ニューラル正則化の幾何

著　佐藤 怜
所属　筑波大学 情報科学類
日付　2017.9.13

1　概要

本稿の目的は, 深層学習の正則化を幾何的に解釈することである.

2　分離超平面解釈から写像解釈へ

本章では, ニューラルネットワークの層間結合の幾何的な解釈として, 分離超平面解釈と写像解釈について述べる.

2.1　分離超平面

本稿では, $l-1$ 層 i ノードと l 層 j ノードの間の重みを $w^l_{i(j)}$ と表記する. また, l 層 j ノードへの重み, すなわち層間重み行列の j 行を $w^l_{(j)}$ と表記する. ここで層間重み行列 W はバイアスを含まないものとする. $w^l_{(j)}$ に付随するバイアスは $b^l_{(j)}$ と表記する. 入力層は第 0 層である.

全結合ニューラルネットワークの l 層 j ノードの活性 h^l_j は, 活性化関数を $f(x)$ とおいて以下で与えられる.

$$h^l_j = f((\sum_i w^l_{i(j)} * h^{l-1}_i) + b^l_{(j)}) \tag{1}$$

ここで, 超平面の方程式を考える.

$$(\sum_i w_i x_i) + b = 0 \tag{2}$$

すると, 式 1の活性化関数 $f(x)$ の引数を見ると, これは $w^l_{(j)}$ と $b^l_{(j)}$ をパラメータとする超平面と, ベクトル h^{l-1}_i のノルム正規化されていない符号付き距離と捉えることが出来る. 従って中間層の各ノードの活性は, 1 つの分離超平面からのノルム倍された距離に活性化関数による非線形性を与えたものであると解釈することが出来る.

図 1に, 2 次元で 2 クラスの人工データに対して 3 ノードの隠れ層を 1 層持つ全結合ニューラルネットワークを学習させた $w^1_{(j)}, b^1_{(j)}$ の例を, 分離超平面として解釈した結果を図示する.

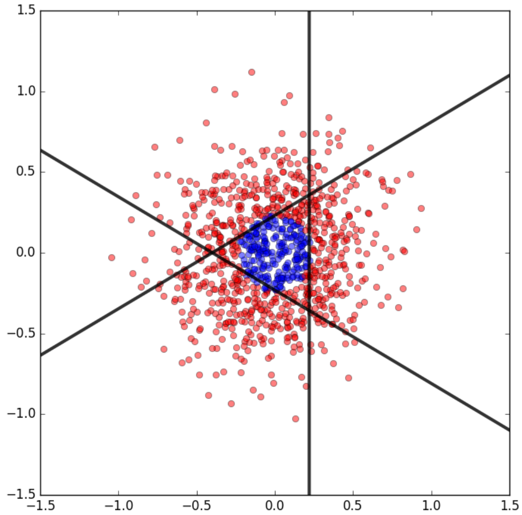

図 1　重み行列の分離超平面解釈. 黒い線が分離超平面

2.2　写像

分離超平面はそれぞれ, 層間重み行列の行ベクトルと入力ベクトルの内積に, バイアスを足したものが 0 になるベクトルの集合であることは既に述べた. この層間重み行列の行ベクトル, すなわち各分離超平面の法線ベクトルを延長したベクトルを図 2に図示する.

また, 図 3には 2 次元のデータ群を重み行列によって写した像空間を図示している. 行列の性質として自明であるが, 図 2で非直交だった重み行列の行ベクトルらは, 図 3で直交基底になっている.

図 4には図 3をバイアスに基づいて平行移動させ, ReLU によって活性化した様子を図示する.

解説

図 1の元空間に分布する 2 クラスのデータ群は, 1 つの分離超平面では仕切れそうにない. そこで 3 つの分離超平面を用意し, これらは学習によって, 図 1のように直感的に 2 クラスを仕切るように配されているとする.

すると写像解釈と活性化関数を適用した像空間 (図 4) では, 原点付近を $w^2_{(j)}, b^2_{(j)}$ による新たな分離超平面で仕切れば, 1 つの分離超平面だけで 2 クラスを仕切れそうに見受けられる.

従って, 分離超平面として解釈した重み行列が直感的に識別的であれば, 重み行列を写像として解釈した際の像空間は, 元空間より識別が容易である (言い換えると, 少ない分離超平面で仕切れる) ということがわかる.

(ただし分離超平面が識別的であるには, 法線ベクトルの表裏も重要である.)

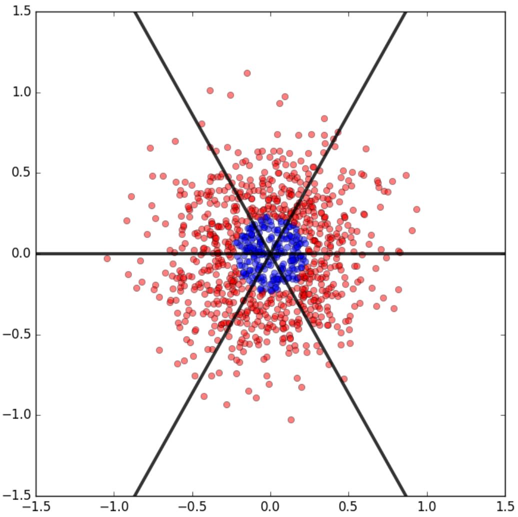

図 2　元と重み行列の行ベクトル

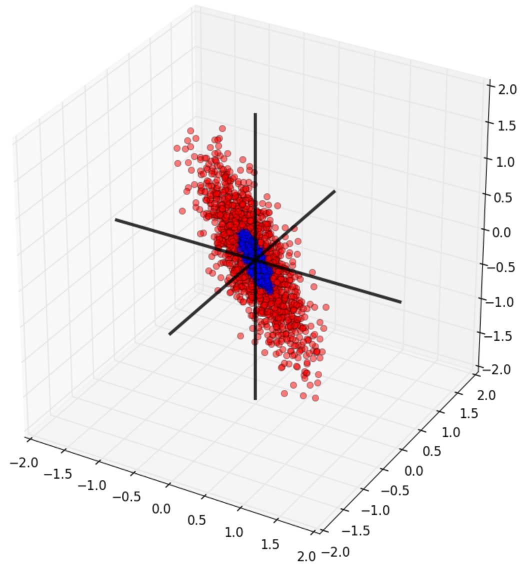

図 3　重み行列による像

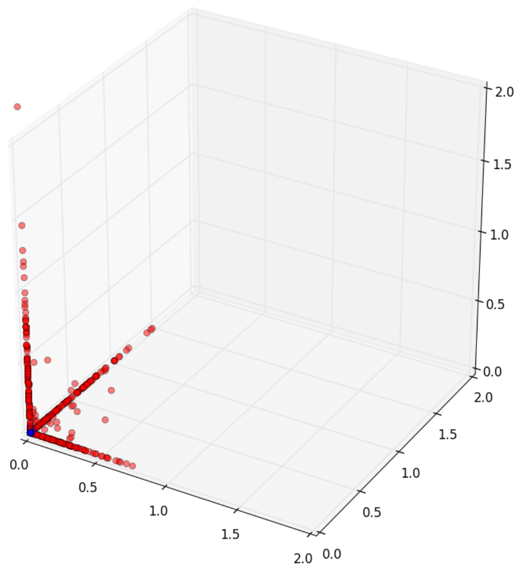

図 4　図 3を活性化させた像

3　スライサーとしての ReLU

既に登場しているが, ReLU(Rectified Linear Unit) は次式で与えられる活性化関数を持つノードである.

$$f(x) = max(0, x) \tag{3}$$

ReLU を採用したニューラルネットワークの動作を図 5と図 6に示す. ReLU により, 原点を通る軸で分布が切

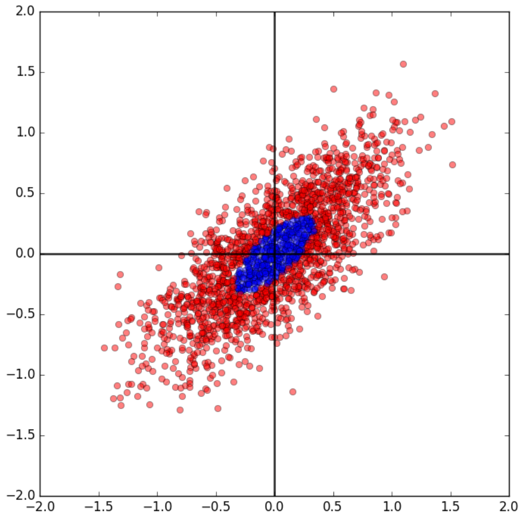

図 5　層間結合による像の例

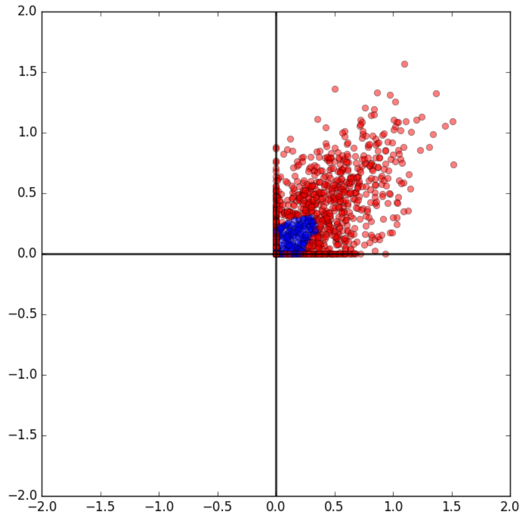

図 6　像に活性化関数を適用した分布

り取られるような変形をしている．さて, 像を切り取る位置は原点を通る各軸に固定されており, 学習の余地は無い. 分布の切り取り方は, バイアス項も含めた層間重み行列によって分布を動かすことによってのみ調節が可能である.

ReLU による切り取りを本稿では**スライス**すると呼ぶことにする.

4　行列の作用

行列による写像には, **Translation**(平行移動),**Rotation**(回転),**Skew**(歪み),**Scale**(拡大縮小) の作用がある．ReLU によるスライス効果を考えると, 層間重み行列とバイアスによって平行移動, 回転, 歪み, 拡大縮小を加えながら固定されたスライサーの位置へ分布を合わせることで非線形性を与えるので, この内 Scale はニューラルネットワークの層間重み行列に不要な作用であることが分かる.

本稿ではこの後, 層間重み行列から Scale の作用を取り除くことが正則化と見なせることを示す.

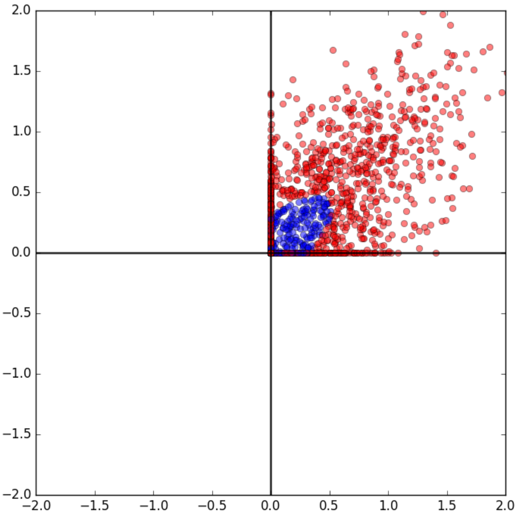

図 7　図 5 の分布を拡大し ReLU を適用した分布. 切り取られ方は図 6 と同じであり, 層間重み行列に Scale 作用は不要であることがわかる.

5　QR 分解と正規化による正則化

層間重み行列 W を直交行列 Q と三角行列 R に QR 分解する. 三角行列 R の対角成分の積が 1 になるように R 全体に係数を掛けて正規化し, これを R' とする.

すると, Q と R' の行列式はそれぞれ 1 になるから, $W' = QR'$ とした時, W' には Scale の作用は無いことが分かる.

これは厳密には重み減衰 (weight decay) や重み上限 (max-norm regularization) と一致しないが, 重みの大きさに対して制約を課しているので, 一種の正則化と見なせる.

6 Appendix

6.1 QR 分解 (QR Decomposition)

QR 分解とは, $m \times n$ 行列 W を $m \times m$ 直交行列 Q と $n \times n$ 三角行列 R の積

$$W = QR \tag{4}$$

に分解することを言う.

6.1.1 グラムシュミットの直交化法による QR 分解

W がフルランクであるとする.W の行ベクトルを $w_1 ... w_j$ とする.

$$\begin{aligned}
q_1 &= w_1 \\
v_1 &= \frac{q_1}{|q_1|} \\
q_2 &= w_2 - (v_1 \cdot w_2)v_1 \\
v_2 &= \frac{q_2}{|q_2|} \\
q_3 &= w_3 - (v_1 \cdot w_3)v_1 - (v_2 \cdot w_3)v_2 \\
v_3 &= \frac{q_3}{|q_3|}
\end{aligned}$$

さて, 上式から下式が導ける.

$$\begin{aligned}
w_1 &= |q_1|v_1 \\
w_2 &= (v_1 \cdot w_2)v_1 + |q_2|v_2 \\
w_3 &= (v_1 \cdot w_3)v_1 + (v_2 \cdot w_3)v_2 + |q_3|v_3
\end{aligned}$$

これを行列で表現すると

$$W = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \begin{pmatrix} |q_1| & (v_1 \cdot w_2) & (v_1 \cdot w_3) \\ 0 & |q_2| & (v_2 \cdot w_3) \\ 0 & 0 & |q_3| \end{pmatrix}$$

今, v_n はそれぞれ正規直交なベクトル群であるから, $(v_1\ v_2\ v_3)$ は直交行列である.

6.2 積の行列式の性質

性質

2 つの行列の積の行列式は, それぞれの行列の行列式の積に等しい.

証明

省略

6.3 直交行列の行列式の性質

性質

直交行列の行列式は ± 1 である.

証明

直交行列の定義により, $M^T = M^{-1}$ であるから, $M^T M = I$ である.

転置行列の行列式が元の行列の行列式と等しいことを用いて, $det\, M^T = det\, M$ である.

積の行列式の性質を用いて, $det\, M^T M = det\, M^T\, det\, M$ である.

以上より, $(det\, M)^2 = det\, I = 1$ である.

従って, 直交行列の行列式は ± 1 である.

6.4 三角行列の行列式の性質

性質

三角行列の行列式は対角成分の積に等しい.

証明

ここでは上三角行列について言及する. $sgn(\sigma) \in \{-1, +1\}$ は互換の回数の偶奇を表し, 奇置換なら -1, 偶置換なら $+1$ とする.

行列式は $\sum_\sigma sgn(\sigma) \prod_i^n r_{i\sigma(i)}$ である.r_{ij} は行列の (i, j) 成分を表す.

$\sigma(i)$ が恒等な時, $\prod_i^n r_{i\sigma(i)}$ は $\prod_i^n r_{ii}$ に等しいから, 対角成分の積である. また, 恒等な σ は n 次の置換で唯一つである.

$\sigma(i)$ が恒等でない時, $\prod_i^n r_{i\sigma(i)}$ は対角成分より下の要素を積に含むから, 0 である.

よって行列式は $sgn(\sigma) \prod_i^n r_{ii} = \prod_i^n r_{ii}$ に等しい.

6.5 行列式の拡大率としての性質

行列式は, 行列による写像の拡大率に等しい. 行列式に関する定理で直感的なものとしては, 「固有値の積は行列式に等しい」や「特異値の二乗の積は行列式の絶対値に等しい」といったものが挙げられるが, 証明は省略する.

参考文献