

『AI・データサイエンスのための数学入門』学習補助資料

次トークンの確率から、学習と生成へ

言語モデルをつなぐ数学

1 モデルの出力は、まず「確率の一覧」

ここでは、生成 AI 全般のうち、文章を順に生成する**自己回帰型言語モデル**を扱います。画像生成などを含むすべての生成 AI が、同じ手順を使うわけではありません。

語彙に V 種類のトークンがあるとします。モデルのパラメータを θ 、それまでの文脈を c とすると、モデルは

$$f_{\theta}(c) = (p_1, \dots, p_V), \quad p_j \geq 0, \quad \sum_{j=1}^V p_j = 1$$

という確率ベクトルを返します。この分布からトークンを選び、文脈へ追加し、次の分布を計算します。

数学の各分野が担当するもの

段階	数学的な見方
トークン化	文字列を、モデルの語彙に属する ID 列に変換する。
埋め込み	ID に対応する学習済みベクトルを取り出す。
文脈の計算	位置の情報を含め、行列演算や非線形関数で表現を更新する。
確率の計算	候補ごとのスコアから softmax で確率分布を作る。
学習	観測されたトークンへの負の対数確率を損失にし、微分する。
生成	分布から選んだトークンを文脈に追加する。

「学習」と「生成」は別の操作

学習では訓練データを使って θ を更新します。通常の生成では θ を固定し、入力と生成済みトークンから出力を計算します。会話の文脈に合わせて応答が変わることと、モデルの重みが更新されることは別です。

「正解トークン」という呼び方について

本節の正解とは、**訓練・評価データに実際に現れたトークン**を指します。それが唯一自然な続きとは限りません。また、文章としてもっともらしいことと、内容が事実として正しいことは同じではありません。

2 連鎖律は、独立性を仮定していない

条件付き確率の定義 $P(A, B) = P(A)P(B | A)$ を繰り返すと、

$$P(w_1, w_2, w_3) = P(w_1)P(w_2 | w_1)P(w_3 | w_1, w_2)$$

を得ます。これは、トークンが独立だから掛け算するものではありません。過去への依存を条件付き確率の中に残したまま、同時確率を分解しています。

文脈が変わると、同じトークンの確率も変わる

「銀行で」の後と「川の」の後では、続く語の分布は異なります。各時点で同じ固定分布から抽選するモデルではなく、文脈ごとに分布を作り直すモデルです。

プロンプトが与えられている場合

入力文脈 c が既に与えられているとき、続きを $w_1, \dots, w_T$ とすれば、

$$P_\theta(w_1, \dots, w_T | c) = \prod_{t=1}^T P_\theta(w_t | c, w_1, \dots, w_{t-1}).$$

入力文脈そのものの確率を掛ける必要はありません。完全な終了文章の確率を表すには、通常、最後に終了トークンの確率も含めます。本文の式は、長さ T の列または接頭列の確率として読めます。

学習時の文脈と、生成時の文脈

訓練で次トークン損失を計算するときは、データ中の直前までの列を文脈として使います。生成するときは、モデル自身が選んだトークンを次の文脈へ入れます。

逐次生成と、学習計算の並列化

自己回帰的に生成する際は、次の文脈が選択結果に依存します。一方、Transformer の学習では、既知の訓練列に対し未来を見ないマスクを使い、多くの位置の予測を並列に計算できます。確率の逐次的な分解と、計算をすべて逐次実行することは別です [1]。

短い確認問題

$P(A) = 0.6$ 、 $P(B | A) = 0.2$ なら $P(A, B) = 0.12$ です。 $P(B)$ の値が分からなくても計算できます。ここで $P(B | A)$ を勝手に $P(B)$ へ置き換えてはいけません。

3 損失は、確率の積を足し算に変える

観測されたトークンへの確率を q_t と略記します。 $q_t > 0$ として、

$$L = -\log \prod_{t=1}^T q_t = -\sum_{t=1}^T \log q_t.$$

対数は単調増加なので、同じ列の確率を最大にすることと L を最小にすることは同値です。本ノートの $\log$ は自然対数です。

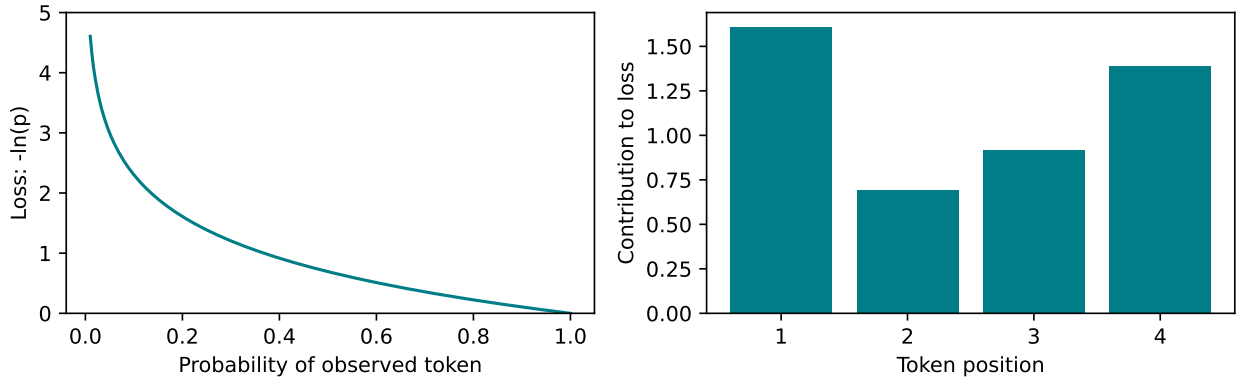


図 1：左は観測トークンへの確率と損失。右は本文例 (0.2, 0.5, 0.4, 0.25) の各項の損失です。小さな確率を与えた箇所ほど大きな損失になります。

なぜ「確信した間違い」を強く罰するのか

$q = 0.5, 0.1, 0.01$ に対する損失は、それぞれ約 0.693, 2.303, 4.605 です。 $q \rightarrow 0$ では $-\log q \rightarrow \infty$ となります。逆に $q = 1$ なら損失は 0 です。

交差エントロピーとの一致

観測トークンが候補 j であることを、 $a_j = 1$ 、他は $a_i = 0$ という one-hot ベクトルで表すと、

$$\ell = -\sum_{i=1}^V a_i \log p_i = -\log p_j.$$

これが 1 時点の交差エントロピー損失です。列では各時点の損失を加算します。

微分が確率の改善につながる仕組み【補足】

ロジット z_i に対する softmax を $p_i = e^{z_i} / \sum_j e^{z_j}$ とすると、

$$\ell = -z_j + \log \sum_i e^{z_i}, \quad \frac{\partial \ell}{\partial z_i} = p_i - a_i.$$

観測トークンでは $p_j - 1 \leq 0$ 、他では $p_i \geq 0$ です。ロジットを直接勾配降下で動かせば、観測トークンのスコアを上げ、他を下げる向きになります。実際には連鎖律で $\nabla_{\theta} L$ を計算し、 $\theta \leftarrow \theta - \eta \nabla_{\theta} L$ と更新します。

4 パープレキシティは幾何平均の逆数

定義を変形すると、

$$\text{PPL} = \exp \left(-\frac{1}{T} \sum_t \log q_t \right) = \left(\prod_t q_t \right)^{-1/T}.$$

したがって、PPL は観測トークンへの確率の**幾何平均の逆数**です。算術平均の逆数ではありません。

「候補数」の感覚をつかむ

毎回 m 候補に等確率 $1/m$ を与え、その中の 1 つが観測されたなら、 $\bar{L} = \log m$ 、 $\text{PPL} = m$ です。この意味で、PPL は実効的な候補数とみなせます。ただし、一般の分布で実際の候補がちょうど PPL 個あるわけではありません。

各時点の観測トークン確率	平均損失	PPL
常に 1/2	$\log 2$	2
常に 1/4	$\log 4$	4
常に 1/10	$\log 10$	10

PPL は 1 以上で、観測トークン確率がすべて 1 なら 1 です。非常に低い確率を観測列へ与えると、語彙数より大きくなる場合もあります。

比較するときにそろえる条件

同じ評価データ、同じトークン化、同じ評価範囲・文脈条件で比較します。トークンの切り方が違えば T も変わるため、PPL をそのまま比べることはできません。また、訓練データへの当てはまりだけでなく、未使用データで評価します。

低い PPL は、万能な品質保証ではない

PPL は評価列への平均的な確率の高さを測ります。事実の正確さ、推論の妥当性、指示への適合性などを単独で保証する指標ではありません。「生成した文章が正しい確率」とも異なります。

数値計算では積を作ってから対数を取らない

長い列の確率の積は、計算機では 0 に丸められることがあります。各時点の対数確率を足し、softmax も最大ロジットを引くなどして安定に計算します。対数の底を 2 にする場合は、PPL も $2^{\bar{L}}$ とし、底を対応させます。

5 確率を作ることと、トークンを選ぶこと

最大値選択は列全体の最大化とは限らない

長さ 2 の架空の列で、最初の候補を A、B、次の候補を X、Y とします。

最初の候補	最初の確率	次の条件	次の確率	列の確率
A	0.6	X A	0.5	AX : 0.30
A	0.6	Y A	0.5	AY : 0.30
B	0.4	X B	0.9	BX : 0.36
B	0.4	Y B	0.1	BY : 0.04

各段階で最大を選ぶと、まず A を選び、AX または AY となります。しかし、列確率が最大なのは BX です。目先の最大値を選ぶ**貪欲法**は、列全体の最大確率を保証しません。

サンプリングでは、同じ文脈でも選択が変わる

本文の天気の場合なら、独立に 100 回選ぶと「晴れ」の選択回数の期待値は 45 回ですが、毎回ちょうど 45 回になるわけではありません。これはモデル出力を用いた選択の例であり、実際の天気の予報確率とは区別します。

本文・問 1 の解答

$$P = 0.5 \cdot 0.4 \cdot 0.25 = \boxed{0.05}, \quad L = -\log 0.05 = \log 20 \approx \boxed{2.996},$$

$$\bar{L} \approx 0.999, \quad \text{PPL} = 20^{1/3} \approx \boxed{2.714}.$$

本文・問 2 の解答

$$P_A = (0.5 \cdot 0.25)^2 = \frac{1}{64}, \quad \text{PPL}_A = 64^{1/4} = 2\sqrt{2} \approx 2.828.$$

$$P_B = 0.4^4 = 0.0256, \quad \text{PPL}_B = 1/0.4 = 2.5.$$

$P_B > P_A$ なので、この列にはモデル B の方が高い確率を与えます。この 1 列だけで、すべての文章に対して B が優れているとは結論できません。

6 追加問題と解答

問 1：同じ平均確率 2 トークンに対してモデル A は (0.9, 0.1)、B は (0.5, 0.5) を観測トークンに与えた。算術平均と PPL を比較せよ。

問 2：損失の改善 平均損失が $\log 8$ から $\log 4$ へ減った。PPL はどう変わるか。

問 3：一様予測 毎回 100 候補すべてに同じ確率を与えると、1 トークン当たりの損失と PPL を求めよ。

問 4：softmax と微分 ロジットが $(\log 2, 0, 0)$ で、観測トークンは第 2 候補とする。確率ベクトル、損失、ロジットに関する勾配を求めよ。

問 5：無関係な条件 $P(A) = 0.4$ 、 $P(B | A) = 0.8$ 、 $P(C | A, B) = 0.5$ から列 ABC の確率を求めよ。独立性は必要か。

問 6：指標の解釈 「PPL が 3 なら、生成文が正しい確率は $1/3$ 」を修正せよ。

解答・解説

- 算術平均はどちらも 0.5。PPL は A で $(0.09)^{-1/2} = 10/3 \approx 3.333$ 、B で $(0.25)^{-1/2} = 2$ です。低確率を与えた箇所が積と損失に影響します。
- $8 \rightarrow 4$ で半分になります。平均損失が $\log 2$ 減ると、PPL は $e^{-\log 2} = 1/2$ 倍になります。
- 確率は 0.01、損失は $\log 100 \approx 4.605$ 、PPL は 100 です。
- 指数を取ると $(2, 1, 1)$ なので $p = (1/2, 1/4, 1/4)$ 。損失は $\log 4$ 、勾配は $p - (0, 1, 0) = (1/2, -3/4, 1/4)$ です。
- $0.4 \cdot 0.8 \cdot 0.5 = 0.16$ 。条件付き確率の連鎖律なので、独立性は必要ありません。
- 観測トークンへの確率の幾何平均が $1/3$ という意味です。文章内容の正しさの確率ではありません。

講義で押さえること

確率の連鎖律で積を作り、対数で損失の和に変える。平均損失の指数が PPL。学習では重みを更新し、生成では分布から選ぶ。もっともらしさと事実の正しさを区別する。

参考[1]A. Vaswani et al., *Attention Is All You Need* (2017), 特に 3.1 節。<https://arxiv.org/abs/1706.03762>。
本ノートの数値例は説明用の架空例です。