Module 2 におけるニューラルネットワーク学習の手順

mol-infer/Cyclic_improved

2021 年 1 月 19 日

1　はじめに

本稿では，本プロジェクト (mol-infer/Cyclic_improved) における Module 2 の手順を解説する．

Module 2 の入力と出力は以下の通りである．

入力: 訓練集合（化合物の特徴ベクトルおよび化学的性質の値の組），ニューラルネットワークの各種パラメータの値（隠れ素子の個数など）．

出力: 入力で指定された構造を持ち，訓練集合の「多くの」化合物に対して化学的性質の値を「良く」推定するようなニューラルネットワーク.

出力の具体的な中身は，学習されたニューラルネットワークにおける各枝の重みと各ノードのバイアスである．

本稿の構成は以下の通りである．

- 第 2 節: 基本的な用語，およびパッケージのファイル構成の説明.
- 第 3 節: 簡単な実行例.
- 第 4 節: プログラムの入出力に関する詳細.

2　準備

2.1　用語の説明

特徴ベクトル　各元素の種類の原子数等の化学物質を説明する数値，あるいはグラフの直径等の化学物質のグラフ表現のトポロジーに基づいて計算される数値のベクトル

ニューラルネットワーク　人工ニューラルネットワーク (artifician neural network, ANN)，または単にニューラルネットワーク (NN) とは，機械学習で最も確立した手法の 1 つである．これらは入力ベクトルに基づいて値を予測するために用いられる．この冊子では，ニューラルネットワークへの入力は，化合物の特徴ベクトルであり，出力は特定の化学的性質の予測値である．

本プロジェクトで用いるのはフィードフォーワード型のニューラルネットワークであり，これは非巡回有向グラフによって表すことができる．図 1 に例を示す．

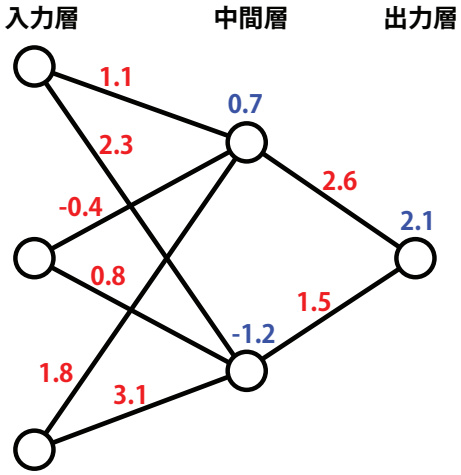

図 1: ニューラルネットワークの具体例．赤色の数値が重み，青色の数値がバイアスを表している．枝はすべて左から右に向いている．

入力層，隠れ層，出力層　人工ニューラルネットワークの多層パーセプトロンモデルを仮定する．このモデルでは，ニューラルネットワークはいくつかの層で構成されている．最初の層は入力層で，入力層は場合によっては特徴ベクトルから数値データを得るため，特徴ベクトルの要素と同じ数のノードがある．次に数値は隠れ層を介して伝播される．隠れ層は 1 つの層の計算が次の層への入力として用いられる．最後に出力層が入力ベクトルに基づいた予測値を与える．

図 1 では，入力層は 3 つのノードを持つ．したがって入力する特徴ベクトルは 3 次元のものでなければならない．また 1 つの隠れ層を有し，この隠れ層は 2 つのノードを持つ．そして出力層は 1 つのノードを持つ．

重み　ニューラルネットワークに含まれているノード間を接続している枝（有向辺）はそれぞれ数値を持っており，その値を重みと呼ぶ．入力層から出力層への値の伝播には，これらの重みに基づく計算が含まれている．

図 1 では，枝の重みは赤によって示される．

バイアス　ニューラルネットワークの隠れ層の各ノードにはバイアスと呼ばれる数値が割り当てられている．この数値は重みと共に，入力ベクトルに基づいて出力値を計算する過程で使用される．

図 1 では，ノードのバイアスは青によって示される．

活性化関数　活性化関数はニューラルネットワークの各ノードに割り当てられており，与えられた入力ベクトルから出力値を計算する際に用いられる．特に各ノードの出力値は，重み付けされた対応する枝の重みと前の層からのノードの出力の線形結合を入力として与えられた活性化関数の値である．

ニューラルネットワークは，与えられた入力ベクトルと目標値の組に基づいて重みとバイアスの組を計算することで「学習」する．

2.2　ファイル構成

パッケージに含まれるファイルとその役割は表 1 の通りである．

表 1: Module 2 のパッケージに含まれるファイルとその役割

ファイル名	役割
`mol-infer_ANN.py`	NN を学習するための Python スクリプト **（Module 3 に進むにはこのスクリプトの実行が必要）** • 使用する非標準ライブラリ: `numpy`，`pandas`，`scikit-learn`
`predict_values.py`	学習した NN を用いて 化学的性質の値を推定するための補助的な Python スクリプト （このスクリプトの実行は必ずしも必要ではない） • 使用する非標準ライブラリ: `numpy`，`pandas`

`Manual_Module_2_BH_cyclic_jp.pdf`
`Manual_Module_2_BH_cyclic_jp.tex`
`fig/ANN_sample_jp.eps`
	マニュアルの PDF，LaTeX ソースファイル
	および画像ファイル（日本語版）

`Manual_Module_2_BH_cyclic_en.pdf`
`Manual_Module_2_BH_cyclic_en.tex`
`fig/ANN_sample_en.eps`
	マニュアルの PDF，LaTeX ソースファイル
	および画像ファイル（英語版）

BP (boiling point; 沸点) に関するデータファイル	
`data/BP.sdf`	化合物に関する SDF ファイル．Module 2 では直接取り扱わない
`data/BP_fv.csv`	Module 1 に `BP.sdf` を入力して生成した特徴ベクトル
`data/BP_value.csv`	化合物の BP 値を記したファイル
`data/BP_ANN.LOG`	`BP_fv.csv`，`BP_value.csv` から成る訓練集合に対して `mol-infer_ANN.py` を実行し，学習したときのログ出力
`data/BP_ANN_biases.txt` `data/BP_ANN_weights.txt`	学習された NN における各ノードのバイアス， および各枝の重み

(表 1 の続き)

ファイル名	役割
HC (heat of combustion; 燃焼熱) に関するデータファイル	
data/HC.sdf	化合物に関する SDF ファイル．Module 2 では直接取り扱わない
data/HC_fv.csv	Module 1 に HC.sdf を入力して生成した特徴ベクトル
data/HC_value.csv	化合物の HC 値を記したファイル
data/HC_ANN.LOG	HC_fv.csv，HC_value.csv から成る訓練集合に対して mol-infer_ANN.py を実行し，学習したときのログ出力
data/HC_ANN_biases.txt data/HC_ANN_weights.txt	学習された NN における各ノードのバイアス， および各枝の重み
KOW (log Kow; オクタノール/水分配係数) に関するデータファイル	
data/KOW.sdf	化合物に関する SDF ファイル．Module 2 では直接取り扱わない
data/KOW_fv.csv	Module 1 に KOW.sdf を入力して生成した特徴ベクトル
data/KOW_value.csv	化合物の KOW 値を記したファイル
data/KOW_ANN.LOG	KOW_fv.csv，KOW_value.csv から成る訓練集合に対して mol-infer_ANN.py を実行し，学習したときのログ出力
data/KOW_ANN_biases.txt data/KOW_ANN_weights.txt	学習された NN における各ノードのバイアス， および各枝の重み
MP (melting point; 融点) に関するデータファイル	
data/MP.sdf	化合物に関する SDF ファイル．Module 2 では直接取り扱わない
data/MP_fv.csv	Module 1 に MP.sdf を入力して生成した特徴ベクトル
data/MP_value.csv	化合物の MP 値を記したファイル
data/MP_ANN.LOG	MP_fv.csv，MP_value.csv から成る訓練集合に対して mol-infer_ANN.py を実行し，学習したときのログ出力
data/MP_ANN_biases.txt data/MP_ANN_weights.txt	学習された NN における各ノードのバイアス， および各枝の重み

3　クイックスタート

ニューラルネットワークの学習　次のコマンドを入力すれば，data/BP_fv.csv を特徴ベクトル，data/BP_value.csv を化学的性質（この場合は BP，すなわち沸点）の値とするような訓練集合に対し，2 つの隠れ層を持ち，それぞれの隠れ層におけるノード数を 20, 10 とするようなニューラルネットワークが学習され，そのニューラルネットワークにおける各枝の重みは output_weights.txt，各ノードのバイアスは output_biases.txt にそれぞれ出力される．

```
$ python mol-infer_ANN.py data/BP_fv.csv data/BP_value.csv output 20 10  (♣)
```

学習されたニューラルネットワークの**重みおよびバイアスに関するファイルは，Module 3 でも使用**する．

化学的性質の値の推定　学習したニューラルネットワークを用いて，化合物の化学的性質の値を推定することができる．

次のコマンド[1]を入力すれば，学習済ニューラルネットワーク（枝の重みは output_weights.txt，ノードのバイアスは output_biases.txt に保持されている）を用いて，data/BP_fv.csv に記述された特徴ベクトル（に対応する化合物）の化学的性質の値が推定され，その結果は predicted.txt に出力される．

```
$ python predict_values.py output_weights.txt output_biases.txt \
    data/BP_fv.csv predicted.txt
```

- 化学的性質の値を推定したい化合物の特徴ベクトルは，Module 1 の特徴ベクトル生成器を用いて生成することができる．
- 上のコマンドは，さらに上のコマンド (♣) で data/BP_fv.csv を特徴ベクトルとした訓練集合から学習したニューラルネットワークに関する output_weights.txt および output_biases.txt を用いるものとみなせば，訓練集合自身の化学的性質の値を推定していることになる．
- この推定機能は補助的なものに過ぎず，Module 3 以降で使用することはない．

[1] 一行目の最後のバックスラッシュ \ は，実際に入力するときには改行してはならないことを示す．

4　プログラムの入出力に関する詳細

4.1　入力

4.1.1　特徴ベクトル

特徴ベクトルは，我々が**FV 形式**と呼ぶフォーマットに基づく csv ファイルに記述されている必要がある．Module 1 の特徴ベクトル生成器は化合物に関する SDF ファイルから FV 形式の csv ファイルを生成するため，当該生成器を用いて生成されたファイルを用いれば問題はない．

以下，FV 形式の記述ルールを簡単に記しておく．

```
CID,n,cs,ch,nsH
244,8,6,2,8
307,10,6,4,8
657014,11,7,1,18
16704,9,9,0,10
```

- FV 形式では先頭行に記述子の名前をコンマ区切りで記述する．
- ただし最初の記述子は CID（化合物識別番号）でなければならない．CID の値が学習に用いられることはない．CID は識別のためのみに用いられる．
- 上記の例では，`n`（水素を除く原子の個数），`cs`（コアサイズ），`ch`（コアハイト），`nsH`（水素原子の個数）と四つの記述子が定められている．
- 二行目以降に，一つの行に一つの化合物の CID および特徴ベクトルを，コンマ区切りで記述する．したがって各化合物は，4 次元の特徴ベクトルで表されることになる．
- 化合物が CID の昇順あるいは降順に並んでいる必要はない．

4.1.2　化学的性質の値

化学的性質の値は，CID と値を羅列した csv ファイルに記述されなければならない．

```
CID,a
307,11.2
244,-0.5
657014,98.124
```

```
16704,-12.8
117,5.3
```

- 先頭行は CID,a でなければならない.
- 二行目以降に，一つの行に一つの化合物の CID および化学的性質の値をコンマ区切りで記述する.
- 化合物が CID の昇順あるいは降順に並んでいる必要はない.

4.2 実行

第 3 節で示したとおり，ニューラルネットワークを学習するには mol-infer_ANN.py を用いる.

4.2.1 引数

第 3 節で示したコマンドを再掲する.

```
$ python mol-infer_ANN.py data/BP_fv.csv data/BP_value.csv output 20 10
```

各引数には以下を指定する.

- 第 1 引数: 特徴ベクトルに関する csv ファイル
- 第 2 引数: 化学的性質の値に関する csv ファイル
- 第 3 引数: 学習されたニューラルネットワークの重み・バイアスを書き込むファイルの名前
- 第 4 引数以降: 隠れ層（中間層）のノードの個数

引数を与えずに実行すれば（もしくは引数が適切に与えられなかった場合），引数に関する説明が英語で出力される.

```
$ python mol-infer_ANN.py
```

引数が適切に与えられると，ニューラルネットワークの学習が始まる．

5 分割交差検定が行われ，試験集合に対して最も高い決定係数 (R^2 値) を実現するニューラルネットワークが採用され，その重みとバイアスが，ファイルに出力される．上記のコマンドの場合，そのファイルの名前は第 3 引数で指定された `output` に基づく，

- `output_weights.txt`（重み）
- `output_biases.txt`（バイアス）

である．これら出力されたファイルは，Module 3 の実行に必要となる．

訓練集合に関する注意　訓練集合は，

- 特徴ベクトルに関する csv ファイル，および
- 化学的性質の値に関する csv ファイル

の二つのファイルから構成されるが，訓練集合は，前者ファイルに記された CID から成るとみなされる．したがって，

前者ファイルに記された CID は，すべて後者ファイルに記されていなければならない．

しかし逆は成り立たなくともよい．つまり，化学的性質の値に関する csv ファイルには，特徴ベクトルの csv ファイルに記されていない，「余計な」CID に関する値が記されていても構わない．そのような値は無視される．

4.2.2　ハイパーパラメータの調整

ニューラルネットワークの学習は `scikit-learn` ライブラリ[2] における `MLPRegressor` を用いて行われる．`mol-infer_ANN.py` の 133 行目以降で `MLPRegressor` インスタンスの初期化が行われるが，ハイパーパラメータの調整はここで行うことができる．一部のパラメータを以下のように設定している．

- `activation`: `'relu'`
- `alpha`: 10^{-5}
- `early_stopping`: `False`
- `hidden_layer_sizes`: 実行時に引数で指定した個数に基づく
- `max_iter`: 10^{10}
- `random_state`: 1
- `solver`: `'adam'`

[2] `https://scikit-learn.org/stable/`

4.3 出力

ふたたび以下のコマンドを取り上げ，その実行によって得られる出力について説明する．

```
$ python mol-infer_ANN.py data/BP_fv.csv data/BP_value.csv output 20 10
```

4.3.1 標準出力

上記コマンドを実行すると端末（標準出力）に計算過程が出力される．この出力例がパッケージ内のファイル data/BP_ANN.LOG に記されている．

```
src/preparation/BP_fv.csv contains 181 vectors for 107 (=CID+106) features.
src/preparation/BP_value.csv contains 230 target values.
n range = [5,30]
a range = [31.5,470.0]
#instances = 181
#features = 106

D1: train 144, test 37
training time: 3.187196731567383
R2 score train = 0.9935319077425955
R2 score test = 0.7855694851759929
R2 score all = 0.9599908495442584
MAE score train = 3.8570127089010384
MAE score test = 19.756408746147212
MAE score all = 7.10716548999556

D2: train 145, test 36
training time: 5.0139079093933105
R2 score train = 0.9930416804444452
R2 score test = 0.6390572625074291
R2 score all = 0.9339056805642507
MAE score train = 3.7281172621746888
MAE score test = 27.996544836634186
MAE score all = 8.554986834995363

D3: train 145, test 36
training time: 4.264358758926392
R2 score train = 0.9961346879653636
R2 score test = 0.8566979846124056
R2 score all = 0.9637772344809027
MAE score train = 2.870368503215682
```

```
MAE score test = 22.925217956024927
MAE score all = 6.8591783391335435

D4: train 145, test 36
training time: 2.9935202598571777
R2 score train = 0.9909067339023991
R2 score test = 0.8390994594153819
R2 score all = 0.9669036994338265
MAE score train = 4.470398694611737
MAE score test = 18.691072072382227
MAE score all = 7.298819918919681

D5: train 145, test 36
training time: 4.9648661613464355
R2 score train = 0.9946933391220482
R2 score test = 0.8479544758088942
R2 score all = 0.9574271159388575
MAE score train = 3.352287445970431
MAE score test = 21.80973582058148
MAE score all = 7.023382150312958
0.9935319077425955 0.7855694851759929 0.9599908495442584 3.187196731567383
0.9930416804444452 0.6390572625074291 0.9339056805642507 5.0139079093933105
0.9961346879653636 0.8566979846124056 0.9637772344809027 4.264358758926392
0.9909067339023991 0.8390994594153819 0.9669036994338265 2.9935202598571777
0.9946933391220482 0.8479544758088942 0.9574271159388575 4.9648661613464355
Average time = 4.08476996421814
Average R2 test score = 0.7936757335040208
Average MAE test score = 22.235795886354005
```

- 冒頭で特徴ベクトルの個数，特徴数が出力される．また水素を除く原子数（特徴 `n`）の最小値と最大値 (`n range`，上の例では 5 と 30)，および化学的性質（この場合は BP; 沸点）の値の最小値と最大値（`a range`，上の例では 31.5 と 470.0）が出力される．
- 続いて 5 分割交差検定における 5 回の学習の概要が出力される．
- 最後に平均計算時間,（試験集合に対する）平均 R^2 値,（試験集合に対する）平均 MAE 値が出力される．
- 上記の例では，3 回目の学習で得られたニューラルネットワークが最も高い（試験集合に対する）R^2 値を達成しているので（0.856697...），当該ニューラルネットワークにおける枝重みが `output_weights.txt`，ノードのバイアスが `output_biases.txt` に出力される．

- なおこれらファイルのコピーを，それぞれdata/BP_ANN_weights.txt，data/BP_ANN_biases.txt として同封している．

4.3.2 枝の重み

mol-infer_ANN.py が出力する枝重みファイルの書式について説明する．

簡単のため，図 1 に示したニューラルネットワークが学習されたとする．このニューラルネットワークの枝重みは以下のようにファイルに出力される．

```
3 2 1
1.1 2.3
-0.4 0.8
1.8 3.1
2.6
1.5
```

- 最初の行はニューラルネットワークの構造，つまり入力層のノード数，各隠れ層のノード数，最後に出力層のノード数である．
- 2 行目以降は枝重みの値である．各行は 1 つのノードから出る枝重みの値を示す．

4.3.3 ノードのバイアス

同じく，mol-infer_ANN.py が出力するノードのバイアスに関するファイルの書式について説明する．

やはり簡単のため，図 1 に示したニューラルネットワークが学習されたとする．このニューラルネットワークのノードのバイアスは以下のようにファイルに出力される．

```
0.7
-1.2
2.1
```

1 行につき 1 つのノードのバイアスの値が示されている．入力層のノードにはバイアスの値がないことに注意．