Assignment 2-Modelling Report

Ομάδα Α:
Νικόλαος-Γεράσιμος Ζαζάτης 2723
Κυριακή Κόφφα 2818
Ηλίας Τσιρώνης 2851

Στο δεύτερο κομμάτι της εργασίας μας ζητήθηκε να υλοποιήσουμε και να εκπαιδεύσουμε ένα νευρωνικό σύστημα το οποίο θα μπορεί να προβλέψει το Engagement το επόμενο λεπτό για τον καθε χρήστη. Αυτά είναι τα πράγματα που παρατηρήσαμε.

Feature selection

Για την είσοδο του νευρωνικού, εκτός από τα κύρια στοιχεία που χρειαζόμαστε, Viewer id και Engagement Sequence, επιλέξαμε να συμπεριλάβουμε και τα στοιχεία Viewer Type, City id και Event Duration. Τα επιπρόσθετα στοιχεία επιλέχθηκαν σε αντίθεση με άλλα, λαμβάνοντας υπόψη correlation που βρήκαμε κατά τη διάρκεια του πρώτου μέρους της εργασίας στο Data Analysis. Τα χαρακτηριστικά που είχαν μεγαλύτερη συσχέτιση με το engagement θεωρήσαμε πως θα ήταν τα πιο σημαντικά για το νευρωνικό έτσι ώστε να μπορεί να βγάλει πιο καλά αποτελέσματα, χωρίς να κινδυνεύουμε από overfitting ή αργό training.

Model Architecture

Το μοντέλο που υλοποιήσαμε για το νευρωνικό έχει 2 embedding layers για τα στοιχεία Viewer id και City id αντίστοιχα για να μπορέσουμε να συσχετίσουμε τις πόλεις και τους χρήστες με κάποια Engagement Sequences, και καθώς οι τιμές τους χωρίς επεξεργασία δεν έχουν κάποια γραμμική σχέση με το επίπεδο του Engagement. Έπειτα προσθέσαμε ένα LSTM layer διότι αποτελεί κοινή επιλογή αρχιτεκτονικής σε περιπτώσεις που θέλουμε να κάνουμε sequence prediction. Το LSTM έχει είσοδο το output του embedding layer όπως και τα υπόλοιπα στοιχεία που έχουν απομείνει. Μετά έχουμε Linear layer, ακολουθούμενο από dropout layer, με πιθανότητα 0.1, για καλύτερο generalisation και τέλος Linear layer με activation function Sigmoid γιατί τα targets παίρνουν τιμές από 0 έως και 1.

Loss function

Για Loss function επιλέξαμε Mean Square Error, ως καλή και κλασική επιλογή σε προβλήματα regression, που είναι και αυτό που αντιμετωπίζουμε εδώ. Συγκεκριμένα

επιλέξαμε αυτό έναντι της εναλλακτικής Mean Absolute Error επειδή και συμπεριφέρεται καλύτερα στο back propagation αλλά και επειδή κάνει μεγάλα λάθη μεγαλύτερα και μικρά σχεδόν αμελητέα.

Train - Validation - Test split

Για να επιλέξουμε ποια δεδομένα θα χρησιμοποιούσαμε για training, validation και testing αντίστοιχα, χρησιμοποιήσαμε μια τεχνική όπου όλες τις πρώτες φορές που εμφανίζεται ένας χρήστης τις μαρκάρουμε για να χρησιμοποιηθούν για testing, τις δεύτερες εμφανίσεις για validation και όλες τις υπόλοιπες για training. Έτσι εξασφαλίζουμε την ποικιλία στο testing για να σιγουρέψουμε πως θα μπορέσουμε να ελέγξουμε την απόδοση σωστά.

Hyperparameter tuning

Για hyperparameter tuning επιλέξαμε και ελέγξαμε τα παρακάτω στοιχεία

Batch size: Χρησιμοποιήσαμε μέγεθος 128 καθώς αποτελεί καλή ισορροπία για να μην υπερφορτώνει την gpu-ram, και για να ολοκληρώνει το training πιο γρήγορα. Αρχικά είχαμε επιλέξει μέγεθος 100, και αν και δεν είναι τεράστια η διαφορά στο χρόνο τρεξίματος, είναι πιο αργό.

Learning rate: Επιλέξαμε αρχικό learning rate 10e-3. Το λέμε αρχικό καθώς ο optimiser που χρησιμοποιούμε έχει την ικανότητα να αλλάζει το learning rate δυναμικά κατά τη διάρκεια της εκπαίδευσης του νευρωνικού.

Optimiser: Διαλέξαμε τον optimiser Adam αφού αποτελεί κλασική επιλογή optimiser. Είναι αποδοτικός και κατάλληλος για το πρόβλημα που θέλουμε να λύσουμε.

Embedding dimension: Μέγεθος 100, καθώς είναι μία επιλογή που δεν προβάλει τα viewer και city id σε μεγάλα διανύσματα, ενώ , παράλληλα, μπορούν αντικατροπτιστούν με μεγάλη ακρίβεια σε κάποιο engagement.

Epoch number: Καταλλήξαμε στα 2 epochs επειδή δεν παρατηρείται μεγάλη διαφορά ανάμεσα στο πρώτο και το δεύτερο στο loss και δεν παίρνει πολύ χρόνο στην ολοκλήρωση του training.

Performance evaluation

Με το πρόβλημά μας να είναι regression τύπου, Mean Square Error είναι η καλύτερη επιλογή για να αξιολογήσουμε την τελική απόδοση του νευρωνικού.

Main observations

Έχουμε εκπαιδεύσει τρία μοντέλα. Το πρώτο έχει ένα LSTM layer με 20 units, και 40 units στο πρώτο linear hidden layer. Το δεύτερο έχει ένα LSTM layer με 200 units, και 400 units στο πρώτο linear hidden layer, ενώ το τρίτο έχει δύο stacked LSTM layers με 200 units, και 400 units στο πρώτο linear hidden layer.

Το test validation score έβγαλε αποτέλεσμα MSE 0.059 στο πρώτο μοντέλο, στο δεύτερο 0.0585 και τέλος στο τρίτο 0.0586. Παρατηρούμε πως επειδή δεν υπάρχει μεγάλη διαφορά στο αποτέλεσμα των τριών αυτών μοντέλων, το 0.058 αποτελεί κάποιο ελάχιστο, αλλα μας είναι δύσκολο να αποφανθούμε αν αυτό είναι τοπικό ή ολικό και στα τρία.

Τέλος και τα τρία μοντέλα παρουσιάζουν train test score 0.04 από το οποίο μπορούμε να συμπεράνουμε πως εμφανίζεται ένα μικρό bias στο training set αλλά αποφεύγουμε περίπτωση overfitting.