Homework 5: Add onto Homework 3

Jiayu Li

March 24, 2021

In homework 3 students studied industry transitions in many different areas including the stock market, the automobile industry, drug discovery, protein structure prediction, consumer marketing, energy, and power systems. There are two steps in this homework 5 building on this work

1. Decide if you want to keep studying the topic you chose in homework 3 and increase the depth of work
2. If the answer to a) is no, then homework 5 is to repeat homework 3 for a different application (transition)
3. If the answer to a) is yes, then take your area and look online for an AI algorithm related to your study that has 1) some open-source data and 2) either a description of analysis with enough detail to identify AI algorithms or an open-source GitHub with relevant software. You can be broad-minded – protein structure and material structure use similar methods and perhaps the stock market acts like earthquakes or other time series. We would like each student to eventually identify a simple illustrative deep learning example that uses networks/architecture similar to “production system”. This study will evolve into the final project

1 Structural Protein Sequences Data set

PDB is a data set dedicated to the three-dimensional structure of proteins and nucleic acids. It has a very long history, dating back to 1971. In 2003, PDB developed into an international organization wwPDB. Other members of wwPDB, including PDBe (Europe), RCSB (United States), and PDBj (Japan) also provide PDB with a center for data accumulation, processing and release. Although PDB data is submitted by scientists from all over the world, each piece of data submitted will be reviewed and annotated by wwPDB staff, and whether the data is reasonable or not. The PDB and the software it provides are now free and open to the public. In the past few decades, the number of PDB structures has grown at an exponential rate.

Structural biologists around the world use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the position of each atom relative to each other in the molecule. Then they will submit this structural information, wwPDB will annotate it and publish it to the database publicly.

You can search for ribosomes, oncogenes, drug targets, and even the structure of the entire virus in the PDB data set. However, the number of structures archived in the PDB is huge, and finding the information you need may be a difficult task.

The information in the PDB data set mainly includes: protein/nucleic acid source, protein/nucleic acid molecule composition, atomic coordinates, experimental methods used to determine the structure.

2 Open-source software

alphafold_casp13: `https://github.com/deepmind/deepmind-research/tree/master/alphafold_casp13`

Kaggle Structural Protein Sequences: `https://www.kaggle.com/shahir/protein-data-set`