

Μέθοδοι Μηχανικής Μάθησης στη Χρηματοοικονομική

Εργασία 1

Καταληκτική ημερομηνία: 29 Μαρτίου. 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.

1. Χρησιμοποιείτε τα δεδομένα του αρχείου *Salary vs. Age Example.xlsx* της ενότητας 1. Το αρχείο *Salary vs. Age Example.xlsx* εξετάζει ποιο από τα παρακάτω μοντέλα (α) γραμμικό, (β) πολυωνμικό μοντέλο 2^{ου} βαθμού και (γ) πολυωνμικό 5^{ου} βαθμού γενικεύει καλύτερα τη σχέση Μισθού (SALARY) σε συνάρτηση με την Ηλικία (AGE) για ένα συγκεκριμένο επάγγελμα σε μια συγκεκριμένη περιοχή. Η απάντηση είναι (όπως είδαμε) ότι το πολυωνμικό μοντέλο 2^{ου} βαθμού (quadratic model) γενικεύει καλύτερα τη σχέση βασιζόμενοι στα training και validation sets. Θεωρείστε πολυωνμικά μοντέλα 3^{ου} και 4^{ου} βαθμού και με τη βοήθεια του excel ελέγξτε αν τα μοντέλα αυτά γενικεύουν καλύτερα τη σχέση Μισθού (SALARY) σε συνάρτηση με την Ηλικία (AGE) από το πολυωνμικό μοντέλο 2^{ου} βαθμού βασιζόμενοι στα training και validation sets.

2. Data Quality Report με Python: Στην παρούσα άσκηση καλείστε να εκτελέσετε τον δοθέντα κώδικα για τη δημιουργία ενός συνθετικού dataset πελατών. Στη συνέχεια, με τη βοήθεια της Python να πραγματοποιήσετε βασική διερευνητική ανάλυση δεδομένων (EDA), υπολογίζοντας για τις αριθμητικές μεταβλητές τις εξής περιγραφικές στατιστικές: count, ποσοστό missing, mean, median, standard deviation, minimum, Q1, Q3, maximum και skewness, και για τις κατηγορικές μεταβλητές: count, ποσοστό missing, cardinality και mode (με ποσοστό). Επιπλέον, να δημιουργήσετε ιστογράμματα για όλες τις αριθμητικές μεταβλητές, bar plots για τις

κατηγορικές, scatter plot μεταξύ *income* και *total_spent*, καθώς και correlation matrix για τις αριθμητικές μεταβλητές. Τέλος, να παραδώσετε σύντομο σχολιασμό (10–15 γραμμές) σχετικά με την ασυμμετρία των κατανομών, την ύπαρξη ακραίων τιμών, τις σημαντικότερες συσχετίσεις και τον τρόπο διαχείρισης των missing values.

3. 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 5 περιλαμβάνει δεδομένα για 121 χώρες για το έτος 2019.

(α) Το αρχείο *5.1 K means_elbow.ipynb* χρησιμοποιεί τα παρακάτω τρία χαρακτηριστικά Peace Index, Legal Risk Index, GDP growth για να ομαδοποιήσει τις χώρες με βάση τον κίνδυνο. Χρησιμοποιείται ο k-means αλγόριθμος, και ο αριθμός των clusters βασίζεται στην μέθοδο elbow. Προσθέστε και το τέταρτο χαρακτηριστικό Corruption Index και επαναλάβετε τον αλγόριθμο k-means. Συγκρίνετε τις χώρες που είναι στο high risk cluster όταν τρέχετε τον αλγόριθμο με τα 3 και με τα 4 χαρακτηριστικά.

(β) Η Βενεζουέλα δεν περιλαμβάνεται στις 121 χώρες της Εφαρμογής: Κίνδυνος χώρας (Country Risk). Οι παρατηρήσεις της είναι σχετικά ακραίες. Το Peace Index, Legal Risk Index, GDP growth και το Corruption Index της Βενεζουέλας είναι ίσο με 2.671, 2.895, -35% και 16, αντίστοιχα. Ενσωματώστε την Βενεζουέλα στο δείγμα σας, χρησιμοποιείστε τα τρία χαρακτηριστικά Peace Index, Legal Risk Index, GDP growth και ομαδοποιείστε τις χώρες με βάση τον κίνδυνο χρησιμοποιώντας τον αλγόριθμο k-means. Θεωρείστε 3 clusters. Τι παρατηρείτε σε σχέση με τα

αποτελέσματα στην περίπτωση που η Βενεζουέλα δε βρίσκεται στο δείγμα; Είναι ο αλγόριθμος k-means ευαίσθητος στις ακραίες τιμές (outliers);

4. Ενδιαφέρεστε να μελετήσετε την επίδραση των **μεταβολών των επιτοκίων** στις **αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α.** Το αρχείο *US interest rates.xlsx* που βρίσκεται στην ενότητα 5 του e-class περιλαμβάνει τα US dollar Treasury interest rates με λήξεις (maturities) 1 έως 30 χρόνια και το αρχείο *FF-factors.csv* τις υπερβάλλουσες αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α. (Mkt-Rf) σε ημερήσια βάση. Επιθυμείτε να «περιορίσετε» τον αριθμό των παραγόντων (δηλαδή τα interest rates) σε έναν μικρότερο αριθμό παραγόντων (factors) που να έχουν πληροφόρηση των μεταβολών από τις διαφορετικές λήξεις. Για τον λόγο αυτό σκοπεύετε να χρησιμοποιήσετε principal component analysis (PCA).

(α) Βρείτε τα factor loadings, τους PCA factors και το ποσόστο της διακύμανσης που εξηγεί ο κάθε PCA factor.

(β) Με πόσους PCA factors θα συνεχίζατε τη μελέτη σας και γιατί;

(γ) Εκτιμήστε το μοντέλο παλινδρόμησης (με τη βοήθεια του excel) με εξαρτημένη μεταβλητή τις αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α. και ανεξάρτητες μεταβλητές τον αριθμό των PCA factors που θα επιλέξετε.

(δ) Θα μπορούσατε να εκτιμήσετε ένα μοντέλο γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τις αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α. και ανεξάρτητες μεταβλητές όλες τις μεταβολές των επιτοκίων με διαφορετικές λήξεις;