Οδηγίες Τελικής Εργασίας για το μάθημα “Python”, BIS-Analytics 2022

(Η συγγραφή της εργασίας μπορεί να ξεκινήσει μετά το μάθημα για τη βιβλιοθήκη “pandas”.)

Αντικείμενο Εργασίας

Για την τελική εργασία, θα επιλέξετε τουλάχιστον δύο διαφορετικά (ή τρία αν προτιμάτε) σύνολα δεδομένων και θα εφαρμόσετε όλους τους αλγόριθμους που παρουσιάστηκαν στις διαλέξεις του μαθήματος. Οι αλγόριθμοι διακρίνονται σε τρεις μεγάλες κατηγορίες: Παλινδρόμηση (Regression), Clustering (Συσταδοποίηση/Ομαδοποίηση), Κατηγοριοποίηση (Classification). Για τους αλγόριθμους Κατηγοριοποίησης και Συσταδοποίησης μπορεί να χρησιμοποιηθεί το ίδιο ή διαφορετικό σύνολο δεδομένων.

Δομή και Περιεχόμενο Εργασίας

Κάθε μεγάλη κατηγορία αλγορίθμων θα πρέπει να υλοποιηθεί σε ένα ξεχωριστό αρχείο *.ipynb, δηλαδή θα παραδοθούν 3 αρχεία τύπου “interactive python notebooks”.

Ο τίτλος κάθε αρχείου θα έχει “snake_case” μορφή, θα αποτελείται από παρόμοιο, αλλά διαφορετικό πρόθεμα (το επίθετο και το πρώτο γράμμα του ονόματος ενωμένο με το επίθετο) και τέλος από κοινό επίθεμα με το πλήρες όνομα της κατηγορίας αλγορίθμων, χωρίς καθόλου κεφαλαία. Δηλαδή, αν ο φοιτητής ονομάζεται Αργυρίου Θανάσης, θα παραδώσει 3 *.ipynb αρχεία με τα εξής ονόματα (χωρίς κεφαλαία):

- argyriou_t_regression.ipynb
- argyriou_t_clustering.ipynb
- argyriou_t_classification.ipynb

Τα αρχεία *.ipynb θα είναι όλα σε ένα φάκελο και μέσα στο φάκελο αυτό **θα υπάρχει ξεχωριστός φάκελος data, όπου θα υπάρχουν οπωσδήποτε και τα αναγκαία αρχεία δεδομένων.**

Τα αρχεία θα έχουν παρόμοιες ενότητες και δομή. Μπορείτε αρχικά να φτιάξετε ένα «πρότυπο» αρχείο μόνο με «κελιά» με markdown και μετά να δημιουργήσετε τα άλλα 2 αρχεία. Πιο συγκεκριμένα, κάθε αρχείο θα έχει τις εξής ενότητες, διακριτά οριοθετημένες με χρήση γλώσσας “markdown”:

1) Εισαγωγή των απαραίτητων βιβλιοθηκών.

2) Διάβασμα (άνοιγμα) του αρχείου δεδομένων και συνοπτική περιγραφή του είδους και των χαρακτηριστικών (features) και της μεταβλητής στόχου (target variable). Δηλ, για κάθε χαρακτηριστικό θα αναφέρετε επιγραμματικά, αν είναι ποσοτική ή κατηγορική και τι μετράει/καταγράφει, τι τιμές παίρνει.

3) Διερευνητική και γραφική παρουσίαση των χαρακτηριστικών των δεδομένων (διερευνητική ανάλυση – Exploratory Data Analysis) με τη χρήση των κατάλληλων διαγραμμάτων ανάλογα με το είδος του χαρακτηριστικού (μεταβλητής). Τα σχόλια μπορεί να είναι συνοπτικά, χωρίς να αποκλείεται η δυνατότητα πιο εκτεταμένου σχολιασμού αν αυτό κρίνεται χρήσιμο για την κατανόηση των δεδομένων.

4) Συνοπτική περιγραφική στατιστική των χαρακτηριστικών (Descriptive Statistics) των δεδομένων και αν κρίνεται απαραίτητο και ορισμένων συσχετίσεων μεταξύ τους.

5) Προετοιμασία των δεδομένων, μετατροπή στον κατάλληλο τύπο όπου χρειάζεται και διάκριση των δεδομένων στο υποσύνολο «εκπαίδευσης» του αλγόριθμου και στο υποσύνολο «δοκιμής» του. Αν το κρίνετε απαραίτητο, θα πρέπει να γίνει μετατροπή των τιμών των δεδομένων στην κατάλληλη κλίμακα, μετονομασία στηλών ή/και δημιουργία «παράγωγων» χαρακτηριστικών.

6) Εφαρμογή του αλγόριθμου για την εκτίμηση προβλέψεων (στο σετ δοκιμής) για την μεταβλητή στόχο, εφαρμόζοντας τουλάχιστον 2 διαφορετικές παραλλαγές, τροποποιήσεις, παραμετροποιήσεις, του κάθε

αλγορίθμου. Π.χ στην Παλινδρόμηση, μπορείτε να αφαιρέσετε χαρακτηριστικά που δεν επηρεάζουν σημαντικά το αποτέλεσμα, στη Συσταδοποίηση να δοκιμάστε περισσότερες ή λιγότερες συστάδες, στο KNN μικρότερο αριθμό γειτνιαζόντων δειγμάτων, κλπ. Η επικρατέστερη/βέλτιστη επιλογή που δίνει το καλύτερο αποτέλεσμα πρέπει να είναι τεκμηριωμένη συνοπτικά στην επόμενη ενότητα «αξιολόγησης».

Στην περίπτωση των αλγορίθμων Κατηγοριοποίησης, η ενότητα θα είναι χωρισμένη σε υπο-ενότητες, στις οποίες θα γίνεται εφαρμογή όλων των αλγορίθμων κατηγοριοποίησης με τη σειρά των διαλέξεων.

Μετά τις προβλέψεις στα δεδομένα του υποσυνόλου του τεστ, ζητείται επίσης η παρουσίαση της πρόβλεψης της τιμής της μεταβλητής στόχου για 1 παρατήρηση τιμών των χαρακτηριστικών (δηλ, να φτιάξετε μια νέα «σειρά» με τιμές για τον πίνακα *X* και να βρείτε την τιμή που προβλέπει ο αλγόριθμος για την *y*).

7) Αξιολόγηση αποτελεσματικότητας του αλγόριθμου, ως προς την ικανότητα πρόβλεψης, (όχι την ταχύτητα του αλγορίθμου). Η αξιολόγηση χρειάζεται να γίνει με την κατάλληλη μέθοδο μέτρησης ανάλογα τον αλγόριθμο. Ο σχολιασμός επίσης θα είναι συνοπτικός.

Στην περίπτωση των αλγορίθμων κατηγοριοποίησης, εκτός από την αξιολόγηση μεταξύ τουλάχιστον δυο εφαρμογών του ίδιου αλγόριθμου θα γίνει και αξιολόγηση μεταξύ όλων των διαφορετικών αλγορίθμων Κατηγοριοποίησης. Αφού παρουσιάσετε τον πιο αποτελεσματικό από άποψη ακρίβειας αλγόριθμο Κατηγοριοποίησης, θα ήταν χρήσιμο να υπάρχουν ορισμένα συνοπτικά σχόλια για τους λόγους για τους οποίους θεωρείτε ότι επιφέρει ο αλγόριθμος πιο «ακριβή» αποτελέσματα σε σχέση με τους υπόλοιπους.

Δεδομένα για την Εργασία

Η επιλογή κατάλληλων δεδομένων είναι μέρος της εργασίας. Μπορείτε να διαλέξετε στους συνδέσμους:

- UCI
- Kaggle datasets
- Kaggle competitions

ή να χρησιμοποιήσετε κάποιο δικό σας σύνολο δεδομένων. Σε όλες τις περιπτώσεις θα υπάρχει σχετική έγκριση για κάθε σύνολο δεδομένων, το συντομότερο δυνατόν πριν την τελική ημερομηνία δήλωσης. *Προσοχή: δεν έχετε διδαχθεί ordered logistic regression, συνεπώς αποφύγετε τα δεδομένα με χαρακτηριστικά ranking (αύξουσα σειρά κατάταξης/ανώτερη ποιότητα) μεταξύ κατηγοριών.*

Προκειμένου τα σύνολα των δεδομένων να είναι όλα διαφορετικά μεταξύ τους, θα σας παρακαλούσα να τα αναρτήσετε στο κοινό αρχείο στον σχετικό σύνδεσμο (Υποβολή Δήλωσης). Σε περίπτωση που υπάρχει κοινή επιλογή δεδομένων, θα τηρηθεί η σειρά προτεραιότητας δήλωσης. Επομένως, **πριν δηλώσετε τα δεδομένα που επιθυμείτε να χρησιμοποιήσετε παρακαλώ να ελέγξετε τι έχουν ήδη δηλώσει οι προηγούμενοι.**

Ζητήστε πρόσβαση ώστε να μπορείτε να επεξεργαστείτε το αρχείο για την Υποβολή Δήλωσης δεδομένων. Επιλογή δεδομένων το αργότερο έως 2 Ιουνίου. **Θα πρότεινα να ξεκινήσετε την εργασία το συντομότερο, θα σας βοηθήσει και για το μάθημα της «Μηχανικής Εκμάθησης».**

Τελική προθεσμία υποβολής των εργασιών: Κυριακή 2 Ιουλίου, ώρα 23:45.

Η προθεσμία υποβολής είναι απόλυτα ανελαστική. Δεν θα δοθεί παράταση για λόγους «φόρτου εργασίας», παρά μόνο για πολύ σοβαρό προσωπικό λόγος.

Είμαι πάντα στη διάθεσή σας για οποιαδήποτε διευκρίνιση ή βοήθεια, για την εργασία, αλλά και μετά το πέρας του μαθήματος και του μεταπτυχιακού. Σας ευχαριστώ για το ενδιαφέρον και τη συμμετοχή σας,

Θανάσης Αργυρίου