CS286: AI for Science and Engineering (Fall 2023)

Homework 3: Prediction of Protein-Protein Interactions

Due time: 11:59 pm, Jan. 3, 2024

Background: Protein-protein interactions (PPIs) are physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by interactions that include electrostatic forces, hydrogen bonding and the hydrophobic effect. Many are physical contacts with molecular associations between chains that occur in a cell or in a living organism.

Task: In this homework assignment, your task is to construct a binary classification model for accurately predicting PPIs. You can use traditional machine learning algorithms such as linear regression and support vector machines (SVM), or deep learning methods based on neural networks with the provided datasets to accomplish the goal. You cannot introduce additional PPI relationships in the training dataset. Finally, we highly recommend Graph Neural Networks (GNNs) in this assignment.

Participation: This homework is held as a Kaggle competition. You need to participate in the competition through the invitation link：
https://www.kaggle.com/t/5db5f5addf274b04a7eea5d4cbdbc7ec
Note: You must first click the above invitation link to participate in the competition. Otherwise, you would not have permission to access the data download page or the submission page.

Dataset and Evaluation Criteria: Please download the training dataset and test dataset from https://www.kaggle.com/competitions/homework-3-cs286-fall-2023/data. The performance of your model will be evaluated based on the Area Under the Receiver Operating Characteristic curve (AUROC).

Submission: Please submit your predictions to the Kaggle platform (https://www.kaggle.com/competitions/homework-3-cs286-fall-2023) before the homework deadline. You are allowed to try at most 5 times per day. You should select two submissions for the final evaluation, of which the one with better performance will be used to calculate the score of your solution. Please upload the final code of your model and a screenshot of your team name on the Kaggle leaderboard to Blackboard before the homework deadline.

Scoring: The final scores of this homework consist of two parts (capped at 100):

- Basic score (70): This score is to evaluate the correctness and completeness of your solution (i.e. complete code and prediction submission).

- Performance score (AUROC − 0.5)×100: This score is for the final evaluation of your model's performance. The higher the AUROC, the higher is your performance score.