EESTech Challenge Thessaloniki

10 Απριλίου 2022

Στο πλαίσιο του διαγωνισμού EESTech Challenge Thessaloniki σας ζητείται να **αναπτύξετε ένα σύστημα μηχανικής μάθησης σε Python για πρόβλεψη της βαρύτητας νόσησης από COVID-19**. Τα δεδομένα που θα χρησιμοποιήσετε για την ανάπτυξη του συστήματος είναι διαθέσιμα στο github repository του διαγωνισμού, στο αρχείο ***covid-data.csv***: https://github.com/EESTechChallengeThessaloniki/EESTechChallengeThess2022

Για τον έλεγχο της απόδοσης των μοντέλων κατηγοριοποίησης (classification) θα πρέπει να χρησιμοποιήσετε το σύνολο δεδομένων ***covid-data-testing.csv***. Για τη δημιουργία των μοντέλων ομαδοποίησης θα πρέπει να χρησιμοποιήσετε τα δεδομένα του αρχείου ***covid-data.csv***. Λεπτομερή επεξήγηση των δεδομένων μπορείτε να βρείτε στο github repository του διαγωνισμού, στο αρχείο ***Dataset Description.pdf***.

Για κάθε παράδειγμα που είναι διαθέσιμο στα δεδομένα υπάρχει μία σειρά από χαρακτηριστικά (features) όπως ηλικία, φύλο, συμπτώματα, κτλ. και μια ετικέτα πρόβλεψης (target) που δηλώνει τη βαρύτητα της νόσησης (ήπια, μέτρια ή σοβαρή). Για τον σκοπο του διαγωνισμού καλείστε:

1) Να **προεπεξεργαστείτε** κατάλληλα τα δεδομένα ώστε να μετατραπούν σε μορφή κατάλληλη για χρήση από αλγορίθμους μηχανικής μάθησης. Δεν υπάρχει περιορισμος στις τεχνικές που μπορείτε να χρησιμοποιήσετε, αλλά θα πρέπει να αιτιολογήσετε σύντομα την επιλογή κάθε τεχνικής που επιλέξατε να εφαρμόσετε. Ενδεικτικά, μπορείτε να επιλέξετε συγκεκριμένα χαρακτηριστικά (features) από το σύνολο δεδομένων ή/και να δημιουργήσετε δικά σας χαρακτηριστικα και να εφαρμόσετε τεχνικές κανονικοποίησης (standardization/normalization) ή/και μείωσης διαστάσεων (dimensionality reduction).
2) **Να διαλέξετε και να εφαρμόσετε έναν ή περισσότερους αλγορίθμους μηχανικής μάθησης για την κατηγοριοποίηση άγνωστων παραδειγμάτων.** Σκοπός του συγκεκριμένου βήματος είναι η δημιουργία μοντέλων ικανών να προβλέπουν τη βαρύτητα νόσησης ενός ασθενή από κορονοϊό. Δεν υπάρχει περιορισμός στον αλγόριθμο που θα επιλέξετε να χρησιμοποιήσετε και σας ενθαρρύνουμε να πειραματιστείτε με διάφορους αλγοριθμους που θεωρείτε ότι ταιριάζουν στο

συγκεκριμένο πρόβλημα. Ενδεικτικοί αλγόριθμοι: δέντρα απόφασης, νευρωνικά δίκτυα, κτλ. **Να αξιολογήσετε την απόδοση του μοντέλου σας σύμφωνα με το σύνολο ελέγχου (βλ. παρακάτω).**

3) **Να διαλέξετε και να εφαρμόσετε έναν ή περισσότερους αλγορίθμους ομαδοποίησης των δεδομένων.** Σκοπός του συγκεκριμένου βήματος είναι η δημιουργία μοντέλων ικανών να ομαδοποιούν ασθενείς που εμφανίζουν παρόμοια συμπτώματα. Δεν υπάρχει περιορισμός στον αλγόριθμο που θα επιλέξετε να χρησιμοποιήσετε και σας ενθαρρύνουμε να πειραματιστείτε με διάφορους αλγοριθμους που θεωρείτε ότι ταιριάζουν στο συγκεκριμένο πρόβλημα. Ενδεικτικοί αλγόριθμοι: k-means, ιεραρχική ομαδοποίηση, κτλ. **Να αξιολογήσετε την απόδοση του μοντέλου σας σύμφωνα με το σύνολο ελέγχου (βλ. παρακάτω).**

Μετρικές Αξιολόγησης: Αφού εκπαιδεύσετε τους αλγόριθμους κατηγοριοποίησης και ομαδοποίησης στο σύνολο εκπαίδευσης (train set), θα πρέπει να αξιολογήσετε την επίδοσή τους στο σύνολο ελέγχου (test set - ***covid-data-testing.csv***). Θα πρέπει να καταγράψετε τις εξής μετρικές (μπορείτε να χρησιμοποιήσετε επιπλέον όσες μετρικές επιθυμείτε):

1) Κατηγοριοποίηση: Accuracy, Precision, Recall και F1 score
2) Ομαδοποίηση: Rand Index, Mutual Information και Silhouette score

Παραδοτέα:

1) Jupyter ή Colab Notebook(s) ή Python αρχεία όπου θα φαίνονται όλα τα στάδια της διαδικασίας προεπεξεργασίας, εκπαίδευσης και αξιολόγησης των αλγορίθμων. Στο τελικό παραδοτέο μπορείτε να συμπεριλάβετε όλες τις προσπάθειες που κάνετε (επιτυχημένες και μη) μέχρι να φτάσετε στην τελική σας επιλογή, καθώς η προσπάθεια σας θα αξιολογηθεί συνολικά και όχι μόνο με βάση τις τιμές των προαναφερθέντων μετρικών. Ενθαρρύνεται ιδιαίτερα να προσθέσετε σχόλια σε κάθε φάση των δοκιμών σας σχετικά με τις αποφάσεις και τις επιλογές σας.
2) Σύντομη αναφορά στην οποία να αναφέρετε τα αποτελέσματα αξιολόγησης των αλγορίθμων σύμφωνα με τις παραπάνω μετρικές

Οδηγίες Υποβολής:

- Μεταβείτε στο github repository του διαγωνισμού, στην ενότητα “Issues”: https://github.com/EESTechChallengeThessaloniki/EESTechChallengeThess2022/issues

- Δημιουργήστε νέο Issue, με τίτλο το όνομα της ομάδας σας. Στο περιεχόμενο του Issue, θα πρέπει να ανεβάσετε ένα αρχείο “.zip” ή “.rar” το οποίο θα πρέπει να περιέχει:
    1) Το(-α) Jupyter ή Colab Notebook(s) ή Python αρχεία με τον κώδικά σας
    2) Σύντομη αναφορά
    3) Οποιοδήποτε άλλο αρχείο θεωρείτε απαραίτητο για την κατανόηση της προσπάθειάς σας.

Κριτήρια Αξιολόγησης:

- Ορθότητα προεπεξεργασίας δεδομένων, εφαρμογής αλγορίθμων και αξιολόγησης. (50%)
- Συνολική προσπάθεια (25%)
- Απόδοση σύμφωνα με τις μετρικές αξιολόγησης (25%)

Χρήσιμες αναφορές:

Προεπεξεργασία δεδομένων

- Data Preprocessing in Machine Learning: 7 Easy Steps To Follow | upGrad blog
- A Simple Guide to Data Preprocessing in Machine Learning (v7labs.com)
- Data Preprocessing in Python - YouTube
- An End-to-End Guide on Data Preprocessing in Machine Learning in Python
- Normalization Techniques in Python Using NumPy
- Introduction to Dimensionality Reduction - GeeksforGeeks
- Dimensionality reduction techniques you should know in 2021 | by Rukshan Pramoditha | Towards Data Science
- Machine Learning - Dimensionality Reduction - Feature Extraction & Selection - YouTube
- Dataset Dimensionality Reduction in Python

Αλγόριθμοι Κατηγοριοποίησης (Classification)

- Supervised Learning: Basics of Classification and Main Algorithms | by Victor Roman | Towards Data Science
- Supervised and Unsupervised Learning In Machine Learning | Machine Learning Tutorial | Simplilearn - YouTube
- GitHub - Awesome-Machine-Learning/Machine-Learning-Classifications

Αλγόριθμοι Ομαδοποίησης (Clustering)

- Clustering — Unsupervised Learning | by Anuja Nagpal | Towards Data Science
- Clustering Algorithms based on centroids namely K-Means Clustering, Agglomerative Clustering and Density Based Spatial Clustering

Μετρικές Αξιολόγησης

- Accuracy Score with scikit-learn
- F1 Score with scikit-learn
- Clustering Evaluation with scikit-learn
- Silhouette Score with scikit-learn
- Accuracy, Precision, Recall or F1?
- Why Accuracy Isn't Everything: Precision and Recall Simply Explained
- Performance Metrics in Machine Learning — Part 3: Clustering

Notebooks

- Welcome To Colaboratory - Colaboratory (google.com)
- Google Colab
- Get started with Google Colaboratory (Coding TensorFlow) - YouTube
- Jupyter/IPython Notebook Quick Start Guide
- Jupyter Notebook