

『AI・データサイエンスのための数学入門』学習補助資料

損失からパラメータ更新まで

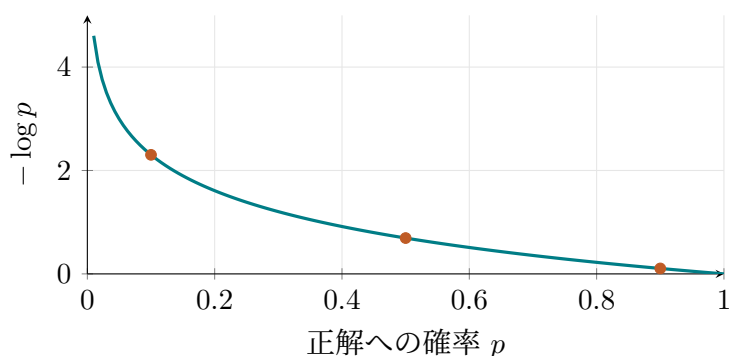
交差エントロピー・連鎖律・勾配降下法

1 損失は「正解に与えた確率」を評価する

ここでいう正解は、学習データに現れたトークンや与えられたラベルです。唯一可能な続きや、事実としての正しさを意味するとは限りません。以下では $\log$ を自然対数とします。

$$L = -\log p, \quad \frac{dL}{dp} = -\frac{1}{p} < 0.$$

正解への確率が増えるほど損失は小さくなります。正解への確率が 0 に近い予測には、大きな損失を与えます。



正解への確率	0.1	0.5	0.9
損失	2.303	0.693	0.105

正解率が同じでも、損失は違う

正解が第 1 クラスのとき、予測 (0.6, 0.4) と (0.9, 0.1) は、どちらも最大確率のクラスを選べば正解です。しかし損失はそれぞれ約 0.511 と 0.105 で異なります。交差エントロピーは、正誤だけでなく確率の割り当てでも評価します。

one-hot 表現で消える項

正解クラスを a とすると、 $y_a = 1$ 、それ以外は 0 です。したがって $-\sum_i y_i \log p_i = -\log p_a$ 。ただし、他クラスのロジットも softmax の分母を通じて p_a に影響します。

2 ロジットの勾配を、重みの勾配へ伝える

本文で求めた $g = \nabla_z L = p - y$ は、ロジットに関する勾配です。通常、学習で保存・更新するのはロジットそのものではなく、それを作る行列やバイアスです。

一般モデルで連鎖律を使う

$x \in \mathbb{R}^d$ 、 $W \in \mathbb{R}^{m \times d}$ 、 $b \in \mathbb{R}^m$ とし、

$$z = Wx + b, \quad z_i = \sum_{j=1}^d W_{ij}x_j + b_i$$

を考えます。入力 x を固定すると、

$$\frac{\partial L}{\partial W_{ij}} = \frac{\partial L}{\partial z_i} \frac{\partial z_i}{\partial W_{ij}} = g_i x_j, \quad \frac{\partial L}{\partial b_i} = g_i.$$

したがって、行列でまとめると

$$\boxed{\nabla_W L = gx^\top}, \quad \boxed{\nabla_b L = g}.$$

一方、前の層へ伝える入力の勾配は、各ロジットからの寄与を足して

$$\frac{\partial L}{\partial x_j} = \sum_i W_{ij} g_i, \quad \boxed{\nabla_x L = W^\top g}$$

です。 $gx^\top$ は $m \times d$ 行列で、 W と同じ形になります。

活性化関数を通る場合

$u = Bx + a$ 、 $h = \phi(u)$ とします。後ろの層から $\nabla_h L$ が届いたら、成分ごとに連鎖律を用いて

$$\nabla_u L = \nabla_h L \odot \phi'(u)$$

を求めます。 $\odot$ は成分ごとの積です。さらに $\nabla_B L = (\nabla_u L)x^\top$ として前へ進めます。ReLU では $u_j < 0$ で微分 0、 $u_j > 0$ で微分 1 です。 $u_j = 0$ では通常の意味で微分できないため、実装では規約を定めます。

逆伝播と更新は別の段階

逆伝播は勾配を計算する方法、勾配降下法は勾配を使ってパラメータを更新する方法です。まず同じ更新前のパラメータで全勾配を計算し、その後に更新します。

3 予測から更新までを 1 回、手で計算する

2 クラス分類で、正解を第 2 クラスとします。初期値を

$$x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad W = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \eta = 0.1$$

とします。これは計算を簡単にするための一段モデルです。

順方向：予測と損失

$$z = Wx + b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad p = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \quad L = \log 2 \approx 0.6931.$$

逆方向：勾配

$$g = p - y = \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}, \quad \nabla_W L = gx^\top = \begin{pmatrix} 1/2 & 1 \\ -1/2 & -1 \end{pmatrix}, \quad \nabla_b L = g.$$

重みとバイアスを同時に更新

$$W_{\text{new}} = W - 0.1 \nabla_W L = \begin{pmatrix} -.05 & -.10 \\ .05 & .10 \end{pmatrix}, \quad b_{\text{new}} = \begin{pmatrix} -.05 \\ .05 \end{pmatrix}.$$

同じ入力でもう一度予測

$$z_{\text{new}} = W_{\text{new}}x + b_{\text{new}} = \begin{pmatrix} -.30 \\ .30 \end{pmatrix}.$$

正解への確率と損失は

$$p_{2,\text{new}} = \frac{e^3}{e^{-.3} + e^3} = \frac{1}{1 + e^{-.6}} \approx 0.6457, \quad L_{\text{new}} \approx 0.4375.$$

正解への確率が 0.5 から約 0.646 へ増え、損失が下がりました。

ロジットを直接更新する場合との違い

もし z を独立なパラメータとして更新するなら、 $z_{\text{new}} = z - 0.1g = (-.05, .05)^\top$ です。上の結果と違うのは、実際には W, b を更新し、入力 x を通してロジットが変わるからです。一般の共有パラメータでは、全データの各ロジットが一律に望む向きへ動くとは限りません。

4 学習率が大きすぎると、損失は増える

微分可能な損失について、十分小さい変化 $\Delta\theta$ では

$$L(\theta + \Delta\theta) \approx L(\theta) + \nabla L(\theta)^\top \Delta\theta.$$

$\Delta\theta = -\eta \nabla L(\theta)$ を代入すると、

$$L(\theta - \eta \nabla L) \approx L(\theta) - \eta \|\nabla L\|^2.$$

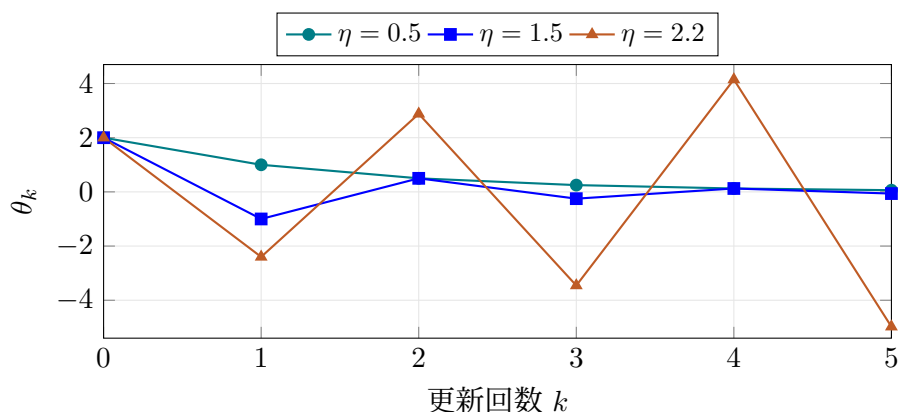
この局所的な近似が、勾配の反対向きへ進む理由です。大きな一歩でも損失が減るという保証はありません。

二次関数なら、条件を正確に計算できる

$L(\theta) = \theta^2/2$ なら $L'(\theta) = \theta$ なので、

$$\theta_{k+1} = (1 - \eta)\theta_k, \quad L(\theta_{k+1}) = (1 - \eta)^2 L(\theta_k).$$

従って $0 < \eta < 2$ なら最小点 0 へ収束します。 $\eta = 2$ では非零の初期値は符号を交互に変え、 $\eta > 2$ では絶対値が増えます。この境界はこの損失に固有の値です。



複数データでは、平均損失を用いる

B 個の例の平均損失 $\bar{L} = B^{-1} \sum_{b=1}^B L_b$ に対し、

$$\nabla_{\theta} \bar{L} = \frac{1}{B} \sum_{b=1}^B \nabla_{\theta} L_b.$$

各例の勾配が互いに打ち消し合うこともあります。一部のデータを使うミニバッチ更新では、各回に全データの損失が必ず下がるとは限りません。訓練損失と、学習に使わないデータでの性能も区別して確認します。

5 本文・章末総合問題の解答

本文の間 1：交差エントロピー

正解への確率は 0.5 です。損失は $-\log 0.5 = \log 2 \approx 0.6931$ 。 $-\log p$ は $p > 0$ で単調減少するので、正解への確率が大きくなるほど損失は小さくなります。

本文の間 2：ロジットの勾配

$$y = (0, 0, 1), \quad L = \log 2, \quad \nabla_z L = p - y = \left(\frac{1}{6}, \frac{1}{3}, -\frac{1}{2}\right).$$

成分の和は 0 です。これは softmax が、すべてのロジットへの共通定数の加算では変わらない性質とも対応します。

本文の間 3：パラメータ更新

$$\theta_{\text{new}} = 2 - 0.1 \times 0.5 = 1.95.$$

総合問題 1：一段モデルの予測

$$z = Wx + b = \begin{pmatrix} 1 \\ 2 + \log 2 \\ 3 \end{pmatrix}.$$

$D = e + 2e^2 + e^3$ とおくと、

$$p = \left(\frac{e}{D}, \frac{2e^2}{D}, \frac{e^3}{D}\right) = \frac{(e^{-2}, 2e^{-1}, 1)}{e^{-2} + 2e^{-1} + 1}.$$

最後の形は、全ロジットから最大値 3 を引いて計算した形です。

$$L = -\log p_3 = \log D - 3 = \log(1 + 2e^{-1} + e^{-2}) = 2\log(1 + e^{-1}) \approx 0.6265.$$

総合問題 2：attention の重み

各行の和は $2/3 + 1/3 + 0 = 1$ 、 $1/4 + 1/4 + 1/2 = 1$ です。第 1 行の出力は、第 3 位置を参照しません。value を v_1, v_2, v_3 と書けば、2 行の出力は

$$c_1 = \frac{2}{3}v_1 + \frac{1}{3}v_2, \quad c_2 = \frac{1}{4}v_1 + \frac{1}{4}v_2 + \frac{1}{2}v_3.$$

重み 0 の読み取りに注意

マスクなしの softmax は、有限のスコアに対して厳密には正の重みを返します。問題の 0 は、マスクなどで参照が除かれた設定として理解できます。ただし、この行列だけから、マスクの理由や query の位置までは決められません。

6 追加問題と学習の全体像

問 A：行列の勾配

$x = (2, -1)^\top$, $p = (0.8, 0.2)^\top$ 、正解が第 2 クラスとします。 $z = Wx + b$ に対する $\nabla_z L, \nabla_W L, \nabla_b L$ を求めてください。

問 B：大きな学習率

$L(\theta) = \theta^2/2$ 、 $\theta = 2$ から学習率 $\eta = 3$ で 1 回更新します。更新前後の損失を比較してください。

問 C：勾配の平均

ある一つのパラメータに対し、2 個のデータの勾配がそれぞれ 2, -1 でした。平均損失の勾配を求め、 $\theta = 1, \eta = 0.2$ から更新してください。

追加問題の解答

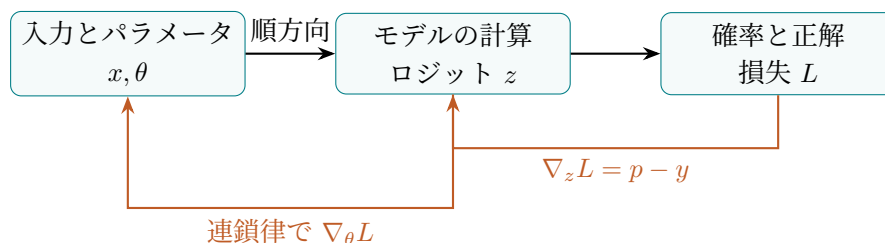
A $g = (0.8, -0.8)^\top$ より、

$$\nabla_W L = gx^\top = \begin{pmatrix} 1.6 & -0.8 \\ -1.6 & 0.8 \end{pmatrix}, \quad \nabla_b L = \begin{pmatrix} 0.8 \\ -0.8 \end{pmatrix}.$$

B $\theta_{\text{new}} = 2 - 3 \cdot 2 = -4$ 。損失は 2 から 8 へ増加します。

C 平均勾配は $(2 - 1)/2 = 0.5$ 。更新後は $1 - 0.2 \cdot 0.5 = 0.9$ です。

順方向と逆方向を対応させる



順方向で損失を計算し、逆方向で勾配を計算した後、 $\theta \leftarrow \theta - \eta \nabla_\theta L$ と更新します。通常の生成時はパラメータを固定し、確率分布からトークンを選びます。生成の計算を行うこと自体は、パラメータの学習ではありません。