き確率、$p(\boldsymbol{x}_j, \boldsymbol{y}_i)$ は統合確率 (同時確率) を表しています。最終的に得られた量は、分布 $p(\boldsymbol{x}, \boldsymbol{y})$ の下での $-\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ の期待値となることが分かります。これを次のように表現することにします。

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\Big[-\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\Big] = \begin{cases} \displaystyle\sum_i \sum_j \Big[-p(\boldsymbol{x}_j, \boldsymbol{y}_i) \log p_{\mathrm{model}}(\boldsymbol{y}_i|\boldsymbol{x}_j;\boldsymbol{\theta})\Big] & \text{(離散確率分布)} \\ \displaystyle\int\int \Big[-p(\boldsymbol{x}, \boldsymbol{y}) \log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\Big]\, d\boldsymbol{x}d\boldsymbol{y} & \text{(連続確率分布)} \end{cases}$$

連続分布の場合は、$\sum$ が $\int$ になるだけです。この量は厳密に言えば「交差エントロピーの統合確率分布に関する期待値」なのですが、一般にはこの量のことを単に「交差エントロピー」と呼ぶことになっています。

※ ここでは少し教科書 “Deep Learning (2016)” とは異なる記法を用いているので、その理由についてコメントしておきます。 教科書の式は、全体的に確率分布の引数が省略されてしまっていて、

$$\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\big[\cdots\big], \quad \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\big[\cdots\big], \quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\big[\cdots\big]$$

は全て、

$$\mathbb{E}_{\boldsymbol{y}\sim p}\big[\cdots\big], \quad \mathbb{E}_{\boldsymbol{x}\sim p}\big[\cdots\big], \quad \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p}\big[\cdots\big],$$

と表記されています。つまり、統合確率、周辺確率、条件付き確率をハッキリとは書き分けておらず、慣れている人からしたらスッキリしていて見やすいですし意味も分かるのですが 、初めて勉強する人には分かりづらい書き方になっていると思います。この解説プリントでは、きちんとこれらを区別して書くことにしました。

♣ 損失関数としての交差エントロピー (2)

ここまでで、交差エントロピー

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})}\Big[-\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\Big]$$

が、真の確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ とモデル $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ の離れ具合を表す量となっていることを説明しました。ただし、この量をそのまま損失関数として用いることはできません。なぜなら、上の量を具体的に計算するためには、真の確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ やそれに対応する統合分布 $p(\boldsymbol{x}, \boldsymbol{y})$ を知っている必要があるからです。しかし、そもそもそれらを知らないからモデルを苦労して作るわけで、$p(\boldsymbol{y}|\boldsymbol{x})$ や $p(\boldsymbol{x}, \boldsymbol{y})$ は知っているはずがありません (知っているならモデルを作る必要がありません)。

では、どうしたらよいのでしょうか？それには、既に実際の測定などにより得られた手元にあるデータに基づいて $p(\boldsymbol{x}, \boldsymbol{y})$ を近似することになります。この近似には様々な手法がありますが、1つの方法として、データから得られる**「経験分布 (empirical distribution)」** $\hat{p}_{\text{data}}(\boldsymbol{x}, \boldsymbol{y})$ を使用する方法があります。つまり、$\hat{p}_{\text{data}}(\boldsymbol{x}, \boldsymbol{y})$ が真の確率分布 $p(\boldsymbol{x}, \boldsymbol{y})$ に近いはずと考え、$p_{\text{model}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})$ を $\hat{p}_{\text{data}}(\boldsymbol{x}, \boldsymbol{y})$ に近づけることで学習を行います。つまり、この2つの分布の交差エントロピー (の期待値)[12]

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\hat{p}_{\text{data}}(\boldsymbol{x},\boldsymbol{y})}\Big[-\log p_{\text{model}}(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta})\Big]$$

が、実際には損失関数として用いられます[13]。現在の多くのニューラルネットワークにおいて、この交差エントロピー誤差が損失関数として用いられています。

ここで出てきた経験関数は、実際に観測されたデータから得られる頻度割合の分布です。例えば、30人のクラスで1週間に読む本の数 x を調べたところ、1冊が15人、2冊が10人、3冊が5人となったとしましょう。経験分布は、それぞれの値の全データ数に対する割合として与えられます。つまり、

$$\hat{p}_{\text{data}}(x=1) = \frac{15}{30} = \frac{1}{2}, \quad \hat{p}_{\text{data}}(x=2) = \frac{10}{30} = \frac{1}{3}, \quad \hat{p}_{\text{data}}(x=3) = \frac{5}{30} = \frac{1}{6}$$

となります。このようにすると、この経験分布 $\hat{p}_{\text{data}}(x)$ は確率の基本ルール「全て足したら1になる」

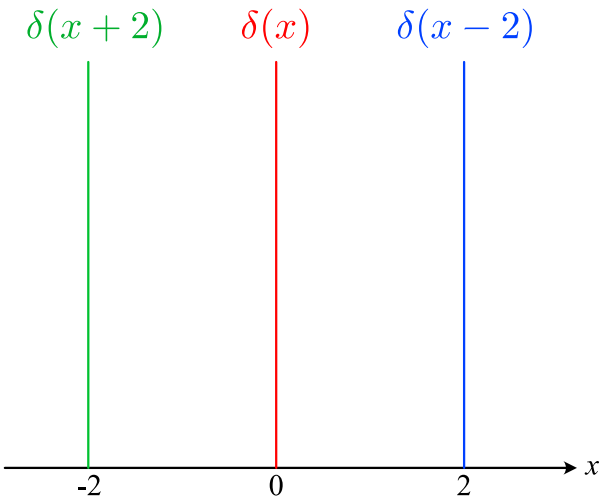
$$\sum_i \hat{p}_{\text{data}}(x_i) = 1$$

を満たすようになり、確率として扱うことができるようになります。教科書 “Deep Learning (2016)”では、経験分布をもう少し数学的に次のように定義しています。

$$\hat{p}_{\text{data}}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

ここで出てきた $\delta(\boldsymbol{x})$ は「デルタ関数」と呼ばれる特殊な関数です[14]。$\delta(x-\mu)$ のグラフ (のイメージ) を描くと次のようになります。

[12] 教科書 “Deep Learning (2016)”の (6.12) 式です。
[13] 一般に経験分布 $\hat{p}_{\text{data}}$ を使って得られる期待値のことを「経験リスク (empirical risk)」といいます。
[14] 「超関数」と呼ばれる関数の1つです



つまり、$\delta(\cdots)$ は、かっこの中が 0 となる時にピークを持ち、それ以外の値の時には 0 となるような関数です。ただしこれはあくまでイメージで、むしろ次の式をデルタ関数の定義として理解する方が正確です。

$$\sum_i f(\boldsymbol{x}_i)\delta(\boldsymbol{x}_i - \boldsymbol{\mu}) = f(\boldsymbol{\mu}) \qquad (\boldsymbol{x} \text{ は離散的な変数})$$

$$\int f(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{\mu})d\boldsymbol{x} = f(\boldsymbol{\mu}) \qquad (\boldsymbol{x} \text{ は連続的な変数})$$

$\sum$ や $\int$ によって、変数 $\boldsymbol{x}$ は様々な値を取ります。しかし、デルタ関数は基本的には 0 なので、ほとんんどの $\boldsymbol{x}$ の値に対して $f(\boldsymbol{x})\delta(\boldsymbol{x} - \boldsymbol{\mu})$ は 0 となります。ただ唯一の例外として、デルタ関数の引数が 0 となる時 ($\boldsymbol{x} = \boldsymbol{\mu}$ の時) 、デルタ関数が値を持ち、その時の関数の値 $f(\boldsymbol{x} = \boldsymbol{\mu})$ が取り出されます。このように、総和 (積分) の中から特定の値を取り出すためにデルタ関数は用いられます。つまり、デルタ関数は単独でボンっと出てくることはなく、基本的には総和記号や積分の中でのみ登場する関数となっています。

2 つの確率変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ の統合確率を表す経験分布も同じようにデルタ関数を使って表すことができます。

$$\hat{p}_{\text{data}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})\delta(\boldsymbol{y} - \boldsymbol{y}^{(i)})$$

ここで、

$$\left(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}\right), \quad \left(\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}\right), \quad \cdots \quad \left(\boldsymbol{x}^{(m)}, \boldsymbol{y}^{(m)}\right)$$

が m 個の手持ちの訓練データを表します。この経験分布を先程の損失関数としての交差エントロピーの式に代入して、計算を進めてみましょう。

$$
\begin{aligned}
J(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\Big[-\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\Big] \\
&= \frac{1}{m}\sum_i\sum_j\left[-\left(\sum_k \delta(\boldsymbol{x}_i-\boldsymbol{x}^{(k)})\delta(\boldsymbol{y}_j-\boldsymbol{y}^{(k)})\right)\log p_{\mathrm{model}}(\boldsymbol{y}_j|\boldsymbol{x}_i;\boldsymbol{\theta})\right] \\
&= \frac{1}{m}\sum_k\left[-\log p_{\mathrm{model}}(\boldsymbol{y}^{(k)}|\boldsymbol{x}^{(k)};\boldsymbol{\theta})\right]
\end{aligned}
$$

このあたりはかなりややこしい計算になっているので、初めての方は何となく分かれば十分です。2行目から3行目に移る際には次のような操作を行なっています。期待値を取るためのループ $\sum_i\sum_j$ によって、デルタ関数の引数に含まれる $\boldsymbol{x}_i$ と $\boldsymbol{y}_j$ が更新されてきます[15]。しかし、デルタ関数は基本的に0なので、ほとんどの $\boldsymbol{x}_i$ と $\boldsymbol{y}_j$ の値に対して、このループの中の関数の値は0となります。ただ唯一の例外として、$(\boldsymbol{x}_i,\boldsymbol{y}_j)=(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(k)})$ となる場合のみ、デルタ関数は値を持ち、先ほどと同様に $\log p_{\mathrm{model}}(\boldsymbol{y}^{(k)}|\boldsymbol{x}^{(k)};\boldsymbol{\theta})$ という値が取り出されています。

色々ややこしい計算をしてきましたが、実用的には次の量 (交差エントロピー誤差) を損失関数として用いれば良いことになります[16]。

$$
J(\boldsymbol{\theta}) = \frac{1}{m}\sum_i\left[-\log p_{\mathrm{model}}(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})\right]
$$

$\frac{1}{m}$ は単なる定数なので、要するに

$$
J(\boldsymbol{\theta}) = -\sum_i \log p_{\mathrm{model}}(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})
$$

が最小となるように学習を行えば良いことになります。お気づきでしょうか、これは前回の「条件付き対数尤度と平均二乗誤差」でやった最尤法 (最尤推定) に他なりません。最尤法では、

$$
\sum_i \log p_{\mathrm{model}}(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{\theta})
$$

という**条件付き対数尤度**を最大化することで、最も適切なモデルのパラメータ $\boldsymbol{\theta}_{\mathrm{ML}}$ (最尤推定量) を決定しました。交差エントロピーは、この条件付き対数尤度にマイナスをつけた、いわば**「負の対数尤度」**になっているのです。交差エントロピーが現在多くのニューラルネットワークで損失関数として用いられている背景には、この量を最小化することが最尤推定に対応しており、統計学的にも最もらしい方法になっているという事情があるのです[17]。

[15] $\sum_k$ は、経験分布を構成するためのループなので、i と j とは異なる添字 (カウンタ変数) を用いています。
[16] 添え字はなんでも良いので、ここでは k でなく i を使っています。
[17] 学習速度の面からも、交差エントロピーは非常に効率のよい損失関数となります。

♣ 損失関数としての平均二乗誤差と平均絶対誤差

ここまでで、損失関数としては交差エントロピー誤差

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\Big[-\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\Big]$$

を用いれば良いことが分かりました。この交差エントロピー誤差では、特に p_{model} に対して具体的な分布を仮定しませでした。しかし、ある特定の分布の形を仮定すると、この交差エントロピー誤差をさらに書き換えることができます。

1 つ目として、モデルの確率分布が次のような Gauss 分布で与えられる場合を考えましょう[18]。

$$p_{\mathrm{model}}(y|x;\boldsymbol{\theta}) = \mathcal{N}(y; f(x;\boldsymbol{\theta}), 1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\Big[y - f(x;\boldsymbol{\theta})\Big]^2\right)$$

分散は何でもよいのですが、ここでは簡単のために 1 としています。平均は入力値 x とパラメータ $\boldsymbol{\theta}$ で表現される関数としています。例えば、最初にあげた「走行速度」と「停車距離」のデータに対しては、おおまかに 1 次関数的な傾向が見て取れるので、

$$f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x$$

とするのが良いでしょう。もちろん 1 次関数でなくて

$$f(x;\boldsymbol{\theta}) = \theta_0 + \theta_1 x + \theta_2 x^2$$

のような 2 次関数でも何でも大丈夫です。モデル分布がこのように与えられた時の交差エントロピー誤差を計算すると、

$$\begin{aligned} J(\boldsymbol{\theta}) =& \mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\left[-\log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\big[y - f(x;\boldsymbol{\theta})\big]^2\right)\right)\right] \\ =& \mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\left[-\log\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\big[y - f(x;\boldsymbol{\theta})\big]^2\right] = -\log\frac{1}{\sqrt{2\pi}} + \frac{1}{2}\mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\big[y - f(x;\boldsymbol{\theta})\big]^2 \end{aligned}$$

となります。何回か出てきた対数 log の性質や、期待値の足し算の性質を使っています。また、定数の期待値に関しては、確率の合計が 1 となることに注意して、

$$\mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\left[-\log\frac{1}{\sqrt{2\pi}}\right] = -\log\frac{1}{\sqrt{2\pi}}\sum_i\sum_j \hat{p}_{\mathrm{data}}(\boldsymbol{x}_i,\boldsymbol{y}_j) = -\log\frac{1}{\sqrt{2\pi}}$$

となることを用いています。ただ、この部分は単なる定数なので、損失関数として使う際には

[18] 簡単のためにここでは 1 変数の場合を考えます。

不要となります。つまり、モデルの分布が Gauss 分布の場合には、次の量を損失関数として使えば良いことが分かります[19]。

$$J(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{x,y\sim\hat{p}_{\text{data}}(\boldsymbol{x},\boldsymbol{y})}\big[y - f(x;\boldsymbol{\theta})\big]^2 = \frac{1}{2m}\sum_{i=1}^{m}\big[y^{(i)} - f(x^{(i)};\boldsymbol{\theta})\big]^2$$

この量は、前回の「条件付き対数尤度と平均二乗誤差」にて説明した、平均二乗誤差 (MSE: mean squared error)

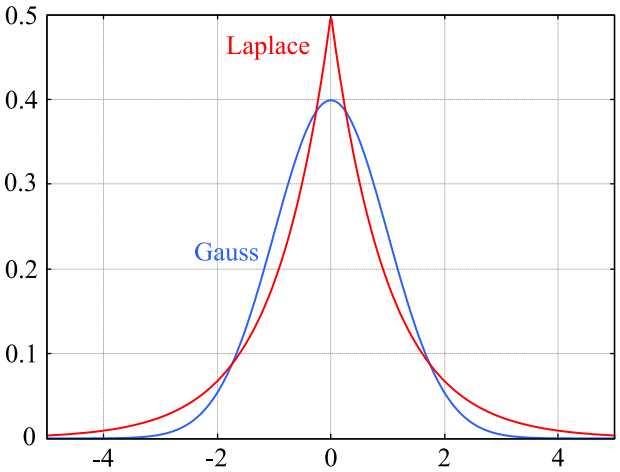
$$J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\big[y^{(i)} - f(x^{(i)};\boldsymbol{\theta})\big]^2$$

と定数倍 $\frac{1}{2}$ の違いを除いて全く同じであることが分かります。つまり、モデルの分布が Gauss 分布であると仮定した場合には、損失関数として交差エントロピーを使うことは平均二乗誤差を最小化することと一致するのです。インターネットで検索すると「損失関数としては平均二乗誤差と交差エントロピー誤差の 2 つが広く使われおり、問題によって使い分ける」といった記述をよく見かけるため、両者は全く別の量に勘違いしてしまいがちですが、今見たように本質的には同じものなのです。

先ほどは、分布が Gauss 分布の場合を考えましたが、今度は次のような Laplace 分布で与えられる場合を考えましょう。

$$p_{\text{model}}(y|x;\boldsymbol{\theta}) = \text{Laplace}(y; f(x;\boldsymbol{\theta}), 1) = \frac{1}{2}e^{-|y-f(x;\boldsymbol{\theta})|}$$

Laplace 分布のグラフは次のようになります。比較のために、分散と平均の値が等しい Gauss 分布も描いています。



(平均 0、分散 1 の Gauss 分布と Laplace 分布)

[19] 教科書 “Deep Learning (2016)” の (6.13) 式に対応する式です。

グラフをみると、Laplace 分布は Gauss 分布と比べて裾野が広い分布になっていることが分かります。ですから、データの値が散らばっている場合には、Gauss 分布を用いるより Laplace 分布を用いた方が精度の良いモデルになる傾向にあります。この時の交差エントロピーを計算すると、

$$\begin{aligned} J(\boldsymbol{\theta}) =& \mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\left[-\log\left(\frac{1}{2}e^{-[y-f(x;\boldsymbol{\theta})]}\right)\right] \\ =& \mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}\left[-\log\frac{1}{2}+|y-f(x;\boldsymbol{\theta})|\right] = -\log\frac{1}{2}+\mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}|y-f(x;\boldsymbol{\theta})| \end{aligned}$$

となります。先ほど同様、$-\log\frac{1}{2}=\log 2$ の部分は単なる定数なので、この項を落として

$$J(\boldsymbol{\theta}) = \mathbb{E}_{x,y\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},\boldsymbol{y})}|y-f(x;\boldsymbol{\theta})| = \frac{1}{m}\sum_{i=1}^{m}\left|y^{(i)}-f(x^{(i)};\boldsymbol{\theta})\right|$$

を損失関数として用いれば良いことになります。この量は、**「平均絶対誤差」(MAE：Mean Absolute Error)** と呼ばれます。この量も、あくまで交差エントロピー誤差の具体例に過ぎません。

以上のように、モデルの確率分布を指定すると、交差エントロピーから平均二乗誤差や平均絶対誤差が導かれることが分かりました。ただ実際には、学習の効率の面から、交差エントロピーをそのまま損失関数として用いることの方が一般的となっています。