



Εργαστήριο MM1 Επιβλεπόμενη Ταξινόμηση

Αντικείμενο - Στόχοι

- ✓ Επιβλεπόμενη ταξινόμηση με νευρωνικά δίκτυα
- ✓ Εισαγωγή στη Pytorch

Ταξινόμηση σε πολλές κατηγορίες

Καλείστε να εκτελέσετε επιβλεπόμενη ταξινόμηση σε πολλές κατηγορίες στο σετ υπερφασματικών δεδομένων “Indian Pines”. Τα δεδομένα αυτά έχουν προκύψει από τον υπερφασματικό δέκτη AVIRIS και οι κατηγορίες στις οποίες καλείστε να ταξινομήσετε τα εικονοστοιχεία είναι οι εξής :

Groundtruth classes for the Indian Pines scene and their respective samples number

#	Class	Samples
1	Alfalfa	46
2	Corn-notill	1428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93

Πηγή : http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

Σημειώνεται ότι στα δεδομένα αληθείας όσα εικονοστοιχεία έχουν την τιμή “0” τότε δεν ανήκουν σε κάποια από τις παραπάνω κατηγορίες, οπότε πρέπει να **αγνοηθούν** κατά το στάδιο της ταξινόμησης.

Οδηγίες

Κατεβάστε τα δεδομένα από τον ακόλουθο σύνδεσμο : ([data](#))

Τα δεδομένα αφορούν δύο αρχεία .npy τα οποία περιέχουν τον υπερφασματικό κύβο και την εικόνα δεδομένων αληθείας αντίστοιχα. Προκειμένου να φορτώσετε τα αρχεία στον κώδικά σας χρησιμοποιήστε την συνάρτηση `numpy.load(filename)`.

Αρχικά, καλείστε να δημιουργήσετε ένα σετ δεδομένων αποτελούμενο από 200-D διανύσματα, αφού τα κανάλια της απεικόνισης είναι 200. Για κάθε σημείο του σετ δεδομένων πρέπει να αντιστοιχίσετε το κατάλληλο δεδομένο αληθείας βάσει του συνοδευτικού αρχείου. Χωρίστε το σετ δεδομένων σε ένα σύνολο εκπαίδευσης (training set) και ένα σύνολο αξιολόγησης (validation set). Ο λόγος διαμοιρασμού (train/validation split ratio) να είναι στο 70%/30%. Βεβαιωθείτε ότι κάθε κατηγορία εκπροσωπείται επαρκώς τόσο στο σετ εκπαίδευσης όσο και στο σετ ελέγχου!

Για λόγους επαναληψιμότητας ορίστε συγκεκριμένη τιμή *random seed* για τον αλγόριθμο διαμοιρασμού της επιλογής σας. Για παράδειγμα αν χρησιμοποιήσετε την συνάρτηση του scikit-learn: `sklearn.model_selection.train_test_split(*arrays, **options)`

Ορίστε την παράμετρο *random_state* : int, RandomState instance or None, optional (default=None) σε μία συγκεκριμένη τιμή (π.χ. τα τελευταία 3 ψηφία του Α.Μ. σας)

Χρησιμοποιήστε ταξινομητή τύπου MLP. Αναπτύξτε πλήρη διαδικασία τροφοδότησης δεδομένων καθώς και κώδικα εκπαίδευσης και αξιολόγησης των μοντέλων.

Τέλος, για κάθε ταξινομητή υπολογίστε στο σετ ελέγχου :

- Δείκτες ακρίβειας (ανά κατηγορία και μέσο όρο) : Accuracy, Recall, Precision, F1-score
- Πίνακα σύγχυσης (Confusion Matrix)

Ζητούμενα (ενδεικτικά)

εκπονήστε τεχνική έκθεση περιγράφοντας τις διαδικασίες που ακολουθήσατε, απαντώντας και στα παρακάτω ερωτήματα

1. Εφαρμόστε τον ταξινομητή στα δεδομένα ελέγχου, αφού πρώτα εκπαιδευτούν στα δεδομένα εκπαίδευσης. Αξιολογήστε τα αποτελέσματά τους.

2. Απαντήστε στα ακόλουθα ερωτήματα σχετικά με τα πειράματά σας :

B3.1. Γιατί παρατηρούνται σημαντικές διαφορές στις μετρικές ακρίβειας ανάμεσα στις κατηγορίες;

B3.2. Στην περίπτωση των MLP παρατηρείτε φαινόμενα overfit ή underfit; Γιατί;

B3.3. Έχετε επιλέξει κατάλληλο ρυθμό εκμάθησης (learning rate) στα MLP; Γιατί;

B3.4. Για κάθε MLP που εκπαιδεύσατε ποιος κατά τη γνώμη σας είναι ο βέλτιστος αριθμός εποχών εκπαίδευσης;

Σημείωση : Στην απάντηση των ερωτημάτων συμπεριλάβετε τα κατάλληλα σχήματα ή διαγράμματα.

