Πανεπιστήμιο Πειραιώς, Τμήμα Πληροφορικής
Προπτυχιακό Πρόγραμμα Σπουδών

Ακαδ. έτος 2020-21 (εαρ. εξάμηνο)

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ

Απαλλακτική Εργασία
(Ομάδες των 1-3 ατόμων)
Ημερομηνία παράδοσης: Παρασκευή 2 Ιουλίου, 23:59μμ

Σκοπός της εργασίας είναι η εξοικείωση με τεχνικές μηχανικής μάθησης και θα χρησιμοποιηθούν τα εργαλεία που παρουσιάστηκαν στο εργαστηριακό κομμάτι του μαθήματος. Αποδεκτές γλώσσες προγραμματισμού είναι οι Python και R, με έμφαση στην Python και τις σχετικές βιβλιοθήκες, όπως αυτές αναφέρονται στα επιμέρους θέματα.

Το τελικό παραδοτέο θα αποτελείται από ένα αρχείο zip, το οποίο θα αποσταλεί ηλεκτρονικά και θα περιέχει τα εξής:

1. Τεχνική αναφορά (report) με αναλυτική περιγραφή των προσεγγίσεων που ακολουθήσατε σε καθένα από τα βήματα (π.χ. παράμετροι αλγορίθμων, προπαρασκευή δεδομένων κοκ.) και ερμηνεία των αποτελεσμάτων που προέκυψαν.
2. Τα αρχεία πηγαίου κώδικα (source code) και τυχόν συμπληρωματικά αρχεία (που είναι απαραίτητα για την εκτέλεση του κώδικα), καθώς και τα αποτελέσματα που παρήχθησαν (π.χ. plots).

Για να γίνει αποδεκτή η παράδοση της εργασίας, απαιτούνται απαραιτήτως:

1. Αποστολή όλου του πακέτου (report + source codes) στο email: evachon@unipi.gr
2. Κοινοποίηση (cc) στα: ytheod@unipi.gr , pikrakis@unipi.gr , hgeorgiou@unipi.gr

Κάθε email θα έχει ως τίτλο "Εργασία DA_under_2021_<ΑΜ μελών ομάδας (να χωρίζονται με backslash)>" και θα περιέχει τα ζητούμενα σε ένα αρχείο zip με τίτλο "DA_under_2021_AM_<ΑΜ μελών ομάδας (να χωρίζονται με backslash)>".

Για οποιαδήποτε απορία σχετικά με την εργασία μπορείτε να απευθύνεστε στους διδάσκοντες. Σε περίπτωση αντιγραφής οι εμπλεκόμενες εργασίες μηδενίζονται. Σε περίπτωση αμφιβολίας ενδέχεται να ζητηθούν συμπληρωματικές διευκρινίσεις.

ΜΕΡΟΣ Α:

Θα εργαστείτε με ένα σύνολο δεδομένων από το UCI ML Repository και συγκεκριμένα το Cardiotocography Data Set, το οποίο μπορείτε να βρείτε στον παρακάτω σύνδεσμο:

https://archive.ics.uci.edu/ml/datasets/Cardiotocography

Το σύνολο δεδομένων Cardiotocography αποτελείται από 2126 καρδιοτογράμματα εμβρύων (CTGs), τα οποία υποβλήθηκαν σε επεξεργασία και μετρήθηκαν 21 χαρακτηριστικά. Τα CTGs ταξινομήθηκαν από τρεις ειδικούς σε σχέση με το μορφολογικό μοτίβο FHR (10 κατηγορίες) και την κατάσταση του εμβρύου NSP (3 κατηγορίες). Για περισσότερες λεπτομέρειες δείτε τις σχετικές δημοσιεύσεις που αναφέρονται στη σελίδα περιγραφής του dataset.

Οι επιμέρους εργασίες που πρέπει να πραγματοποιηθούν είναι:

Ερώτημα 1: Προπαρασκευή δεδομένων – υλοποίηση με εργαλεία στατιστικής επεξεργασίας, π.χ. R/Python (0,5 μονάδες).

Από το παραπάνω dataset θα επιλέξετε τα δεδομένα που θα χρησιμοποιήσετε για αναλυτική επεξεργασία και θα προχωρήσετε στην όποια προπαρασκευαστική εργασία (επιλογή, καθαρισμό, μετασχηματισμό, δειγματοληψία, κλπ.) θεωρείτε απαραίτητη, ώστε: α) να «καθαρίσετε» τα δεδομένα (από ελλιπείς ή εσφαλμένες - μη λογικές - τιμές), β) να κανονικοποιήσετε – διακριτοποιήσετε τα δεδομένα (π.χ. για αντιμετώπιση των συνεχών πεδίων τιμών), γ) να μειώσετε τον όγκο των δεδομένων (μείωση διαστάσεων, μείωση πλήθους εγγραφών). Ειδικά για τη μείωση του πλήθους των εγγραφών, επειδή το πλήθος εξαρχής δεν είναι μεγάλο, απλά θα πειραματιστείτε με τις διάφορες τεχνικές, αλλά τελικά θα χρησιμοποιήσετε το αρχικό πλήθος για τα περαιτέρω βήματα.

Ερώτημα 2: clustering – υλοποίηση με scikit-learn (2,5 μονάδες).

Χρησιμοποιήστε το dataset με στόχο να πραγματοποιήσετε συσταδοποίηση (clustering). Αυτό σημαίνει ότι θα παραλειφθούν εντελώς τα χαρακτηριστικά στόχοι (FHR και NSP), ώστε να πραγματοποιηθεί ταξινόμηση χωρίς επίβλεψη. Στο τελικό στάδιο, οι πραγματικές τιμές του χαρακτηριστικού-στόχου (ή FHR ή NSP) μπορούν να συσχετιστούν με τις συστάδες που σχηματίζονται. Με τον τρόπο αυτό θα ελεγχθούν η ακρίβεια και τα σφάλματα της διαδικασίας, συγκρίνοντας την «κατά πλειοψηφία» τιμή κατηγορίας κάθε συστάδας με τις πραγματικές τιμές του χαρακτηριστικού-στόχου για κάθε μέλος της (cluster labeling).

Χρησιμοποιήστε τουλάχιστον 3 διαφορετικές τεχνικές που παρέχονται από τη βιβλιοθήκη scikit-learn (μπορείτε να χρησιμοποιήσετε και τεχνικές clustering που δεν διδαχθήκατε στο εργαστήριο). Σκοπός είναι να μεγιστοποιήσετε την απόδοση κάθε αλγορίθμου ξεχωριστά κάνοντας δοκιμές με την προπαρασκευή του dataset και τις παραμέτρους του. Συγκρίνετε τις διάφορες προσεγγίσεις. Περιγράψτε τη διαδικασία και εξηγήστε τα αποτελέσματα που προκύπτουν. Τέλος, θα πρέπει να παραχθούν τα αντίστοιχα διαγράμματα (π.χ. scatter plots) και τα confusion matrices, για την καλύτερη δυνατή παρουσίαση των αποτελεσμάτων.

Ερώτημα 3: classification – υλοποίηση με keras/tensorflow (2,0 μονάδες).

Χρησιμοποιήστε το dataset με στόχο να εκπαιδεύσετε ένα νευρωνικό δίκτυο MLP, το οποίο θα μαθαίνει από τα 21 χαρακτηριστικά και θα πραγματοποιεί κατηγοριοποίηση σε 10

μορφολογικά μοτίβα (FHR). Θα πρέπει να χρησιμοποιήσετε τον παρακάτω χωρισμό των δεδομένων σε σύνολα:

- Σύνολο Εκπαίδευσης: 60%
- Σύνολο Αξιολόγηση: 10%
- Σύνολο Ελέγχου: 30%

Επιλέξτε τουλάχιστον δύο στατιστικούς δείκτες για την αξιολόγηση του μοντέλου στα σύνολα αξιολόγησης και ελέγχου.

Το νευρωνικό δίκτυο θα πρέπει να έχει τη στοιβάδα εισόδου, 2 κρυφές στοιβάδες και τη στοιβάδα εξόδου. Θα πρέπει να πειραματιστείτε για να βρείτε τον αριθμό των νευρώνων στις κρυφές στοιβάδες. Παρατηρείτε σημαντικές αποκλίσεις στις τιμές των στατιστικών δεικτών όταν αλλάζετε τον αριθμό των νευρώνων; Εάν ναι, πού πιστεύετε ότι οφείλεται αυτό; Περιγράψτε τη διαδικασία και εξηγήστε τα αποτελέσματα που προκύπτουν. Τέλος, θα πρέπει να παραχθούν τα αντίστοιχα διαγράμματα (π.χ. history loss function) για την καλύτερη δυνατή παρουσίαση των αποτελεσμάτων.

Τα δεδομένα θα πρέπει να δίνονται στο νευρωνικό δίκτυο κανονικοποιημένα.

ΜΕΡΟΣ Β:

Θα εργαστείτε με ένα σύνολο δεδομένων από το UCI ML Repository και συγκεκριμένα το ISTANBUL STOCK EXCHANGE dataset, το οποίο μπορείτε να βρείτε στον παρακάτω σύνδεσμο:

https://archive.ics.uci.edu/ml/datasets/ISTANBUL+STOCK+EXCHANGE

Το εν λόγω σύνολο δεδομένων αποτελείται από 536 τιμές δεικτών χρηματιστηρίου, 7 δείκτες χρηματιστηρίου (SP, DAX, FTSE, NIKKEI, BOVESPA, EU, EM) και στόχος είναι η πρόβλεψη του δείκτη Istanbul stock exchange (ISE) (είτε TL BASED, είτε USD BASED). Για περισσότερες λεπτομέρειες δείτε τις σχετικές δημοσιεύσεις που αναφέρονται στη σελίδα περιγραφής του dataset.

Οι επιμέρους εργασίες που πρέπει να πραγματοποιηθούν είναι:

Ερώτημα 1: Προπαρασκευή δεδομένων – υλοποίηση με εργαλεία στατιστικής επεξεργασίας, π.χ. R/Python (0,5 μονάδες).

Από το παραπάνω dataset επιλέξετε τα πρώτα 530 δεδομένα (δηλ, dataset.iloc[0:529]), διαγράψτε τη στήλη date και μετασχηματίστε τη χρονοσειρά σε πρόβλημα supervised και στη μορφή [samples, steps, features].

Ερώτημα 2: Time-series prediction – υλοποίηση με keras/tensorflow (4,5 μονάδες)

a) Time-series prediction με MLP

Χρησιμοποιήστε το dataset με στόχο να εκπαιδεύσετε ένα νευρωνικό δίκτυο MLP, το οποίο θα μαθαίνει χρησιμοποιώντας όσα steps κρίνετε εσείς απαραίτητο. Μπορείτε να δοκιμάσετε διάφορες τιμές για steps.

b) Time-series prediction με RNN

Χρησιμοποιήστε το dataset με στόχο να εκπαιδεύσετε ένα νευρωνικό δίκτυο SimpleRNN, το

οποίο θα μαθαίνει χρησιμοποιώντας όσα steps κρίνετε εσείς απαραίτητο. Λόγω της ανατροφοδότησης, το δίκτυο αυτό «θυμάται», οπότε προσέξτε πώς θα φτιάξετε τα δεδομένα στη μορφή [samples, steps, features]. Μπορείτε να δοκιμάσετε διάφορες τιμές για steps.

Επιλέξτε τουλάχιστον έναν στατιστικό δείκτη. Συγκρίνετε τις δύο προσεγγίσεις. Περιγράψτε τις δύο διαδικασίες και εξηγήστε τα αποτελέσματα που προκύπτουν. Το μοντέλα που θα φτιάξετε μπορεί να μην δίνουν καλές προβλέψεις. Εξηγήστε γιατί.

Τα δεδομένα θα πρέπει να δίνονται στο νευρωνικό δίκτυο κανονικοποιημένα.