ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΑΚΑΔΗΜΑΙΚΌ ΕΤΟΣ 2020-2021

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Καταληκτική ημερομηνία 15/01/2022

Σκοπός της εργασίας είναι η υλοποίηση ενός ολοκληρωμένου project μηχανικής μάθησης. Για τον λόγο αυτό οι φοιτητές καλούνται να εφαρμόσουν τις γνώσεις που έχουν αποκομίσει κατά την διάρκεια του εξαμήνου πάνω σε ένα σύνολο δεδομένων.

Εκφώνηση

Μέρος 1ο (10 Μονάδες)

1) Μέσω της σελίδας Yahoo Finance (https://finance.yahoo.com/quote/BTC-USD?p=BTC-USD&.tsrc=fin-srch) αποθηκεύστε τα δεδομένα σχετικά με τις ιστορικές τιμές μετοχών του Bitcoin των τελευταίων 5 ετών.

Σημείωση: Η αποθήκευση πραγματοποιείται σε αρχείο .csv αλλά μπορείτε, εναλλακτικά, να εργαστείτε πάνω στα δεδομένα κάνοντας χρήση της βιβλιοθήκης pandas και yfinance για απ' ευθείας επεξεργασία/χρήση των δεδομένων χωρίς αποθήκευση

2) Επεξεργαστείτε κατάλληλα τα δεδομένα και οπτικοποιήστε τα.
3) Εφαρμόστε πάνω στις τιμές κλεισίματος όλες τις φάσεις

- Preprocessing: normalization
- Learning: training, cross-validation
- Diagnostics: testing, accuracy, loss

για τις μεθόδους:

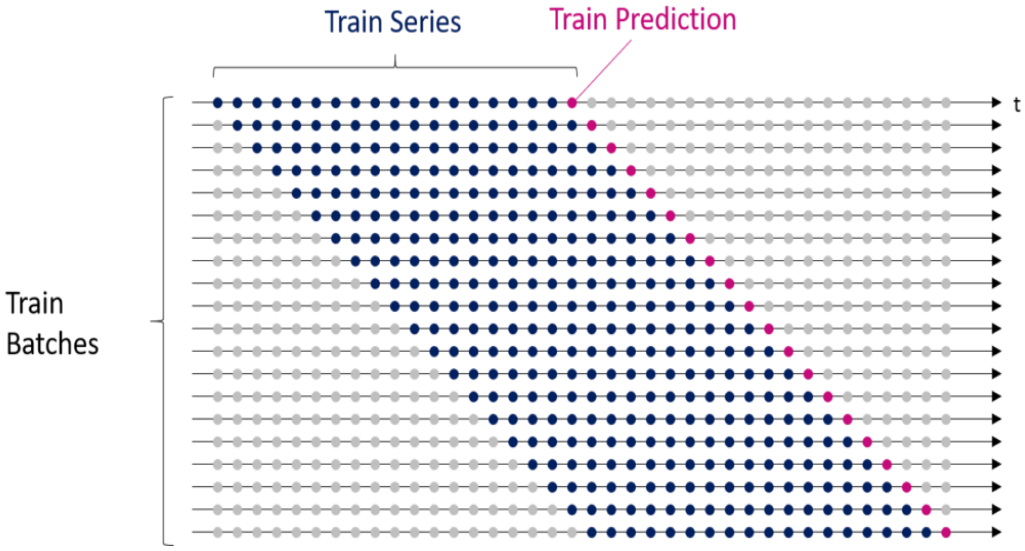
1) **Γραμμικής παλινδρόμησης**
2) **Λογιστικής παλινδρόμησης**

Σημείωση 1: Προσοχή! τα δεδομένα μας αποτελούν χρονοσειρές και για το λόγο αυτό τα dataset της κάθε φάσης θα πρέπει να επιλεγούν κατάλληλα

3) Εφαρμόστε πάνω στις τιμές κλεισίματος ένα **Neural Network έτσι ώστε να εκπαιδεύσετε το δίκτυο στο σύνολο των δεδομένων και στην συνέχεια να πραγματοποιήσετε μια πρόβλεψη για την επόμενη μέρα βασισμένοι στις τιμές κλεισίματος των τελευταίων 50 ημερών.**

Προεραιτικά ακολουθήστε τα παρακάτω:

- Χωρίστε τα δεδομένα σας σε 80% training data και 20% test data
- Σαν βέλτιστη πρακτική normalization, μπορείτε να χρησιμοποιήσετε την συνάρτηση **MinMaxScaler** (από την sklearn.preprocessing) για να ομαλοποιήσετε τις τιμές στα δεδομένα σας σε ένα εύρος μεταξύ 0 και 1.
- Τα νευρωνικά δίκτυα μαθαίνουν σε μια επαναληπτική διαδικασία και πάνω στο σύνολο των δεδομένων. Για το λόγο αυτό, θα πρέπει να δημιουργήσετε τα δεδομένα εκπαίδευσης με βάση τα οποία θα εκπαιδεύσετε το νευρωνικό δίκτυο. Προτείνεται να δημιουργήσετε πολλαπλές πτυχές των δεδομένων εκπαίδευσης, τις λεγόμενες mini-batches που θα περιέχουν 50 συνεχόμενες τιμές η κάθεμια. Έτσι, κατά τη διάρκεια της εκπαιδευτικής διαδικασίας, το νευρωνικό δίκτυο θα είναι σε θέση να επεξεργαστεί τις πτυχές μία προς μία και θα δημιουργεί ξεχωριστή πρόβλεψη για κάθε mini-batch (βλέπε εικόνα)



- Επιπλέον το μοντέλο χρειάζεται μια δεύτερη λίστα για να αξιολογήσει την ποιότητα της πρόβλεψης, και η οποία θα περιέχει τις πραγματικές τιμές. Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο συγκρίνει τις προβλέψεις με τις πραγματικές τιμές και υπολογίζει το σφάλμα εκπαίδευσης για να το ελαχιστοποιήσει με την πάροδο του χρόνου.
- Σαν αρχιτεκτονική μοντέλου προτείνεται να χρησιμοποιήσετε 2 κρυφά επίπεδα ανάμεσα στο input (των 50 νευρώνων) και το output (του ενός νευρώνα)
- Για την αξιολόγηση της απόδοσης του μοντέλου προτείνεται ο υπολογισμός του mean squared error (RMSE)

4) Για όλες τις μεθόδους παράξτε ένα συγκριτικό διάγραμμα και αποφανθείτε για την καταλληλότητα της κάθεμίας.

Μέρος 2ο (Προαιρετικό -- Bonus 1 Μονάδα)

Καλείστε να **υλοποιήσετε και να εφαρμόσετε, όπως στο Μέρος 1,** ένα **μοντέλο κινητού μέσου.**

Ένα **μοντέλο κινητού μέσου,** αντί να χρησιμοποιεί προηγούμενες τιμές της μεταβλητής πρόβλεψης σε μια παλινδρόμηση, **χρησιμοποιεί προηγούμενα σφάλματα πρόβλεψης σε ένα μοντέλο που μοιάζει με παλινδρόμηση.**

Προηγούμενα χρονικά σημεία των δεδομένων χρονοσειρών μπορούν να επηρεάσουν τα τρέχοντα και μελλοντικά χρονικά σημεία. Θα θεωρήσουμε λοιπόν ένα μοντέλο που λαμβάνει υπόψη αυτήν την έννοια όταν προβλέπει τρέχουσες και μελλοντικές τιμές. Το μοντέλο θα χρησιμοποιεί έναν αριθμό παλιότερων παρατηρήσεων για να προβλέψει τις παρατηρήσεις. Εφαρμόζεται ένα βάρος σε κάθε έναν από τους προηγούμενους όρους και οι σταθμίσεις μπορεί να διαφέρουν ανάλογα με το πόσο πρόσφατες είναι.

Ένα μοντέλο κινητού μέσου q βαθμού έχει τον τύπο

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q},$$

Όπου:

a) ο παράγοντας ε_t ονομάζεται λευκός θόρυβος ή σφάλμα (κανονικά κατανεμημένος με μέσο μηδέν και διακύμανση 1)
b) κάθε τιμή του y_t μπορεί να θεωρηθεί ως ένας σταθμισμένος κινητός μέσος όρος των τελευταίων **σφαλμάτων πρόβλεψης**
c) θ_i τα βάρη

Για παράδειγμα το μοντέλο 1ου βαθμού είναι το

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1}.$$

Δλδ, το σταθμισμένο σφάλμα την χρονική στιγμή t υπολογίζεται ως το σφάλμα την χρονική στιγμή t αυξημένο κατά το σταθμισμένο σφάλμα την στιγμή t-1

Από την άλλη το πιο πρόσφατο σφάλμα μπορεί να γραφτεί ως γραμμική συνάρτηση τρεχουσών και προηγούμενων παρατηρήσεων, δλδ στη μορφή:

$$\varepsilon_t = \sum_{j=0}^{\infty} (-\theta)^j y_{t-j}.$$

Επίσης, όταν |θ|>1 τα βάρη αυξάνονται όσο αυξάνονται οι αποστάσεις, επομένως όσο πιο απομακρυσμένες είναι οι παρατηρήσεις τόσο μεγαλύτερη είναι η επιρροή τους στο τρέχον σφάλμα. Όταν |θ|=1 τα βάρη είναι σταθερά σε μέγεθος, και οι μακρινές παρατηρήσεις έχουν την ίδια επιρροή με τις πρόσφατες παρατηρήσεις. Καθώς καμία από αυτές τις καταστάσεις δεν έχει πολύ νόημα, απαιτούμε οι πιο πρόσφατες παρατηρήσεις να έχουν μεγαλύτερο βάρος από τις παρατηρήσεις από το πιο μακρινό παρελθόν. Έτσι, η διαδικασία είναι χρήσιμη όταν |θ|< 1.

Αφού δημιουργήσετε το μοντέλο σας, δοκιμάστε το για βαθμό q=1 και q=2

Σημείωση: το παραδοτέο πρέπει να περιέχει τον κώδικα και να συνοδεύεται από αρχείο στο οποίο οι φοιτητές θα στοιχειοθετούν όλες τις αποφάσεις τους σε σχέση με την υλοποίηση και τα συμπεράσματα τους