ディープラーニングと物理学

4.2 再帰的ニューラルネットワークと誤差逆伝播法

須賀勇貴

茨城大学大学院 理工学研究科 量子線科学専攻 2 年

April 16, 2023

目次

時系列データについて

系列データ

個々の要素が順序付きの集まりとして与えられるデータのこと

(ex)

- 動画データ $\rightarrow$ 順序付きの自然画像データ
- 文章データ $\rightarrow$ 順序付きの文字画像データ
- 会話データ $\rightarrow$ 順序付きの音声データ

長さが T の系列データは以下のように表現できる

$$\begin{pmatrix} x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(T) \end{pmatrix} = |x(t)\rangle \quad (t = 1, 2, 3, \cdots, T)$$

時系列データについて

(ex) "This is an apple ." という文章データを系列データとして扱う場合

$$
\begin{aligned}
|x(1)\rangle &= |\text{This}\rangle \\
|x(2)\rangle &= |\text{is}\rangle \\
|x(3)\rangle &= |\text{an}\rangle \\
|x(4)\rangle &= |\text{apple}\rangle \\
|x(5)\rangle &= |.\rangle
\end{aligned}
$$

文字データなどは **1-of-K** ベクトルなどにより数値ベクトルとして表現される

$$
\boldsymbol{x}^t = \left(\overset{\text{a}}{0}\ \ \overset{\text{aah}}{0}\ \ \cdots\ \ \overset{\text{applause}}{0}\ \ \overset{\text{apple}}{1}\ \ \overset{\text{applet}}{0}\ \ \cdots\ \ \overset{\text{zzz}}{0}\right)^\top
$$

図: "apple" の 1-of-k ベクトル表示

再帰的ニューラルネットワーク (RNN) の考え方

これまで、扱ってきたデータはデータ間につながりがないものだった

$$\Downarrow$$

系列データをニューラルネットワークで扱えるようにしたい

$$\Downarrow$$

データ間のつながりを表現できるようなニューラルネットワークを構築すればよい！

再帰的ニューラルネットワーク (RNN) の考え方

素朴な考え方

⇒ 前の時刻の出力を次の時刻の入力に加えるようなニューラルネットワーク

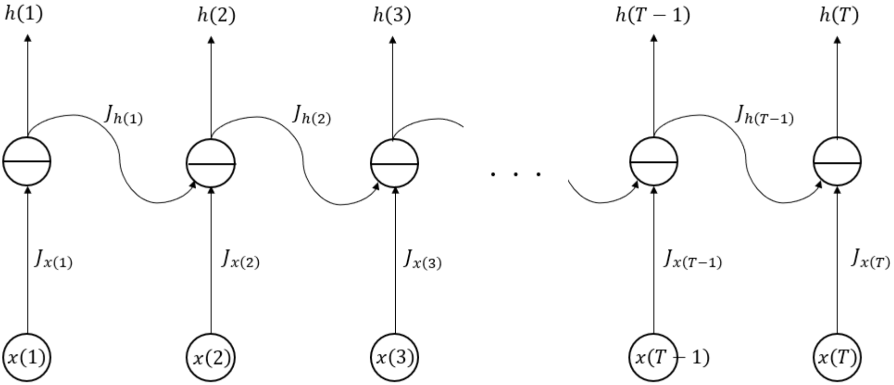
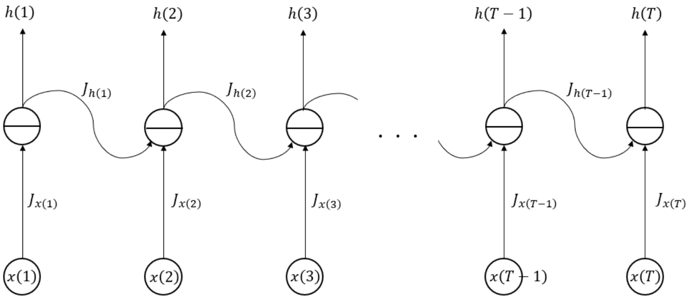


図: 最もシンプルな形の RNN

再帰的ニューラルネットワーク (RNN) の考え方



以下のようにベクトルや行列で表現しておく

$$|x(t)\rangle = (x(1), x(2), \cdots, x(T))^\top$$
$$|h(t)\rangle = (h(1), h(2), \cdots, h(T))^\top$$
$$\mathbb{J}_x = \mathrm{diag}(J_{x(1)}, J_{x(2)}, \cdots, J_{x(T)})$$
$$\mathbb{J}_h = \mathrm{diag}(J_{h(1)}, J_{h(2)}, \cdots, J_{h(T)})$$

再帰的ニューラルネットワーク (RNN) の考え方

各時刻での出力は以下のように簡単に求められる

$$\begin{aligned} h(1) &= \sigma_\bullet(J_{x(1)}x(1)) \\ h(2) &= \sigma_\bullet(J_{x(2)}x(2) + J_{h(1)}h(1)) \\ &\vdots \\ h(t) &= \sigma_\bullet(J_{x(t)}(t)x(t) + J_{h(t-1)}h(t-1)) \end{aligned}$$

ブラケット表記より，まとめて以下のように書ける

$$\begin{aligned} |h(t)\rangle &= \sigma_\bullet(\mathbb{J}_x |x(t)\rangle + \mathbb{J}_h |h(t-1)\rangle) \\ &= \sum_m |m\rangle \sigma_\bullet\Big(\langle m| \mathbb{J}_x |x(t)\rangle + \langle m| \mathbb{J}_h |h(t-1)\rangle \Big) \end{aligned}$$

再帰的ニューラルネットワーク (RNN) の考え方

「ループ」のような構造 (=再帰的) を持つニューラルネットワーク
⇒ 再帰的ニューラルネットワーク

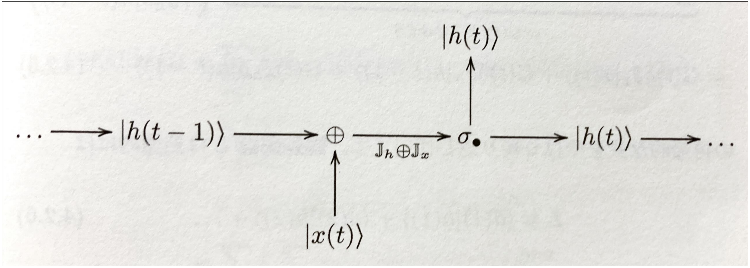


図: 素朴な再帰的ニューラルネットワークの模式図

$$|h(t)\rangle = \sum_m |m\rangle\, \sigma_\bullet\Big(\langle m|\, \mathbb{J}_x\, |x(t)\rangle + \langle m|\, \mathbb{J}_h\, |h(t-1)\rangle \Big)$$

RNNにおける誤差逆伝播

誤差関数 L は典型的に以下のようになっているとする

$$L = \langle d(1)|h(1)\rangle + \langle d(2)|h(2)\rangle + \cdots + \langle d(T)|h(T)\rangle = \sum_t \langle d(t)|h(t)\rangle$$

t 番目の変化量に着目 (計算↓)

$$\begin{aligned}
\delta \langle d(t)|h(t)\rangle &= \langle d(t)|\, \delta\, |h(t)\rangle \\
&= \cdots \\
&= \langle \delta_t(t)|\, \delta\mathbb{J}_x\, |x(t)\rangle + \langle \delta_t(t)|\, \delta\mathbb{J}_h\, |h(t-1)\rangle \\
&+ \langle \delta_t(t-1)|\, \delta\mathbb{J}_x\, |x(t-1)\rangle + \langle \delta_t(t-1)|\, \delta\mathbb{J}_h\, |h(t-2)\rangle \\
&+ \cdots \\
&+ \langle \delta_t(1)|\, \delta\mathbb{J}_x\, |x(1)\rangle + \langle \delta_t(1)|\, \delta\mathbb{J}_h\, |h(0)\rangle \\
&= \sum_{\tau \le t} \Big(\langle \delta_t(\tau)|\, \delta\mathbb{J}_x\, |x(\tau)\rangle + \langle \delta_t(\tau)|\, \delta\mathbb{J}_h\, |h(\tau-1)\rangle \Big)
\end{aligned}$$

RNN における誤差逆伝播

ここで，

$$\delta \mathbb{J} = \sum_{m,n} |m\rangle \langle n| \, \delta J \ mn$$

とすると

$$\begin{aligned}
\delta \langle d(t)|h(t)\rangle &= \sum_{\tau \leq t} \Big(\langle \delta_t(\tau)| \, \delta \mathbb{J}_x \, |x(\tau)\rangle + \langle \delta_t(\tau)| \, \delta \mathbb{J}_h \, |h(\tau-1)\rangle \Big) \\
&= \sum_{\tau \leq t} \sum_{m,n} \Big(\langle \delta_t(\tau)|m\rangle \langle n|x(\tau)\rangle \, \delta J_x^{mn} + \langle \delta_t(\tau)|m\rangle \langle n|h(\tau-1)\rangle \, \delta J_h^{mn} \Big) \\
&= \sum_{m,n} \Big(\sum_{\tau \leq t} \langle \delta_t(\tau)|m\rangle \langle n|x(\tau)\rangle \, \delta J_x^{mn} + \sum_{\tau \leq t} \langle \delta_t(\tau)|m\rangle \langle n|h(\tau-1)\rangle \, \delta J_h^{mn} \Big)
\end{aligned}$$

RNN における誤差逆伝播

以上の結果から誤差関数 L の変化量は

$$
\begin{aligned}
\delta L &= \sum_t \delta \langle d(t)|h(t)\rangle \\
&= \sum_{m,n} \Bigg(\sum_t \sum_{\tau \le t} \langle \delta_t(\tau)|m\rangle \langle n|x(\tau)\rangle \, \delta J_x^{mn} \\
&\qquad + \sum_t \sum_{\tau \le t} \langle \delta_t(\tau)|m\rangle \langle n|h(\tau-1)\rangle \, \delta J_h^{mn} \Bigg)
\end{aligned}
$$

と表すことができる．この結果より，パラメータ J_x^{mn} と J_h^{mn} はそれぞれ

$$
\begin{aligned}
\delta J_x^{mn} &= -\epsilon \sum_t \sum_{\tau \le t} \langle \delta_t(\tau)|m\rangle \langle n|x(\tau)\rangle \\
\delta J_h^{mn} &= -\epsilon \sum_t \sum_{\tau \le t} \langle \delta_t(\tau)|m\rangle \langle n|h(\tau-1)\rangle
\end{aligned}
$$

という更新ルールにすれば，誤差関数の値が小さくなるようになる．

RNN における誤差逆伝播

(check)

$$
\delta L = -\epsilon \sum_{n,m} \times \underbrace{\left[\left(\sum_t \sum_{\tau \le t} \langle \delta_t(\tau) | m \rangle \langle n | x(\tau) \rangle \right)^2 + \left(\sum_t \sum_{\tau \le t} \langle \delta_t(\tau) | m \rangle \langle n | h(\tau - 1) \rangle \right)^2 \right]}_{>0} < 0
$$

今回の場合，逆伝播の式は以下

$$
\langle \delta_t(\tau - 1) | = \langle \delta_t(\tau) | \, \mathbb{J}_h \mathbb{G}(\tau - 1)
$$

$$
\langle \delta_t(t) | = \langle d(t) | \, \mathbb{G}(t)
$$

勾配爆発/勾配消失

ここで考えた再帰的ニューラルネットワークでは勾配爆発か勾配消失のどちらかが起こってしまい，うまく学習することができないという問題がある．

逆伝播のより

$$
\begin{aligned}
\langle \delta_t(1)| &= \langle \delta_t(2)|\,\mathbb{J}_h \mathbb{G}(1) = \langle \delta_t(3)|\,\mathbb{J}_h \mathbb{G}(2) \mathbb{J}_h \mathbb{G}(1) \\
&= \cdots \\
&= \langle d(T)|\,\mathbb{G}(T) \mathbb{J}_h \mathbb{G}(T-1) \cdots \mathbb{J}_h \mathbb{G}(2) \mathbb{J}_h \mathbb{G}(1)
\end{aligned}
$$

$\Rightarrow \mathbb{J}_h$ が右から $T-1$ 回右から作用することになる

文章データを扱うのであれば T は文章の単語数などに対応する
⇒1000 単語ある文章を学習するとなると $\mathbb{J}_h$ がほぼ 1000 回かかることになる

勾配爆発/勾配消失

$$\langle \delta_t(1)| = \langle d(T)|\, \mathbb{G}(T)\underline{\mathbb{J}_h}\mathbb{G}(T-1)\cdots\underline{\mathbb{J}_h}\mathbb{G}(2)\underline{\mathbb{J}_h}\mathbb{G}(1)$$

$|\mathbb{J}_h| > 1 : \langle \delta_t(1)|$ が巨大すぎる $\Rightarrow$ 勾配爆発

$|\mathbb{J}_h| < 1 : \langle \delta_t(1)|$ が小さすぎる $\Rightarrow$ 勾配消失

勾配爆発はなぜいけないのか $\cdots$
パラメータの更新量が大きくなることになるので適切な極値を見つけられなくなる (更新量はある程度小さい値でなければならない)

勾配消失はなぜいけないのか $\cdots$
勾配がほぼ 0 になってしまうと，パラメータが更新されないということになってしまう (=記憶の忘却)

$\delta\,|h(t)\rangle$ の計算

パラメータに依存するのは $\mathbb{J}_{x,h}$ と $|h(t-1)\rangle$ なので

$$
\begin{aligned}
\delta\,|h(t)\rangle &= \underbrace{\sum_m |m\rangle\,\sigma'_\bullet\Big(\,\langle m|\,\mathbb{J}_x\,|x(t)\rangle + \langle m|\,\mathbb{J}_h\,|h(t-1)\rangle\,\Big)\,\langle m|}_{=:\mathbb{G}(t)} \\
&\quad\times\delta\Big(\mathbb{J}_x\,|x(t)\rangle + \mathbb{J}_h\,|h(t-1)\rangle\,\Big) \\
&= \mathbb{G}(t)\Big(\delta\mathbb{J}_x\,|x(t)\rangle + \delta\mathbb{J}_h\,|h(t-1)\rangle + \mathbb{J}_h\delta\,|h(t-1)\rangle\,\Big) \\
&= \mathbb{G}(t)\delta\mathbb{J}_x\,|x(t)\rangle + \mathbb{G}(t)\delta\mathbb{J}_h\,|h(t-1)\rangle + \mathbb{G}(t)\mathbb{J}_h\delta\,|h(t-1)\rangle
\end{aligned}
$$

(戻る)