Silène Gobeil & Carissa Abbas

Half-semester project, Data Sciences Fundamentals

Part 2

1. Which packages are available for ML? Describe the pros and cons and document the availability.
   Scikit-learn is one of the most used packages for ML, it has a simple interface and many functionalities. However, the processing of large datasets is not very efficient compared to other packages. Scikit-learn is available for Python. Another example of a Python package used for ML is **TensorFlow**, it can handle large and complex datasets. However, its syntax and debugging are more complex, and there is no support for Windows [1]. **Weka** is a Java based software used for ML, it is simple and free to use but it can only handle small datasets, and it cannot really be used for industrial purposes [2].

2. What is Chembl? How do you access it?
   Chembl is a manually organised database of bioactive molecules. Chembl is free to use, the database is accessible from the Chembl website [3].

3. What is machine learning, and how does it differ from traditional programming?
   ML stands for machine learning; it is the capacity of a computer to learn and make predictions from data, without following explicit instructions. Traditional programming follows explicit instruction to achieve specific tasks.

4. What are the key concepts and techniques in machine learning?
   There are 3 main concepts in ML. Representation: how the data are represented and encoded. Evaluation: assesses the performance of ML models to determine how well they apply to new data. Optimization: process for finding good models to improve their efficiency and performance [4].

5. What are the different types of machine learning algorithms?
   Supervised learning trains the computer on known input and output data so it can predict future output. There are classification techniques which predicts discrete responses (for ex, if an email is genuine or spam) and regression techniques which predict continuous responses.
   Unsupervised learning involves learning patterns and structures from unlabeled data. Clustering is the most common technique [5].

6. What are the common applications of machine learning?
   Image recognition (surveillance system, augmented reality, autonomous vehicles…), natural language processing (translation, text summarization…), fraud detection, medicine…

7. How do you evaluate the performance of a machine learning model?
   There are several methods to assess the performance of an ML model. All of them being based on the same model with 4 different parameters: True positives (predicted positive and actually positive), false positives (predicted positive and actually negative), true negatives (predicted negative and actually negative), false negatives (predicted negative and are actually positive). Accuracy, precision, sensitivity, specificity, F1 score, PR curve, ROC curve are all examples of model used to describe the performance of an ML model, all of them being based on the parameters mentioned above [6].

8. How do you prepare data for use in a machine learning model?
   - Data collection
   - Data cleaning: identify missing values, inconsistencies…
   - Data transformation: scaling and encoding values so that they are in a suitable format for the ML algorithms
   - Data reduction: simplifying the data so that the machine can spot patterns more easily
   - Data splitting: divide the dataset into subset (training, validation & test sets) so that the machine can generalize well to new data [7]

9. What are some common challenges in machine learning, and how can they be addressed?
   - Poor data quality (for ex: missing, unreliable values), can be corrected during the data cleaning step
   - Overfitting and underfitting (not generalize well to new data) can be solved by using more complex algorithms
   - The need of significant computational resources can be lowered by using distrusted computing framework [8]

10. What are some resources and tools available to help you learn and practice machine learning?
    Online resources (courses, tutorial, YouTube videos…), books

Part 3

1. What is in the training set, how big is it?
   This training set contains the "machine readable" information of 179826 molecules, necessary to train an artificial neuronal network (ANN, subset of ML) predicting the kinase activity. It contains the index of the molecule (from 1 to 179826), the chembl_id (to identify the molecule in the Chembl database), the standard_value (IC50 of the molecule, use to describe its efficacity, this values indicates the amount of substance needed to inhibates half of the kinases), the standard_units (units of the IC50 value), the target_chembl_id (identifies the target molecule in the Chmbl database), the smiles (structure of the molecule).

2. What modifications do you need to do to the data set to perform the tutorial?
   - Keep only the important columns
   - Calculate the pIC50 value: pIC50 = -log(IC50)
   - Convert the smiles strings into numerical data

3. What is a test set? Any other types of set?
   A portion of the dataset is emitted from the training process and used solely for evaluating the performance of a trained machine learning model. By keeping the test data separate from the training data, you can estimate of how well the model generalizes to unseen examples of data.

4. Before starting, describe with 1-2 sentences, in your own words, what is done in each of the cells.

Practical

1. The libraries that are required to complete the task are imported.
2. The file path is set up.

Data preparation

3. The data is loaded into the notebook from its path and indexed.
4. The shape of the data frame and information about the columns are displayed.
5. The head (first 5 lines) of the data frame is displayed.
6. The columns necessary for the machine learning are copied into a separate data frame and the head is displayed.
7. The function which converts the smiles values to fingerprints is defined. Corresponding molecular objects to the smiles values are created.
8. The function defined in 7. is applied in this step and the head is displayed.
9. The data is split into training and test sets (70% training and 30% test data), and the shape of the data is printed.

Define neural network

10. A function is defined to create a neural network with two hidden layers.

Train the model

11. The neural network is tested with different batch sizes.
12. The respective test losses are plotted so that they can be compared.
13. The trained model is saved and fitted.

Evaluation & prediction on test set

14. The model is evaluated by calculating the score of the mean absolute error.
15. The pIC50 values are predicted on the test data and the first 5 are displayed.

Scatter plot

16. A scatter plot is created which lets one compare the true pIC50 values to the predicted pIC50 values.

Prediction on external/unlabelled data

17. A test file is read into the notebook and the head of the data frame is displayed.
18. The function from 7. is used on the new data and the head is displayed.
19. The pre-trained model is loaded but not compiled.
20. The model is run with the new data and the head is displayed.
21. The predicted values are saved into a new file.

Select the top 3 compounds

22. The 3 compounds with the highest predicted pIC50 are selected and displayed.
23. The chemical structures of the 3 compounds with the highest predicted pIC50 are drawn.

Part 5

1. What is Ubelix?
   Ubelix stands for University of Bern Linux Cluster, it is HPC (high performance computing) cluster.

2. How do you gain access?
   To gain access to Ubelix, we must make a request to activate our campus account for Ubelix.

3. How do you submit a job?
   Jobs can be submitted using sbatch (using a batch script), srun (directly running the executable), or salloc (interactive submission). Slurm processes job submission [9].

4. Who can have access?
   The members of the university of Bern as well as external coworkers can have access to Ubelix.

5. What resources are available there?
   Software stack, SLURM (resource manager / batch-queuing system), parallel jobs, 320 compute nodes (featuring ~12k CPU cores and 160 GPUs), and data storage (~3.5 PB of disk storage net) [9].

References

ChatGPT was also used to answer the questions, complementary to the sources listed below.

[1]: https://www.analyticsvidhya.com/blog/2023/08/scikit-learn-and-tensorflow/

[2]: https://libstore.ugent.be/fulltxt/RUG01/000/842/101/RUG01-000842101_2010_0001_AC.pdf

[3]: https://www.ebi.ac.uk/chembl/

[4]: https://www.domo.com/glossary/what-are-machine-learning-basics#:~:text=There%20are%20three%20main%20elements,models%3B%20how%20programs%20are%20generated

[5]: https://www.mathworks.com/discovery/machine-learning.html#:~:text=Tutorials%20and%20examples-,How%20Machine%20Learning%20Works,intrinsic%20structures%20in%20input%20data

[6]: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

[7]: https://www.pecan.ai/blog/data-preparation-for-machine-learning/

[8]: https://iabac.org/blog/issues-in-machine-learning

[9]: https://hpc-unibe-ch.github.io/general/ubelix-overview.html