1 έως το πλήθος των στοιχείων του X (train_features), εντός της οποίας αν το στοιχείο y(i)

υπάρχει στο διάνυσμα $digits = \{0, 1, ..., 9\}$, τότε σχεδιάζεται το αντίστοιχο ψηφίο και στη συνέχεια αφαιρείται από το διάνυσμα digits, μέχρι να μην υπάρχει κανένα στοιχείο πλέον εντός του digits. Η ενσωμάτωση των εικόνων και των 10 διαφορετικών ψηφίων σε μία εικόνα γίνεται με τη χρήση της συνάρτησης $plt.subplots()$ και παρουσιάζονται στην εικόνα 2.



Εικόνα 2: Απεικόνιση του 131ου ψηφίου των train_features.

Στη συνέχεια, για την υλοποίηση του 4ου βήματος, κατασκευάστηκε η συνάρτηση $mean\_value\_at\_pixel\_of\_digit(X, y, digit, pixel = (10, 10))$, η οποία υπολογίζει τη μέση τιμή όλων των χαρακτηριστικών ενός συγκεκριμένου pixel (συγκεκριμένα του 10ου) για το training set και δέχεται σαν input τα train_features, τα train_labels και το ψηφίο που μας ενδιαφέρει (στη συγκεκριμένη περίπτωση το 0). Ο τύπος τις μέσης τιμής που χρησιμοποιήθηκε είναι η 1:

$$\mu = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right), \tag{1}$$

όπου n το πλήθος των χαρακτηριστικών του pixel (10,10) για το ψηφίο 0.

Στο βήμα 5 η λογική που ακολουθήθηκε είναι όμοια με αυτή του προηγούμενου βήματος, μόνο που σε αυτή την περίπτωση το function που δομήθηκε ($mean\_value\_at\_pixel\_of\_digit(X, y, digit, pix$ $(10, 10))$) κάνει τους υπολογισμούς με βάση της διασπορά των χαρακτηριστικών, δηλαδή γίνεται χρήση της σχέσης 2:

$$Var(x) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \mu)^2\right). \tag{2}$$

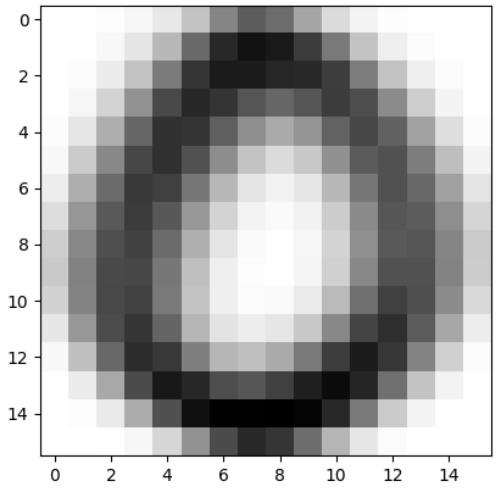
Εφαρμόζοντας τις δύο παραπάνω συναρτήσεις για το 10 pixel όλων των χαρακτηριστικών του ψηφίου μηδέν προκύπτουν οι εξής τιμές για τη μέση τιμή και διασπορά του:

	Μέση Τιμή	Διασπορά
10ο pixel του 0	-0.0826	0.115

Πίνακας 2: Αποτελέσματα μέσης τιμής και διασποράς του 10ου pixel όλων των στοιχείων για το ψηφίο 0.

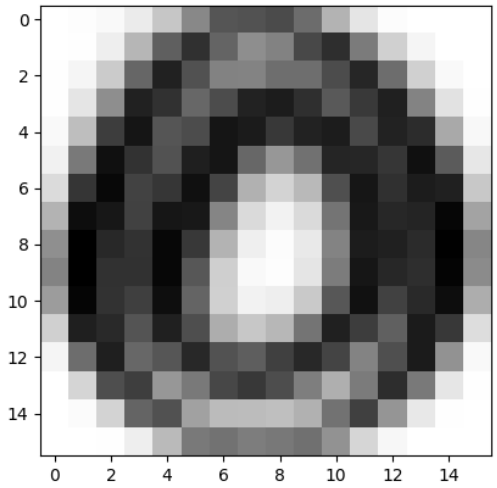
Στη συνέχεια (Βήμα 6) για τον υπολογισμό της μέσης τιμής και της διασποράς των χαρακτηριστικών όλων των pixel ενός οποιουδήποτε ψηφίου, δομήθηκαν οι συναρτήσεις $digit\_mean(X, y, digit)$ και $digit\_variance(X, y, digit)$. Αυτές οι δύο συναρτήσεις σε πρώτο επίπεδο υπολογίζουν τη μέση τιμή και τη διασπορά κάθε ενός pixel του επιθυμητού ψηφίου και στη επιστρέφει αυτές τις τιμές σε ένα διάνυσμα.

Οι τελευταίες δύο συναρτήσεις χρησιμοποιούνται ως εσωτερικές της νέας συνάρτησης $plot\_digit\_mean(X, y, digit)$, η οποία χρησιμοποιείται για το σχεδιασμό ενός επιθυμητού ψηφίου χρησιμοποιώντας τη μέση τιμή όλων των pixel όλων των χαρακτηριστικών του συγκεκριμένου ψηφίου. Η εικόνα που παράγεται με τη χρήση του συγκεκριμένου function για το ψηφίο 0, χρησιμοποιώντας το training set είναι η εικόνα 3.



Εικόνα 3: Απεικόνιση του ψηφίου 0 με τη χρήση της μέσης τιμής των pixel όλων των χαρακτηριστικών του training set.

Το επόμενο βήμα (**Βήμα 8**) είναι να σχεδιάσουμε ένα από τα ψηφία που επιθυμούμε χρησιμοποιώντας όχι την μέση τιμή, αλλά την διασπορά των pixel των εικονων που αντιπροσωπεύουν το επιθυμητό ψηφίο. Αυτό γίνεται με την συνάρτηση $plot\_digit\_variance$. Για παράδειγμα, για το ψηφίο 0 παίρνουμε την εξής εικόνα:

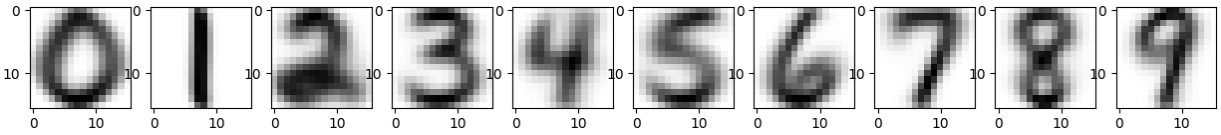


Εικόνα 4: Απεικόνιση του ψηφίου 0 μέσω της διασποράς των pixel

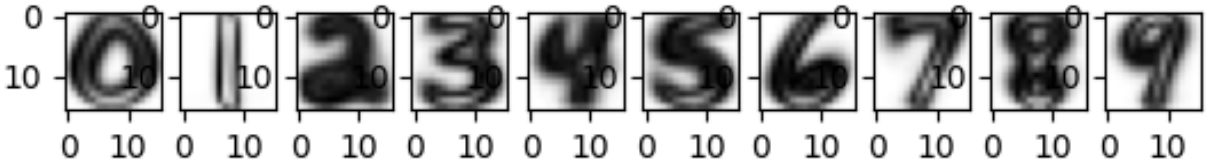
Αν συγκρίνουμε την εικόνα που πήραμε από το ψηφίο 0 μέσω της μέσης τιμής με αυτήν που πήραμε μέσω τις διασποράς θα παρατηρήσουμε κάποιες σχετικά αναμενόμενες διαφορές. Η εικόνα που αντιστοιχεί στην διασπορά έχει τα πιο σκούρα της pixels αυτά που τα αντίστοιχα στην εικόνα της μέσης τιμής είναι τα λιγότερο σκούρα. Για παράδειγμα τα pixels της εικόνας της μέσης τιμής για το 0 είναι περισσότερο σκούρα στις δύο "κορυφές" του 0. Αντίστοιχα, τα pixels της εικόνας της διασποράς είναι λιγότερο σκούρα (περισσότερο γκρί παρά μαύρα) στις κορυφές αυτές. Αυτό είναι αναμενόμενο, καθώς η διασπορά είναι στην ουσία η απόσταση μιας μεταβλητής από την μέση τιμή της.

Για το **Βήμα 9** υπολογίζουμε την μέση τιμή και την διασπορά όλων των pixel για όλα τα δεδομένα εκπαίδευσης. Αυτό μπορεί να γίνει πολύ απλά χρησιμοποιωντας list-comprehension της python και τις συναρτήσεις $digit\_mean$ και $digit\_variance$ αντίστοιχα. Παίρνουμε έτσι δύο λίστες 10 γραμμών και 256 στηλών ($mean\_X$ και var_X). Κάθε γραμμή αντιστοιχεί σε ένα ψηφίο και κάθε στήλη στην μέση τιμή και την διασπορά του εκάστοτε pixel.

Για να σχεδιάσουμε όλα τα ψηφία χρησιμοποιούμε τις συνάρτησεις $plot\_digits\_by\_mean$ και $plot\_digits\_by\_variance$, οι οποίες χρησιμοποιούν 10 φορές η καθε μία τις συναρτήσεις $digit\_mean$ και $digit\_variance$ αντίστοιχα.



Εικόνα 5: Απεικόνιση όλων των ψηφίων μέσω της μέσης τιμής των pixel



Εικόνα 6: Απεικόνιση όλων των ψηφίων μέσω της διασποράς των pixel

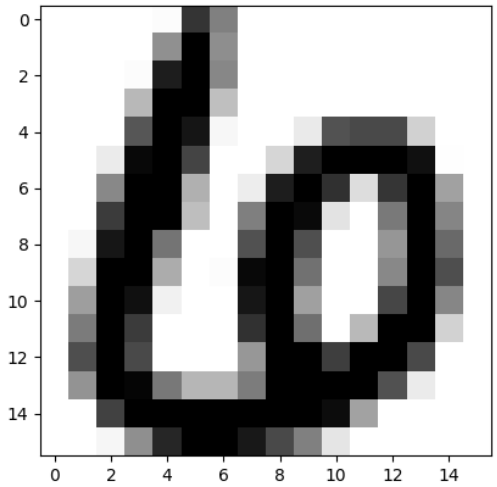
Από το **Βήμα 10** και μετά ξεκινούν οι δοκιμές του ταξινομητή της Ευκλείδιας απόστασης. Αυτός υλοποιείται με την βοήθεια δύο συναρτήσεων. Η συνάρτηση $euclidean\_distance$ υπολογίζει την Ευκλείδια απόσταση μεταξύ δύο διανυσμάτων (ή πινάκων) μέσω της σχέσης:

$$d(A,B) = \sqrt{(A[1]-B[1])^2 + (A[2]-B[2])^2 + \cdots + (A[N]-B[N])^2} \qquad (3)$$

Στην συνέχεια για την ταξινόμηση των στοιχείων του test set καλούμε την συνάρτηση $euclidean\_distance\_classifier$. Η συνάρτηση αυτή παίρνει ως ορίσματα τα χαρακτηριστικά του test set (test features) καθώς και την λίστα που περιλαμβάνει τις μέσες τιμές όλων

των ψηφίων του training set. Υπολογίζει για το κάθε στοιχείο του test set τις ευκλείδιες αποστάσεις του στοιχείου αυτού από τις μέσες τιμές. Επιλέγεται η μικρότερη από αυτές, η οποία ελάχιστη απόσταση είναι αντιπροσωπευτική της κλάσης στην οποία ταξινομήθηκε τελικά το στοιχείο που επιλεξαμε.

Ως πρώτη δοκιμή του ταξινομητή, διαλέγουμε το στοιχείο υπ'αριθμόν 101 των test δεδομένων και το δίνουμε ως είσοδο στον ταξινομητή μας. Για να έχουμε μία καλύτερη εικόνα, το στοιχείο αυτό είναι το εξής:



Εικόνα 7: Το ψηφίο υπ'αριθμόν 101 των test δεδομένων

Αν τρέξουμε τον ταξινομητή μας θα διαπιστώσουμε ότι η ταξινόμηση δεν είναι σωστή. Όπως μπορούμε να διακρίνουμε στην παραπάνω εικόνα, το ψηφίο που αναπαριστά είναι το 6. Παρόλα αυτά ο ταξινομητής δίνει την πρόβλεψη 0. Οπότε "μπερδεύει"το παραπάνω εξάρι με μηδενικό. Αυτό δεν πρέπει να μας αποθαρρύνει αφού ένα μόνο test δεδομένο δεν αρκεί για να βγάλουμε συμπέρασμα για τον ταξινομητή μας. Εξάλλου η παραπάνω εικόνα είναι πράγματι ένα ασυνήθιστο 6 για τα τυπικά χειρόγραφα δεδομένα.

Ερχόμαστε τώρα στο **Βήμα 11**, όπου τροφοδοτούμε τον ταξινομητή μας με όλα τα test δεδομένα. Η συνάρτηση $euclidean\_distance\_classifier$ θα μας γυρίσει ένα % ποσοστό επιτυχίας (δηλαδή τον αριθμό των δεδομένων που ταξινομήθηκαν επιτυχώς διαιρεμένο με τον συνολικό αριθμό των δεδομένων) και μία λίστα με όλες τις προβλέψεις του ταξινομητή (δηλαδή μία λίστα που θα μας λέει σε ποια κατηγορία ταξινομήθηκε το κάθε test δεδομένο). Αν τρέξουμε τον κώδικα και τυπώσουμε τις τιμές τότε θα λάβουμε στην οθόνη μας:

The accuracy of the euclidean distance classifier is: 81.42 %

Το ποσοστό αυτό είναι αρκετά ικανοποιητικό για τον περιορισμένο αυτό όγκο δεδομένων που είχαμε στην διάθεσή μας.

Στο **Βήμα 12** υλοποιούμε τον ταξινομητή της Ευκλείδιας απόστασης σαν έναν scikit-learn estimator. Ουσιαστικά δημιουργούμε μία κλάση EuclideanDistanceClassifier της python, μέσα στην οποία θα καλούμε τις κατάλληλες συναρτήσεις που έχουμε ήδη σχηματίσει προηγουμένως στον κώδικα. Η μόνη μεταβλητή που θα αρχικοποιείται στην κλάση είναι η μέση τιμή των κατηγοριών (των ψηφίων) και αυτή θα υπολογίζεται στην συνέχεια από την συνάρτηση fit μέσω των train δεδομένων. Η συνάρτηση predict θα έχει ως είσοδο τα test features και θα επιστρέφει την λίστα με τις προβλέψεις του ταξινομητή, ενώ η συνάρτηση score θα έχει ως είσοδο τα test features και τα test labels επιστρέφει το ποσοστό επιτυχίας.

Αν θέλουμε να καλέσουμε τώρα τον ταξινομητή μας σαν να ήταν ταξινομητής του scikit-learn θα πρέπει απλά να καλέσουμε την κλάση μας ώς:

clf = EuclideanDistanceClassifier()

Στην συνέχεια θα πρέπει να τροφοδοτήσουμε τα train δεδομένα στην κλάση για να υπολογίσουμε τις μέσες τιμές, μέσω της συνάρτησεις fit:

clf.fit(out training features)

Τέλος, αν θέλουμε να μας επιστραφούν οι προβλέψεις ή η ακρίβεια του ταξινομητή καλούμε τις συναρτήσεις predict ή score αντίστοιχα:

clf.predict(our test features)
clf.score(our test features, our test labels)

Προφανώς αν τυπώσουμε το score μέσω της παραπάνω κλάσης θα λάβουμε το ίδιο ακριβώς αποτέλεσμα με την συνάρτηση $euclidean\_distance\_classifier$.

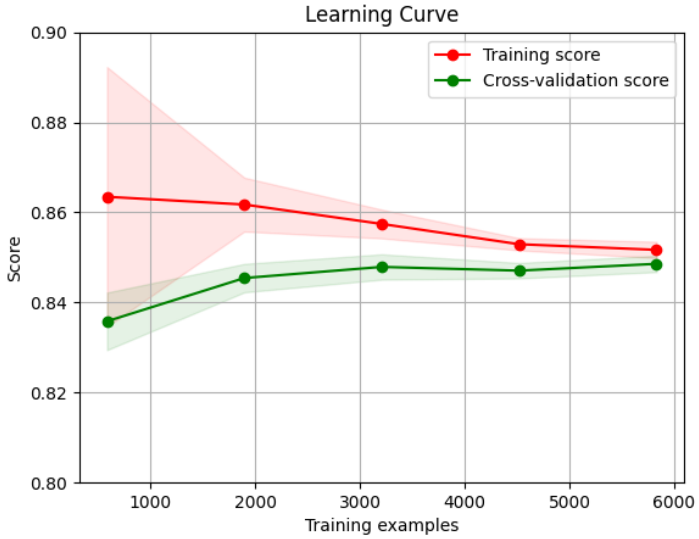
Στο **Βήμα 13** εξερευνούμε λίγο παραπάνω τον ταξινομητή που δημιουργήσαμε υπολογίζοντας το cross-validation-score με την χρήση της τεχνικής K-Folding καθώς επίσης σχεδιάζουμε την περιοχή απόφασης και την καμπύλη εκμάθησης του ταξινομητή.

Για τον υπολογισμό του cross-validation-score αντλούμε από το scikit-learn τις κλάσεις KFOLD και $cross\_val\_score$. Στην συνέχεια καλούμε την συνάρτηση k_fold_cv η οποία παίρνει ως είσοδο τα train features και τα train labels και τυπώνει το cross-validation-score και το cross-validation-error. Αναμένουμε σαν cross-validation-score να λάβουμε κάτι λίγο παραπάνω από το score του ταξινομητή μας. Πράγματι, αν καλέσουμε την συνάρτηση τότε θα λάβουμε στην οθόνη:

The 5-fold cross-validation score is 84.858036 +-0.181618
The 5-fold cross-validation error is 15.141964 +-0.181618

Το ποσοστό του cross-validation είναι περίπου 85 %, λίγο μεγαλύτερο από το 82 % του ταξινομητή.

Τέλος σχεδιάζουμε την καμπύλη εκμάθησης του ταξινομητή μας. Αυτό γίνεται μέσω της συνάρτησης *plot_learning_curve*. Παίρνει ως είσοδο τον ταξινομητή (την κλάση που τον αντιπροσωπεύει) και όλα τα train δεδομένα (features, labels).



Εικόνα 8: Η καμπύλη εκμάθησης του Ευκλείδιου ταξινομητή

Όπως αναμέναμε, το cross-validation-score αρχικά αυξάνεται και, όσο προχωρά η εκπαίδευση, τείνει να εξισωθεί με το training score.