MAGICIAN 養成講座

MAterials Genome/Informatics and Chemo-Informatics Activate Networks

第 6a 回 ニューラルネットワーク法の基礎

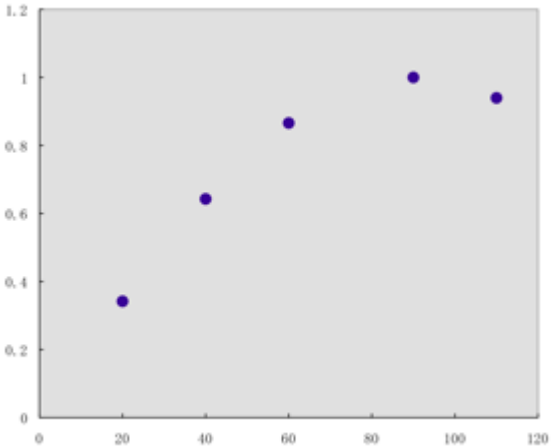
2018.9.25　横浜国大非常勤講師　山本博志

Materials Genome を使って何か素材を設計したいとしよう。Genome というのは遺伝子のことで、素材を遺伝子解析してより優れた素材を探索するのが Materials Genome の語源だ。それでは、素材のある部分の遺伝子をいじった時に、結果がどうなるか？　は、どうしたら解るだろう？　実際にハイスループットで実験を行う。Net, 論文, 特許を検索してデータを集める。そうして集まったデータを統計解析して、予測式を立てて、より良い素材になるように改良していく。　？？？　昔から行われているこうした統計解析の方法と、Materials Genome の方法はどこが違うのだろうか？　実は全く変わらないのである。筆者は 20 年前から、「物性推算(Properties Estimation)と逆設計(Reverse Design)」と呼んでいた。筆者が多用している物性推算法はニューラルネットワーク法であり、多用している逆設計法は遺伝的アルゴリズム法である。ニューラル（神経細胞）、遺伝などの言葉が入っているので、まるで生物の進化に関係があるように聞こえるが、単なる統計解析法の 1 種とも言える。画像解析や自然言語解釈などは、ネットの発達によって非常に容易にビッグデータが入手できるようになった。「猫のヒゲとはなんぞや？」という問いに膨大なデータが集まる。しかし、素材のある部分をいじったときの結果はどこを探しても見つからない。そこで、なるべくデータは集めるにしても、足りない部分は合理的な方法で補完することが必要になる。一つの方法は、分子軌道計算や分子動力学計算によってデータ数を増やすことだ。しかしこれらの方法は、第 3 回で紹介したような、市販されているカーボンブラックの品番によってゴムの性能が変わるなどには適用できない。残念ながら、現時点では王道はないのだと思う。開発したい素材に合わせて、化学の知識を総動員して Materials Genome を行っていれば、AI に仕事を奪われるようなことはないだろう。そう言い切れるためには、まず、ニューラルネットワーク法による物性推算の基礎を学ぼう。第 7 回で、遺伝的アルゴリズムについて解説しよう。

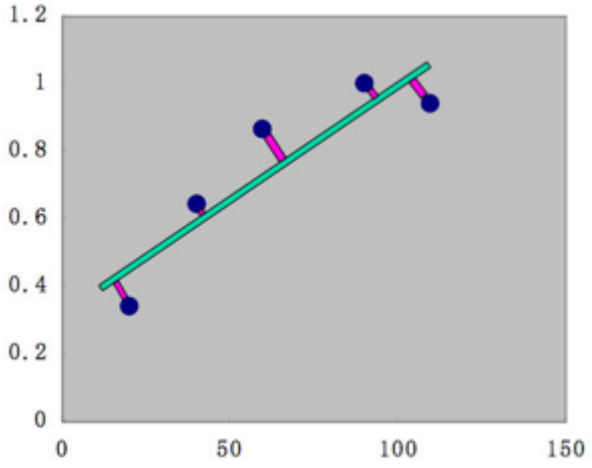
ニューラルネットワーク法の基礎

ニューラルネットワーク法の基礎については、筆者の HP、
http://www.pirika.com/
に多分 2005 年ぐらいに記載している。
https://www.pirika.com/JP/ChemInfo/NNbasic.html
https://www.pirika.com/JP/ChemInfo/NNMechanism.html
https://www.pirika.com/JP/ChemInfo/OverLearn.html
これらのページを既に読んでいる学生は、前半は読み飛ばしても良い。ただし、HP は 15 年近く前の記載なので、細かいところは変更になっている。（ちなみに、pirika.com の HP は来年で 20 周年になる。物好きの趣味のページがこれだけ長続きしているのも、付き合ってくれる読者があってのこと。だいたい 1 日あたり 1000 ページ、世界中から訪ねてくれる。誰かお祝いでもしてくれないものだろうか？　閑話休題）

下の図のような X-Y 関係のデータポイントがあったとする。化学系の研究者が最も一般的に用いる解析方法は、（重）回帰計算であろう。これは最小二乗法とも呼ばれる技術である。

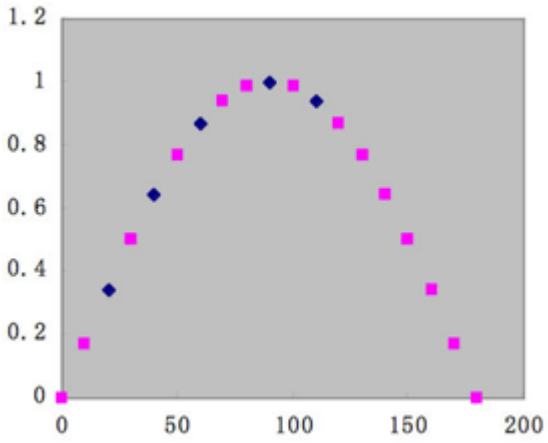


重回帰計算を簡単に説明すれば、下の図に示すように、緑色の直線、Y=aX+b を仮定して、この直線からの誤差の 2 乗が最低になる a と b の組を見つける計算だ。計算自体は非常に簡単で、Excel などの表計算ソフトにもその機能が搭載されている。



多くの場合、化学の現象を扱うのはこの重回帰計算で十分である。

しかし、計算によっては非線形性を導入しなくてはならない場合がある。実は、上の図の青い点は下に示すようにSin カーブの一部であった。



これを正しく扱いたい場合には、非線形回帰計算を行う必要がある。そこで、先ほどの重回帰計算と同様に、

f(y)=A1* / [1+exp(-(B1*X1+B2*1))＋
A2* / [1+exp(-(B3*X1+B4*1))＋
A3* / [1+exp(-(B5*X1+B6*1))

という曲線を仮定して、誤差が最小になる A1-A3, B1-B6 の組を見つければ良い。曲線としてどんな曲線を仮定するは各人の自由である。

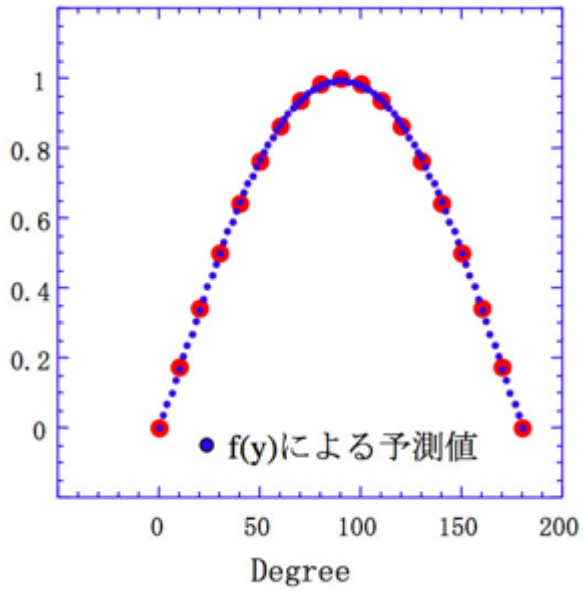
例えば、

A1= -2.14　B1= -1.84　B2= 0.30

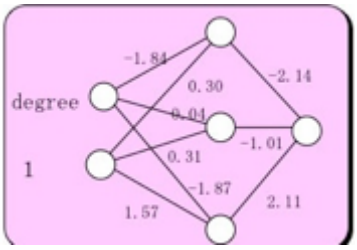
A2= -1.01　B3= 0.04　B4= 0.31

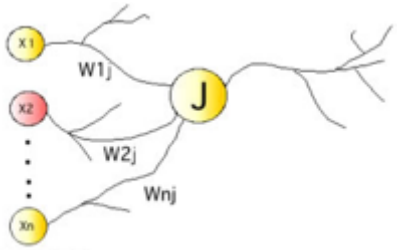
A3= 2.11　B5= -1.87　B6= 1.57

だとすると、下図に示すように良好に Sin カーブを再現する。



この f(y)をを模式的に書くと、生体の神経細胞（ニューロン）がつながったような図になる。左の 2 つを入力層ニューロン、真ん中の 3 つを中間層（隠れ層）ニューロン、右の 1 つを出力層ニューロンと呼ぶ。



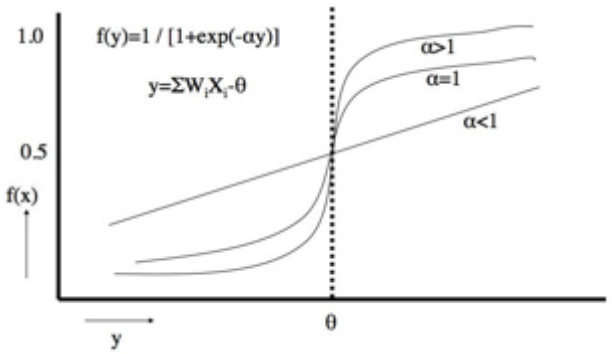


神経細胞

式中の変数、A1-A3, B1-B6 は模式図中でニューロンを同士の結合の強さとして表わされている。

また、f(y)=A1* / [1+exp(-(B1*X1+B2*1))という関数は、シグモイド関数と呼ばれ、αの値によって、ある閾値θで急に値が変わる関数になる。これは、生物の刺激一応答曲線に似ているため、ニューラルネットワーク法と呼ばれている。

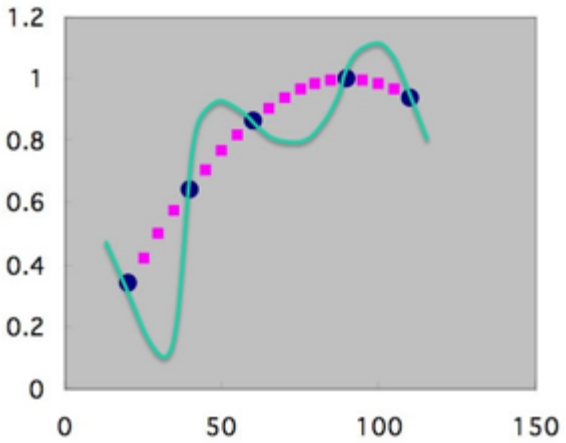
このシグモイド関数をいくつ（Sin の場合は 3 つ）足し合わせるかで、複雑な現象に対しても良好にフィティングを行うことができるようになる。



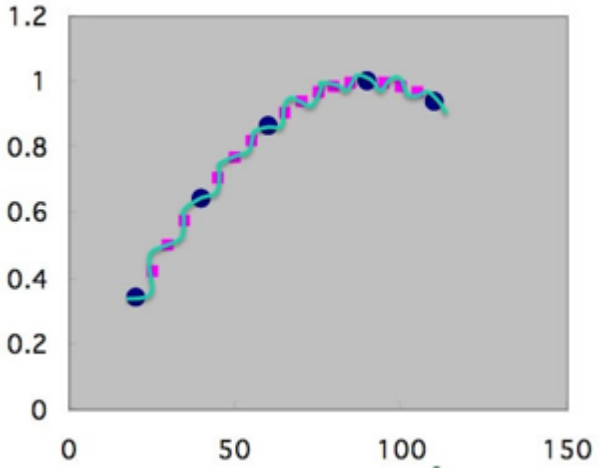
ただし、シグモイド関数をむやみに増やすと、後述するように、様ざまな弊害ももたらされる。

それでは、この、A1-A3, B1-B6 の組はどうしたら求まるであろうか？　重回帰法の場合は、連立方程式の解法なので、方程式を行列表現にして、逆行列を求めれば係数はすぐに求まる。非線形回帰式は、数学的な解の式が求まらないので、収束計算を行う。最初の重回帰の式、緑色の直線、Y=aX+b の a,b を収束計算で求める事を考えてみよう。データ点は 5 点ある。それから 2 点選ぶには、${}_5C_2$ の組み合わせがあるので 10 種類の a,b の組みが求まる。その 10 種類の直線のうち、一番トータル誤差の小さな直線がどれかは、簡単に解る。5 点全部の誤差の２乗を足し合わせ、最小のものが初期値になる。次に初期値の直線で一番誤差の大きい点を特定して、誤差が少し小さくなる方向に、Δa, Δb 動かす。そしてトータル誤差を計算する。これを繰り返していくと、（途中で最大誤差を与えるポイントは入れ替わるかもしれないが）一つの a, b に収束していく。
ニューラルネットワーク法の収束計算も、これと同じ方法をとる。単なる収束計算であるが、誤差逆伝播学習法(Error Back Propagation)というもったいぶった名前がつけられている。20 年前にニューラルネットワーク法がブームを迎えたのは、この EBP 学習法の性能がよかったからと言われている。しかし、その後廃れてしまったのは、EBP 学習法の問題点を解決できなかったからだと思う。20 年の時を経て、ビッグデータがあればという条件付きでディープラーニングというニューラルネットワーク法が復権したのは感慨深い。

コンピュータは人間と違い、飽きることを知らないので、命令すれば一晩中でも、1 年でも学習を続ける。すると、本来欲しい点はピンクの点であるのに、最終的には学習した点をなめらかにつなぐ緑の線を答えとして提案する。学習に使ったデータでは相関係数 R=1.000、”おおー、ものすごく良くモデルを再現できる式が構築できた”と喜んで論文に投稿する。でも、下のグラフを見れば明らかなように、学習した近傍以外は非常に悪い答えを返すモデルである。これなら、重回帰のほうがまだまし、っていうレベルである。



このような、学習状態を過学習という。過学習は学習のデータポイントが少ない、NN の中間ニューロン数が多すぎるときに起こりやすい。
そこで、ピンクの点も含めて学習させてしまうことを考える。

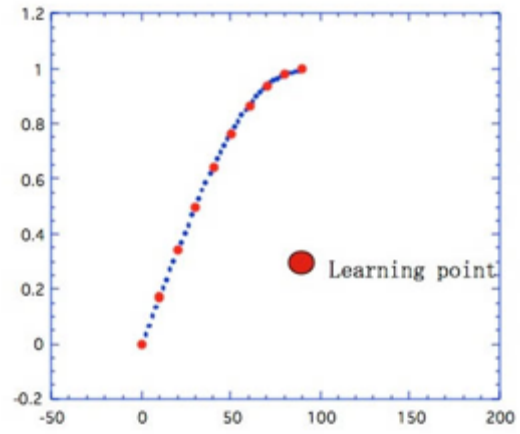


このようにデータ数が増えてくれば大きくハズレる曲線は取りようがなくなる。つまり、ビッグデータがあればニューラルネットワーク法の過学習は抑えられるのは、昔から分かっていた。問題は、新素材を考えなくてはならない化学の領域では、なかなかビッグデータが集まらないことだ。それは今も昔も状況は変わっていない。

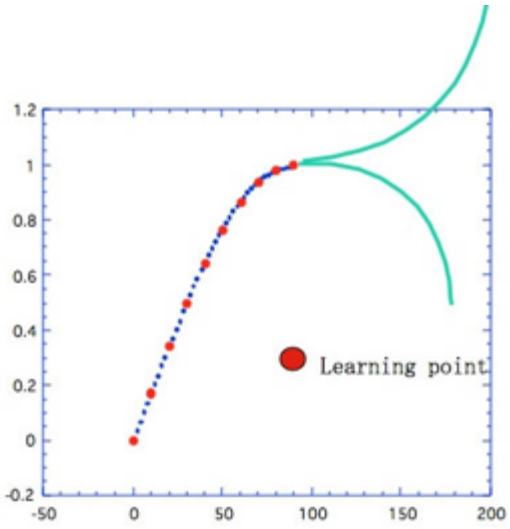
中間層のニューロン数は、何個シグモイド関数を足し算して、この現象を表現しようとしているかを表している。当

然ニューロン数を増やせば複雑な現象にも追随できるが、その分、過学習という、やっかいも背負いこむ。
こうした、中間層のニューロン数の最適化とデータポイントの数、これが考慮されずにニューラルネットワークを組んでも、無駄である。それを評価するのに赤池情報量基準（AIC）というものも知られているが、難しすぎて自分には理解出来できていない。それに、化学の現象はそれ以前に多くの、誤差、あやふやさを含んでいる。極端なことをいえば、熟練工の A さんと新人の B さんが同じ手順書にしたがって実験を行っても結果が異なる。AIC を考えるよりも、実験者 A,B をニューラルネットワークに入れる方が結果はよくなる。

もう一つ、ニューラルネットワーク法が廃れた理由は、予測性の欠如である。
先程の Sin カーブの学習で、Sin の頂点までしか学習しなかったとする。



そうした場合は、学習した内側については、中間ニューロンの数が適当であれば、青い点が示すように良好に予測することができる。これは、”内挿問題に対する予測性”は高いという言い方をする。
しかし、θ が 90 を超えた先がどこに行くのかは、NN は判らない。したがって 1 点でもよいから 90 度以上のデータを入れる必要がある。これは実は先程の過学習よりも、よっぽどたちが悪い。例えば、化合物の官能基を定義して沸点を推算する NN を構築しようとする。ある官能基、例えば水酸基を持つ化合物の沸点のデータが、水酸基が 1, 2, 3 個もつ化合物の沸点データがあり、NN に取り込んだとしよう、その場合は分子中に水酸基を 4 つもつものは、”外挿” 問題になり、予測性は高くないだろうと見当がつく。ところが、カルボン酸を 1-2 個もつ化合物、アミノ基を 1-2 個持つ化合物を NN に組み込んでも、分子中に、カルボン酸とアミノ基を両方もつアミノ酸を NN に組み込んでいなければ、これはやはり外挿になってしまう。こうした”組み合わせ”外挿はどれだけあるのかは、一般的に言えないし、非常にむずかしい問題である。また、アミノ酸のように、ひとつの炭素にカルボキシル基とアミノ基が付いている場合、遠く離れて付いている場合でも物性値は大きく異なる。



従って、NN で物性推算式を構築した場合には、どのような母集団から NN を構築したのかが非常に重要な問題になる。
この場合でも、ビッグデータが存在すれば問題は解決する。しかし、”組み合わせ”外挿を全て網羅するようなデータベースは残念ながら存在しない。画像解析や自然言語解釈では上手くいったディープ・ラーニングであるが、マテリアルに適用するのであれば、アプローチを変えなくてはならない。

実際の適用例

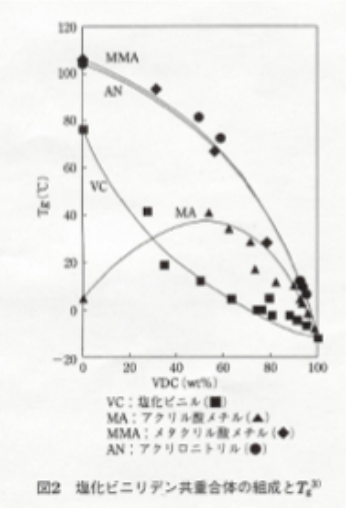
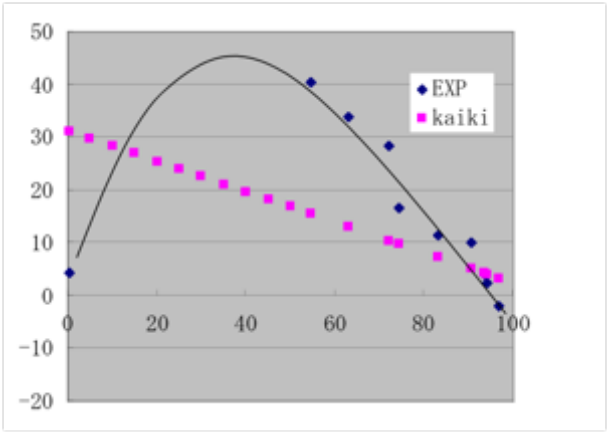


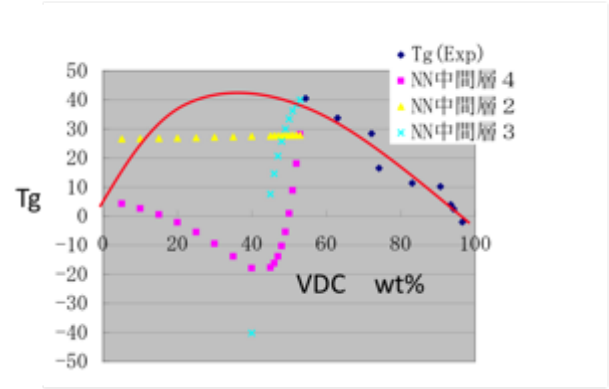
図2 塩化ビニリデン共重合体の組成とT_g[3]

第 4b 回でバリアー性素材のガラス転移温度(Tg)の予測

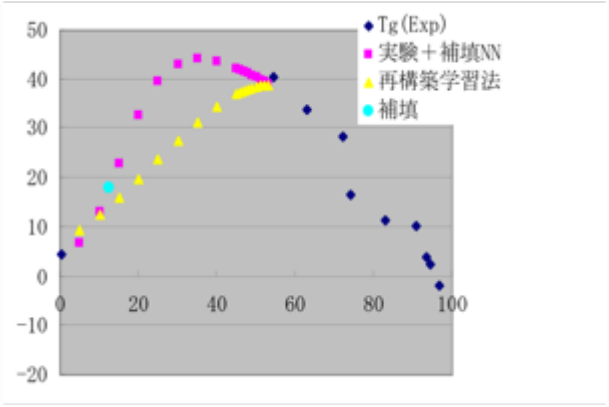
を取り上げた。ビニリデンクロライド と他のモノマーの共重合を考えた場合、アクリル酸メチルとの共重合体の場合だけ、Tg が上に凸になる現象があった。この曲線は見ようによると Sin カーブの一部のように見えなくもない。そして、先ほど述べた Sin 曲線の頂点の先の１点が与えられている。



当然のことながら、重回帰法では直線しか表現できないのでフィティングすることはできない。
そこで 2003 年当時用いていたニューラルネットワーク法のプログラムを使って解析したところ、中間層数が２では収束せず、3 とか4 では過学習を起こしてしまうことが判明した。

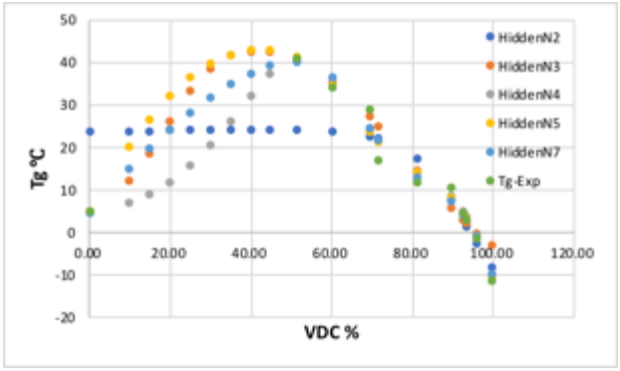


当時のニューラルネットワークの実力など、この程度であった。



そこで、こうしたデータ数が少なく、過学習になりやすい系を扱うときは、様ざまな工夫を駆使したものであった。例えば、VDC10%の値をデータに補填してあげると、過学習が抑えられるであるとか、再構築学習法をとれば、データ補填をしなくても上に凸を表現できるであるとかだ。

「中間層が多い場合、過学習を起こし、データを滑らかにつないでしまう」これを確認するために、最新の NN シミュレータで再計算してみた。



中間層が２の時には全く収束しない。しかし３以上の時に過学習を起こす事もない。何故だかは不明だが、中間層が 4 の時だけ挙動が異なる。15 年かけて、シミュレータを改良してきたので、データ数が少なくても過学習は起こしにくいシミュレータに進化したようだ。

シミュレータを進化させるには、自分でプログラムが書ける必要がある。しかし、化学系の学生にはそれはきついかもしれない。でも、プログラムを書くなどというのは、そのうち、AI が幾らでもやってくれるので心配することはない。問題は、「そのうち」というのが何時かということだ。それまでは、今できることをやっておこう。

シミュレータをいじらなくても今できることは、入力データを工夫することである。Sin カーブを予測する例で入力データの工夫のしかたを示そう。
Sin カーブは、シグモイド関数を用いれば、

f(y)=A1* / [1+exp(-(B1*X1+B2*1))＋
　　A2* / [1+exp(-(B3*X1+B4*1))＋
　　A3* / [1+exp(-(B5*X1+B6*1))

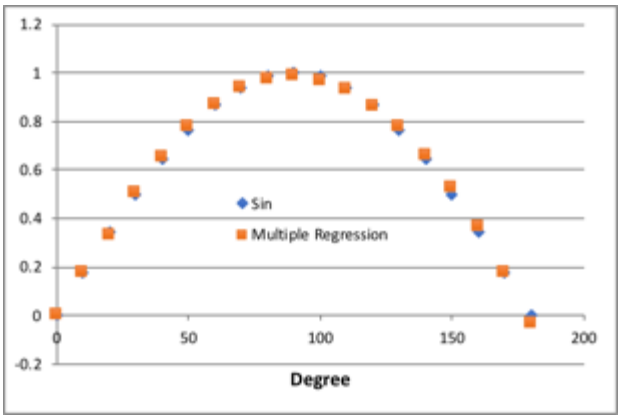
A1= -2.14　B1= -1.84　B2= 0.30
A2= -1.01　B3= 0.04　B4= 0.31
A3= 2.11　B5= -1.87　B6= 1.57

とすれば、精度高く予測可能であった。
その時の説明では、シグモイド関数を選択するか、他の関数を選択するかは自由だと述べた。
それでは、

f(y)=A1*(1/X1)
+A2*log(X1)
+A3*(X^2)
+A4*(sqrt(X1))
とした時の、A1-A4 を求めてみよう。
この係数を求めるのは非常に簡単で、普通の重回帰計算を使えば良い。

X:dgree	Sin	1/x	logX	x^2	sqrt(X)
0.01	0.0002	100.0000	-2.0000	0.0001	0.1000
10	0.1736	0.1000	1.0000	100	3.1623
20	0.3419	0.0500	1.3010	400	4.4721
30	0.4998	0.0333	1.4771	900	5.4772
40	0.6425	0.0250	1.6021	1600	6.3246
50	0.7658	0.0200	1.6990	2500	7.0711
60	0.8658	0.0167	1.7782	3600	7.7460
70	0.9395	0.0143	1.8451	4900	8.3666
80	0.9847	0.0125	1.9031	6400	8.9443
90	1.0000	0.0111	1.9542	8100	9.4868
100	0.9850	0.0100	2.0000	10000	10.0000
110	0.9400	0.0091	2.0414	12100	10.4881
120	0.8666	0.0083	2.0792	14400	10.9545
130	0.7668	0.0077	2.1139	16900	11.4018
140	0.6437	0.0071	2.1461	19600	11.8322
150	0.5011	0.0067	2.1761	22500	12.2474
160	0.3434	0.0063	2.2041	25600	12.6491
170	0.1751	0.0059	2.2304	28900	13.0384
180	0.0016	0.0056	2.2553	32400	13.4164

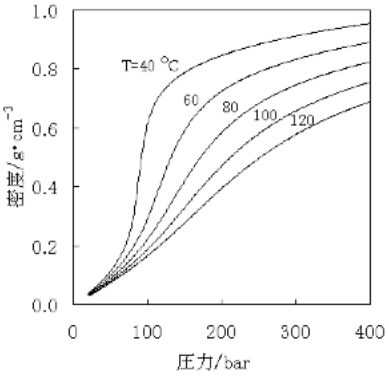
こうしたテーブルを作り、重回帰計算を行えば、重回帰係数が求まる。



厳密な Sin 計算値に対して、重回帰計算結果はほぼ完全に重なる。つまり、入力値を工夫すれば、非線形の曲線であっても重回帰の形で表すことができる。
ちなみに
A1:-0.0295014488567471
A2:-1.39531945518547
A3:-0.000099576191739178
A4:0.464330591108506
定数 : 0.113247988857779
と求まる。
（ただし、X1=0.0 の時には、1/X1 がエラーとなるので X1=0.0 は計算できない。）

このように、過学習や予測性の欠如の問題を抱えながら、無理やりニューラルネットワーク法を使う必要はない。かといって、無理やり 1/X だの log(X)などの非線形関数を使う必要もない。
例えば、ある温度での蒸気圧をフィティングする式として、Antoine 式というものがある。
$\log P(\mathrm{mmHg}) = A - B/(T℃ + C)$
A, B, C は化合物特有の定数としてフィティングされる。
この、Antoine 定数には物理化学的な意味がある。
例えば　Antoine B は化合物の蒸発潜熱と相関がある。
Antoine C は沸点と相関がある。
ところが、最近開発されている蒸気圧式は、
$\log P = A + B/T + C{:}\log_{10}T + D*T + E*T^2$
などとなっている。この場合の A-E は単なるフィティング・パラメータであって物理化学的意味はない。
しかし、Antoine A, B, C は素材を設計する際に非常に有用なパラメータとして重用される。
また、何故ある化合物では Antoine 式では不十分なのかを考えることによって、その化合物特有の分子間力を定量的に見積もることができるようになる。

宿題
超臨界炭酸ガスの密度は、温度圧力によって急激に変化する。



例えば、密度が 0.18 から 0.5 の間にはいる領域は、Miracle Density と呼ばれ、化学反応が特異的に促進される密度として知られる。温度と圧力は計器の測定値としてわかるとして、ある温度、圧力の時の密度の予測値を求める式を作ってみよう。データー点をデジタイザーで取り、ニューラルネットワーク法で計算すれば、すぐに収束するだろう。
次に、Sin カーブの時と同じように、入力値を工夫することによって、重回帰法で予測する式を作成し提出するように。

雑感

シミュレーションは、よく合うものについては、全く面白くない。あとは Ai にでも任せれば良い。ある理論でシミュレーションしたのに合わないものにこそ、研究する価値のある未知の何かが隠されている。

ニューラルネットワークは、中がブラックボックスで、何故だかわからないが合ってしまうので、気付きがなくなりがちになる。気をつけないと、中身のない薄っぺらな研究者になってしまう。シミュレータを工夫するか、入力を工夫するかして、化学の知識を総動員して、かつ、AI にアシストしてもらい、何故、VDC—MA の Tg が上に凸になるかを理解しなければ、次にどんなモノマーを使ったら良いかはわからない。

かといって、無視できる技術かといえば、そんな甘いものでもない。

化学の地力も持っている者が AI アシストを受けて、気づきを得られるなら鬼に金棒なのだと思う。

それまでは, 実験科学者とよく相談しながら協業するのだろう。

Pirika マテリアル・ゲノム のページ

以下 PDF

第 1 回 イントロダクション 2018.8.23

第 2 回 データ収集と昔ながらのやり方 2018.8.24

第 0 回 物性推算と逆設計と呼んでいた時の話 2000.8.28

なんと 18 年前！

第 3a 回 ポリマー設計と 3 つの MI（その 1） 2018.9.3

第 3b 回 ポリマー設計と 3 つの MI（その 2） 2018.9.3

第 4a 回 MI に適した簡単なデータベースの利用法 2018.9.4

第 4b 回 複雑なポリマーのデータベース化 2018.9.7

プレゼン用：MI を使う時のデータベース構築法 2018.9.11

プレゼン用：複雑なポリマーの設計と DB 2018.9.15

第 5 回 データのクレンジング 2018.8.28

第 6a 回 ニューラルネットワーク法の初歩 2018.9.25

第 6b 回 ニューラルネットワーク法を使った Drug Design 2018.9.22

第 7 回 遺伝的アルゴリズム(GA)を理解しよう