

Μηχανική Μάθηση

1^η Εργασία – Regression problems

Καλείστε να γράψετε κώδικα σε Python στον οποίο:

1ο μέρος: προσέγγιση κατάστασης υποθέτοντας θόρυβο στις μετρήσεις και γνώση του παραμετρικού μοντέλου (γνώση μορφής εξίσωσης, άγνωστες οι τιμές των συντελεστών στους όρους).

1. Θα δημιουργήσετε 150 τυχαίους πραγματικούς αριθμούς που ακολουθούν ομοιόμορφη κατανομή. Φροντίστε ώστε η διάταξη με τους αριθμούς να βρίσκεται σε αύξουσα σειρά. Όλοι οι αριθμοί πρέπει να ανήκουν στο διάστημα $[-4, 4]$.
2. Θα ορίσετε την ακόλουθη συνάρτηση (function): $y = \lambda_1 \cdot \frac{1}{\exp(x)} + \lambda_2 \cdot \sin(x)$, με όνομα συνάρτησης `myCustFunc()`.
Η συνάρτηση `myCustFunc(x, λ_1 , λ_2)` πρέπει να δέχεται σαν όρισμα μια διάταξη αριθμών σαν και αυτή του ερωτήματος 1 και τις τιμές για τις παραμέτρους λ_1 και λ_2 . Θα επιστρέφει μια διάταξη(array) ίδιου μεγέθους με αυτή στην είσοδο.
3. Θα δημιουργήσετε ένα (1) νέο array μεγέθους 100×1 , δίνοντας τιμές για τα λ_1, λ_2 και το διάνυσμα x που προέκυψε από το ερώτημα 1.
4. Δημιουργήστε μία γραφική παράσταση στην οποία θα φαίνονται τα δεδομένα που δημιουργήσατε στα ερωτήματα 1 και 3. Τα `outputs` των συναρτήσεων πρέπει να διαφοροποιούνται: χρησιμοποιήστε διαφορετικά χρώματα και σύμβολα. Μην ξεχάσετε να βάλετε τίτλο, ονόματα αξόνων, και λεζάντα, για να ξέρουμε ποια συνάρτηση δημιούργησε τα εκάστοτε δεδομένα.
5. Εισάγετε θόρυβο στα δεδομένα εισόδου που να ακολουθεί μια κατανομή της επιλογής σας [**εκτός** της $N(0,1)$]. Περιγράψτε την κατανομή που επιλέξατε, στην αναφορά που θα υποβάλετε.
6. Θα κάνετε `fit` τις παραμέτρους στα δεδομένα **με θόρυβο**, δηλ. `[x_input, y_noisy]`, της `myCustFunc`, χρησιμοποιώντας την `scipy.optimize.curve_fit`.
7. Θα ορίσετε μια νέα συνάρτηση, ονόματι `poly4thDegree(x, a, b, c, d, e)`: ένα πολυώνυμο 4^{ου} βαθμού. Προσοχή: η συνάρτηση θα πρέπει να οριστεί με τέτοιο τρόπο ώστε να μπορεί να χρησιμοποιηθεί από την `curve_fit()`, για την εκτίμηση των παραμέτρων.
8. Χρησιμοποιώντας το σύνολο τιμών που έχετε διαθέσιμες, `[x_input, y_noisy]`, προσεγγίστε τις παραμέτρους του πολυωνύμου.

9. Θα κάνετε 2 διαφορετικά plots που θα δείχνουν τις προβλεφθείσες τιμές ($y_{\text{predicted}}$) όταν χρησιμοποιείτε τις `myCustFunc` (ερωτ. 2) και `poly4thDegree` (ερωτ. 7).
 - a. Στο 1^ο θα φαίνονται οι τιμές της `myCustFunc` με τις παραμέτρους που προέκυψαν από το ερώτημα 6.
 - b. Στο 2^ο θα κάνετε ακριβώς τα ίδια για την συνάρτηση `poly4thDegree`.
10. Θα υπολογίσετε τα στατιστικά σφάλματα `mean absolute error` και `root mean squared error`, για κάθε ένα από τους ακόλουθους συνδυασμούς:
 - a. `original_values` - `noisy_values`
 - b. `noisy_values` - `myCustFunc` (παραμέτροι από ερώτημα 6)
 - c. `noisy_values` - `poly4thDegree` (παραμέτροι από ερώτημα 8)

Τα παραπάνω σφάλματα πρέπει να εμφανίζονται στην οθόνη μαζί με μια πρόταση που να εξηγεί σε τι αφορούν οι αριθμοί που μόλις τυπώθηκαν.

Πόσο καλή προσέγγιση σας δίνει το πολυώνυμο για το συγκεκριμένο πρόβλημα;

2ο μέρος: προσέγγιση κατάστασης βασιζόμενοι μόνο στις μετρήσεις που έχετε λάβει (υποθέτετε ότι υπάρχει θόρυβος στις μετρήσεις, δεν έχετε κάποια γνώση σχετική με το μοντέλο που δημιουργεί τα δεδομένα).

1. Χρησιμοποιείτε τις 150 τιμές, *με θόρυβο*, που δημιουργήσατε στο 1^ο μέρος.
2. Διαχωρίστε τα διαθέσιμα δεδομένα, `[x_input, y_noisy]`, σε τρία επιμέρους υποσύνολα: `train`, `validation`, και `test`.
3. Εκπαιδεύστε τρεις διαφορετικούς regressors (π.χ. `kNN`, `SVR`, κ.λπ.) της επιλογής σας, πάνω στο `train set`.
4. Αξιολογήστε την απόδοση των regressors πάνω στο `test set` και δημιουργείτε κατάλληλες γραφικές παραστάσεις που να φαίνονται τα αποτελέσματα.
5. Επαναλάβετε τα βήματα 2 έως 4, αλλά αυτή την φορά κανονικοποιείτε τις τιμές εισόδου στο διάστημα `[0,1]`

Το σύνολο των αποτελεσμάτων θα τα συγκεντρώσετε και θα τα υποβάλετε σε μορφή αναφοράς. Πληροφορίες για την δομή της αναφοράς θα βρείτε στην ενότητα «Οδηγίες».

Οδηγίες:

A. Οι εργασίες είναι σε ομάδες μέχρι τέσσερα (4) άτομα. Κάθε άτομο μπορεί να υποβάλει εργασία σε μία μόνο ομάδα κάθε φορά.

B. Οι εργασίες θα πρέπει να αναρτώνται στο eClass σε ένα αρχείο zip (όχι rar) εντός της προβλεπόμενης προθεσμίας. Δεν θα δοθεί παράταση. **Προσοχή:** Κάθε ομαδική εργασία θα υποβάλλεται μόνο από ένα μέλος της ομάδας (εσείς επιλέγετε ποιος/ποια).

Γ. Κάθε εργασία πρέπει να συνοδεύεται από:

- Ένα και μόνο ένα αρχείο .py θα περιέχει τις απαντήσεις στα ερωτήματα
- Μια **αναφορά** σε pdf με τα ακόλουθα στοιχεία:
 - Εξώφυλλο: 1 σελίδα, περιλαμβάνει τα στοιχεία των φοιτητών της ομάδας, όνομα μαθήματος, ημερομηνία, τμήμα και λοιπά σχετικά στοιχεία.
 - Συγκεντρωτικός πίνακας περιεχομένων, εικόνων, και λοιπών γραφημάτων που παραθέτετε στην αναφορά.
 - Ενότητα εισαγωγή: 1 σελίδα, περιγράφετε το πρόβλημα (*χωρίς* να αντιγράψετε αυτούσια την εκφώνηση της άσκησης)
 - Μέθοδοι που εφαρμόστηκαν: από 2 μέχρι 10 σελίδες, περιγράφετε τις μεθόδους που χρησιμοποιήσατε και παραθέτετε τα σχετικά αποτελέσματα. Φροντίστε να είναι ξεκάθαρο στο ποιο ερώτημα αναφέρεστε.
 - Συμπεράσματα: 1 σελίδα, με βάση τα αποτελέσματα τι προτείνετε, ποιο μοντέλο αποδίδει καλύτερα, τι θα μπορούσε να γίνει για περαιτέρω βελτίωση στην απόδοση.
 - Η αναφορά θα περιέχει γραφικές παραστάσεις κάθε είδους και πίνακες αξιολόγησης των αποτελεσμάτων που πρέπει να συνοδεύονται (έκαστο) από μια τουλάχιστον παράγραφο με σχολιασμό.

Φροντίστε ώστε:

- Ο κώδικας να συνοδεύεται απαραίτητως από κατάλληλα σχόλια.
- Να έχει γίνει συντακτικός και ορθογραφικός έλεγχος στην αναφορά που θα υποβάλετε.
- Οι προτάσεις να είναι κατανοητές και μικρές σε έκταση.
- Οι εικόνες να ***μην*** έχουν προκύψει από print screen. Αν το πρόγραμμα δημιουργεί μια εικόνα αποθηκεύστε την κανονικά (jpg ή png), πριν την χρησιμοποιήσετε.
- Οι γραφικές παραστάσεις να περιλαμβάνουν ονόματα στους άξονες και λεζάντα. Σκοπός είναι να γίνεται κατανοητό τι δείχνει, με μια ματιά.
- Αν κάτι δεν διευκρινίζεται, έχετε το δικαίωμα να κάνετε όποια υλοποίηση σας βολεύει. Φροντίστε να μπορείτε να εξηγήσετε τι ακριβώς κάνατε στον κώδικα.
- Οι βιβλιοθήκες που θα χρησιμοποιήσετε ***πρέπει*** να μπορούν να εγκατασταθούν μέσω του pip.
- Ο κώδικας ***πρέπει*** να τρέχει σε Google Colab.

Καταληκτική Ημερομηνία Παράδοσης: 12 Δεκεμβρίου 2022. Δεν θα δοθεί παράταση.