Μέθοδοι Μηχανικής Μάθησης στη Χρηματοοικονομική

Εργασία 2

Καταληκτική ημερομηνία: 5 Μαΐου. Καλείστε να παραδώσετε ένα pdf αρχείο με τις απαντήσεις σας σε όλες τις ερωτήσεις. Στις απαντήσεις σας παρακαλείστε να έχετε βασικά screenshots από το excel/python output. Το pdf αρχείο θα πρέπει να συνοδεύεται και από ένα zip αρχείο με τους excel/python κώδικές σας.

Ενότητα 4: Unsupervised Learning

Εφαρμογή: Κίνδυνος χώρας (Country Risk)

Θέλετε να κατανοήσετε τον κίνδυνο των χωρών προτού προβείτε σε κάποια επένδυση. Θεωρείτε ότι τα παρακάτω χαρακτηριστικά (features) είναι σημαντικά για την ανάλυσή σας: Peace Index (scale 1 (very peaceful) – 5 (not at all peaceful)), Legal Risk Index (scale 0-10 with high values being favorable), GDP growth και Corruption Index (scale 0 (highly corrupt) – 100 (no corruption)). Το αρχείο *Country_risk_2019_data.csv* της ενότητας 4 του e- class περιλαμβάνει δεδομένα για 121 χώρες για το έτος 2019.

1. Χρησιμοποιείστε το Hierarchical Clustering βασιζόμενοι στα παρακάτω χαρακτηριστικά Peace Index, Legal Risk Index και GDP growth για να ομαδοποιήσετε τις χώρες με βάση τον κίνδυνο σε τρία clusters. Συγκρίνετε τις χώρες που βρίσκονται στο high risk cluster όταν χρησιμοποιείται ο αλγόριθμος k-means (όπως είδαμε στη διάλεξη) και όταν χρησιμοποιείται το Hierarchical Clustering. Συμβουλευτείτε το AgglomerativeClustering documentation της Python

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Χρησιμοποιείστε δύο μέτρα εγγύτητας , (α) τη μέση απόσταση και (β) τη ward method. Τα αντίστοιχα αρχεία για να τρέξετε τους αλγορίθμους είναι τα 4.3 hierarchical_clustering_averagemethod.ipynb και 4.4 hierarchical_clustering_wardmethod.ipynb.

Ποια είναι η διαφορά των δύο μέτρων εγγύτητας; Εξηγείστε τη διαφορά με δικά σας λόγια.

Ενότητα 5: Supervised Learning: Linear Regression

2. Χρησιμοποιείστε το training και validation set του αρχείου *Salary vs. Age Example.xlsx* που είναι αναρτημένο στην ενότητα 3 του e-class. Αφού κανονικοποιείστε τα δεδομένα σας υπολογίστε τα biases, weights και mean squared errors (mse) για τα παρακάτω μοντέλα:

(α) Πολυμεταβλητό γραμμικό μοντέλο με εξαρτημένη μεταβλητή το Salary και ανεξάρτητες μεταβλήτες τις ακόλουθες AGE, AGE^2, AGE^3, AGE^4 και AGE^5 .

(β) Ridge μοντέλο με εξαρτημένη μεταβλητή το Salary και ανεξάρτητες μεταβλήτες τις ακόλουθες AGE, AGE^2, AGE^3, AGE^4 και AGE^5 με λ =0.02, 0.05 και 0.1.

(γ) LASSO μοντέλο με εξαρτημένη μεταβλητή το Salary και ανεξάρτητες μεταβλήτες τις ακόλουθες AGE, AGE^2, AGE^3, AGE^4 και AGE^5 με λ =0.02, 0.05 και 0.1.

Καλείστε η ανάλυση να γίνει με τη χρήση του excel, βασιζόμενοι στο αρχείο 5. *Salary vs. Age Example - Regression.xlsx* που είναι ανερτημένο στην ενότητα 5 του eclass. Ποιο μοντέλο θα επιλέγατε και γιατί; Υπολογίστε το mse για το test set (που βρίσκεται στο αρχείο *Salary vs. Age Example.xlsx* που είναι αναρτημένο στην ενότητα 3 του e-class.) του μοντέλου που επιλέγετε.

3. Χρησιμοποιείστε τα δεδομένα του αρχείου *Original_Data.xlsx* της ενότητας 5. Τα δεδομένα περιλαμβάνουν τιμές πώλησης και διάφορα χαρακτηριστικά για τα σπίτια στην Αϊόβα. Επιλέξτε τουλάχιστον 20 χαρακτηριστικά και υπολογίστε τα mse για τα training και validation sets για τα παρακάτω 4 μοντέλα

(α) πολυμεταβλητή γραμμική παλινδρόμηση, (β) Ridge μοντέλο, (γ) LASSO μοντέλο και (δ) ENET μοντέλο με εξαρτημένη μεταβλητή τις τιμή πώλησης και ανεξάρτητες τα χαρακτηριστικά που επιλέξατε. Θυμηθείτε ότι τα χαρακτηριστικά θα πρέπει να κανονικοποιηθούν. Χρησιμοποιήστε 1800 παρατηρήσεις στο training set, 600 στο validation set και 508 στο test set.

Για την ανάλυσή σας μπορείτε να βασιστείτε στα python αρχεία της ενότητας 5 *5.1 linear_regression.ipynb, 5.2 ridge_regression.ipynb, 5.3 lasso_regression.ipynb* και *5.4 elasticnet_regression.ipynb*. Η επιλογή του λ μπορεί να είναι είτε δική σας είτε να βασιστεί στο k-fold cross-validation. Ποιο μοντέλο θα επιλέγατε και γιατί. Υπολογίστε το mse για το test set του μοντέλου που επιλέγετε.