

『AI・データサイエンスのための数学入門』学習補助資料

ベクトルが確率になるまで

埋め込み・非線形変換・softmax

1 埋め込みは「番号」から「特徴の組」への変換

トークン ID が 1、2、3 であっても、ID の大小や差に意味の近さがあるわけではありません。埋め込みは、各 ID に学習可能な実数ベクトルを対応させます。本文の数値例は計算を説明するための仮の値であり、各成分に最初から「知性」「便利さ」などの意味が付いているわけではありません。

行の取り出しも行列積で表せる

以下では個々のベクトルを列ベクトルとして書きます。語彙数 $V = 4$ 、次元 $d = 2$ で、「便利」が 3 番目なら、対応する one-hot ベクトルは

$$q = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad E^T q = \begin{pmatrix} .80 & .10 & .30 & .70 \\ .10 & .60 & .90 & .40 \end{pmatrix} q = \begin{pmatrix} .30 \\ .90 \end{pmatrix}.$$

つまり、 $E^T q$ は E の第 3 行を列にしたものです。実際の計算では、巨大な one-hot ベクトルを作らず、対応する行を直接取り出せます。

量	形	役割
E	$V \times d$	語彙全体の埋め込み表
q	$V \times 1$	トークン ID の選択
$e = E^T q$	$d \times 1$	トークンのベクトル
$x_t = e_t + p_t$	$d \times 1$	位置情報を加えた入力

同じ語でも、位置や文脈が違ふ

通常の埋め込み表では、同じ ID は同じ行を参照します。異なる位置では p_t が変わり、その後の文脈処理では周囲のトークンに応じて表現が変わります。埋め込み表から取り出した表現と、文脈を処理した後の表現を区別しましょう。

位置ベクトルの加算について

$e_t + p_t$ は位置情報を組み込む一つの方法です。この和だけから e_t と p_t をいつでも一意に分離できる、という意味ではありません。また、位置の扱いには加算以外の方法もあります。

2 なぜ活性化関数が必要なのか

アフィン変換を 2 回続けると、

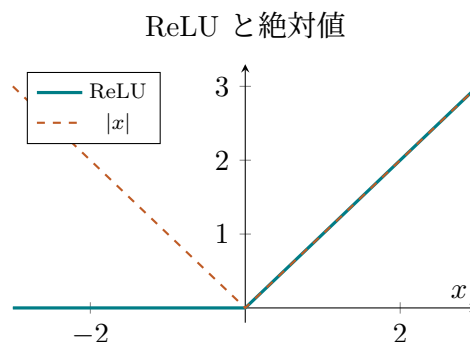
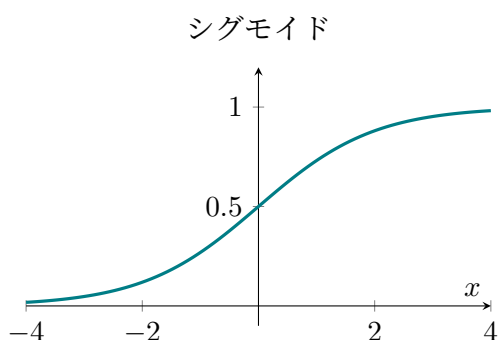
$$W_2(W_1x + b_1) + b_2 = (W_2W_1)x + (W_2b_1 + b_2)$$

です。深く重ねても、全体は一つのアフィン変換にまとめられます。途中に非線形関数を入れると、この制限を超えられます。

ReLU を 2 個組み合わせるだけでも形が変わる

$$\text{ReLU}(x) + \text{ReLU}(-x) = \begin{cases} x & (x \geq 0), \\ -x & (x < 0) \end{cases} = |x|.$$

1 本の直線では表せない折れ曲がりだが、単純な部品の組合せで作れます。



左：出力は $(0, 1)$ に入る。右：正の側では 2 本が重なる。

シグモイドの値と「確率分布」は別

ベクトルの各成分にシグモイドを適用しても、総和は一般に 1 になりません。例えば入力 $(0, 0, 0)$ では出力は $(1/2, 1/2, 1/2)$ です。互いに排他的な候補の確率分布を作るには、候補全体で正規化する softmax が適しています。

普遍近似定理が保証する範囲

本文の定理は、連続関数を有界な閉領域で任意の精度に近似できる**重みが存在する**という主張です。必要なユニット数が小さいこと、学習がその重みを発見すること、訓練範囲外でも正確なことまでは保証しません。

3 小さなモデルで、一連の計算を追う

計算のつながりを見るため、入力から 3 候補のスコアを作る模型を考えます。文脈を処理する attention などは省略しています。

埋め込みに位置を加える

$$e = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad p = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad x = e + p = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

特徴を変換し、ReLU を適用する

$$W = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad Wx + b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

$$h = \text{ReLU}(Wx + b) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

ReLU はベクトルの各成分に作用します。

語彙の候補ごとにロジットを作る

特徴の次元と、候補の個数は同じでなくても構いません。

$$A = \begin{pmatrix} 0 & 1 \\ \log 2 & -1 \\ \log 3 & 2 \end{pmatrix} \in \mathbb{R}^{3 \times 2}, \quad z = Ah = \begin{pmatrix} 0 \\ \log 2 \\ \log 3 \end{pmatrix}.$$

指数を取り、合計で割る

$$(e^{z_1}, e^{z_2}, e^{z_3}) = (1, 2, 3), \quad (p_1, p_2, p_3) = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6}\right).$$

計算の形を確認する

入力 $x \in \mathbb{R}^2$ 、特徴 $h \in \mathbb{R}^2$ 、ロジット $z \in \mathbb{R}^3$ 、確率 $p \in \mathbb{R}^3$ です。一般には、最終的なロジットの個数が語彙数になります。softmax は候補の順位を保ちますが、どのトークンを実際に選ぶかは、その後の選択方法で決まります。

確率 $1/2$ は、このモデルが文脈から候補 3 に割り当てた値です。「そのトークンを含む文の内容が 50% の確率で事実」という意味ではありません。

4 softmax の性質を式で確かめる

以下、ロジットは有限の実数とし、 $\log$ は自然対数を表します。

正の値になり、和が 1 になる

$e^{z_i} > 0$ なので $p_i > 0$ です。また、

$$\sum_i p_i = \frac{\sum_i e^{z_i}}{\sum_j e^{z_j}} = 1.$$

共通の定数を加えても変わらない

$$\frac{e^{z_i+c}}{\sum_j e^{z_j+c}} = \frac{e^c e^{z_i}}{e^c \sum_j e^{z_j}} = \frac{e^{z_i}}{\sum_j e^{z_j}}.$$

例えば (1002, 1001, 1000) は、最大値 1002 を引けば (0, -1, -2) になります。結果の確率は同じですが、指数計算のオーバーフローを避けられます。丸め誤差までなくなるわけではありません。

確率の「比」はロジットの「差」で決まる

$$\boxed{\frac{p_i}{p_j} = e^{z_i - z_j}}, \quad \boxed{\log \frac{p_i}{p_j} = z_i - z_j}.$$

$z_i - z_j = \log 2$ なら、 p_i は p_j の 2 倍です。ただし p_i の値そのものは、他の候補のロジットにも依存します。

ロジット	p_1	p_1/p_2
(log 2, 0)	2/3	2
(log 2, 0, 0)	1/2	2

候補が 2 個ならシグモイドにつながる

$$p_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} = \frac{1}{1 + e^{-(z_1 - z_2)}} = \sigma(z_1 - z_2), \quad p_2 = 1 - p_1.$$

学習への橋渡し【発展】

正解候補を a とすると、損失は $\ell = -\log p_a = -z_a + \log \sum_j e^{z_j}$ です。微分すると

$$\frac{\partial \ell}{\partial z_i} = p_i - \mathbf{1}_{\{i=a\}}.$$

正解候補では $p_a - 1 \leq 0$ 、他の候補では $p_i > 0$ となり、ロジットに関する勾配が求められます。

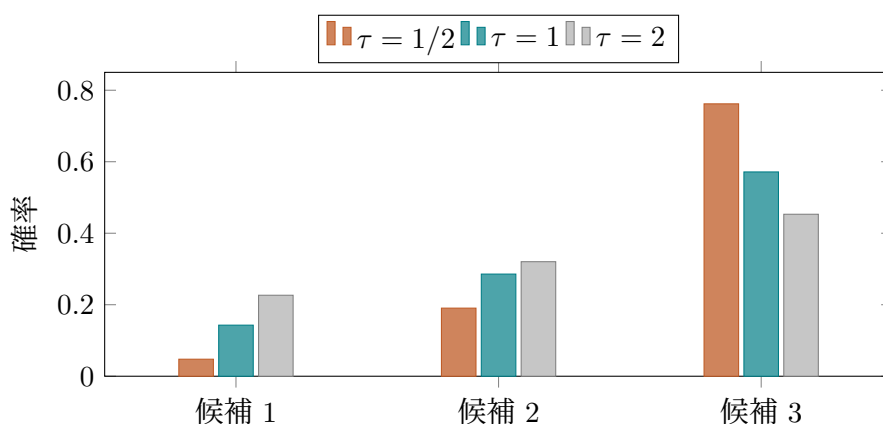
5 温度は順位を保ちながら、確率の差を変える

温度 $\tau > 0$ では、確率比は

$$\frac{p_i}{p_j} = \exp\left(\frac{z_i - z_j}{\tau}\right)$$

です。ロジット差が正なら、温度を下げるほど比が大きくなります。

$z = (0, \log 2, \log 4)$	候補 1	候補 2	候補 3
$\tau = 1/2$	$1/21$	$4/21$	$16/21$
$\tau = 1$	$1/7$	$2/7$	$4/7$
$\tau = 2$	$1/(3 + \sqrt{2})$	$\sqrt{2}/(3 + \sqrt{2})$	$2/(3 + \sqrt{2})$



同じロジットから計算した確率。温度を変えても最大の候補は変わらない。

- $\tau \rightarrow \infty$ では、各 $e^{z_i/\tau} \rightarrow 1$ なので一様分布 $p_i \rightarrow 1/n$ になります。
- $\tau \rightarrow 0^+$ では、最大ロジットが一つなら、その候補の確率が 1 に近づきます。同率最大が m 個なら、それぞれ $1/m$ に近づきます。

本文の練習問題：解答

問 1 指数を取ると $(1, 2, 3)$ 。確率は $(1/6, 1/3, 1/2)$ で、候補 3 が最大です。

問 2 $z = Wx = (0, \log 3)^\top$ 。 $\tau = 1$ では $(1/4, 3/4)$ 、 $\tau = 2$ では

$$\left(\frac{1}{1 + \sqrt{3}}, \frac{\sqrt{3}}{1 + \sqrt{3}} \right) \approx (0.366, 0.634).$$

共通の定数を足すと分子・分母に同じ因子 $e^{c/\tau}$ が掛かり、約分されます。

6 追加問題と解答

問 A：ベクトルの形を追う

$E \in \mathbb{R}^{100 \times 4}$ 、 $W \in \mathbb{R}^{6 \times 4}$ とします。 $h = \text{ReLU}(Wx + b)$ を作り、 $z = Ah + c$ で 100 候補のロジットを出すとき、 x, b, h, A, c, z の形を教えてください。

問 B：足し算と掛け算の違い

$z = (0, \log 2)$ について、 $\text{softmax}(z)$ 、 $\text{softmax}(z + 101)$ 、 $\text{softmax}(2z)$ を求めてください。

問 C：確率分布からロジットを作る

$p = (0.2, 0.3, 0.5)$ を出力するロジットを一組求めてください。また、その組が一意でない理由を説明してください。

問 D：非線形性を確認する

$f(x) = \text{ReLU}(x)$ として、 $f(1) + f(-1)$ と $f(1 + (-1))$ を比較し、加法性が成り立つか教えてください。

問 E：最大値選択と温度

有限のロジットに温度 $\tau > 0$ を使うとき、最大確率の候補を常を選ぶ方法では、温度によって選ぶ候補が変わるでしょうか。同率最大の扱いは固定するとします。

解答・考え方

A $x: 4 \times 1$ 、 $b, h: 6 \times 1$ 、 $A: 100 \times 6$ 、 $c, z: 100 \times 1$ 。

B 順に $(1/3, 2/3)$ 、 $(1/3, 2/3)$ 、 $(1/5, 4/5)$ 。共通の加算では変わりませんが、倍率を掛けるとロジット差が変わります。

C $z = (\log 0.2, \log 0.3, \log 0.5)$ 。指数の和は 1 なので、指定した分布になります。任意の共通定数を足しても同じ結果です。

D $1 + 0 = 1$ に対し $f(0) = 0$ 。加法性を満たさず、線形写像ではありません。

E 変わりません。正の数で割る操作と指数関数は順位を保つためです。温度が選択に影響するのは、主に分布に従ってサンプリングする場合です。

学習の要点：行列の形を確かめ、アフィン変換と非線形変換を区別し、ロジット差が確率比を決めることを理解する。