

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ 2020-2021

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Καταληκτική ημερομηνία 9/01/2022

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 1^ο

- 1) Μέσω της σελίδας Google ή Yahoo Finance 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 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 (learning-cross validation-testing-normalization-accuracy) για τις μεθόδους:
 - 1) Γραμμικής παλινδρόμησης
 - 2) Λογιστικής παλινδρόμησης
 - 3) **Support Vector Machine με Radial Basis Function Kernel**

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

- 4) Εφαρμόστε πάνω στις τιμές κλεισίματος ένα **Neural Network**. Λόγω του μεγάλου όγκου δεδομένων κατά την εφαρμογή του NN, τα δεδομένα θα πρέπει να διαμοιραστούν και να εκπαιδευτούν σε «εποχές». Για να επιλέξετε τον καλύτερο συνδυασμό στοιχείων στην είσοδο/εποχών θα πρέπει να υλοποιήσετε μια συνάρτηση που θα σας βοηθήσει να

αποφασίσετε πόσες εισόδους πρέπει να έχει το δίκτυο. Αυτή η συνάρτηση θα λαμβάνει τον αριθμό των εισόδων στο δίκτυο για να ελέγξει τον αριθμό των εποχών για εκπαίδευση. Η συνάρτηση θα δημιουργήσει ένα δίκτυο, θα προετοιμάσει δεδομένα για αυτό, στη συνέχεια θα εκπαιδεύσει το δίκτυο και θα αξιολογήσει την απόδοσή του στο δοκιμαστικό σύνολο παράγοντας σαν output ένα συγκριτικό πίνακα για το λάθος υπολογισμού σε σχέση με τον αριθμό των στοιχείων στο input και των αριθμό των εποχών.

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

Μέρος 2^ο

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

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

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 1^{ου} 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 $|\theta| > 1$ τα βάρη αυξάνονται όσο αυξάνονται οι αποστάσεις, επομένως όσο πιο απομακρυσμένες είναι οι παρατηρήσεις τόσο μεγαλύτερη είναι η επιρροή τους στο τρέχον σφάλμα. Όταν $|\theta| = 1$ 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 $|\theta| < 1$.

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