ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΙ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΑΘΗΜΑ: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΚΑΘΗΓΗΤΕΣ : ΚΩΣΤΑΣ ΔΙΑΜΑΝΤΑΡΑΣ, ΚΩΣΤΑΣ ΓΟΥΛΙΑΝΑΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ 5

ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΤΟ ΜΟΝΤΕΛΟ SVM

Σκοπός της άσκησης: Η εκτίμηση της επίδοσης ενός ταξινομητή τύπου **Support Vector Machine (SVM)** χρησιμοποιώντας τον πυρήνα **Gaussian (RBF)**. Θα γίνει χρήση της μεθόδου διασταύρωσης (Cross-Validation) και τα κριτήρια επίδοσης:

1. Ακρίβεια (accuracy)
2. Ευστοχία (precision)
3. Ανάκληση (recall)
4. F-Measure
5. Ευαισθησία (Sensitivity)
6. Προσδιοριστικότητα (Specificity)

Βήματα υλοποίησης:

1. Χρησιμοποιήστε το σύνολο δεδομένων IRIS από το προηγούμενο εργαστήριο, καθώς και τον κώδικα από το εργαστήριο αυτό.
2. Με τη χρήση της εντολής plot δημιουργήστε τη **γραφική παράσταση** των προτύπων **των 3 κλάσεων** με **διαφορετικό σύμβολο και χρώμα για την κάθε κλάση** χρησιμοποιώντας την $1^{η}$ και $3^{η}$ στήλη του πίνακα x, ώστε να τα απεικονίσετε στο χώρο των 2 διαστάσεων και εμφανίστε τα στο ίδιο γράφημα, ώστε να πάρετε μια ιδέα για το πώς είναι η διασπορά των προτύπων στο χώρο των 4 διαστάσεων. Μη ξεχνάτε ότι η αρίθμηση ξεκινάει απ' το 0.
3. Χρησιμοποιώντας τη συνάρτηση zeros από τη βιβλιοθήκη numpy αρχικοποιήστε τον πίνακα `t` ώστε να είναι γεμάτος μηδενικά και να έχει διάσταση `NumberOfPatterns`.
4. **Εκχωρήστε** στη μεταβλητή ans την τιμή "y'.
5. Για όσο (ans = "y')
   - **Εμφανίστε** το παρακάτω **menu** επιλογών :

     1 **Διαχωρισμός `Iris-setosa`** από `Iris-versicolor και Iris-virginica`

     2 **Διαχωρισμός `Iris-virginica`** από `Iris-setosa και Iris-versicolor`

     3 **Διαχωρισμός `Iris-versicolor`** από `Iris-setosa και Iris-virginica`

     Διαβάστε την επιλογή (1/2/3)

     Αν επιλογή = 1

     Δημιουργήστε ένα dictionary `map_dict` με τα εξής ζευγάρια key/values:

     - `"Iris-setosa": 1`
     - `"Iris-versicolor": 0`
     - `"Iris-virginica": 0`

     Κατόπιν, χρησιμοποιώντας loop θέστε για κάθε pattern την τιμή στόχου `t[pattern]` ως εξής:

`t[pattern] = 1` αν η 5^{ο} στήλη για το pattern είναι "`Iris-setosa`"

`t[pattern] = 0` σε διαφορετική περίπτωση

Αν επιλογή = 2

Δημιουργήστε ένα dictionary `map_dict` με τα εξής ζευγάρια key/values:

- `"Iris-setosa": 0`
- `"Iris-versicolor": 0`
- `"Iris-virginica": 1`

Κατόπιν, χρησιμοποιώντας loop θέστε για κάθε pattern την τιμή στόχου `t[pattern]` ως εξής:

`t[pattern] = 1` αν η 5^{ο} στήλη για το pattern είναι "`Iris-virginica`"

`t[pattern] = 0` σε διαφορετική περίπτωση

Αν επιλογή = 3

Δημιουργήστε ένα dictionary `map_dict` με τα εξής ζευγάρια key/values:

- `"Iris-setosa": 0`
- `"Iris-versicolor": 1`
- `"Iris-virginica": 0`

Κατόπιν, χρησιμοποιώντας loop θέστε για κάθε pattern την τιμή στόχου `t[pattern]` ως εξής:

`t[pattern] = 1` αν η 5^{ο} στήλη για το pattern είναι "`Iris-versicolor`"

`t[pattern] = 0` σε διαφορετική περίπτωση

Μπορείτε να το κάνετε αυτό χρησιμοποιώντας το `map_dict` και να αποφύγετε εντολή `if-else`.

6. Χωρισμός προτύπων σε πρότυπα εκπαίδευσης και ανάκλησης

Τεμαχίστε τα δεδομένα των πινάκων x και t σε 4 πίνακες:

- `xtrain` πίνακας με τα πρότυπα που θα χρησιμοποιηθούν στην εκπαίδευση, τα 40 πρώτα πρότυπα της κάθε κλάσης.
- `xtest` πίνακας με τα πρότυπα που θα χρησιμοποιηθούν στον έλεγχο, τα 10 τελευταία πρότυπα της κάθε κλάσης.
- `ttrain` διάνυσμα με τους στόχους που θα χρησιμοποιηθούν στην εκπαίδευση, οι 40 πρώτοι στόχοι της κάθε κλάσης.
- `ttest` διάνυσμα με τους στόχους που θα χρησιμοποιηθούν στον έλεγχο, οι 10 τελευταίοι στόχοι της κάθε κλάσης.

- Χρησιμοποιώντας τη συνάρτηση `plot` από τη βιβλιοθήκη `matplotlib.pyplot` σχεδιάστε

  - τα διανύσματα `xtrain[:,0]` → άξονας x, `xtrain[:,2]` → άξονας y, χρησιμοποιώντας τελείες με μπλε χρώμα και
  - τα διανύσματα `xtest[:,0]` → άξονας x, `xtest[:,2]` → άξονας y, χρησιμοποιώντας τελείες με κόκκινο χρώμα.

Δημιουργήστε τις λίστες :

– glist = [0.01, 0.03, 0.1, 0.3, 1]
– Clist = [1, 10, 100, 1000]

Για gamma in glist
  Για C in Clist

- Δημιουργήστε ένα δίκτυο SVM με πυρήνα RBF χρησιμοποιώντας την κλάση SVC (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)
- SVC(*C, kernel='rbf', gamma, degree, coef0, …*) όπου:
  - *C = παράμετρος C του SVM*
  - *kernel = επιλογή συνάρτησης πυρήνα μεταξύ* 'linear', 'poly', 'rbf', 'sigmoid'. Επιλέξτε 'rbf'
  - *gamma = η παράμετρος γ του πυρήνα RBF*

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$$

- Εκπαιδεύστε το δίκτυο που φτιάξατε χρησιμοποιώντας τη συνάρτηση
  - `fit()` : με εισόδους το μοντέλο, τον πίνακα των προτύπων εκπαίδευσης (`xtrain`), και το διάνυσμα των στόχων εκπαίδευσης (`ttrain`)
- Αφού εκπαιδεύσετε το μοντέλο κάνετε ανάκληση χρησιμοποιώντας τη συνάρτηση
  - `predict()` : με εισόδους το εκπαιδευμένο μοντέλο και τον πίνακα των προτύπων ελέγχου (`xtest`)
- Ονομάστε $predict_{test}$ το διάνυσμα που πήρατε.
- Καλέστε τη συνάρτηση `evaluate()` από το προηγούμενο εργαστήριο να υπολογίσετε το Accuracy, Precision, Recall, F-measure, Sensitivity και Specificity.
- Στο figure(1) τυπώστε το εξής γράφημα:
  - δείξτε με μπλε τελείες τους πραγματικούς στόχους $t_{test}\,(i)$ για όλα τα πρότυπα του test set
  - δείξτε με κόκκινους κύκλους τους εκτιμώμενους στόχους $predict_{test}\,(i)$ για όλα τα πρότυπα του test set

Μετά το τέλος των loops υπολογίστε και τυπώστε στην οθόνη τα εξής:

1. τη μέση τιμή του Accuracy
2. τη μέση τιμή του Precision
3. τη μέση τιμή του Recall
4. τη μέση τιμή του F-Measure
5. τη μέση τιμή του Sensitivity
6. τη μέση τιμή του Specificity

Βρείτε για ποιο συνδυασμό gamma και C πετυχαίνετε την καλύτερη μέση τιμή test-accuracy και τυπώστε την στην οθόνη. Γι' αυτές τις 2 τιμές του gamma και C :

- Δημιουργήστε ένα δίκτυο SVM με πυρήνα RBF χρησιμοποιώντας την κλάση SVC (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)
- SVC(*C, kernel='rbf', gamma, degree, coef0, …*) όπου:
  - *C = παράμετρος C του SVM*

- *kernel = επιλογή συνάρτησης πυρήνα μεταξύ* 'linear', 'poly', 'rbf', 'sigmoid'. Επιλέξτε 'rbf'
- *gamma = η παράμετρος γ του πυρήνα RBF*

$$\mathrm{k}(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$$

- Εκπαιδεύστε το δίκτυο που φτιάξατε χρησιμοποιώντας τη συνάρτηση
  - `fit()` : με εισόδους το μοντέλο, τον πίνακα των προτύπων εκπαίδευσης (`xtrain`), και το διάνυσμα των στόχων εκπαίδευσης (`ttrain`)
- Αφού εκπαιδεύσετε το μοντέλο κάνετε ανάκληση χρησιμοποιώντας τη συνάρτηση
  - `predict()` : με εισόδους το εκπαιδευμένο μοντέλο και τον πίνακα των προτύπων ελέγχου (`xtest`)
- Ονομάστε $predict_{test}$ το διάνυσμα που πήρατε.
- Στο figure(2) τυπώστε το εξής γράφημα:
  - δείξτε με μπλε τελείες τους πραγματικούς στόχους $t_{test}\,(i)$ για όλα τα πρότυπα του test set
  - δείξτε με κόκκινους κύκλους τους εκτιμώμενους στόχους $predict_{test}\,(i)$ για όλα τα πρότυπα του test set

7. Θα εφαρμοστεί η μέθοδος `train_test_split()` για K=9 folds.
8. Στο Cross-Validation loop θα πρέπει να κάνετε τα εξής:

Για gamma in glist

Για C in Clist

`Για κάθε` *fold*

- Έχετε ήδη δημιουργήσει τους αρχικούς πίνακες προτύπων `xtrain` και `xtest` (χωρίς επαύξηση) καθώς και τα διανύσματα στόχων `ttrain` και `ttest`. Οι τιμές των στόχων θα είναι 0/1.
- Δημιουργήστε ένα δίκτυο SVM με πυρήνα RBF χρησιμοποιώντας την κλάση SVC (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)
- SVC(*C, kernel='rbf', gamma, degree, coef0, …*) όπου:
  - *C = παράμετρος C του SVM*
  - *kernel = επιλογή συνάρτησης πυρήνα μεταξύ* 'linear', 'poly', 'rbf', 'sigmoid'. Επιλέξτε 'rbf'
  - *gamma = η παράμετρος γ του πυρήνα RBF*

$$\mathrm{k}(\mathbf{x},\mathbf{y}) = \exp(-\gamma\|\mathbf{x}-\mathbf{y}\|^2)$$

- Εκπαιδεύστε το δίκτυο που φτιάξατε χρησιμοποιώντας τη συνάρτηση
  - `fit()` : με εισόδους το μοντέλο, τον πίνακα των προτύπων εκπαίδευσης (`xtrain`), και το διάνυσμα των στόχων εκπαίδευσης (`ttrain`)
- Αφού εκπαιδεύσετε το μοντέλο κάνετε ανάκληση χρησιμοποιώντας τη συνάρτηση
  - `predict()` : με εισόδους το εκπαιδευμένο μοντέλο και τον πίνακα των προτύπων ελέγχου (`xtest`)

- Ονομάστε $predict_{test}$ το διάνυσμα που πήρατε.
- Καλέστε τη συνάρτηση `evaluate()` από το προηγούμενο εργαστήριο όσες φορές χρειάζεται έτσι ώστε για το συγκεκριμένο fold να υπολογίσετε το Accuracy, Precision, Recall, F-measure, Sensitivity και Specificity.
- Χρησιμοποιώντας κατάλληλο subplot σε grid 3x3 στο figure(3) τυπώστε το εξής γράφημα:
  - δείξτε με μπλε τελείες τους πραγματικούς στόχους $t_{test}\,(i)$ για όλα τα πρότυπα του test set
  - δείξτε με κόκκινους κύκλους τους εκτιμώμενους στόχους $predict_{test}\,(i)$ για όλα τα πρότυπα του test set

9. Μετά το τέλος του loop υπολογίστε και τυπώστε στην οθόνη τα εξής:
   1. τη μέση τιμή του Accuracy για όλα τα folds
   2. τη μέση τιμή του Precision για όλα τα folds
   3. τη μέση τιμή του Recall για όλα τα folds
   4. τη μέση τιμή του F-Measure για όλα τα folds
   5. τη μέση τιμή του Sensitivity για όλα τα folds
   6. τη μέση τιμή του Specificity για όλα τα folds

Βρείτε για ποιο συνδυασμό gamma και C πετυχαίνετε την καλύτερη μέση τιμή test-accuracy για όλους τους συνδυασμούς και για όλα τα folds και τυπώστε την στην οθόνη.

Διαβάστε την απάντηση ans του χρήστη, αν θέλετε να συνεχίσετε.

Οδηγίες κατάθεσης ασκήσεων

1. Συνδεθείτε στο URL: http://aetos.it.teithe.gr/s
1. Επιλέξτε το μάθημα “Μηχανική Μάθηση – Εργαστήριο Χ” (Όπου Χ ο αριθμός του εργαστηρίου του οποίου τις ασκήσεις πρόκειται να καταθέσετε) και πατήστε επόμενο.
2. Συμπληρώστε τα στοιχεία σας. Πληκτρολογείστε USERNAME 00003 και PASSWORD 30000 (Επώνυμο και Όνομα με ΛΑΤΙΝΙΚΟΥΣ ΧΑΡΑΚΤΗΡΕΣ).
3. Αν θέλετε να καταθέσετε μόνο ένα αρχείο μη το βάζετε σε zip file. Αντίθετα, αν θέλετε να καταθέσετε περισσότερα από ένα αρχεία, τοποθετήστε τα σε ένα zip ή rar file.
4. Επιλέξτε το αρχείο που θέλετε να στείλετε επιλέγοντας “choose file” στο πεδίο FILE1 και πατήστε “Παράδοση”