定式化チュートリアル 1

西山大輝

2022/05/09

背景情報

- ある分布からの i.i.d サンプルで構成される、IMDB データセットと AG's ~データセットの 2 種類
- ワードの種類の数を d とする。
  - IMDB

$$\mathcal{D}_{\mathrm{IMDB}} = \{\boldsymbol{X}_i, Y_i\}_{i=1}^{N}, \tag{1}$$

    ここで $N = 50000$ であり、$\boldsymbol{X}_i \in \mathbb{R}^{L\times d}$ を長さ L の映画のレビュー文を意味する確率変数、$Y_i \in \{0, 1\}$ をラベルを意味する確率変数とする。
  - AG's ~

$$\mathcal{D}_{\mathrm{AG}} = \{\boldsymbol{X}_i, Y_i\}_{i=1}^{M}, \tag{2}$$

    ここで $M = 196000$ であり、$\boldsymbol{X}_i \in \mathbb{R}^{L\times d}$ を長さ L のニュース記事を意味する確率変数、$Y_i \in \{0, 1, 2, 3\}$ をラベルを意味する確率変数とする。

登場人物

- 学習者
- 敵対者

学習者

- ２つのモデルそれぞれに対して、次のモデル f を所持

$$f(\boldsymbol{X}|\Theta) = \underset{k\in\{0,\cdots,K\}}{\arg\max}\ (P(Y|\boldsymbol{X}, \Theta)_k) \tag{3}$$

  ここで $\boldsymbol{X}$ はモデルに入力される単語列/文字列、Θ はモデルのパラメータである。なお K は、IMDB に対して $K = 1$、AG's ~に対して $K = 3$ である。
- 目標は、モデル f に対する敵対的サンプルの分布 $A(\boldsymbol{X}|f_\Theta)$ の構築。
- 次の予測精度の最大化

$$\sup_f \inf_{A\in A(\epsilon)} E_{\mathrm{ACC}}(f, A) = \sup_f \inf_{A\in A(\epsilon)} \mathbb{P}[f(A(\boldsymbol{X}|f_\Theta)) \neq f(\boldsymbol{X}|\Theta)] \tag{4}$$

  ただし ϵ は摂動

敵対者

- ２つのモデルそれぞれに対して、学習者が構築したモデル f_Θ と予測対象となるデータ $(\boldsymbol{x}, y)$ を所持している
- 摂動 ϵ は小さい値とし、次の制限が課せられる

$$\|\boldsymbol{x} - A(\boldsymbol{x}|f_\Theta)\| \leq \epsilon \tag{5}$$

- 目標はモデルの出力に対する予測を変えさせ、その回数を最大化

$$f(A(\boldsymbol{x}|f_\Theta)|\Theta) \neq f(\boldsymbol{x}|\Theta) \tag{6}$$

- 攻撃の成功回数の割合を計算する関数を $\mathbb{P}$ として、

$$\sup_{A \in A(\epsilon)} \mathbb{P}[f(A(\boldsymbol{x}|f_\Theta)|\Theta) \neq f(\boldsymbol{x}|\Theta)] \tag{7}$$