CAPSTONE PROJECT

Stage I

by

Sarita Charde

Predicting Molecular Properties

Problem Statement

Predicting the Molecular Properties

Is data science smart enough to make big predictions at a molecular level?

Goals

To apply predictive analytics to chemistry and chemical biology and develop an algorithm that can predict the magnetic interaction between two atoms in a molecule (i.e., the scalar coupling constant)..

Potential Audience

Pharmaceutical company, Environmental and Materials Science are benefited

Success Metrics

based on scalar coupling constant

- ROI
- Data Source(s): Kaggle

What is the challenge?

This challenge aims to predict interactions between atoms to finally predict the type of molecule . Imaging technologies like MRI enable us to see and understand the molecular composition of tissues. Nuclear Magnetic Resonance (NMR) is a closely related technology which uses the same principles to understand the structure and dynamics of proteins and molecules. Researchers around the world conduct NMR experiments to further understanding of the structure and dynamics of molecules, across areas like environmental science, pharmaceutical science, and materials science.

About Scalar Coupling Constant

What is scalar coupling constant?

Using NMR to gain insight into a molecule's structure and dynamics depends on the ability to accurately predict so-called "scalar couplings". These are effectively the magnetic interactions between a pair of atoms. The strength of this magnetic interaction depends on intervening electrons and chemical bonds that make up a molecule's three-dimensional structure.

Why this study?

Using state-of-the-art methods from quantum mechanics, it is possible to accurately calculate scalar coupling constants given only a 3D molecular structure as input. However, these quantum mechanics calculations are extremely expensive (days or weeks per molecule), and therefore have limited applicability in day-to-day workflows.

Application

- A fast and reliable method to predict these interactions will allow medicinal chemists to gain structural insights faster and cheaper, enabling scientists to understand how the 3D chemical structure of a molecule affects its properties and behavior.
- Ultimately, such tools will enable researchers to make progress in a range of important problems, like designing molecules to carry out specific cellular tasks, or designing better drug molecules to fight disease.

Application

An important application was in my research on materials characterization studies. We carried out modification of nanoparticles to improve interaction between the polymer matrix and nanoparticles. After this we tested using FTIR technique to confirm the change on the surface of the molecules. The peer reviewers were objecting to this. It was suggested to go for NMR which easily confirmed the results but the cost and the time consumed was around one month for one sample.

Expected Predictions

- will be predicting the scalar_coupling_constant between atom pairs in molecules, given the two atom types (e.g., C and H), the coupling type (e.g., 2JHC), and any features you are able to create from the molecule structure (xyz) files.

Data Set

The training and test splits are by molecule, so that no molecule in the training data is found in the test data.

Data Set

train.csv (46lacs, 4)

the training set, where the first column (molecule_name) is the name of the molecule where the coupling constant originates (the corresponding XYZ file is located at ./structures/.xyz), the second (atom_index_0) and third column (atom_index_1) is the atom indices of the atom-pair creating the coupling and the fourth column (scalar_coupling_constant) is the scalar coupling constant that we want to be able to predict

Additionl Data

scalar_coupling_contributions.csv (4658147, 8)

The scalar coupling constants in train.csv (or corresponding files) are a sum of four terms. scalar_coupling_contributions.csv contain all these terms. The first column (molecule_name) are the name of the molecule, the second (atom_index_0) and third column (atom_index_1) are the atom indices of the atom-pair, the fourth column indicates the type of coupling, the fifth column (fc) is the Fermi Contact contribution, the sixth column (sd) is the Spin-dipolar contribution, the seventh column (pso) is the Paramagnetic spin-orbit contribution and the eighth column (dso) is the Diamagnetic spin-orbit contribution.

Data exploration, analysis and visualization

- checking the shape of data in train.csv . It is around 46 lacs
- study the additional data which is available only for training
- check if information provided by each feature is nearly same
- Merge the data for a better analysis
- Check the train and test data for the molecule types and size
- Cleaning the data
- Analyze the influence of additional features
- Visualization of the results using 'ase', which is a python package that allows one to work with atoms and molecules.

Apply Feature Engineering

- Create new features based on distance and angle between the bonds
- visualize them using new methods
- Generate the final dataset

Aim is to be different in analyzing the same topic

- using different visualization libraries
- using different application framework
- using cloud computing(Azure)