Blue text on a black background

Description automatically generated

Data Mining

Course Project

2023/2024

Group 10

André Borgas, 60436 - 12 Hours

Bernardo Godinho, 51497 - 12 Hours

Pedro Sousa, 54729 - 12 Hours

### Report on Latent Factors for Protein-Ligand Activity Prediction

#### Introduction

Latent factor models are a popular technique used in recommendation systems to predict user preferences based on past interactions. In this project, we apply a latent factor model to predict the activities of protein-ligand interactions. The dataset consists of activities measured between various proteins and ligands (ChEMBL IDs). This report details the implementation, evaluation, and results of the latent factor model approach using Singular Value Decomposition (SVD) and Stochastic Gradient Descent (SGD) for optimization.

The code for this report can be seen in <https://github.com/bgbigberna/DP_2324/tree/main/FinalProjectG10>

## Dataset

The dataset used for training the model is activity\_train.csv, which contains the following columns:

* ProteinID: Identifier for the protein.
* ChemblID: Identifier for the chemical.
* Activity: Activity value representing the interaction strength between the protein and the chemical.

The test dataset, activity\_test\_blanked.csv, contains the same columns but with some activity values missing, which need to be predicted.

## Data Preparation

The first step involves reading the datasets and creating a matrix where rows represent proteins, columns represent chemicals, and the values represent activity levels. This matrix is then used for matrix factorization. A function create\_rankingMatrix is defined to transform the dataset into a matrix format. Ranking Matrix can be observed below.

#### Uma imagem com texto, captura de ecrã Descrição gerada automaticamente

#### Matrix Factorization using SVD

The core of the prediction model involves matrix factorization using SVD. The function make\_init\_PQ initializes the matrices P and Q, which will be factorized iteratively.

A mask matrix is created to identify the locations of the known values, which helps in calculating the error during training.

The function calc\_error computes the Mean Squared Error (MSE) for the predictions.

#### Optimization using Stochastic Gradient Descent (SGD)

SGD is used to iteratively update the matrices P and Q to minimize the prediction error. The function run\_epoch\_reg performs a single epoch of SGD with regularization.

The function RecSys\_SGD runs the SGD algorithm for a specified number of epochs, regularization parameters, and learning rate.

## Training the Model

The model is trained with different learning rates, and the MSE for each epoch is plotted to observe the convergence of the algorithm.

Uma imagem com texto, Gráfico, captura de ecrã, file

Descrição gerada automaticamenteUma imagem com texto, Gráfico, file, diagrama

Descrição gerada automaticamente

#### Hyperparameter Tuning

Hyperparameters for regularization (Lp and Lq) are tuned by trying different combinations and selecting the one with the lowest MSE.

The best combination found was: Lp = 0.5 and Lq = 0.1 for which we got an mse of 0.151.

Then we got the number of epochs to get the minimum MSE value. It was the last (epochs = 99) meaning it still hadn’t reached a minimum, but since it was already very stable we found that value to be sufficient.

Uma imagem com texto, captura de ecrã, file, Gráfico

Descrição gerada automaticamente

#### Evaluation and Results

Finally with the parameters that resulted in the lowest MSE we could finally predict the activity for the test dataset. The test data matrix was created similarly to the training data matrix.

Below you can observe the image of the activity\_test\_blanked.csv but with the activity already predicted.

Uma imagem com texto, captura de ecrã, Tipo de letra, número

Descrição gerada automaticamente

#### Conclusion

The Latent Factors approach using SVD and SGD successfully predicted protein-ligand activities with a reasonably low MSE. The hyperparameter tuning process was crucial in improving the model's performance. The final model can be used to predict activities for new protein-ligand pairs, aiding in drug discovery and related fields.

Further improvements could involve calculating the similarity of the molecules and implementing it in the model, experimenting with different factorization techniques, incorporating additional features, or using more advanced optimization algorithms.