

Εργασία 3: Σύγκριση τριών αλγορίθμων μηχανικής μάθησης σε προβλήματα δυαδικής ταξινόμησης

Στόχος της εργασίας αυτής είναι η σύγκριση των τριών μοντέλων νευρωνίων ADALINE (Linear Neuron), Logistic Regressor* και Perceptron ως προς τις δυνατότητές τους στην επίλυση προβλημάτων δυαδικής ταξινόμησης για την περίπτωση α) γραμμικά διαχωρίσιμων δεδομένων και β) μη-γραμμικά διαχωρίσιμων δεδομένων. Για λόγους οπτικοποίησης των ορίων απόφασης τα δεδομένα θα είναι δισδιάστατα.

Δεδομένα

α) Υποσύνολο #1 του Iris dataset γραμμικά διαχωρίσιμο (αρχείο subset1.mat).

Το σύνολο δεδομένων Iris αποτελείται από 150 δείγματα λουλουδιών ίριδας από τρία διαφορετικά είδη (τα πρώτα 50 από το πρώτο είδος, τα επόμενα 50 από το δεύτερο και τα τελευταία 50 από το τρίτο είδος). Κάθε δείγμα περιλαμβάνει 4 χαρακτηριστικά (μήκη και πλάτη πετάλων και σεπάλων) αλλά για λόγους οπτικοποίησης, επιλέγονται τα εξής δύο: το 1^ο και το 3^ο. Στο υποσύνολο αυτό επιλέγονται τα δύο πρώτα είδη (δείγματα 1 έως 100) με την κωδικοποίηση της εξόδου να είναι η διπολική (+1/-1).

β) Υποσύνολο #2 του Iris dataset μη-γραμμικά διαχωρίσιμο (αρχείο subset2.mat).

Στο υποσύνολο αυτό επιλέγονται το δεύτερο και τρίτο είδος (δείγματα 51 έως 150) με διπολική κωδικοποίηση της εξόδου.

γ) Συνθετικό Υποσύνολο #3 με οριακή γραμμική διαχωρισιμότητα (αρχείο subset3.mat).

Το υποσύνολο αυτό είναι τροποποίηση του πρώτου με προσθήκη δεδομένων και από τις δύο κατηγορίες έτσι ώστε να πλησιάσουν τα δείγματα των δύο κατηγοριών χωρίς όμως να παραβιαστεί η γραμμική διαχωρισιμότητα. Ο σκοπός είναι να αυξηθεί η δυσκολία εύρεσης του βέλτιστου ορίου των κατηγοριών.

Βήματα εργασίας.

Θα πρέπει να γίνουν τρία πειράματα, ένα για κάθε υποσύνολο του Iris dataset, στα οποία θα συγκριθούν τα τρία μοντέλα νευρωνίων.

Για κάθε πείραμα τα βήματα είναι τα εξής:

1. Προεπεξεργασία δεδομένων

Φόρτωση και οπτικοποίηση του συνόλου δεδομένων.

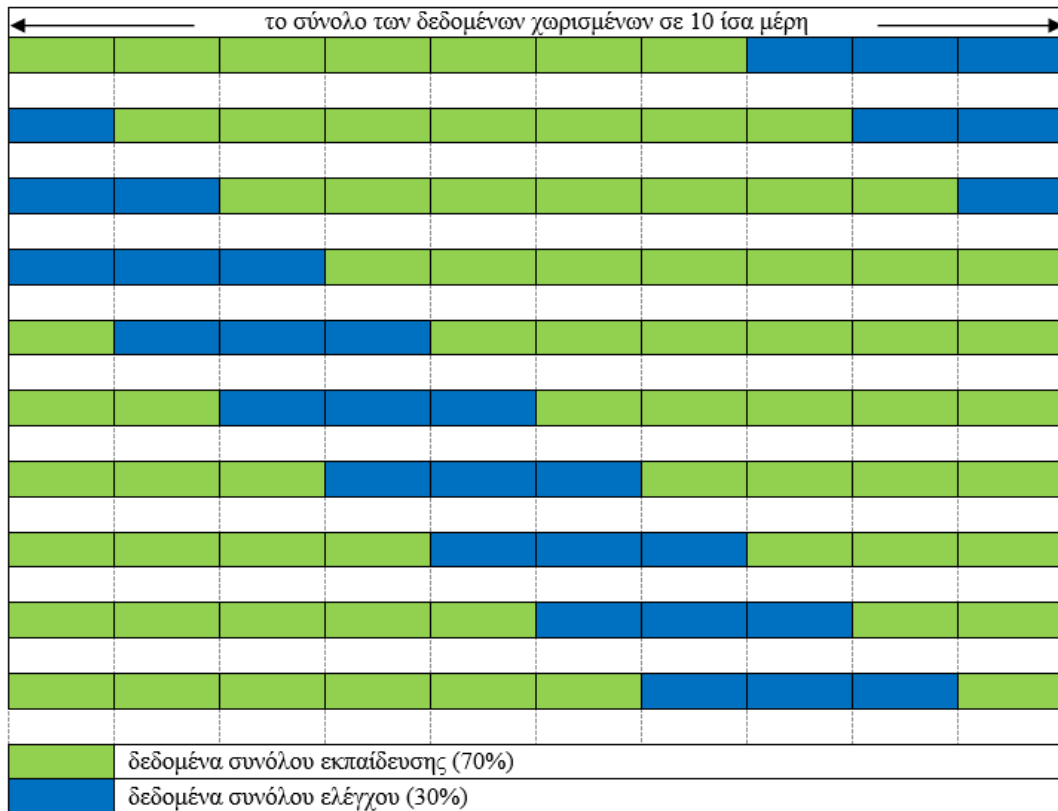
Κανονικοποίηση των χαρακτηριστικών ώστε να έχουν μηδενική μέση τιμή και μοναδιαία διακύμανση.

Χωρισμός των δεδομένων σε δύο υποσύνολα, ένα για κάθε κατηγορία (από 50 δείγματα).

Εφαρμογή τυχαίας μετάθεσης (ανακάτεμα) των δεδομένων της κάθε κατηγορίας.

Διάρθρωση των δεδομένων της κάθε κατηγορίας σε 10 ίσα μέρη (από 5 δείγματα) και συνένωση με το αντίστοιχο μέρος της άλλης κατηγορίας. Μετά τις συνενώσεις, το κάθε μέρος περιλαμβάνει 10 δείγματα (5 ανά κατηγορία).

Υλοποίηση μεθόδου 10-fold cross-validation (βλ. Εικ. 1) σύμφωνα με την οποία τα αποτελέσματα ταξινόμησης θα είναι μέσοι όροι 10 επιμέρους πειραμάτων, ένα για κάθε κυκλικό χωρισμό σε συνόλα δεδομένων εκπαίδευσης (70%) και ελέγχου (30%). Με την διαδικασία που ακολουθήθηκε στο προηγούμενο βήμα, εξασφαλίζουμε ίση εκπροσώπηση των κατηγοριών τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα ελέγχου (class balancing). Τα ίδια κυκλικά δημιουργημένα σύνολα δεδομένων θα χρησιμοποιηθούν για την εκπαίδευση και αποτίμηση της γενικευτικής ικανότητας όλων των μοντέλων.



Εικ 1: Η τεχνική της 10-πλής διασταυρούμενης επικύρωσης (10-fold cross-validation).

2. Υλοποίηση αλγορίθμων

Υλοποιήστε τα μοντέλα Adaline, Logistic regressor και Perceptron στο MATLAB χρησιμοποιώντας την μέθοδο στοχαστικής κατάβασης κλίσης (stochastic gradient descent) δηλαδή την online παραλλαγή που παρουσιάστηκε στις διαλέξεις του μαθήματος όσον αφορά στα δύο πρώτα μοντέλα και τον αλγόριθμο μάθησης του perceptron όσον αφορά το μοντέλο Perceptron. Μην χρησιμοποιήσετε έτοιμες συναρτήσεις από κάποια βιβλιοθήκη. Έχετε τη δυνατότητα είτε να γράψετε τις δικές σας συναρτήσεις (π.χ. ADALINE(x_train, y_train, lr, epochs)) που θα επιστρέφουν τα τελικά συνναπτικά βάρη που βρήκε ο αλγόριθμος εκπαίδευσης είτε να ενσωματώσετε τον κώδικα υλοποίησης των μοντέλων στο κυρίως πρόγραμμα.

Η εκπαίδευση των μοντέλων να γίνει σε προκαθορισμένο μέγιστο πλήθος εποχών όπου σε κάθε εποχή τα δεδομένα θα παρουσιάζονται από μια φορά το καθένα με τυχαία σειρά (να γίνεται μετάθεση των δεδομένων στην αρχή της κάθε εποχής) και τα βάρη θα προσαρμόζονται μετά από κάθε παρουσίαση δεδομένων. Να επιλέξετε μέγιστο πλήθος 50 εποχών. Επίσης, να κάνετε δοκιμές με τρεις ρυθμούς εκμάθησης: 0.1, 0.05 και 0.01. Στο τέλος κάθε εποχής να υπολογίζετε το μέσο τετραγωνικό σφάλμα (MSE) μεταξύ της πραγματικής και της επιθυμητής εξόδου και να το αποθηκεύετε σε πίνακα σφαλμάτων.

3. Σύγκριση αλγορίθμων ως προς τη φάση εκπαίδευσης

Για κάθε νευρωνικό μοντέλο και για κάθε μια επιλογή του ρυθμού εκμάθησης να δείξετε το MSE στα δεδομένα εκπαίδευσης ως προς τις εποχές για το 10-fold επιμέρους πείραμα με το μικρότερο τελικό MSE (το καλύτερο μοντέλο) και για αυτό με το μεγαλύτερο MSE (το χειρότερο μοντέλο). Συνολικά να δείξετε 18 plots (6 ανά μοντέλο).

Επίσης, για κάθε νευρωνικό μοντέλο και για κάθε μια επιλογή του ρυθμού εκμάθησης να δείξετε το scatter plot του συνόλου εκπαίδευσης με το διαχωριστικό όριο των κατηγοριών για τα τελικά βάρη που βρήκε ο αλγόριθμος εκμάθησης για το καλύτερο και το χειρότερο μοντέλο αντίστοιχα.

4. Αξιολόγηση αλγορίθμων στο σύνολο ελέγχου

Για κάθε νευρωνικό μοντέλο και για κάθε μια επιλογή του ρυθμού εκμάθησης να δείξετε:

- α) την μέση ακρίβεια ταξινόμησης (ως προς τα 10 πειράματα) στα δεδομένα ελέγχου,
- β) την ακρίβεια ταξινόμησης για το καλύτερο και το χειρότερο μοντέλο εκπαίδευσης,
- γ) το scatter plot του συνόλου ελέγχου με το διαχωριστικό όριο των κατηγοριών για τα συνναπτικά βάρη του καλύτερου και του χειρότερου μοντέλου αντίστοιχα.

Να συγκρίνετε και να σχολιάσετε τις επιδόσεις και τα όρια απόφασης των τριών μοντέλων.

Παραδοτέα εργασίας

A) Μία τεχνική αναφορά η οποία θα περιέχει ένα listing του αλγόριθμου και όλων των συναρτήσεων που υλοποιήσατε με πλήρη/λεπτομερή τεκμηρίωση και απαντήσεις στα παραπάνω ερωτήματα. Να περιλάβετε συγκριτική αξιολόγηση των μοντέλων και να σχολιάσετε τα τυχόν πλεονεκτήματα/μειονεκτήματα της κάθε μεθόδου για κάθε ένα από τα τρία σύνολα δεδομένων.

B) ένα αρχείο zip με το όνομα σας το οποίο θα ανεβάσετε στο e-class (και θα περιέχει την τεχνική αναφορά του (A) και τον κώδικα σε Matlab.

Βαθμολογία

Θα υπάρξει bonus 1 μονάδας για αποδοτικό διανυσματικό κώδικα σε Matlab.

Η δημιουργία των Υποσυνόλων 1 & 2 του Iris dataset πρέπει να γίνει από εσας ενώ το Υποσύνολο 3 θα το φορτώσετε από το αρχείο subset3.mat. Θα υπάρξει ποινή 1 μονάδας αν χρησιμοποιήσετε τα έτοιμα αρχεία subset1.mat και subset2.mat.

*Προκειμένου να συγκρίνουν όλα τα μοντέλα την έξοδό τους με το μηδέν για τη λήψη απόφασης ως προς την κατηγορία στην οποία ανήκει η είσοδος, τροποποιείται η συνάρτηση ενεργοποίησης από τη λογιστική στην υπερβολική εφαιπτομένη:

$$y = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{με παράγωγο} \quad \frac{\partial y}{\partial x} = \tanh'(x) = 1 - \tanh^2(x)$$

ώστε η εξίσωση προσαρμογής των βαρών να γίνει:

$$w_i(k+1) = w_i(k) - a(k)(y_k - d_k) \frac{\partial y_k}{\partial v_k} x_{ki}$$

ή

$$w_i(k+1) = w_i(k) - a(k)(y_k - d_k)(1 + y_k)(1 - y_k)x_{ki}$$

ΠΡΟΣΟΧΗ: ΟΙ ΑΝΤΙΓΡΑΦΕΣ ΘΑ ΜΗΔΕΝΙΖΟΝΤΑΙ