ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

ΕΡΓΑΣΙΑ 1: Πρόβλεψη Τιμών Μετοχών με Γραμμική Παλινδρόμηση

Φλουρής Παντελής, up1093507

A. partA.py

Σε αυτο το υποερώτημα ο κωδικας χρησιμοποιεί ενα γραμμικο μοντέλο παλινδρώμησης για να βρει την σχεση μεταξυ προηγούμενων τιμών κλεισίματως και της επόμενης.

Αρχικά διαβάζεται το csv αρχειο με τα δεδομένα, και φιλτράρεται το πεδιο Close με sigma=1.

Στην συνέχεια επιλέγονται οι παράμετροι του μοντέλου, που μετά απο πειραματισμό, κατέληξα στους

- close_t-1
- close_t-2
- close_t-3
- weeklyAvg

καθώς με αυτες τις παραμέτρους οι μετρικές σφάλματος εχουν τις μικρότερες τιμές.

	Train Data	Validation Data
RMSE	0.6589731515886386	0.8537146124455577
MAE	0.49463881717007363	0.6288538915355353
MSE	0.4342456145146624	0.7288286395030692

Η διαφορά μεταξύ των δυο σετ δεδομένων απεικονίζει υπερεκπαίδευση του μοντέλου σε κάποιο βαθμό.

Αξίζει να σημειωθεί οτι ο κώδικας λαμβάνει υποψιν τις εγγραφές μετα το 2020, καθώς ετσι **ελαχιστοποιείται η διαφορά μεταξύ των μετρικών σφάλματων αναμεσα στα συνολα εκπαίδευσης και επικύρωσης**. Κατ’ επέκταση ελαχιστοποιείται η υπερεκπαίδευση του μοντέλου.

Αναλυτικά αποτελέσματα του πειραματισμού βρισκονται στο αρχείο partA_experiment_results.txt.

Τα δεδομένα έπειτα χωρίζονται, και τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του μοντέλου.

Στην συνέχεια γίνονται δυο σετ προβλέψεων, ένα πάνω στα δεδομένα εκπαίδευσης και ένα πάνω στα δεδομένα επικύρωσης. Υπολογίζονται και εκτυπώνονται στην οθόνη οι τιμές των σφαλμάτων για τα δυο σετ.

Τυπώνεται στην οθόνη ενα γράφημα που απεικονίζει τη διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών κλεισίματος για τα δεδομένα του 2024. Η γραμμή x=y αντιπροσωπεύει την περίπτωση όπου οι προβλέψεις ταιριάζουν ακριβώς με τις πραγματικές τιμές. Οι κόκκινες τελείες αντιπροσωπεύουν τις προβλεπόμενες τιμές, και η μαύρη γραμμή δείχνει την ιδανική γραμμή x=y, η οποία χρησιμεύει ως αναφορά για την αξιολόγηση της ακρίβειας του μοντέλου.

Το επόμενο γράφημα δείχνει τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών για το 2024, στον άξονα του χρόνου. Οι προβλεπόμενες τιμές εμφανίζονται με κόκκινο χρώμα, ενώ οι πραγματικές τιμές με μπλε διακεκομμένη γραμμή.

Το τρίτο γράφημα απεικονίζει τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών για το σύνολο των δεδομένων από το 2020 έως σήμερα, στον άξονα του χρόνου. Οι προβλεπόμενες τιμές είναι κόκκινες, ενώ οι πραγματικές τιμές φαίνονται με μπλε διακεκομμένη γραμμή.

Στο τέλος του κώδικα βρίσκεται η συνάρτηση που μας δίνεται predict_linear_regression, και μια συναρτηση predict_tomorrow, που τυπώνει την πρόβλεψη για την επόμενη μερα απο την τελευταία των δεδομένων.

Για την υλοποίηση χρησιμοποιείται ο κώδικας που ειναι αναρτημένος στο eclass (Coding_linear_regression.ipynb), αλλαγμένος για να λειτουργεί με το dataset.

B. partB.py

Σε αυτο το υποερώτημα ο κωδικας χρησιμοποιεί ένα πολυωνυμικό μοντέλο παλινδρόμησης με L1, L2 νόρμες κανονικοποίησης για να βρει την σχεση μεταξυ προηγούμενων τιμών κλεισίματως και της επόμενης.

Οπως προηγουμένως, τα δεδομένα φιλτράρονται, και έπειτα χωριζονται σε 60% εκπαίδευσης, 20% επικύρωσης και 20% ελέγχου.

Στην συνέχεια επιλέγονται οι παράμετροι του μοντέλου, που μετά απο πειραματισμό, κατέληξα στους

- close_t-1
- close_t-2
- close_t-3
- weeklyAvg

Η υπερπαράμετρος alpha ειναι 1.0 για την L2 κανονικοποίηση και 0.25 για την L1. Η επιλογή αυτη εγινε πειραματικά, συμφωνα με την τιμή που μειώνει οσο περισσοτερο γίνεται το MSE.

```
Testing different degrees of polynomial without regularization
MESA-INTEL: warning: ../mesa-24.2.7/src/intel/vulkan/anv_formats.c:763: FINISHME: support YUV colorspace with DRM format modifiers
MESA-INTEL: warning: ../mesa-24.2.7/src/intel/vulkan/anv_formats.c:794: FINISHME: support more multi-planar formats with DRM modifiers
Best degree based on validation error: 2
Test MSE with degree 2 and None regularization: 0.139
Testing different degrees of polynomial with L2 regularization
Best degree based on validation error: 2
Test MSE with degree 2 and L2 regularization: 0.139
Testing different degrees of polynomial with L1 regularization
Best degree based on validation error: 3
Test MSE with degree 3 and L1 regularization: 0.172
```

Παρατηρούμε οτι την καλύτερη απόδοση εχει το μοντέλο χωρις νόρμα κανονικοποίησης, μαζι με το μοντέλο Ridge, ενώ το μοντέλο Lasso εχει χειρότερη απόδοση.

Η χειρότερη απόδοση του L1 (Lasso) μπορεί να ευθύνεται στον μηδενισμό συντελεστών που στην πραγματικότητα βοηθούν το μοντέλο.

Η ελλειψη βελτίωσης – μείωση απόδοσης απο τις νόρμες κανονικοποίησης μπορει να σημαίνει οτι το μοντέλο δεν υπερεκπαιδεύεται εξ’αρχής, ή οτι οι υπερπαράμετροι που επέλεξα ειναι λανθασμένες.

Για το κάθε μοντέλο τυπώνεται στην οθόνη η διαφορά σφαλματος για διαφορετικά degree, επιλέγεται αυτόματα το μοντέλο με το μικρότερο σφάλμα και τυπώνεται ενα γράφημα με τα δεδομένα εκπαίδευσης (μπλέ κουκίδες), ταδεδομένα ελέγχου (κόκκινες κουκίδες) και μια πράσινη γραμμη με τις προβλέψεις.

Για την υλοποίηση χρησιμοποιείται ο κώδικας που ειναι αναρτημένος στο eclass (training_L1_L2.ipynb), αλλαγμένος για να λειτουργεί με το dataset.