



§ 3.10 Sample Estimation of Cost Function

PDF 版はここ↑

伊冉 (Andre YI) - 21122200512

2025 年 9 月 11 日

1 Mean Squared Error

Def. 1.1 (平均二乗誤差 (mean squared error)): ターゲット Z とネットワークの出力 $Y = f(X; \theta)$ の二乗差の期待値は、平均として推定できる：

$$\mathbb{E}[(Z - f(X; \theta))^2] \approx \frac{1}{N} \sum_{j=1}^N (z_j - f(x_j; \theta))^2.$$

$\mathbb{E}[(Z - f(X; \theta))^2]$ の値が小さいほどモデルの予測精度が高いと言える。その近似を取る理由はデータは有限個からである、大数の法則に従えば、 N が十分に大きくなると、サンプル平均は期待値に収束するからである。

2 Quadratic Renyi Entropy

Def. 2.1 (二次 Renyi エントロピー (Quadratic Renyi Entropy)): この推定では Parzen window 法を使用する。まず密度関数 $p(x)$ を window W_σ を用いたサンプルベース密度関数で置き換える：

$$\hat{p}(x) = \frac{1}{N} \sum_{k=1}^N W_\sigma(x, x_k).$$

簡単のため、window を一次元ガウシアンと仮定する：

$$W_\sigma(x, x_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_k|^2}.$$

二次ポテンシャルエネルギー $U(p) = \int p(x)^2 dx$ を考える。二次 Renyi エントロピーは $H_2(p) = -\ln \int p(x)^2 dx = -\ln U(p)$ であるため、 $U(p)$ を推定すれば十分である。推定は以下のように与えられる：

$$\begin{aligned} U(\hat{p}) &= \int \hat{p}(x)^2 dx = \int \hat{p}(x)\hat{p}(x)dx \\ &= \int \frac{1}{N} \sum_{k=1}^N W_\sigma(x, x_k) \frac{1}{N} \sum_{j=1}^N W_\sigma(x, x_j) dx \\ &= \frac{1}{N^2} \sum_{k=1}^N \sum_{j=1}^N \int W_\sigma(x, x_k) W_\sigma(x, x_j) dx. \end{aligned}$$

ここで、データサンプルは有限なので、 N は有限のサンプル数を表し、積分の線形性により、その積分と求和を交換することができる。

window がガウシアンの場合、 $W_\sigma(x, x') = \phi_\sigma(x - x')$ であり、 $\phi_\sigma(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}$ である。変数変換により畳み込みに変換することで、前の積分を明示的に計算できる：

$$\begin{aligned} \int W_\sigma(x, x_k) W_\sigma(x, x_j) dx &= \int \phi_\sigma(x - x_k) \phi_\sigma(x - x_j) dx \\ &= \int \phi_\sigma(t) \phi_\sigma(t - (x_j - x_k)) dt \\ &= (\phi_\sigma * \phi_\sigma)(x_j - x_k) \end{aligned}$$

ここで、 $(\phi_\sigma * \phi_\sigma)(x_j - x_k)$ はガウシアンの畳み込み (convolution)¹を表す。正規分布の和の再生性²を利用すると、以下の式を得る：

$$(\phi_\sigma * \phi_\sigma)(x_j - x_k) = W_{\sigma\sqrt{2}}(x_j, x_k)$$

最後の等式では、ガウシアンとそれ自身の畳み込みがスケールされたガウシアンになることを用いた。次に二次ポテンシャルエネルギーに代入すると、以下の推定が得られる：

$$U(\hat{p}) = \frac{1}{N^2} \sum_{k=1}^N \sum_{j=1}^N W_{\sigma\sqrt{2}}(x_j, x_k).$$

したがって、二次 Renyi エントロピーの推定は以下のように与えられる：

$$H_2(\hat{p}) = -\ln \left(\frac{1}{N^2} \sum_{k=1}^N \sum_{j=1}^N W_{\sigma\sqrt{2}}(x_j, x_k) \right)$$

3 Integrated squared error

Def. 3.1 (積分二乗誤差 (Integrated Squared Error)): p_Z と p_Y がそれぞれ目標密度関数と出力密度関数を表すとき、コスト関数

$$C(p_Z, p_Y) = \int |p_Z(u) - p_Y(u)|^2 du$$

は二次ポテンシャルエネルギーを用いて次のように書ける：

$$\begin{aligned} C(p_Z, p_Y) &= \int p_Z(u)^2 du + \int p_Y(u)^2 du - 2 \int p_Z(u)p_Y(u) du \\ &= U(p_Z) + U(p_Y) - 2 \int p_Z(u)p_Y(u) du. \end{aligned}$$

ここで $U(p)$ は前に定義された密度 p の二次ポテンシャルエネルギーを表す。そこで推定は次の形をとる：

¹ 一般に、独立な 2 つの確率変数 X_1, X_2 の確率密度関数を $p_1(x), p_2(x)$ とすると、その和 $X_1 + X_2$ の確率密度関数 $p(x)$ は、その畳み込み (convolution) $p_1 * p_2(x)$ 、i.e.、 $p(x) = p_1 * p_2(x) = \int_{-\infty}^{\infty} p_1(y)p_2(x-y) dy$ となる、その証明は A.1 を参照してください。

² $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$ とするとき、 $a_1 X_1 + a_2 X_2 \sim N(a_1 \mu_1 + a_2 \mu_2, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2)$ となることが知られています。これを、正規分布の再生性といいます。

$$C(\hat{p}_Z, \hat{p}_Y) = U(\hat{p}_Z) + U(\hat{p}_Y) - 2 \int \hat{p}_Z(u) \hat{p}_Y(u) du.$$

上の式によって、その積分項の計算を行えばよい。これは *Renyi* クロスエントロピーとも呼ばれる。したがって、以下のようになる：

$$\begin{aligned} \int \hat{p}_Z(u) \hat{p}_Y(u) du &= \int \frac{1}{N} \sum_{j=1}^N W_\sigma(u, z_j) \frac{1}{N'} \sum_{k=1}^{N'} W_\sigma(u, y_k) du \\ &\stackrel{N \leq \infty}{=} \frac{1}{NN'} \sum_{j=1}^N \sum_{k=1}^{N'} \int W_\sigma(u, z_j) W_\sigma(u, y_k) du \\ &= \frac{1}{NN'} \sum_{j=1}^N \sum_{k=1}^{N'} W_{\sigma\sqrt{2}}(z_j, y_k) \\ &= \frac{1}{NN'} \sum_{j=1}^N \sum_{k=1}^{N'} W_{\sigma\sqrt{2}}(z_j, f(x_k; \theta)), \end{aligned}$$

ここで $y = f(x; \theta)$ はニューラルネットの入力-出力写像である。したがって、次の推定を得る：

$$\begin{aligned} C(p_Z, \hat{p}_Y) &= \frac{1}{N^2} \sum_{j=1}^N \sum_{k=1}^N W_{\sigma\sqrt{2}}(z_j, z_k) + \frac{1}{N'^2} \sum_{j=1}^{N'} \sum_{k=1}^{N'} W_{\sigma\sqrt{2}}(f(x_j; \theta), f(x_k; \theta)) \\ &\quad - \frac{2}{NN'} \sum_{j=1}^N \sum_{k=1}^{N'} W_{\sigma\sqrt{2}}(z_j, f(x_k; \theta)). \end{aligned}$$

4 Maximum Mean Discrepancy(MMD)

Def. 4.1 (最大平均乖離 (MMD)): maximum mean discrepancy は 2 つの分布間の差異を測る指標です。これは 2 つの分布に対応する期待値間の差異を表す関数空間上の上界です。すべての実用的な目的において、確率変数 X は分布 $p(x)$ から無作為に抽出された n 個の観測値 x_1, \dots, x_n のサンプルから知られる、 $X \stackrel{i.i.d.}{\sim} p(x)$ 。同様に、確率変数 Y は分布 $q(y)$ から無作為に抽出された m 個の観測値 y_1, \dots, y_m のサンプルから知られる、 $Y \stackrel{i.i.d.}{\sim} q(y)$ 。期待値は次のように不偏平均として推定される：

$$\mathbb{E}_{X \sim p}[\phi(X)] \approx \frac{1}{n} \sum_{i=1}^n \phi(x_i),$$

$$\mathbb{E}_{Y \sim q}[\phi(Y)] \approx \frac{1}{m} \sum_{i=1}^m \phi(y_i),$$

次に、大数の法則によって、

$$\frac{1}{n} \sum_{i=1}^n \phi(x_i) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{X \sim p}[\phi(X)], \quad \frac{1}{m} \sum_{i=1}^m \phi(y_i) \xrightarrow[m \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{Y \sim q}[\phi(Y)]$$

そして p と q の間の最大平均乖離は、前の 2 つのサンプルを用いて次のように推定できる：

$$\begin{aligned}
d_{MMD}(p, q) &= \|\mathbb{E}_{X \sim p}[\phi(X)] - \mathbb{E}_{Y \sim q}[\phi(Y)]\|^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{i=1}^m \phi(y_i) \right\|^2 \\
&= \left(\frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{i=1}^m \phi(y_i) \right)^T \left(\frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{i=1}^m \phi(y_i) \right) \\
&= \frac{1}{n^2} \sum_{i,j} K(x_i, x_j) + \frac{1}{m^2} \sum_{i,j} K(y_i, y_j) - \frac{2}{mn} \sum_{i,j} K(x_i, y_j),
\end{aligned}$$

ここで、カーネル記法 $K(x, y) = \phi(x)^T \phi(y)$ を用いた。

Appendix A

証明問題

A.1

畳み込みの証明

Ex 1: 畳み込みの証明

$W_{\sigma(x,x_k)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_k|^2} \sim \mathcal{N}(x_k, \sigma^2)$, $W_{\sigma}(x, x_j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_j|^2} \sim \mathcal{N}(x_k, \sigma^2)$ とし、window がガウシアンの場合、 $W_{\sigma}(x, x') = \phi_{\sigma}(x - x')$ であり、 $\phi_{\sigma}(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{2\sigma^2}}$ である。
 $\int \phi_{\sigma}(x - x_k) \phi_{\sigma}(x - x_j) dx \stackrel{t=x-x_k}{=} \int \phi_{\sigma}(t) \phi_{\sigma}(t - (x_j - x_k)) dt = (\phi_{\sigma} * \phi_{\sigma})(x_j - x_k)$ はなぜ $x_j - x_k$ の確率密度関数になるのか？

解答

Proof. 確率変数 (Random Variance) W の値を $x_j - x_k$, 確率変数 (R.V.) X の値を $x - x_k$, 確率変数 (R.V.) Y の値を $x_j - x$, W の密度関数を $f_W(w)$ とすると、 $W = ((x - x_k) + (x_j - x)) = X + Y$ となる。したがって、

$$F_W(w) = P(W \leq w) = \int_{-\infty}^w f_W(w) dw \implies f_W(w) = \frac{d}{dw} F_W(w)$$

$f_W(w)$ は微積分学の基本原理を用いて $F_W(w)$ を微分する式である。

次に、 $F_W(w)$ は同時確率密度関数、 X と Y は独立な確率変数であるため、 $F_W(w) = P(X + Y \leq w) = \iint_{\{x+y \leq w\}} f_{XY}(x, y) dx dy \stackrel{\text{indep.}}{=} \iint_{\{x+y \leq w\}} f_X(x) f_Y(y) dx dy$ となる。 $w \in \mathbb{R}$ であるので、一番外の y の積分の範囲は $-\infty$ から ∞ までであり、内の x の積分の範囲は $x \leq w - y$ により、 $-\infty$ から $w - y$ までである。

したがって、以下のように書き換えることができる：

$$F_W(w) = \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{w-y} f_X(x) dx \right) dy$$

ここで、 $z = x + y$ とおくと、 $-\infty < z - y \leq w - y \implies -\infty < z \leq w$ となる。一方で、 $\frac{dx}{dz} = 1$ である。
 以下のように書ける：

$$\int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{w-y} f_X(x) dx \right) dy = \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^w f_X(z-y) dz \right) dy$$

したがって、

$$\begin{aligned}
 f_W(w) &= \frac{d}{dw} F_W(w) = \int_{-\infty}^{\infty} f_Y(y) \left[\frac{d}{dw} \int_{-\infty}^w f_X(z-y) dz \right] dy \\
 &= \int_{-\infty}^{\infty} f_Y(y) f_X(w-y) dy \\
 &= \int_{-\infty}^{\infty} \phi_{\sigma}(x_j - x) \phi_{\sigma}(x - x_k) dx \\
 &= \int_{-\infty}^{\infty} \phi_{\sigma}(x - x_k) \phi_{\sigma}(x - x_j) dx \quad \Leftarrow (\text{正規分布の密度関数は偶関数なので}) \\
 &= (\phi_{\sigma} * \phi_{\sigma})(x_j - x_k) = f_{x+y}(w)
 \end{aligned}$$

ここで、 $t = x - x_k$ とおくと、 $\frac{dt}{dx} = 1$ であるため、 $f_W(t) = \int_{-\infty}^{\infty} \phi_{\sigma}(t) \phi_{\sigma}(t - (x_j - x_k)) dt = (\phi_{\sigma} * \phi_{\sigma})(x_j - x_k)$ となる。i.e. $(\phi_{\sigma} * \phi_{\sigma})(x_j - x_k) \sim \mathcal{N}(x_j - x_k, 2\sigma^2)$

したがって、 $(\phi_{\sigma} * \phi_{\sigma})(x_j - x_k) = \phi_{\sigma\sqrt{2}}(x_j - x_k) = W_{\sigma\sqrt{2}}(x_j, x_k)$

■

A.2 推定関数は確率密度関数の証明

Ex 2: 推定関数は確率密度関数の証明

$W_{\sigma}(x, x_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_k|^2}$ とし、その推定関数を $\hat{p}(x) = \frac{1}{N} \sum_{k=1}^N W_{\sigma}(x, x_k)$ とする。 $\hat{p}(x)$ は密度関数になることを証明せよ。

解答

Proof. 非負化と正規化から証明します。

非負化の証明:

$W_{\sigma}(x, x_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_k|^2} \sim \mathcal{N}(x_k, \sigma^2)$ なので、正規分布の密度関数の定義によって、 $W_{\sigma}(x, x_k) \geq 0$ が成り立つことがわかる。したがって、 $\hat{p}(x) = \frac{1}{N} \sum_{k=1}^N W_{\sigma}(x, x_k) \geq 0$ となる。

正規化の証明:

$$\int_{-\infty}^{\infty} \hat{p}(x) dx = \int_{-\infty}^{\infty} \left(\frac{1}{N} \sum_{k=1}^N W_{\sigma}(x, x_k) \right) dx \stackrel{N \leq \infty}{=} \frac{1}{N} \sum_{k=1}^N \int_{-\infty}^{\infty} W_{\sigma}(x, x_k) dx$$

ここで、 $\int_{-\infty}^{\infty} W_{\sigma}(x, x_k) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}|x-x_k|^2} dx = 1$ であるので、 $\int_{-\infty}^{\infty} \hat{p}(x) dx = \frac{1}{N} \sum_{k=1}^N \int_{-\infty}^{\infty} W_{\sigma}(x, x_k) dx =$

$\frac{1}{N} \cdot N = 1$ となる。*i.e.* $\int_{-\infty}^{\infty} \hat{p}(x)dx = 1$ である。

以上より、推定関数 $\hat{p}(x)$ は確率密度関数の性質を満たすことがわかる。 ■