名古屋工業大学 先進セラミックス研究センター　井田 隆　　2023 年 2 月 2 日更新

１６．損失函数　Loss Function

ニューラルネットワーク（人工神経網）の動作の仕方は「重みパラメータ」によって決まります。ニューラルネットワークに何か理想的な（好ましい）判断・動作をさせようとする場合に，「理想的な判断からのずれ」を数値化して，その数値がなるべく小さくなるように「重みパラメータ」を変化させる自動計算の仕組みを作ります。「重みパラメータ」を変数 variable としてこの「ずれの大きさ」を数値化する函数を**損失函数** loss function と呼びます。

損失函数を最小化することは，従来の数値計算の分野で**最適化** optimization と呼ばれていたことと同じことです。最適化の方法には，いろいろなタイプのものがありました（計算科学基礎，第 8 章，第 9 章）。一方で機械学習では**逆伝播法** backpropagation と呼ばれる最適化手法が比較的多く用いられるようです。逆伝播法はパラメータを変化させたときの損失函数の変化（勾配 gradient）を求め，最も勾配の降下の大きい方向にパラメータを一定の割合で変化させるもので，ニュートン法と少し似た性格を持ちます。パラメータを変化させる割合は**学習率** learning rate と呼ばれ，あらかじめユーザが決めなければいけない**ハイパー・パラメータ** hyper-parameter と呼ばれます。

数値計算の分野で最適化の対象となる函数にはノルム norm やモード mode がありますが，機械学習の分野で用いられる損失函数としては L_2（エル・ツー）ノルムに相当する「偏差の平方和 sum of squared deviation（自乗和誤差，残差平方和 residual sum of squares）」と「交差エントロピー cross entropy」とが代表的なものです。

１６−１　偏差の平方和　Sum of Squared Deviation

損失函数として偏差（ずれ）の平方の和を選び，それを最小にしようとする考え方は，多くの場合に合理的でもあり，受け入れられやすいでしょう。

回帰問題で，訓練データセットが N 組あり，そのうち j 番目のデータセットが，データ（例題） $\mathbf{x}_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ と正解（**解答例**，supervisory signal） $\mathbf{z}_j = \{z_{j1}, z_{j2}, \cdots, z_{jn}\}$ の組として表されるとします。

ニューラル・ネットワークの出力 $\mathbf{y}$ が，入力データに対応する変数 $\mathbf{x}$ と重み $\mathbf{W}$ の函数として

$$\mathbf{y} = f(\mathbf{x}; \mathbf{W}) \tag{16.1.1}$$

と表されるとします。一組のデータ $(\mathbf{x}_j, \mathbf{z}_j)$ に対して，$\mathbf{y}_j = f\left(\mathbf{x}_j; \mathbf{W}\right)$ として，$\left|\mathbf{z}_j - \mathbf{y}_j\right|$ を偏差と呼び，

$$S(\mathbf{W}) = \sum_{j=1}^{N} \left|\mathbf{z}_j - \mathbf{y}_j\right|^2 = \sum_{j=1}^{N} \left|\mathbf{z}_j - f(\mathbf{x}_j;\, \mathbf{W})\right|^2 \tag{16.1.2}$$

を「偏差の平方和」と呼びます。

偏差の平方和を最小にするような重み $\mathbf{W}$ を求めれば良いと考えます。この考え方は**最小自乗法（最小平方法）** least-squares method （第 10 章）と呼ばれる方法を使う考え方と同じことです。

分類問題の場合に，訓練データセットが N 組あり，そのうち j 番目のデータセットが，データ（例題） $\mathbf{x}_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ と，選択肢$\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \dots\}$ のうちのどれかの正解の組として表されるとします。A が正解の場合 $\mathbf{z}_j = \{1, 0, 0, \cdots, 0\}$，B が正解の場合 $\mathbf{z}_j = \{0, 1, 0, \cdots, 0\}$ のように，正解を 1，それ以外の選択肢を 0 とする表現を使います。このような表現のしかたは **one-hot 表現**と呼ばれます。

One-hot 表現を用いれば，分類問題であっても損失函数として回帰問題と同じ式 (16.1.2) の偏差平方和をそのまま使えます。ニューラル・ネットワークの出力層にソフトマックス函数を使うことにすれば，データ $\mathbf{x}_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ に対する出力 $\mathbf{y}_j = \{y_{j1}, y_{j2}, \cdots, y_{jn}\}$ は，選択肢$\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \dots\}$ のそれぞれが正解であるとみなす確率を表す値と解釈することもできます。

分類問題では，偏差の平方和：

$$S(\mathbf{W}) = \sum_{j=1}^{N} \left|\mathbf{z}_j - \mathbf{y}_j\right|^2 = \sum_{j=1}^{N} \left|\mathbf{z}_j - f(\mathbf{x}_j;\, \mathbf{W})\right|^2 \tag{16.1.2}$$

は，正しい選択肢を選択する確率が高く，誤った選択肢を選択する確率が低いほど小さい値になります。

１６−２　交差エントロピー　Cross Entropy

機械学習の分野では損失函数として**交差エントロピー** cross entropy （補足 16.2.A）の用いられる場合があります。

j 番目の訓練データセットのデータ（例題） $\mathbf{x}_j = \{x_{j1}, x_{j2}, \cdots, x_{jm}\}$ に対する解答例が $\mathbf{z}_j = \{z_{j1}, z_{j2}, \cdots, z_{jn}\}$ と表され，ニューラル・ネットワークの出力が $\mathbf{y}_j = \{y_{j1}, y_{j2}, \cdots, y_{jn}\}$ と表されるとして，交差エントロピーは

$$E = -\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{n} z_{jk} \log_2 y_{jk} \tag{16.2.4}$$

と表されます。交差エントロピーは２つの分布 $\mathbf{y}_j = \{y_{j1}, y_{j2}, \cdots, y_{jn}\}$ と $\mathbf{z}_j = \{z_{j1}, z_{j2}, \cdots, z_{jn}\}$ とが「近い分布」とみなせる場合に小さい値をとる性質を持ちます。

（補足 16.2.A）交差エントロピーに関連すること（↩）

確率空間 probability space $\mathcal{X}$ （$\mathcal{X}$ は X のカリグラフ字体）で定義される**離散確率分布**について，確率変数が x という値をとる確率を $P(x)$ と表せば，**シャノンの情報エントロピー** Shannon entropy は

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\,\log_2 P(x) \tag{16.2.A.1}$$

と表されます。シャノンの情報エントロピーは「**情報量**」と呼ばれることもあります。熱力学的なエントロピーとの関係については，あまりはっきりとしていないようですが，近いものであるとする解釈が有力です。（補足 16.2.A.1）

２つの確率分布の違いの大きさの尺度として，**カルバック・ライブラー発散** Kullback-Leibler divergence（カルバック ライブラー ダイヴァージェンス）が定義されます。情報発散（インフォーメイション） information divergence，情報利得（ゲイン） information gain，相対エントロピー（レラティヴ） relative entropy とも呼ばれます。

同じ確率空間 probability space $\mathcal{X}$ で定義される**離散確率分布** P，Q について，P と Q のカルバック・ライブラー発散は

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x\in\mathcal{X}} P(x)\,\log_2 \frac{Q(x)}{P(x)} = \sum_{x\in\mathcal{X}} P(x)\,\log_2 \frac{P(x)}{Q(x)} \tag{16.2.A.2}$$

と表されます。この値は，$P(x)$ を確率分布として，２つの確率分布の対数の差

$$\log_2 P(x) - \log_2 Q(x) = \log_2 \frac{P(x)}{Q(x)} \tag{16.2.A.3}$$

の「期待値」を計算した値と見ることもできます。カルバック・ライブラー発散は $Q(x) = 0$ となるようなすべての x について $P(x) = 0$ となる（絶対連続の）場合にだけ定義されます。なお，

$$\lim_{x\to+0} x\,\log_2 x = 0 \tag{16.2.A.4}$$

の関係から，式 (16.2.A.2) の中で $P(x) = 0$ となる項はすべてゼロと解釈されます。

連続確率分布 P，Q の確率密度函数が $p(x)$，$q(x)$ と表されるとき，カルバック・ライブラー発散は積分形式で定義され，

$$D_{\mathrm{KL}}(P \parallel Q) \equiv \int_{-\infty}^{\infty} p(x)\,\log_2 \frac{p(x)}{q(x)}\,\mathrm{d}x \tag{16.2.A.5}$$

と書けます。

（補足 16.2.A.1）熱力学的なエントロピー（↩）

Boltzmann（ボルツマン）によれば，熱力学的な**エントロピー**は，微視的な状態の取りうる場合の数（ヴァーシャインリヒカイト） (Wahrscheinlichkeit) W と Boltzmann（ボルツマン） 定数 $k = 1.380\,649 \times 10^{-23}$ J/K （国際単位系 SI では 2019 年に「定義値」とされた）に対して

$$S = k \ln W \tag{16.2.A.1.1}$$

と表されます。

「微視的な状態の取りうる場合の数」とは，例えば「単原子分子理想気体」「自由粒子」の場合には，それぞれの粒子が「（位置）空間の中でとりうる位置」と「速度空間でとりうる速度」（あるいは「運動量空間でとりうる運動量」）を意味します。一つの粒子については 6 次元の位相空間の中での位置をあらわし，N 個の粒子があれば $6N$ 次元の位相空間での位置を表すと考えます。

単原子分子理想気体の持つ内部エネルギーは，粒子の運動エネルギーの総和と同じで，粒子数に比例します。

等しい温度を保ちながら単原子分子理想気体を封入した容器の体積を半分に（等温圧縮）するとします。容器の器壁の単位面積・単位時間あたりに粒子が衝突する頻度は体積に比例するはずなので，圧力は 2 倍になります。このことはボイル Boyle の法則として知られています。このときに気体は圧力を受けながら体積を縮めているので，外部から力学的なエネルギー（仕事）を受けていることになります。この過程では温度は変化せず気体の内部エネルギーは変化しません。外部から受けた力学的なエネルギーはすべて熱として放出します。体積が半分になれば，粒子のとりうる位置の場合の数が半分になるので，エントロピーの変化 ΔS は未知の係数 β を使って $\Delta S = N\beta \ln \frac{1}{2} = -N\beta \ln 2$ のように表されます。

断熱状態で気体を圧縮すれば，外部から力学的なエネルギー（仕事）を受けますが，そのエネルギーは熱として放出されず，気体分子の運動速度を高くし，内部エネルギーを大きくするために使われます。速度空間でとりうる速度の場合の数が増えた分と，圧縮によって位置空間でとりうる場合の数の減少する分が打ち消し合い，エントロピーは変化しません。

(↩)