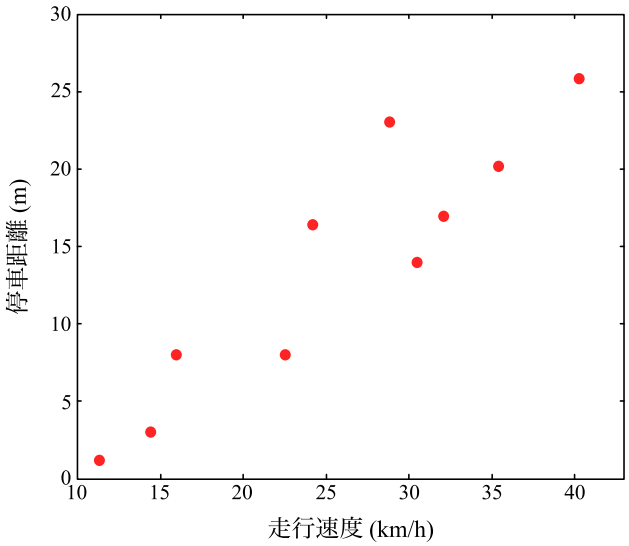
損失関数

ここでは、損失関数 (cost function) と呼ばれる関数について見ていきます。ここでの解説は、I. Goodfellow, Y. Bengio, and A. Courville, “Deep Learning” (2016) の 6.2.1 節の内容や数式を噛み砕いて、初めての人にも分かりやすくしたものです。関数の定義や記法 (notation) は、基本的にこの教科書に従っていますので、必要に応じて教科書の方も参照して下さい。

♣ 機械学習における損失関数とは？

具体的な数式の説明に入る前に、ざっくりとここで扱う損失関数とはどのようなものなのか、そのイメージを説明しておきます。前回の「条件付き対数尤度と平均二乗誤差」の解説において、車の走行速度と停車距離に関する次のデータを考えました。



もう 1 度、この 10 点のデータと同じ走行速度に対する停車距離を測定したとしましょう。上と完璧に同じデータが得られるでしょうか？ そのようなことは稀でしょう。実際の測定には誤差がつきものですから、データは揺らぎます。 このような揺らぎを表現するための道具として確率分布が導入されます。統計の分野 (そして統計をベースにしている機械学習の分野) では、「データはある確率分布から生成される」と考えるのが基本です。つまり、上のデータも、走行速度を表す変数を x、停車距離を表す変数を y とした時に、何らかの (条件付き) 確率分布 $p(y|x)$ が背後に存在し、10 個のデータはこの確率分布に従って生成されたものと解釈されます。

今回考えるのは、与えられたデータ (上の例では 10 個の走行速度と停車距離のデータのセット) から、どのようにして背後にあるこの確率分布 $p(y|x)$ を得るか、ということです。もし仮に、完璧な $p(y|x)$ を知ることができれば、任意の走行速度に対する停車距離を計算で簡単に予測することができるようになるため、非常に便利です。しかし、世の中そんな単純でなく、このよ

うなデータを生成する「真の確率分布」$p(y|x)$ は、一般に非常に複雑で取り扱いは困難 (不可能) であると考えられます。そこで、この真の確率分布 $p(y|x)$ をなるべく高い精度で再現でき、かつ比較的シンプルで取り扱いが容易な「確率モデル」$p_{\rm model}(y|x;\boldsymbol{\theta})$ を代わりに構築することを目指します。例えば、次のような Gauss 分布を採用するのは 1 つの手です[1]。

$$p_{\rm model}(y|x;\boldsymbol{\theta}) = \mathcal{N}(y;\theta_0+\theta_1 x,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\big[y-(\theta_0+\theta_1 x)\big]^2}{2\sigma^2}\right)$$

$\boldsymbol{\theta}$ はモデルを特徴づけるパラメータのベクトル (配列) で、この Gauss 分布のモデルの場合には、分布の平均を与える 1 次関数の切片 θ_0 と傾き θ_1 と、分布の分散 σ^2 の 3 つとなります。

$$\boldsymbol{\theta} = \big(\theta_0,\ \theta_1,\ \sigma^2\big)$$

もちろん、実際のデータは、こんなに単純な確率分布に従って生成されてはいないでしょう。しかし、パラメータ $\boldsymbol{\theta}$ を適切に調整して、確率モデル $p_{\rm model}(y|x;\boldsymbol{\theta})$ を真の確率分布 $p(y|x)$ に十分近づけることが出来れば、高い精度での予測が可能となり実用上の問題はほとんどありません。機械学習における学習 (訓練) とは、ざっくり言えば、このパラメータ $\boldsymbol{\theta}$ を適切に調整して、確率モデル $p_{\rm model}(y|x;\boldsymbol{\theta})$ を真の確率分布 $p(y|x)$ に近づける作業のことを指します。

このような学習を行うためには、確率モデルが真の確率分布からどのくらいズレているかを測る尺度が必要となります。この分布のズレ (=損失) を表す尺度が**「損失関数 (cost function)」**なのです。機械学習では、学習によりこの損失関数をなるべく小さく (最小化) することで、精度の高い予測ができるようになるのです。

♣ カルバック・ライブラー情報量と交差エントロピー

では、そのような分布のズレを表す損失関数とは、具体的にはどのような関数なのでしょうか。一般に、2 つの確率分布 $P(x)$ と $Q(x)$ のズレ (離れ具合) を表す量としては、次の**カルバック・ライブラー情報量 (Kullback-Leibler divergence)** と呼ばれる量が知られています[2]。

$$D_{\rm KL}(P||Q) = \mathbb{E}_{x\sim P}\big[\log P(x) - \log Q(x)\big]$$

上の式で出てくる $\mathbb{E}_{x\sim P}\big[f(x)\big]$ は、確率分布 $P(x)$ に関する $f(x)$ の期待値を表しています。期待値は「Bernoulli 分布」のところでも少し説明したのですが、この記号は初めて出てくるので、

[1] 同じようなことを「条件付き対数尤度と平均二乗誤差」の解説でも行っているので、参考にして下さい。
[2] 縦棒 1 本だと条件付き確率と紛らわしいので、P と Q の間に 2 本縦棒を入れるのが慣例です。

もう1度おさらいしておきましょう。$\mathbb{E}_{x\sim P}\big[f(x)\big]$ を具体的に数式で表すと、

$$\mathbb{E}_{x\sim P}\big[f(x)\big] = \sum_i P(x_i) f(x_i)$$

となります[3]。確率変数 x の取りうる全ての値 x_i に対して、関数の値 $f(x_i)$ と x_i が現れる確率 $P(x_i)$ を掛け合わせたものを足す、という量になっています[4]。ちょっと分かりづらいかもしれないので具体例を考えましょう。次のようなゲームを考えます。100円支払うと、サイコロ1個を1回振ることができ、サイコロの目に応じて次の金額がもらえます。

サイコロの目 x	1	2	3	4	5	6
賞金 $f(x)$	20円	50円	100円	100円	150円	150円

出るサイコロの目を表す確率変数を x として、x_1 を1の目、x_2 を2の目、$\cdots$ と表すことにしましょう。賞金は出る目の数によって変化しますから、x の関数となっていて $f(x)$ と表します。

$$f(x_1) = 20\text{円}, \quad f(x_2) = 50\text{円}, \cdots$$

そして出る目の確率は、(イカサマサイコロでなければ) どの目も $\frac{1}{6}$ となるため、

$$P(x_1) = \frac{1}{6}, \quad P(x_2) = \frac{1}{6}, \quad \cdots$$

となります。この時、貰える賞金の期待値 $\mathbb{E}_{x\sim P}[f(x)]$ を計算してみると、

$$\begin{aligned}\mathbb{E}_{x\sim P}[f(x)] &= \sum_{i=1}^{6} P(x_i) f(x_i) \\ &= \frac{1}{6}\times 20 + \frac{1}{6}\times 50 + \frac{1}{6}\times 100 + \frac{1}{6}\times 100 + \frac{1}{6}\times 150 + \frac{1}{6}\times 150 = 95\text{円}\end{aligned}$$

となります。つまり、このゲームで賞金として貰えると期待される金額は95円であり、参加費が100円であることを考えると、やや参加者が損するようなゲームになっていることが分かります。

ここまではサイコロの目のように、離散的 (とびとび) な値を取る確率変数 x を想定してきましたが、重さ、温度、速度などのように連続的な値を取る変数 x を考える場合も当然あります。このような場合には、取りうる値が無限にあるため、先ほどのように総和記号を用いて期待値を表現することはできません。このような場合には、「積分」と呼ばれる計算によって期待値を

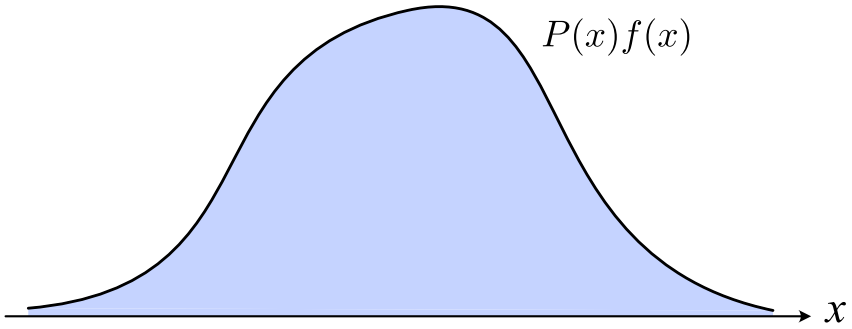
[3] Σ は総和を表す記号です。softmax 関数のところで説明しています。

[4] 一般に x の取りうる値の数がいくつあるか分からないので、和の範囲は書かず、単に Σ_i としています

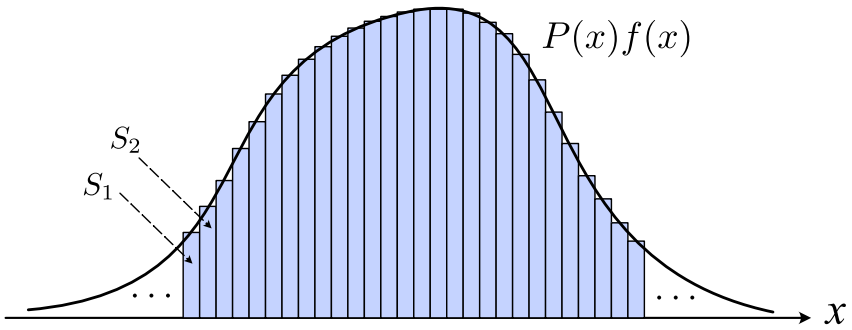
表すことができます。

$$\mathbb{E}_{x\sim P}\big[f(x)\big] = \int P(x)f(x)\,dx$$

積分は、総和 Σ_i の計算を、x が連続的な値を取る場合に拡張した、いわば「特殊な総和」です。一説によると、この積分を表すニョロ $\int$ は、総和を表す Σ の記号を引き伸ばして変形 (進化？) させたものと言われています。具体的に積分とは「面積」を求める計算です。例えば、上の期待値の式での積分は、図形的には $P(x)f(x)$ という関数と x 軸とで挟まれた青い領域の面積を求めることに対応しています。



長方形や三角形の面積は、小学校で習う有名な公式を使えば簡単に求められますが、上の図のようなイビツな領域の面積を求めるためには、積分が必要となるのです。積分の基本的な考え方は、このイビツな図形を細長い長方形の短冊に分解して足し合わせるというものです。



短冊は長方形なので、その面積は縦 $\times$ 横で簡単に計算できます。ですから、イビツな図形でも細長い長方形の短冊に分けて、それぞれの短冊の面積 S_i を計算して、各々の面積を全て足し合わせれば、面積を求めることができるのです[5]。その意味で、

$$\int P(x)f(x)\,dx = \sum_i S_i$$

になっており、確かに積分 $\int$ は、総和 Σ を連続的な変数の場合に拡張したことになっているこ

[5]正確に面積を求めるためには、短冊は物凄く細かくする必要があります。

とが何となくわかるかと思います。具体的にこのような積分をどのように計算するかは、高校数学の範囲となります。ただし、特に具体的な積分の計算が分からなくても以下の内容は十分理解できるので、分からない方は、ひとまず、総和記号 Σ を連続的な変数の場合に拡張したもの程度に思っておけば十分です。

期待値に関しては、いくつか重要な性質が成り立ちます。以下で使う性質として「期待値の和と差」に関する性質を確認しておきましょう。期待値に関しては、次の式が成り立ちます。

$$\mathbb{E}_{x\sim P}\big[f(x)+g(x)\big] = \mathbb{E}_{x\sim P}\big[f(x)\big] + \mathbb{E}_{x\sim P}\big[g(x)\big]$$
$$\mathbb{E}_{x\sim P}\big[f(x)-g(x)\big] = \mathbb{E}_{x\sim P}\big[f(x)\big] - \mathbb{E}_{x\sim P}\big[g(x)\big]$$

つまり、2 つの関数の和 (差) の期待値は、それぞれの関数の期待値の和 (差) になっています。このことは、実際に総和 (または積分) の形で期待値を表現してみると確かめられます。x が離散的な値を取る場合には、

$$\begin{aligned}\mathbb{E}_{x\sim P}\big[f(x)+g(x)\big] &= \sum_i P(x_i)\big[f(x_i)+g(x_i)\big] \\ &= \sum_i \big[P(x_i)f(x_i)+P(x_i)g(x_i)\big] \\ &= \sum_i P(x_i)f(x_i) + \sum_i P(x_i)g(x_i) = \mathbb{E}_{x\sim P}\big[f(x)\big] + \mathbb{E}_{x\sim P}\big[g(x)\big]\end{aligned}$$

といった感じで示せます。連続的な値を取る場合も同じ要領です。

だいぶ話が逸れてしまいましたが、期待値の確認をしたところで、再び Kullback-Leibler 情報量の式を眺めてみましょう[6]。

$$D_{\mathrm{KL}}(P||Q) = \mathbb{E}_{x\sim P}\big[\log P(x) - \log Q(x)\big]$$

期待値の差に関する性質を用いれば、

$$D_{\mathrm{KL}}(P||Q) = \mathbb{E}_{x\sim P}\big[\log P(x)\big] - \mathbb{E}_{x\sim P}\big[\log Q(x)\big]$$

ということになり、$\log P(x)$ と $\log Q(x)$ の確率分布 $P(x)$ に関する期待値の差を表していること

[6] log の底は何でもいいですが、普通はネイピア数 e の自然対数を考えます。

が分かります。上の式の 1 項目と 2 項目の量にはそれぞれ名前がついています。1 項目の

$$\mathbb{E}_{x\sim P}\big[\log P(x)\big] = \begin{cases} \displaystyle\sum_i P(x_i)\log P(x_i) & \text{(離散確率分布の場合)} \\ \displaystyle\int P(x)\log P(x)\,dx & \text{(連続確率分布の場合)} \end{cases}$$

は、**「エントロピー (entropy)」**と呼ばれる、確率分布 $P(x)$ の乱雑さ (予測のしにくさ) を表す量となっています。そして 2 項目の

$$\mathbb{E}_{x\sim P}\big[\log Q(x)\big] = \begin{cases} \displaystyle\sum_i P(x_i)\log Q(x_i) & \text{(離散確率分布の場合)} \\ \displaystyle\int P(x)\log Q(x)\,dx & \text{(連続確率分布の場合)} \end{cases}$$

は、**「交差エントロピー (cross entropy)」**と呼ばれます[7]。ですから、Kullback-Leibler 情報量 $D_{\mathrm{KL}}(P||Q)$ は、分布 $P(x)$ のエントロピーから、分布 $Q(x)$ との交差エントロピーを引いた量となっています。

このセクションの最後に、具体的な 2 つの分布を考えて、確かに Kullback-Leibler 情報量が 2 つの分布の離れ具合を表す量になっていることを実感しておきましょう。ここでは例として、2 つの Gauss 分布 $P(x)$、$Q(x)$ の間の Kullback-Leibler 情報量を計算してみます[8]。

$$P(x) = \mathcal{N}(x;\mu_p,\sigma_p^2) = \frac{1}{\sqrt{2\pi}\sigma_p}\exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)$$

$$Q(x) = \mathcal{N}(x;\mu_q,\sigma_q^2) = \frac{1}{\sqrt{2\pi}\sigma_q}\exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)$$

この場合の Kullback-Leibler 情報量は、次のようになります。

$$D_{\mathrm{KL}}(P||Q) = \int P(x)\log P(x)dx - \int P(x)\log Q(x)dx = \int P(x)\log\frac{P(x)}{Q(x)}dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma_p}\int \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)\left[\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right]dx$$

かなり複雑な式変形に見えるかもしれませんが、単に、log の性質を使っているだけです[9]。

[7] 日本語でも「クロスエントロピー」と呼ぶこともありますが、ここでは交差エントロピーと呼びます。
[8] Gauss 分布については、以前の解説プリントを参照して下さい。
[9] -1 乗に関しては、softmax 関数の解説を見て下さい。

$$\log a - \log b = \log a + \log b^{-1} = \log a + \log \frac{1}{b} = \log \frac{a}{b}$$

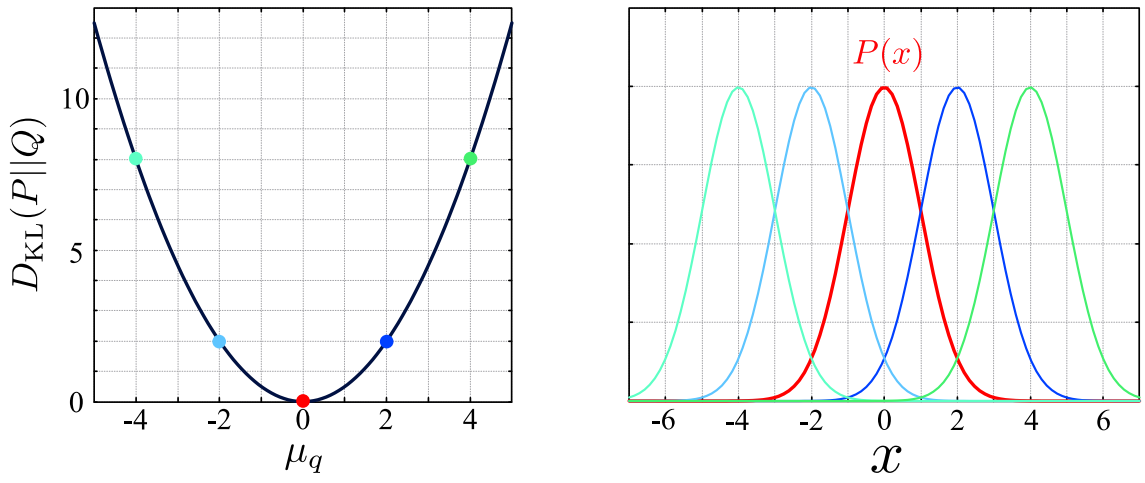
$$\log e^x = \log_e e^x = x \log_e e = x$$

ここでの対数 log の底は、ネイピア数 e であるとして計算しています。Gauss 分布同士の Kullback-Leibler 情報量の計算に出てくる積分は簡単に計算することができて、計算結果は次のようになります。

$$D_{\mathrm{KL}}(P||Q) = \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

基本的な積分の計算は知っていて、さらにここでの積分の詳細が気になる方は「ガウス積分」などのキーワードで検索すれば色々出てくるので参考にして下さい。ただし単なる計算なので、ここでは結果を受け入れてもらえればそれで結構です[10]。

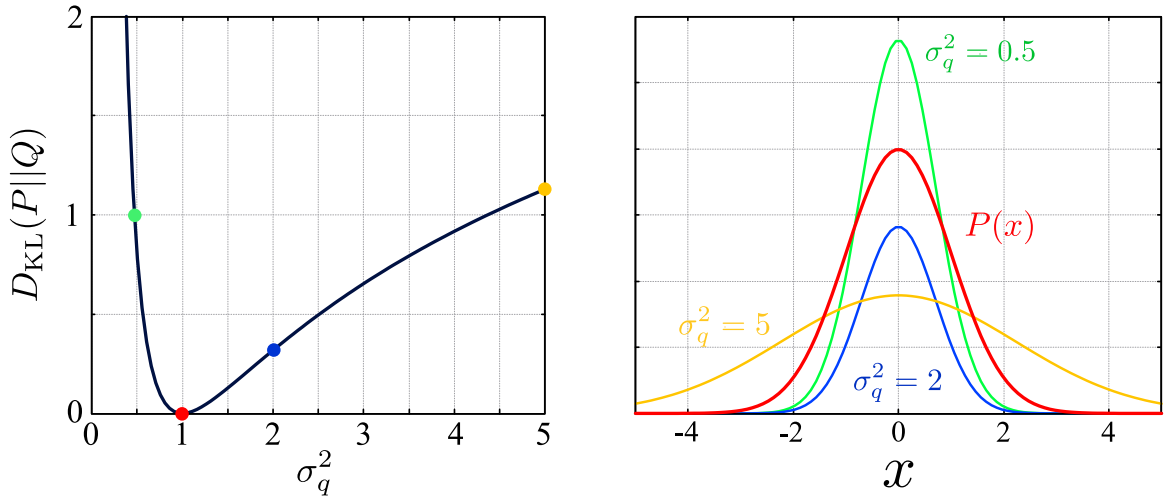
以下では、$P(x)$ の分布の平均を $\mu_p = 0$、$\sigma_p^2 = 1$ として、$Q(x)$ の分布の平均 μ_q と分散 σ_q^2 を変化させた時の、Kullback-Leibler 情報量の値の振る舞いを見ていきましょう。まず、平均 μ_q を変化させると次のようになります。



$\mu_q = 0$ の時、2 つの分布は完璧に重なり、この時 $D_{\mathrm{KL}}(P||Q) = 0$ となります。μ_q を大きくしていくと、2 つの分布は離れていき、それに伴って $D_{\mathrm{KL}}(P||Q)$ の値も大きくなっていく様子が見て取れます。

次に、分散 σ_q^2 を変化させた時の $D_{\mathrm{KL}}(P||Q)$ の変化を調べると、次のようになります。

[10] Wolfram Alpha (https://www.wolframalpha.com) という検索エンジンに式を入力すれば、積分の計算を一瞬でやってくれます。何か別の分布で計算してみたいけど、積分の計算が分からないという時に活用して下さい。



やはりこちらも、2 つの分布が完全に一致する $\sigma_q^2 = 1$ の時に $D_{\mathrm{KL}}(P||Q) = 0$ となり、そこから 2 つの分布がズレていくと $D_{\mathrm{KL}}(P||Q)$ の値も大きくなっていく様子が見て取れます。このことから、確かに Kullback-Leibler 情報量は、2 つの分布の離れ具合を表す量になっていることが実感できるかと思います。

♣ 損失関数としての交差エントロピー (1)

さて、Kullback-Leibler 情報量が 2 つの分布の離れ具合を表す量であることが分かったため、この量は損失関数として用いることができます。入力値 $\boldsymbol{x}$ が与えられた時の出力値 $\boldsymbol{y}$ を生成する真の確率分布を $p(\boldsymbol{y}|\boldsymbol{x})$ とします。そして、出力値 $\boldsymbol{y}$ を予想するためにパラメータ $\boldsymbol{\theta}$ を適切に調整して構築していく確率モデルを $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ とします。

この 2 つの確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ と $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ を近づけることが、学習の目標です。条件付き確率であっても先ほどと同様に Kullback-Leibler 情報量を考えることができて、2 つの分布の離れ具合は、

$$
\begin{aligned}
D_{\mathrm{KL}}(\boldsymbol{p}||\boldsymbol{p}_{\mathrm{model}}) &= \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\big[\log p(\boldsymbol{y}|\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\big] \\
&= \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\big[\log p(\boldsymbol{y}|\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\big[\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\big]
\end{aligned}
$$

と表すことができます。ここで 1 項目 (エントロピー) をみると、そこには $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ が一切含まれていないことが分かります。つまり、学習においてパラメータ $\boldsymbol{\theta}$ を調整する際に、この部分は学習には何も影響しない、単なる定数になります。言い換えると、$\boldsymbol{\theta}$ を変更、すなわちモデル $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ を変化させてもこの部分は全く変化しないため、学習の際には計算する必要がない部分となっています。計算する量はなるべく少ない方が効率的に学習できますから、

通常は2項目の

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

の交差エントロピーの部分のみを、モデルと真の分布との差を表す損失関数として用います。ただし、交差エントロピーをそのままま損失関数に使おうとすると問題が生じます。上の量は、$\boldsymbol{y}$ に関しては総和 (または積分) が取れらて、全ての場合が足し合わされているため、$\boldsymbol{y}$ に依存することはありません。ただし、$\boldsymbol{\theta}$ や $\boldsymbol{x}$ を変化させると、その値は変化します。$\boldsymbol{\theta}$ に依存することは全然問題でなく、むしろ $\boldsymbol{\theta}$ を変化させてこの量を小さくすることで学習を行います。では、$\boldsymbol{x}$ の値はどのようにすれば良いでしょうか？ 入力値 $\boldsymbol{x}$ の値はサンプルによってバラバラで、そのどれを使えば良いかは明らかではありません。そこで、$\boldsymbol{x}$ も確率分布に従って生成されるデータと考えて、それによって期待される交差エントロピーの値を考えることになります。この操作を理解するためには、結合確率、条件付き確率、周辺確率といった確率に関する基本事項を理解する必要がありますので、以下で見ていきましょう。

色々な確率 (結合確率、条件付き確率、周辺確率)

2つの確率変数 x、y が絡む確率には、様々な種類があります。例として、グー、チョキー、パーをランダムで出すジャンケンマシーンを2回動かした時に、グーがでる回数を x、チョキがでる回数を y という確率変数で表すことにしましょう。x は0 (1回もグーが出ない) か1か2の値を取ります (y も同じです)。

この時、「グーが1回 ($x=1$)、チョキが1回 ($y=1$) となるような確率」のような、x と y の値の組み合わせが生じる確率のことを**「結合確率 (joint probability)」**と呼び[11]、$p(x,y)$ で表現します。例えば、2回ともグーが出て1回もチョキが出ないような確率は、

$$p(x=2,y=0) = \frac{1}{3}\times\frac{1}{3} = \frac{1}{9}$$

となります。一方、グーが2回、チョキが1回ということは、2回しかマシーンは動かさないのでありえないため、

$$p(x=2,y=1) = 0$$

となります。これらを表にまとめると (統合確率分布といいます)、次のようになります。

[11] 同時確率と呼ぶこともあります

x \ y	0	1	2
0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$
1	$\frac{2}{9}$	$\frac{2}{9}$	0
2	$\frac{1}{9}$	0	0

場合によっては、y の値は何でも良いけど x が 0 となる確率が知りたい、ということがあります。このような確率を**「周辺確率 (marginal probability)」**と呼び、$p(x=0)$ で表現することになります。$x=0$ の時、y は 0~2 の 3 通りの場合があるため、それらの確率を足し合わせると、

$$p(x=0) = p(x=0, y=0) + p(x=0, y=1) + p(x=0, y=2) = \frac{1}{9} + \frac{2}{9} + \frac{1}{9} = \frac{4}{9}$$

と計算できます。一般に、周辺分布は総和記号を使えば

$$p(x) = \sum_{i=1}^{m} p(x, y=y_i), \qquad p(y) = \sum_{i=1}^{m} p(x=x_i, y)$$

と表せます。この周辺分布を先程の統合確率分布の表に書き足すと次のようになります。

x \ y	0	1	2	$p(x)$
0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{4}{9}$
1	$\frac{2}{9}$	$\frac{2}{9}$	0	$\frac{4}{9}$
2	$\frac{1}{9}$	0	0	$\frac{1}{9}$
$p(y)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$	

右端にあるのが x に関する周辺確率 $p(x)$、下端にあるのが y に関する周辺確率 $p(y)$ となっています。お気づきだと思いますが、このように書いた時に表の周辺にくることから、「周辺」確率という名前がついています。

最後に、これは何度もこれまでに出てきましたが、確率変数 x の値に依存して確率変数 y の確率が決まる場合、**「条件付き確率 (conditional probability)」**と呼び、その分布を $p(y|x)$ と書きます。この条件付き分布と統合確率、周辺確率の間には次の関係式が成り立ちます。

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

これは条件付き確率の意味を考えれば納得できます。$p(y=b|x=a)$ は、$x=a$ となる確率のうち、$y=b$ である可能性が占める割合を意味します。すなわち、統合確率 $p(x=a,y=b)$ が周辺分布 $p(x=a)$ に対して占める割合によって求めることができるため、上の式のように表されます。

さて、確率の基本事項を確認したところで、元の話に戻りましょう。2 つの確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ と $p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ の交差エントロピー

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

を、特定の $\boldsymbol{x}$ への依存しないようにするために、確率分布 $p(\boldsymbol{x})$ についての期待値を考えましょう。ここで、$p(\boldsymbol{x})$ は先ほど説明した $\boldsymbol{x}$ の周辺分布を表します。離散確率分布の場合に、この期待値を計算してみると次のようになります。

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\Big[-\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\big[\log p_{\mathrm{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})\big]\Big] &= \sum_j \left[p(\boldsymbol{x}_j)\left(-\sum_i p(\boldsymbol{y}_i|\boldsymbol{x}_j)\log p_{\mathrm{model}}(\boldsymbol{y}_i|\boldsymbol{x}_j;\boldsymbol{\theta})\right)\right] \\
&= \sum_i \sum_j \Big[-p(\boldsymbol{x}_j)p(\boldsymbol{y}_i|\boldsymbol{x}_j)\log p_{\mathrm{model}}(\boldsymbol{y}_i|\boldsymbol{x}_j;\boldsymbol{\theta})\Big] \\
&= \sum_i \sum_j \Big[-p(\boldsymbol{x}_j,\boldsymbol{y}_i)\log p_{\mathrm{model}}(\boldsymbol{y}_i|\boldsymbol{x}_j;\boldsymbol{\theta})\Big]
\end{aligned}$$

難しい計算に感じるかもしれませんが、落ち着いて各ステップを考えればそこまで難しくありません。2 つの期待値 $\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}$ と $\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}$ を表す総和記号の添字を i と j で区別していることに注意してください。プログラミングに慣れている方なら直ぐにわかると思いますが、これは 2 重ループになっているため、内側のループカウンタ変数と外側のループカウンタ変数が同じだとマズいですよね。そこで i と j という異なる添字 (カウンタ変数) を用いて区別しています。また、2 行目から 3 行目に移るところで、種々の確率の間に成り立つ関係式

$$p(\boldsymbol{x}_j)p(\boldsymbol{y}_i|\boldsymbol{x}_j) = p(\boldsymbol{x}_j,\boldsymbol{y}_i)$$

を用いています。ややこしいのでもう 1 度確認しておくと、$p(\boldsymbol{x}_j)$ は周辺分布、$p(\boldsymbol{y}_i|\boldsymbol{x}_j)$ は条件付

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 $-\log p_{\text{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ 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。

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

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

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

が、真の確率分布 $p(\boldsymbol{y}|\boldsymbol{x})$ とモデル $p_{\text{model}}(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$ 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」

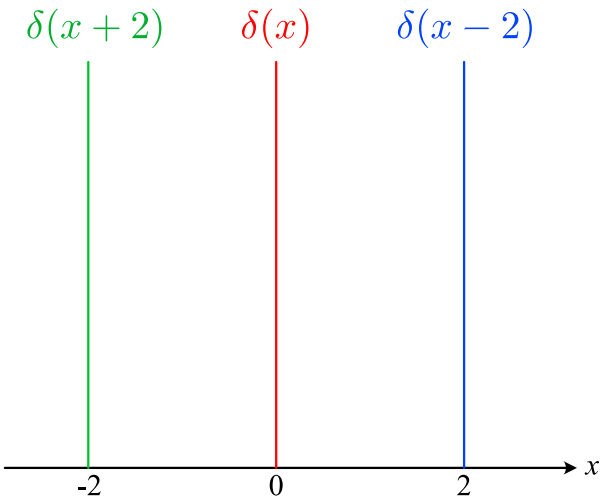
$$\sum_i \hat{p}_{\text{data}}(x_i) = 1$$

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)

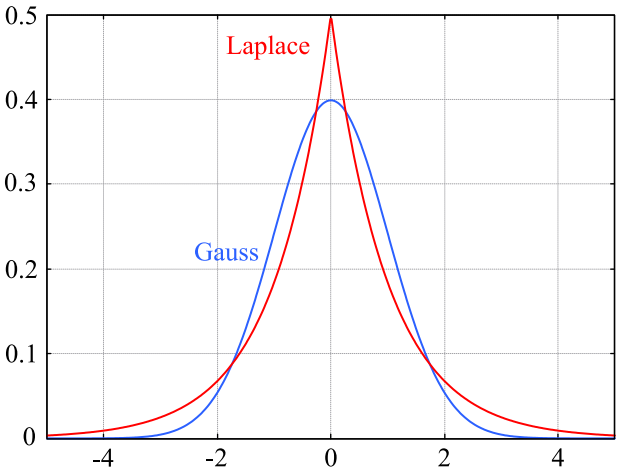
$$J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\big[y^{(i)} - f(x^{(i)};\boldsymbol{\theta})\big]^2$$

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、

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

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

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

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。