Final project

Jonathan Alvarsson

2023-04-23

ULLA online course “Artificial Intelligence in Drug Discovery”

UPPSALA
UNIVERSITET

Final project

Jonathan Alvarsson

April 14, 2023

{⊘}pharmb.io

The project

Your dataset

The project

You will receive a dataset extracted from ChEMBL with SMILES strings and binding affinities then you will:

1. Investigate the dataset in order to become familiar with it

The project

Your dataset

The project

You will receive a dataset extracted from ChEMBL with SMILES strings and binding affinities then you will:

1. Investigate the dataset in order to become familiar with it
2. Perform data preprocessing

The project

Your dataset

The project

You will receive a dataset extracted from ChEMBL with SMILES strings and binding affinities then you will:

1. Investigate the dataset in order to become familiar with it
2. Perform data preprocessing
3. Do supervised machine learning models to be able to predict binding affinities for your target

The project

Investigate the dataset

Investigate the dataset

- Make a plot of the distribution of the binding affinities in your dataset
- Calculate and plot the distributions of molecular weight and log P for your dataset
- Make a plot of the chemical space spanned up by your dataset. Choose a molecular descriptor and make a, e.g., a PCA or UMAP plot.

The project

Data preprocessing

Data preprocessing

- Choose and calculate features for one or more types of molecular descriptors.
- Scale features to unit variance.
- Discretize your dataset to a binary classification problem, *i.e.*, create a label for each object to be 0 or 1. You can for example use a binding affinity of 10 µM as threshold.

The project

Supervised machine learning modelling

Supervised machine learning modelling

- Split the dataset into 10 % for hyperparameter tuning and 90 % for modeling.
- Train supervised models for KNN and Random Forest and perform hyperparameter tuning on the 10 % dataset.
- Perform k-fold crossvalidation on the 90 % dataset using the tuned hyperparameters.
- Make a table with accuracy, AUROC and F1 score for each modeling method.
- Plot AUROC and make a confusion matrix of F1 scores.

Deliverables

What you hand in

Deliverables

You will write all code, make all plots and write your text in a Jupyter notebook which you will hand in. Think of it as a report so write plenty of `markdown` boxes where you explain and discuss your material!

Getting help

Getting help

We will be available tomorrow around 9:00 and we will have Q&A session Thursday at 9:00 as ususal.

But if you need help mail us and we will setup more Zoom meetings

jonathan.alvarsson@farmbio.uu.se
staffan.arvidsson@farmbio.uu.se

Thank you