Practice 02

RDKit – Python library for cheminformatics
Support Vector Machine

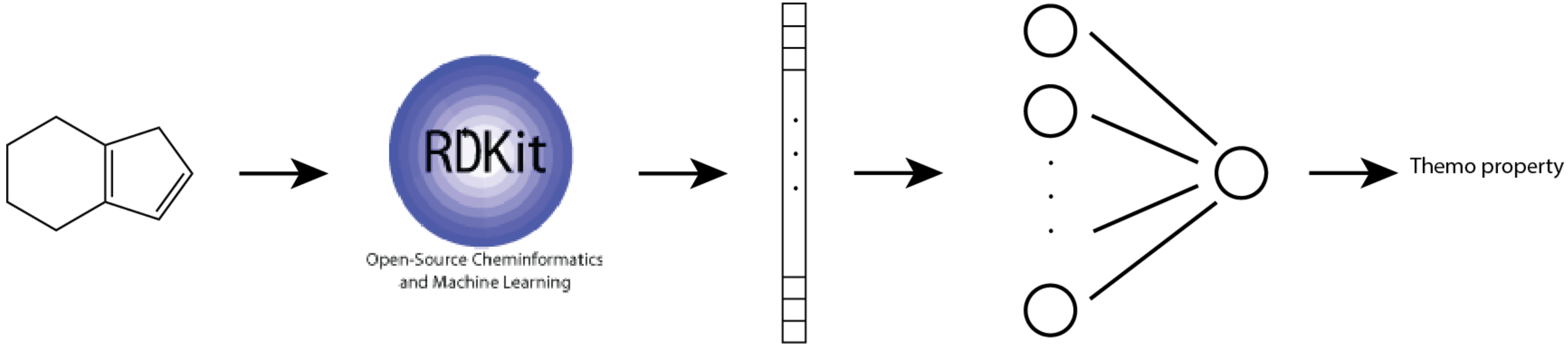
Seongok Ryu

Department of Chemistry, KAIST

Contents

- RDKit
- Support Vector Machine (SVM)

RDKit and SVM



- Becoming accustomed to using RDKit and support vector machine (SVM).
- Prediction of logP with molecular fingerprint and support vector regression.
- https://github.com/SeongokRyu/CH485---Artificial-Intelligence-and-Chemistry/blob/master/Practice%2002/practice_rdkit.ipynb
- https://github.com/SeongokRyu/CH485---Artificial-Intelligence-and-Chemistry/blob/master/Practice%2002/prediction_logP.ipynb
- Assignment : Toxicity classification using a support vector classification.

Assignment #2

- **Toxicity classification using SVM**

  1. Obtain molecular fingerprints for molecules in a tox-21 dataset.
     - I prepared the tox-21 dataset, and implemented a function in ‘utils.py’, which reads SMILES and labels.
     - RDKit cannot convert some molecules to the molecular fingerprints. You should handle this exception.
     - Dataset is in the ‘tox21’, or you can download at this link - https://tripod.nih.gov/tox21/challenge/

  2. Split the dataset to a training set and a test set.

  3. Train a SVM model. Use the kernel function implemented in scikit-learn.

  4. Validate the trained model using a test set.
     Report an accuracy and an auc-roc score. (Use functions implemented in sklearn).

  5. Think about how you can improve the model.

Assignment #2

- **References**

  - http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

  - https://www.rdkit.org/docs/Cookbook.html

  - http://members.cbio.mines-paristech.fr/~jvert/talks/070907aix/aix.pdf

  - Koutsoukas, Alexios, et al. "Predictive toxicology: Modeling chemical induced toxicological response combining circular fingerprints with random forest and support vector machine." *Frontiers in Environmental Science* 4 (2016): 11.