§ 3　Cost Function

デジタル版はここ

伊 冉 (Andre YI) - 21122200512

2025 年 6 月 4 日

3　Cross-Entropy

Def. 3.1 (交差エントロピー (クロスエントロピー)): p と q を $\mathbb{R}$ 上の 2 つの確率密度変数とする。負の尤度関数 $-\ell_q(x) = -\ln q(x)$ が $q(x)$ によって与えられる情報を測定する。q の p に対する交差エントロピーは、次のように定義される。

$$S(p,q) = \mathbb{E}^p[-\ell_q] = -\int_{\mathbb{R}} p(x)\ln q(x)\,dx \text{[1]} \tag{1}$$

p と q が離散確率変数なら、

$$S(p,q) = \mathbb{E}^p[-\ell_q] = -\sum_x p(x)\ln q(x) \tag{2}$$

Problem. 3.1

離散確率変数 p と q に対して、交差エントロピー $S(p,q) \geq 0$ を示せ

Proof. まず、確率密度変数の定義によって、$p(x) \geq 0$ かつ $\sum_x p(x) = 1$ が成り立つ
次に、$-\ell_q(x) \geq 0$ であり、***i.e.*** $-\ln q(x) \geq 0\,, q(x) \in (0,1]$ が成り立つ
したがって、$-\sum_x p(x)\ln q(x) = \sum_x -p(x)\ln q(x) \geq 0$ が成り立つ。 ■

Prop. 3.2: シャノンエントロピー (Shannon-entropy) [2] $H(p)$ と交差エントロピー $S(p,q)$ により、$S(p,q) \geq H(p)$ が成り立つ。$H(p)$ は次のように表される。

$$H(p) = -\mathbb{E}^p[\ell_p] = -\int_{\mathbb{R}} p(x)\ln p(x)\,dx$$

[1] **[1]** 式は連続確率密度変数に対する定義である。

[2] シャノンエントロピー $H(p)$ は、 $p(x)$ 内の情報量を測定するものです。したがって、前の命題は、分布 $q(x)$ によって与えられる情報が、 $p(x)$ から評価された場合、常に $p(x)$ によって定義される情報よりも大きいことを示しています。

Proof. $\ln u \leq u-1\,(for\, u>0)$ を用いると、次の **[3]** のように変形できる。

$$\begin{aligned} S(p,q)-H(p) &= -\int_{\mathbb{R}} p(x)\ln q(x)\,dx + \int_{\mathbb{R}} p(x)\ln p(x)\,dx \\ &= -\left\{\int_{\mathbb{R}} p(x)\left[\ln q(x)-\ln p(x)\right]\,dx\right\} \\ &= -\int_{\mathbb{R}} p(x)\ln\left(\frac{q(x)}{p(x)}\right)\,dx \geq \int_{\mathbb{R}} p(x)\left(\frac{q(x)}{p(x)}-1\right)\,dx \qquad \textbf{[3]} \\ &= \int_{\mathbb{R}} p(x)\,dx - \int_{\mathbb{R}} q(x)\,dx = 1-1=0 \end{aligned}$$

したがって、$S(p,q)-H(p)\geq 0$ であり、$S(p,q)=H(p)$ となるのは $p=q$ の場合のみです。 ■

Cor. 3.3: p を $\mathbb{R}$ 上の一次元ランダム変数 X の密度とします。次のようになる：

$$H(p)\leq \frac{1}{2}\ln(2\pi Var(X))$$

Proof. $\mu=E[X]$ と $\sigma^2=\mathrm{Var}(X)$ とし、正規密度を次のように考えます：

$$q(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$$

これは $p(x)$ と同じ平均と分散を持っています。 p の q に対する交差エントロピーは次のように計算できます。

$$\begin{aligned} S(p,q) &= -\int_{\mathbb{R}} p(x)\ln q(x)dx = -\int_{\mathbb{R}} p(x)\left(\ln\frac{1}{\sigma\sqrt{2\pi}}+\ln e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}\right)dx \\ &= -\left\{\int_{\mathbb{R}} p\,(x)\ln\frac{1}{\sigma\sqrt{2\pi}}dx+\int_{\mathbb{R}} p\,(x)\ln e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}dx\right\} \\ &= -\left\{\ln\frac{1}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} p\,(x)\,d^3x-\frac{1}{2\sigma^2}\int_{\mathbb{R}} p\,(x)\,(x-\mu)^2\,dx\right\} \\ &= -\ln\frac{1}{\sigma\sqrt{2\pi}}+\frac{1}{2\sigma^2}\int_{\mathbb{R}} p\,(x)\,(x-\mu)^2\,dx \\ &= -\ln\left[\left(\sigma\sqrt{2\pi}\right)^2\right]^{-\frac{1}{2}}+\frac{1}{2\sigma^2}\int_{\mathbb{R}} p\,(x)\,(x-\mu)^2\,d^4x \\ &= -\frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{1}{2} \\ &= -\frac{1}{2}\ln\left(2\pi e\sigma^2\right) \end{aligned}$$

上の Prop.3.2 から $H(p)\leq S(p,q)=\frac{1}{2}\ln(2\pi\sigma^2)$ となる。等式は、正規分布を持つ確率変数 X の場合に成り立つ。 ■

Rem. 3.4: 交差エントロピー (Cross-entropy) は、分類問題でよく使用されますが、回帰問題でも一般的に使用されます。

[3] $p(x)$ は確率密度変数であるため、$\int_{\mathbb{R}} p\,(x)\,dx=1$ が成り立つ。
[4] $\int_{\mathbb{R}} p\,(x)\,(x-\mu)^2\,dx$ は、X の分散である。***i.e.*** $\sigma^2=\int_{\mathbb{R}} p\,(x)\,(x-\mu)^2\,dx$

4 Kullback-Leibler Divergence

Def. 4.1 (Kullback-Leibler Divergence): Kullback-Leibler(KL) Divergence(または Relative-entropy) とは、Cross-entropy と Shannon-entropy の差である。***i.e.*** ある確率分布 p を基準にし、別の分布 q がどれだけ異なるかを測る指標であり、その式は以下のように表示される。

$$D_{KL}(p||q) = S(p,q) - H(p)$$

[3] 式により、

$$D_{KL}(p||q) = -\int_{\mathbb{R}} p(x) \ln \frac{q(x)}{p(x)} \, \mathrm{d}x$$

Prop.3.2 の $S(p,q) \geq H(p)$ により、 $D_{KL}(p||q) \geq 0$ が成り立つ。しかしながら、これは距離[5]ではない。なぜなら、対称性がなく、三角不等式を満たさないからである。
Cross-entropy と Kullback-Leibler Divergence の両方は、神経ネットワーク (neuronal networks) のコスト関数として考えることができる。これについては以下で説明する。

まず X を入力ランダム変数とし、$Y = f_\theta(X, \xi)$ $\left(\theta = (w,b), \xi\text{: ネットワークのノイズを示すランダム変数}\right)$ とし, Z をターゲットランダム変数とし、 $P_{X,Z}(x,z)$ を (X,Z) の同時確率密度関数 (joint probability density function)[6]とし、training distribution と見なされる。$P_Y(y|x)$[7]を入力 X を与えられたときの出力 Y の条件付き確率密度関数 (条件付きモデル密度関数) とする。

次に、パラメータ θ を調整することで、密度を比較することを説明する。
与えられたトレーニング分布 $p = p_{X,Z}(x,z)$ に対して、クロスエントロピー p と q をできるだけ小さくする条件付きモデル分布 $q = p_\theta(\cdot|x)$ を得ることができる。その値 $\theta^* = \arg\min_\theta S(p_{X,Z}, p_\theta(Z|X))$ は以下のコスト関数の最小値である。

$$C(\theta) = S(p_{X,Z}, p_\theta(Z|X))$$

上の Prop.3.2により、$S(\theta) \geq H(p_{X,Z})$ が成り立つ。したがって、調整された最小値はトレーニング分布のシャノンエントロピー $H(p_{X,Z})$ と等しくなる。

---

[5] 距離の定義として代表的なものは「距離空間」における距離の性質です。距離 $d(x,y)$ が距離空間において満たすべき条件は以下の通りです。1. 非負性: $d(x,y) \geq 0$ であり、$d(x,y) = 0$ の場合は $x = y$ である。2. 対称性: $d(x,y) = d(y,x)$ 3. 三角不等式: $d(x,z) \leq d(x,y) + d(y,z)$

[6] n 個の確率変数 $X_1, X_2, \ldots, X_n$ の同時確率分布とは、確率変数の組 $(X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ に確率を対応させる関数のことである。同時確率分布は $\mathbb{R}^n$ 上の測度であり、$P_{X_1,X_2,\ldots,X_n}(\cdot)$ と書かれる。確率変数 X, Y が連続型であるとき、$p(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$ を満たす f を同時確率密度関数 (probability density function) という。なお、f は次の 2 つの条件を満たす。すべての x, y に対して、$f(x,y) \geq 0$, $\int_S f(x)\,dx = 1$ (ただし、S は標本空間 (sample space) であるとする。)

[7] $p_Y(y|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)} = \dfrac{p_Y(y) \cdot p_X(x|y)}{p_X(x)}$, その $p_{X,Y}(x,y)$ は X と Y の同時確率密度関数, $p_X(x)$ は X の周辺確率密度関数。確率変数 X,Y が離散型であるとき、 $X = x_i$ が与えられたときの Y となる条件付き期待値 (conditional expectation of Y given$X = x_i$) を次のように定める。$\mathbb{E}[Y|X] = \mathbb{E}[Y|X = x_i] = \sum_{j=1}^{\infty} y_j p_{Y|X}(y_j|x_i) = \sum_{j=1}^{\infty} y_j \frac{p(x_i,y_j)}{p_X(x_i)}$

次に、 $p_X(x)$ を入力変数 X の密度とする。条件付き密度の特性を利用すると、

$$\begin{aligned} C(\theta) &= S(p_{X,Z}, p_\theta(Z|X)) = -\iint p_{X,Z}(x,z) \ln p_\theta(z|x)\, dx\, dz \\ &= -\iint p_{X,Z}(x,z) \ln\left(\frac{p_{\theta(x,z)}}{p_X(x)}\right) dx\, dz \\ &= -\iint p_{X,Z}(x,z) \ln p_\theta(z|x)\, dx\, dz + \iint p_{X,Z}(x,z) \ln p_X(x)\, dx\, dz \\ &= S(p_{X,Z}, p_\theta(X,Z)) + \int \left(\int p_{X,Z}(x,z)\, dz\right) \ln p_X(x)\, dx \qquad \textbf{[4]} \\ &= S(p_{X,Z}, p_\theta(X,Z)) + \int p_X(x) \ln p_X(x)\, dx \\ &= S(p_{X,Z}, p_\theta(X,Z)) - H(p_X) \end{aligned}$$

PROBLEM. 4.1

[4] 式の $p_X(x) = \int p_{X,Z}(x,z)\, dz$ を証明せよ

Proof. 確率論の基本原理から、結合分布は次のように分解できる：

$$p_{X,Z}(x,z) = p_X(x) \cdot p_Z(z|x)$$

次に、$p_{X,Z}(x,z)$ を z に関して積分すると、

$$\int p_{X,Z}(x,z)\, dz = \int p_Z(z|x) p_X(x)\, dz$$

ここで、 $p_X(x)$ は定数なので、 $\int p_Z(z|x) p_X(x)\, dz p_X(x) \int p_Z(z|x)\, dz$ となる。そして、条件付き確率密度関数の性質により、$\int p_Z(z|x)\, dz = 1$ が成り立つ。したがって、

$$\int p_{X,Z}(x,z)\, dz = p_X(x)$$

■

$H(p_X)$ は入力エントロピー、 ***i.e.*** 入力変数 X のシャノンエントリピーである。 $H(p_X)$ はモデルパラメータ θ に依存しないため、新しいコスト関数を以下のように定義する。

$$\overline{C}(\theta) = S(p_{X,Z}, p_\theta(X,Z))$$

これはモデル密度のトレーニング密度のクロスエントロピーであり、同じパラメータ θ に対して $C(\theta)$ 最小値を取れる。すなわち、

$$\theta^* = \arg\min_\theta C(\theta) = \arg\min_\theta \overline{C}(\theta)$$

結論として、トレーニング密度関数 $p_{X,Z}$ とモデル密度関数 $p_\theta(X,Y)$ または条件付きモデル密度関数 $p_\theta(Y|X)$ が与えられた場合、コスト関数 $\overline{C}(\theta)$ または $C(\theta)$ のいずれかが最小となるパラメータ値 θ を見つけることができる。

実際の応用では、ランダム関数 (X, Z) は n 回の測定、$(x_1, z_1), (x_2, z_2), \cdots, (x_n, z_n)$ から得られる。この場合、(X, Z) の同時確率密度関数がその経験トレーニング分布関数[8] $\widehat{p}_{X,Z}(x, z)$ によって近似されると仮定する。この場合、新しいコスト関数はトレーニングセットによって定義された経験的トレーニング分布と、モデルによって定義された確率分布との間のクロスエントロピーである。平均で期待値を近似すると、次のようになる。

$$\tilde{C}(\theta) = S\left(\widehat{p}_{X,Z}, p_\theta(X, Z)\right) = \mathbb{E}^{\widehat{p}_{X,Z}}\left[-\ln p_\theta(Z|X)\right] = -\frac{1}{n}\sum_{j=1}^{n} \ln p_\theta(z_j|x_j)$$

同様の測定に基づく誤差は次のように定義できる。

$$\widehat{C}(\theta) = S\left(\widehat{p}_{X,Z}, p_\theta(X, Z)\right) = \mathbb{E}^{\widehat{p}_{X,Z}}\left[-\ln p_\theta(X, Z)\right] = -\frac{1}{n}\sum_{j=1}^{n} \ln p_\theta(x_j, z_j) \quad \textbf{[5]}$$

したがって、コスト関数 $C(\theta) = D_{KL}\left(p_{X,Z}||p_\theta(X, Z)\right)$ はトレーニング密度関数とモデル密度関数との Kullback-Leibler Divergence によって与えられる。シャノンエトロピー $H(p_{X,Z})$ がパラメータ θ に依存しないため、コスト関数は次のように表される。

$$\theta^* = \arg\min_\theta D_{KL}\left(p_{X,Z}||p_\theta(X, Z)\right) = \arg\min_\theta S\left(p_{X,Z}, p_\theta(X, Z)\right)$$

トレーニング分布とモデルの分布が一致するとき、その前の最小値は 0 になる。
トレーニングセット $(x_1, z_1), (x_2, z_2), \cdots, (x_n, z_n)$ が提供される場合で、コスト関数は経験的密度関数 $\widehat{p}_{X,Z}$ を用いて次のように書ける。

$$C(\theta) = \mathbb{E}^{\widehat{p}_{X,Z}}\left[-\ln \frac{p_\theta(X, Z)}{\widehat{p}_{X,Z}}\right] = -\frac{1}{n}\sum_{i=1}^{n}\left(\ln p_\theta(x_i, z_i) - \ln \widehat{p}_\theta(x_i, z_i)\right)$$

[8] 経験トレーニング分布 (Empirical Training Distribution) は、訓練データに基づいて統計的に計算された確率分布を指します。この分布は、実際のデータの特性を近似するために使用され、機械学習モデルの訓練過程で最適化を行います。

4.1 Maximum Likelihood Estimation

Def. 4.2 (最大尤度推定法 (Maximum Likelihood Estimation)): 最尤法とは、統計学において、与えられたデータからそれが従う確率分布の母数を点推定する方法である。最尤法が解く基本的な問題は「パラメータ θ が不明な確率分布 f_D に従う母集団から標本が得られたとき、データを良く説明する良い θ は何か」である。

[5] の経験的コスト関数の最小値 $\widehat{C}(\theta)$ により、$\theta^* = \arg\min_\theta \widehat{C}(\theta)$ は n 回の独立した測定に基づく最大尤度推定量であるという特異な統計的性質がある。これは、次の計算から明らかであり、対数の性質を利用すると、

$$\begin{aligned}
\theta^* &= \arg\min_\theta \widehat{C}(\theta) = \arg\max_\theta \frac{1}{n}\sum_{j=1}^{n} \ln p_\theta(x_j, z_j) \\
&= \arg\max_\theta \sum_{j=1}^{n} \ln p_\theta(x_j, z_j) = \arg\max_\theta \ln [9] \left(\prod_{j=1}^{n} p_\theta(x_j, z_j) \right) \\
&= \arg\max_\theta \prod_{j=1}^{n} p_\theta(x_j, z_j) = \arg\max_\theta p_\theta(X = \mathbf{x}, Z = \mathbf{z}) \\
&= \theta_{ML}
\end{aligned}$$

要するに、経験的クロスエントロピーとクルバック・ライブラー発散がコスト関数として人気があるのは、最大尤度法との関係によるものである。さらに、これらのコスト関数をニューラルネットワークに使用することで、二乗和コスト関数の場合よりもプラトーが少ないコスト面が得られ、ネットワークの学習時間が改善されることが期待されている。

[9] 対数の性質により、積の対数は和に変換される。
したがって、$\sum_{j=1}^{n} \ln p_\theta(x_j, z_j) = \ln p_\theta(x_1, z_1) + \ln p_\theta(x_2, z_2) + \cdots + \ln p_\theta(x_n, z_n) = \ln \left(\prod_{j=1}^{n} p_\theta(x_j, z_j) \right)$ となる。