

# Μέθοδοι Μηχανικής Μάθησης στη Χρηματοοικονομική

## Εργασία 1

**Καταληκτική ημερομηνία: Κυριακή 30 Απριλίου.** 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 3: Εισαγωγή στη Μηχανική Μάθηση Μέθοδοι**

1. Χρησιμοποιείστε τα δεδομένα του αρχείου *Salary vs. Age Example.xlsx* που είναι αναρτημένο στην ενότητα 3 του e-class. Το αρχείο *Salary vs. Age Example.xlsx* εξετάζει ποιο από τα παρακάτω μοντέλα (α) γραμμικό, (β) πολυωνυμικό μοντέλο 2<sup>ου</sup> βαθμού και (γ) πολυωνυμικό 5<sup>ου</sup> 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 πολυωνυμικό μοντέλο 2<sup>ου</sup> βαθμού (quadratic model) γενικεύει καλύτερα τη σχέση βασιζόμενοι στα training και validation sets.

Θεωρείστε πολυωνυμικά μοντέλα 3<sup>ου</sup> και 4<sup>ου</sup> βαθμού και με τη βοήθεια του excel ελέγξτε αν τα μοντέλα αυτά γενικεύουν καλύτερα καλύτερα τη σχέση Μισθού (SALARY) σε συνάρτηση με την Ηλικία (AGE) από το πολυωνυμικό μοντέλο 2<sup>ου</sup> βαθμού βασιζόμενοι στα training και validation sets.

### **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.

2. Το αρχείο *4.1 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.

3. 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, χρησιμοποιήστε τα τρία χαρακτηριστικά Peace Index, Legal Risk Index, GDP growth και ομαδοποιήστε 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);

#### Εφαρμογή: Interest rates movements

4. Ενδιαφέρεστε να μελετήσετε την επίδραση των μεταβολών των επιτοκίων στις αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α. Το αρχείο *interest rates.csv* που βρίσκεται στην ενότητα 4 του e-class περιλαμβάνει τις μεταβολές των US dollar Treasury interest rates με λήξεις (maturities) 2, 3, 5, 7, 10, και 30 χρόνια και τις αποδόσεις του Δείκτη χρηματιστηρίου Η.Π.Α. για την περίοδο

Μάιος 1983 - Μάρτιος 2023 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;