



補足資料 情報理論



松尾・岩澤研究室

MATSUO-IWASAWA LAB UTOKYO

許諾なく撮影や第三者
への開示を禁止します

• 目的

- ニューラルネットワークで重要となる情報理論の基礎の中で、「確率分布の乖離」に関する内容を理解する.

• 目標

- クロスエントロピーの定義と考え方について説明できる.
- 多値分類におけるクロスエントロピー誤差の式の導出を説明できる.
- KLダイバージェンスについて説明できる.
- KLダイバージェンスとエントロピーの関係をつかむ.
- JSダイバージェンスについて説明できる.

・クロスエントロピーの定義

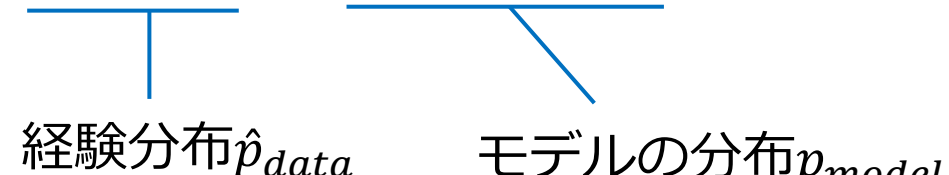
- 二つの確率分布 (p, q) がどれくらい離れているかを表す尺度

$$H(p, q) = - \sum_i p(x_i) \log q(x_i)$$

・クロスエントロピーの活用場面

(例) 識別モデルの目的関数の構築 (経験分布 \hat{p}_{data} とモデルの分布 p_{model} の違い)

$$-\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim \hat{p}_{data}} \log p_{model}(\mathbf{y} | \mathbf{x})$$



経験分布 \hat{p}_{data} モデルの分布 p_{model}

• K クラス分類のクロスエントロピー誤差 (Cross Entropy Error)

- 各クラス k ($1 \leq k \leq K$) に関する予測確率 \hat{y}_k と正解ラベル y_k の「近さ」の総和

- 予測確率 \hat{y}_k と正解ラベル y_k はいずれも「 K 次元ベクトル」

- 正解ラベル y_k はワンホットベクトル

$$\begin{array}{ll} \text{予測確率(推論結果)} & \text{正解ラベル)} \\ \hat{\mathbf{y}} = \begin{Bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_K \end{Bmatrix} = p_{\text{model}}(\hat{\mathbf{y}}|\mathbf{x}) & \mathbf{y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{Bmatrix} = \hat{p}_{\text{data}}(\mathbf{y}|\mathbf{x}) \end{array}$$

- $-\mathbb{E}_{\mathbf{x}, \mathbf{y} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\mathbf{y}|\mathbf{x})$ に代入すると以下が得られる

$$K: \text{クラス数} \quad E(\mathbf{y}, \hat{\mathbf{y}}) \equiv - \sum_{k=1}^K y_k \log \hat{y}_k$$

- N 個のデータで学習する場合に拡張すると

$$E(\mathbf{y}, \hat{\mathbf{y}}) \equiv - \sum_{n=1}^N \sum_{k=1}^K y_{nk} \log \hat{y}_{nk}$$

• ダイバージェンス

- 2つの確率分布がどれだけ異なるかを定量的に測る指標.
- カルバック・ライブラー(KL)ダイバージェンス:
 - 常に非負であり, 2つの分布が同一の場合のみ0になる.
- ジェンセン・シャノン(JS)ダイバージェンス:
 - 常に0以上1以下の値をとり, 2つの分布が同一の場合のみ0になる.
 - KLダイバージェンスに比べて**平滑性・対称性・有界性**の特長がある.

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right] = \mathbb{E}_{x \sim P} [\log P(x) - \log Q(x)]$$

- 相対エントロピーとも呼ばれる.
- 常に非負であり, 2つの分布が同一の場合のみ0になる.
- ユークリッド距離などとは違い, 対称性はない
- (P から Q へのKLダイバージェンスと Q から P へのKLダイバージェンスは一致しない)

- 離散確率変数 x にKLダイバージェンスの式を適用すると

$$\begin{aligned} D_{KL}(P \parallel Q) &= \mathbb{E}_{x \sim P}[\log P(x) - \log Q(x)] \\ &= \sum_{x \in D_X} P(x) \{\log P(x) - \log Q(x)\} \\ &= \sum_{x \in D_X} P(x) \log P(x) - \sum_{x \in D_X} P(x) \log Q(x) \\ &= -H(P) + H(P, Q) \end{aligned}$$

$$D_{JS}(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

$$M = \frac{1}{2}(P + Q) \text{ は } P, Q \text{ の混合分布}$$

- 常に0以上1以下の値をとり, 2つの分布が同一の場合のみ0になる.
- KLダイバージェンスに比べて平滑性・対称性・有界性の特長がある.
 - 平滑性: 極端な確率の差に対してより穏やかに反応する.
 - 対称性: 2つの分布の順序に依存しない.
 - 有界性: 常に定義され, 有限の値を取る.

