交差エントロピー J_{MLE} の導出方法をはっきりさせておきたい。まず多くの場合、多クラス分類において、$\boldsymbol{y}$ は $\left[y^{(1)},\ldots,y^{(m')}\right]^\top$ のような列ベクトルである。各成分 $y^{(i)}$ はクラス $k=0,\ldots,K$ のうち、データ $\boldsymbol{x}^{(i)}$ が属するクラスを表しているものとする。例えば $\boldsymbol{x}^{(i)}$ が手書き数字の「6」の画像のデータであるならば、$K=9,\ y^{(i)}=6$ である。$U^{(2)}$ の正体についてよく考え直したところ、$\boldsymbol{y}$ の推定ではなく、次のような行列になるだろうという結論に至った。

$$U^{(2)}=\begin{bmatrix} u_0^{(1)} & \cdots & u_K^{(1)} \\ \vdots & & \vdots \\ u_0^{(m')} & \cdots & u_K^{(m')} \end{bmatrix}=\begin{bmatrix} {\boldsymbol{u}^{(1)}}^\top \\ \vdots \\ {\boldsymbol{u}^{(m')}}^\top \end{bmatrix},\qquad \boldsymbol{u}^{(i)}=\begin{bmatrix} u_0^{(i)} \\ \vdots \\ u_K^{(i)} \end{bmatrix},\qquad i=1,\ldots,m'. \tag{1}$$

このベクトル $\boldsymbol{u}^{(i)}$ が、教科書の式 6.28 の $\boldsymbol{z}=\boldsymbol{W}^\top\boldsymbol{h}+\boldsymbol{b}$ に対応している。

なぜ $U^{(2)}$ が式 (1) のような行列になるかを計算によって説明する。簡単のため手書き数字認識の例を用いる。

$$X=\begin{bmatrix} x_1^{(1)} & \cdots & x_{784}^{(1)} \\ \vdots & & \vdots \\ x_1^{(m')} & \cdots & x_{784}^{(m')} \end{bmatrix}=\begin{bmatrix} {\boldsymbol{x}^{(1)}}^\top \\ \vdots \\ {\boldsymbol{x}^{(m')}}^\top \end{bmatrix}.$$

隠れ層の中にノードが n 個あると仮定すれば、$W^{(1)}=\left[\boldsymbol{w}_1^{(1)},\ldots,\boldsymbol{w}_n^{(1)}\right]$ のように、パラメータ $W^{(1)}$ は n 個の列ベクトル $\boldsymbol{w}_l^{(1)},\ l=1,\ldots,n$ を並べて作ることができる。よって $U^{(1)}$ を具体的に計算すると

$$U^{(1)}=XW^{(1)}=\begin{bmatrix} {\boldsymbol{x}^{(1)}}^\top \\ \vdots \\ {\boldsymbol{x}^{(m')}}^\top \end{bmatrix}\left[\boldsymbol{w}_1^{(1)},\ldots,\boldsymbol{w}_n^{(1)}\right]=\begin{bmatrix} {\boldsymbol{x}^{(1)}}^\top\boldsymbol{w}_1^{(1)} & \cdots & {\boldsymbol{x}^{(1)}}^\top\boldsymbol{w}_n^{(1)} \\ \vdots & & \vdots \\ {\boldsymbol{x}^{(m')}}^\top\boldsymbol{w}_1^{(1)} & \cdots & {\boldsymbol{x}^{(m')}}^\top\boldsymbol{w}_n^{(1)} \end{bmatrix}$$

となる。すなわち $U^{(1)}$ は $m'\times n$ の行列で、その成分は

$$U_{ij}^{(1)}={\boldsymbol{x}^{(i)}}^\top\boldsymbol{w}_j^{(1)},\qquad i=1,\ldots,m',\qquad j=1,\ldots,n$$

である。次に各成分に対する活性化関数の値を並べた行列 H を計算するが、このサイズは $U^{(1)}$ と同じで $m'\times n$ である。H を m' 個の行ベクトル ${\boldsymbol{h}^{(i)}}^\top=\left[\varphi\left({\boldsymbol{x}^{(i)}}^\top\boldsymbol{w}_1^{(1)}\right),\ldots,\varphi\left({\boldsymbol{x}^{(i)}}^\top\boldsymbol{w}_n^{(1)}\right)\right],\ i=1,\ldots,m'$ を縦に並べたものと解釈する。またパラメータ $W^{(2)}$ は、クラスについての列ベクトル $\boldsymbol{w}_k^{(2)},\ k=0,\ldots,K$ を並べたものとすれば、行列積 $U^{(2)}=HW^{(2)}$ は列ベクトルである必要はない。

$$H=\begin{bmatrix} {\boldsymbol{h}^{(1)}}^\top \\ \vdots \\ {\boldsymbol{h}^{(m')}}^\top \end{bmatrix},\qquad W^{(2)}=\left[\boldsymbol{w}_0^{(2)},\ldots,\boldsymbol{w}_K^{(2)}\right]$$

であるから、

$$U^{(2)}=HW^{(2)}=\begin{bmatrix} {\boldsymbol{h}^{(1)}}^\top \\ \vdots \\ {\boldsymbol{h}^{(m')}}^\top \end{bmatrix}\left[\boldsymbol{w}_0^{(2)},\ldots,\boldsymbol{w}_K^{(2)}\right]=\begin{bmatrix} {\boldsymbol{h}^{(1)}}^\top\boldsymbol{w}_0^{(2)} & \cdots & {\boldsymbol{h}^{(1)}}^\top\boldsymbol{w}_K^{(2)} \\ \vdots & & \vdots \\ {\boldsymbol{h}^{(m')}}^\top\boldsymbol{w}_0^{(2)} & \cdots & {\boldsymbol{h}^{(m')}}^\top\boldsymbol{w}_K^{(2)} \end{bmatrix}$$

となる。すなわち $U^{(2)}$ は $m' \times (K+1)$ の行列で、その成分は

$$U_{ik}^{(2)} = \boldsymbol{h}^{(i)\top} \boldsymbol{w}_k^{(2)}, \qquad i = 1, \ldots, m', \qquad k = 0, \ldots, K$$

である。$U^{(2)}$ の成分を横方向に見ると、ミニバッチの各データに関する情報が並んでいて、縦方向に見ると各クラスのスコアが並んでいるような構造をしている。それを表して書き直したのが式 (1) である。

さて式 (1) の行列を用いて、交差エントロピーがどのように作られるのかを思い出したい。6.2.2.3 節「マルチヌーイ出力分布のためのソフトマックスユニット」の復習である。

i 番目のデータがクラス k に属する確率は、ソフトマックス関数を用いて次のように与えられるのだった。

$$p_k^{(i)} = \mathrm{softmax}\left(\boldsymbol{u}^{(i)}\right)_k = \frac{\exp\left(u_k^{(i)}\right)}{\sum_{k=0}^{K} \exp\left(u_k^{(i)}\right)}. \tag{2}$$

これは規格化されているので、あえてチルダ記号はつけていない。記号を少し整理しておく。

- $u_k^{(i)}$：i 番目のデータに対するクラス k の「スコア」
- $\boldsymbol{u}^{(i)}$：i 番目のデータに対する各クラスのスコアを格納したベクトル
- $p_k^{(i)}$：各クラスのスコアから推計される、i 番目のデータがクラス k に属する確率

ちなみに式 (2) の対数をとることで、行列 $U^{(2)}$ の各成分 $u_k^{(i)}$ が規格化されていない対数尤度になっていることを示すことができる。

$$u_k^{(i)} = \log p_k^{(i)} + \log\left(\sum_{k=0}^{K} \exp\left(u_k^{(i)}\right)\right) = \log \tilde{p}_k^{(i)}. \tag{3}$$

訓練データのラベル $y^{(i)}$ が k であるとすると、その確率は

$$\mathbf{1}_{k=y^{(i)}} = \begin{cases} 1, & k = y^{(i)}, \\ 0, & k \neq y^{(i)} \end{cases}$$

と表せる。これと式 (2) の推計された確率を用いれば、交差エントロピーのコスト関数は次のように作ることができる。

$$J_{\mathrm{MLE}} = -\frac{1}{m'} \sum_{i=1}^{m'} \sum_{k=0}^{K} \mathbf{1}_{k=y^{(i)}} \log p_k^{(i)}. \tag{4}$$

なお式 (3) の規格化されていない対数尤度 $u_k^{(i)} = \log \tilde{p}_k^{(i)}$ をそのまま交差エントロピーに用いると、式 (4) によるものとは異なる結果が得られると考えられる。なぜならば式 (3) の定数に思える部分 $\sum_{k=0}^{K} \exp\left(u_k^{(i)}\right)$ は実際には訓練データ i とパラメータ $W^{(1)}$, $W^{(2)}$ の関数であり、$u_k^{(i)} = \log \tilde{p}_k^{(i)}$ を用いた交差エントロピーは式 (4) のそれに $W^{(1)}$, $W^{(2)}$ の関数を付け足したものになるからである。

参考文献

[1] Aurélien Géron, 長尾高弘訳.「scikit–learn と TensorFlow による実践機械学習」, 株式会社オライリージャパン，2018 年.