Αναγνώριση Προτύπων

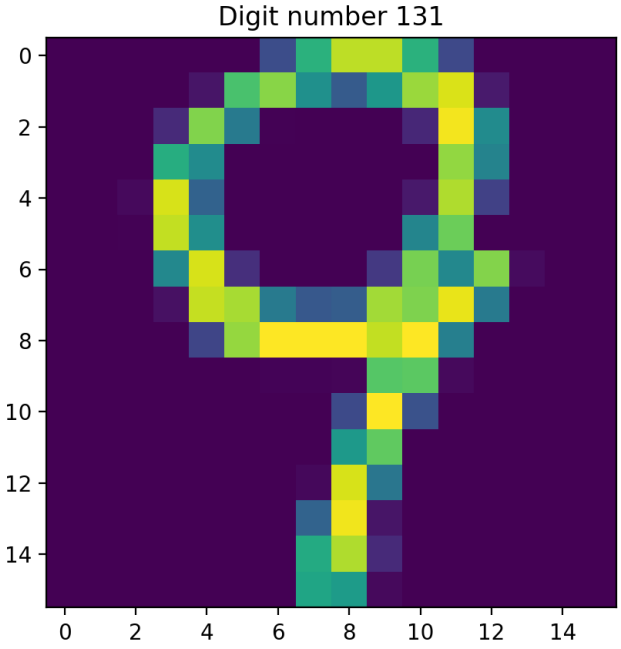
Προπαρασκευή 1ης Εργαστηριακής Άσκησης
Θέμα: Οπτική Αναγνώριση Ψηφίων
ΣΧΟΛΗ: ΣΗΜΜΥ

Ονοματεπώνυμο	Αριθμός Μητρώου
Γιάννης Πιτόσκας	03115077
Αντώνης Παπαοικονόμου	03115140

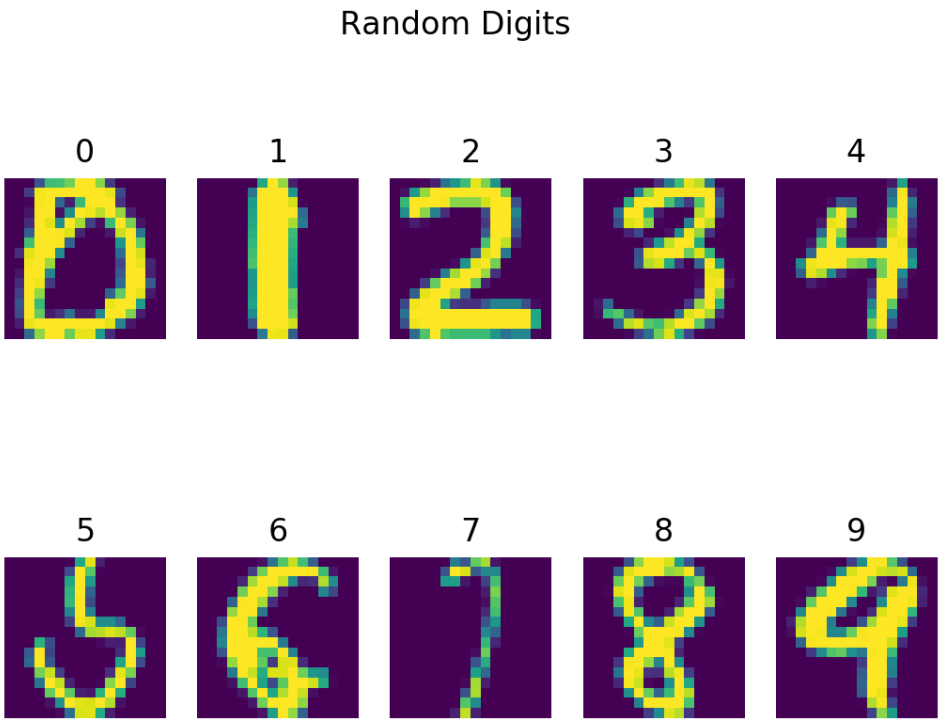
Βήμα 1ο:

Μας δίνονται τα δεδομένα ήδη χωρισμένα σε Train και Test και ακόμη μας δίνονται και τα αντίστοιχα labels μιας και έχουμε πρόβλημα supervised learning. Τα διαβάζουμε λοιπόν από τα δοθέντα αρχεία ως X_train, X_test και τα αντίστοιχα labels ως y_train, y_test.

Βήμα 2ο:



Βήμα 3ο:



Βήμα 4ο:

Ομαδοποιώντας τα δεδομένα που έχουμε στο train set με βάση το label τους σε ένα dictionary της python και με τον τύπο της μέσης τιμής $E[X_{0\,(10,10)}] = \frac{1}{N}\sum_{i=1}^{N} X_{0\,(10,10)}[i]$ έχουμε:

```
The mean value of pixel (10,10) of 0 is: -0.5041884422110553
```

Βήμα 5ο:

Με την παραπάνω ομαδοποίηση και σύμφωνα με τον τύπο $Var(X_{0\,(10,10)}) = E[X_{0\,(10,10)} - \mu^2_{0\,(10,10)}] = \frac{1}{N}\sum_{i=1}^{N} (X_{0\,(10,10)}[i] - E[X_0])^2$ έχουμε:

```
The variance value of pixel (10,10) of 0 is: 0.5249618093885225
```

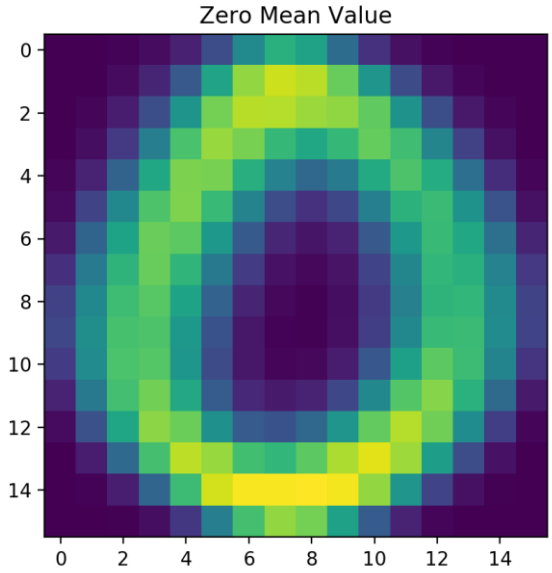
Βήμα 6ο:

Τυπώνοντας τις τιμές για τα πρώτα 4 features έχουμε:

```
The mean value of 0 is: (first 4 numbers) [-0.99862814 -0.99539782 -0.98492295 -
0.94125126]
The variance value of 0 is: (first 4 numbers) [0.00224711 0.00661218 0.01960704
0.06295393]
```

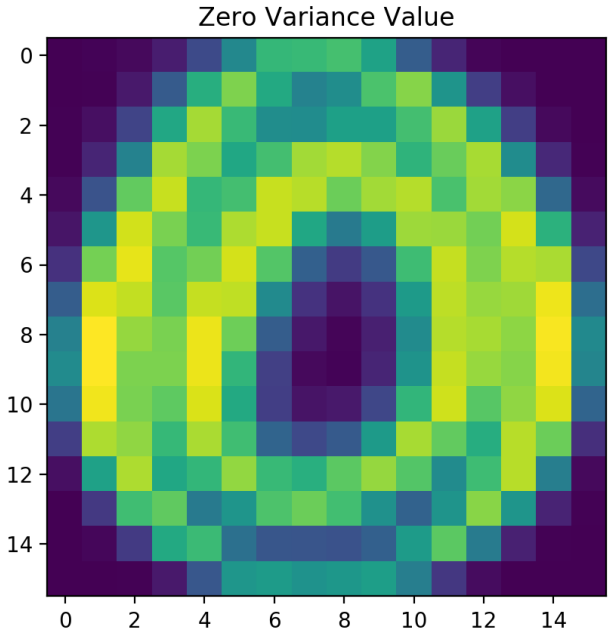
Βήμα 7ο:

Παρακάτω απεικονίζεται το ψηφίο ‘0’ με βάση τις τιμές της μέσης τιμής:



Βήμα 8ο:

Παρακάτω απεικονίζεται το ψηφίο ‘0’ με βάση τις τιμές της διασποράς:



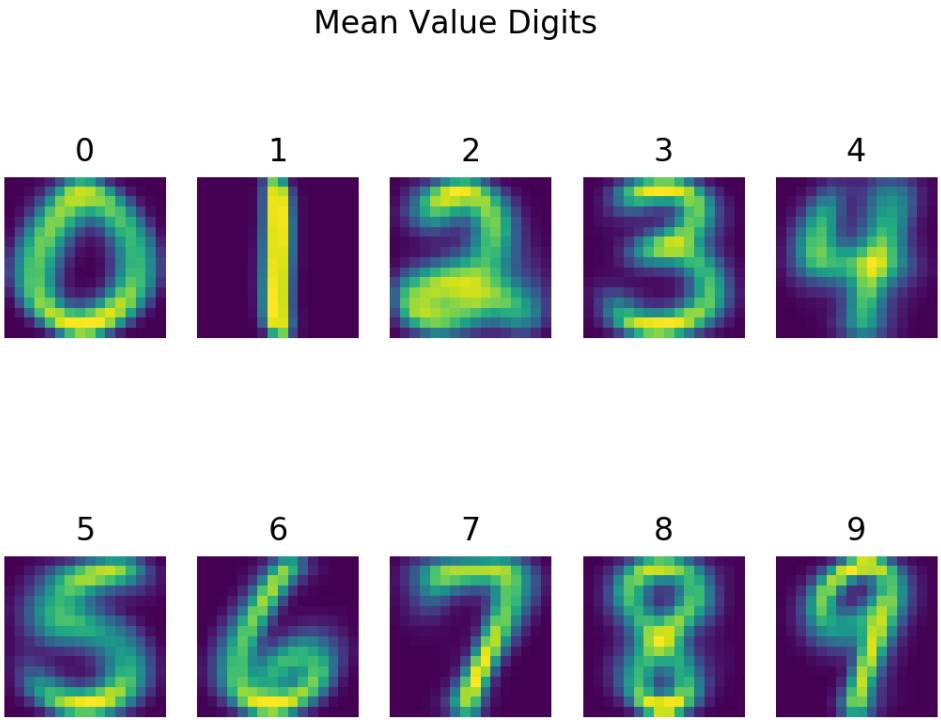
Παρατηρούμε ότι η διασπορά μας δίνει πληροφορία για το σε ποια σημεία γίνεται συνήθως όμοια ο σχεδιασμός του ψηφίου (πχ πάνω και κάτω), ενώ φαίνεται ότι δεξιά και αριστερά η διασπορά έχει μεγάλη τιμή που υποδεικνύει ότι ο σχεδιασμός του μηδενός ως προς το πόσο «πλατύ» είναι ποικίλει.

Βήμα 9ο:

(α) Τυπώνοντας μόνο το την τιμή για το πρώτο feature έχουμε:

```
The mean value of 0 is: (first number only) -0.9986281407035177
The variance value of 0 is: (first number only) 0.002247105527638201
The mean value of 1 is: (first number only) -1.0
The variance value of 1 is: (first number only) 0.0
The mean value of 2 is: (first number only) -0.9924883720930233
The variance value of 2 is: (first number only) 0.005090387193373648
The mean value of 3 is: (first number only) -0.9975151975683889
The variance value of 3 is: (first number only) 0.0012149350807067123
The mean value of 4 is: (first number only) -1.0
The variance value of 4 is: (first number only) 0.0
The mean value of 5 is: (first number only) -0.9994586330935253
The variance value of 5 is: (first number only) 0.0001629514388489166
The mean value of 6 is: (first number only) -1.0
The variance value of 6 is: (first number only) 0.0
The mean value of 7 is: (first number only) -0.9749162790697675
The variance value of 7 is: (first number only) 0.017963918445760705
The mean value of 8 is: (first number only) -0.9985018450184502
The variance value of 8 is: (first number only) 0.0005089713459426635
The mean value of 9 is: (first number only) -0.9998990683229813
The variance value of 9 is: (first number only) 6.560559006210882e-06
```

(β)



Βήμα 10ο:

Σύμφωνα με τον τύπο: $pred(\boldsymbol{X}) = arg \min_{0 \le i \le 9} \|\boldsymbol{\mu} - \boldsymbol{X}\|_2$ έχουμε:

```
Actual value of digit 101: 6
The result of the Euclidean Classifier on digit 101 is: 0
```

Παρατηρούμε πως η πρόβλεψη του ταξινομητή μας είναι λανθασμένη.

Βήμα 11ο (α,β):

Βρίσκοντας την ελάχιστη Ευκλείδεια απόσταση και συγκρίνοντας την με το label y για κάθε δείγμα στο test set μας έχουμε:

```
The accuracy of the Euclidean Classifier on the test set is: 0.8141504733432985
```

Βήμα 12ο:

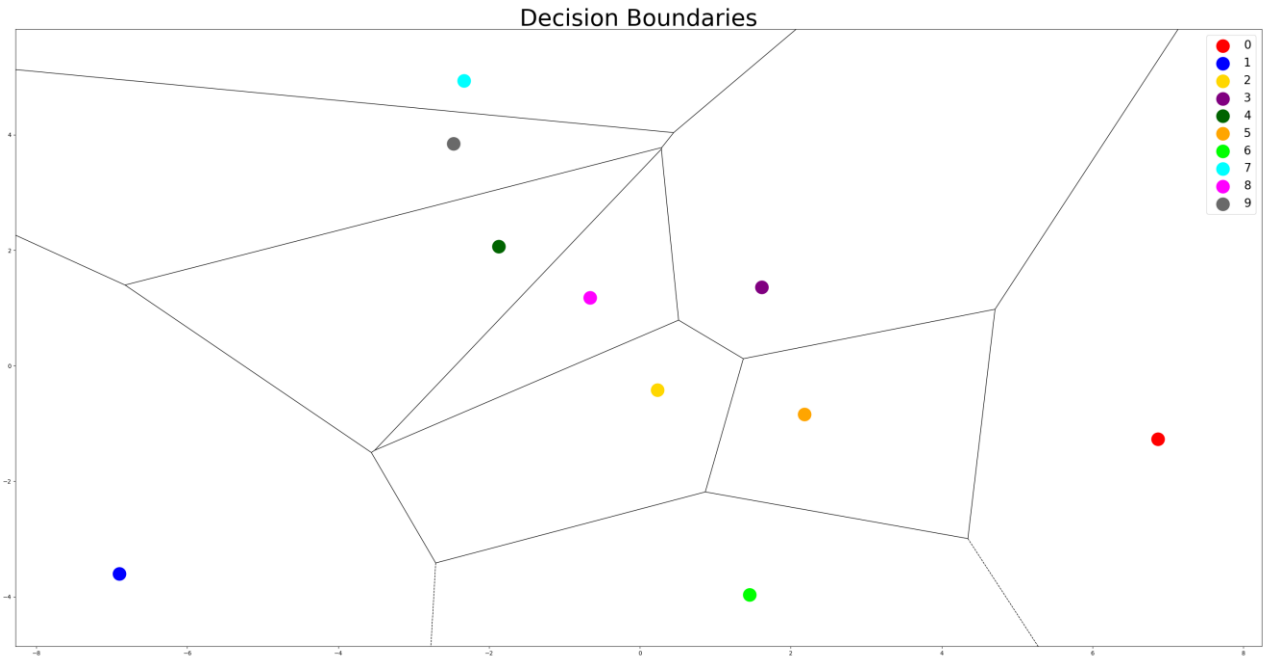
Καλούμαστε να υλοποιήσουμε τον ταξινομητή ευκλείδιας απόστασης σαν ένα scikit-learn estimator. Δημιουργούμε την κλάση `EuclideanClassifier` και εν συνεχεία υλοποιούμε τις μεθόδους `fit`, `predict` και `score` ως εξής:

- `fit()`: Υπολογίζει τα features για την κάθε κατηγορία βάσει του μέσου όρου των χαρακτηριστικών των δεδομένων εκπαίδευσης της εκάστοτε κατηγορίας
- `predict()`: Δεδομένου ενός δείγματος κάνει μια πρόβλεψη βάσει της ελάχιστης από τις ευκλείδιες αποστάσεις από τα διανύσματα χαρακτηριστικών της κάθε κατηγορίας.
- `score()`: Υπολογίζει το ποσοστό επιτυχίας του ταξινομητή πάνω σε κάποιο Test Set.

Η υλοποίηση της κλάσης που περιγράφεται παραπάνω βρίσκεται στο αρχείο EuclideanClassifier.py

Βήμα 13ο:

(α) `The average score using 5-fold cross-validation is: 0.841773724173982`
(β) Τα features που αντιπροσωπεύουν κάθε ψηφίο αποτελούν διανύσματα 256 διαστάσεων και καλούμαστε να βρούμε τις περιοχές απόφασης (decision boundaries). Ωστόσο, προκειμένου να μπορέσει να γίνει οπτικοποίηση των περιοχών απόφασης θα εφαρμόσουμε PCA ώστε να μειώσουμε τον αριθμό των διαστάσεων απο 256 σε 2. Οι περιοχές απόφασης που θα υπολογίσουμε για τις δύο διαστάσεις δεν μας δίνουν το ίδιο καλή πληροφορία με τις 256, αλλά μπορούμε να έχουμε μια κάποια εικόνα των περιοχών απόφασης του εκτιμητή μας. Παρακάτω παρατίθεται το διάγραμμα αυτών των περιοχών απόφασης σε δύο διαστάσεις με τα χρωματισμένα σημεία των περιοχών να αποτέλουν τα δισδιάστατα διανύσματα χαρακτηριστικών για το εκάστοτε ψηφίου:



(γ) Για όλα μας τα δεδομένα (test και train μαζί) και χρησιμοποιώντας 5-fold cross-validation έχουμε τις παρακάτω καμπύλες εκμάθησης (learning curves):