text_based_adversarial

We introduce a general adversarial learning problem between two players, the *learner* who aims to train some classifier that detects adversarial data, and the *adversary* who seeks to modify their data in an attempt to evade detection by the classifier. Let $D \in \mathbb{R}^{n \times p}$ be a collection of data containing n samples of some p features. Let the class labels of the data be $\gamma \in \{0,1\}^n$ where,

$$\begin{cases} \gamma_i = 0 & \text{if } D_i \text{ is legitimate} \\ \gamma_i = 1 & \text{if } D_i \text{ is adversarial.} \end{cases}$$

For example, suppose the *learner* wishes to construct a classifier which detects spam emails, then $\gamma_i = 0$ would correspond to the i^{th} sample being a legitimate email, whereas $\gamma_i = 1$ would correspond to spam. We then introduce a second set of data $X \in \mathbb{R}^{m \times p}$ which may be modified by the *adversary*. The corresponding set of class labels for this data is given by $Y \in \{0,1\}^m$, although we can typically assume that $Y_i = 1 \, \forall \, i \in \{1, \dots, m\}$.

In the upper level, the *learner* seeks to find the optimal weights $w \in \mathbb{R}^p$ of some prediction function $\sigma : \mathbb{R}^p \times \mathbb{R}^p \to (0,1)$, defined as

$$\sigma(w, x) := \frac{1}{1 - e^{-w^T x}},$$

where $x \in \mathbb{R}^p$ is a sample of data, for example x might represent an email. We identify the optimal weights by minimising the logistic loss function $\mathcal{L} : \mathbb{R}^p \times \mathbb{R}^p \times \{0,1\} \to \mathbb{R}$, defined as

$$\mathcal{L}(w, x, y) := -y \log(\sigma(x, y)) + (1 - y) \log(1 - \sigma(w, x)),$$

where $y \in \{0,1\}$ is the class of x. The *learner*, in the upper-level, minimises the loss over both the static data, D, and the *adversary's* data, X.

In the lower-level, the *adversary* modifies their data to evade detection by optimising the loss function $\ell : \mathbb{R}^p \times \mathbb{R}^p \times \{0,1\}$ towards the opposite class. Assuming that $Y_i = 1 \, \forall \, i \in \{1, \dots, m\}$, then this is defined as

$$\ell(w, x, y) := \log(1 - \sigma(w, x)).$$

Finally, we introduce some constraints on the lower-level to ensure that the *adversary* does not change their so much that it loses its original message. We measure the cosine similarity between the data and its initial position and constrain this value to be greater than some $\epsilon \in (-1, 1)$,

$$g(x) = \delta - \frac{x \cdot x^0}{||x|| ||x^0||},$$

where x^0 is the original position of x. The complete bilevel optimisation problem

is then given as

$$\begin{aligned} \underset{w}{\text{minimise}}\ \underset{X}{\text{maximise}} \quad & \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(w, X_i, Y_i) + \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(w, D_i, \gamma_i) \\ \text{subject to} \quad & X \in \arg\min_{X} \begin{cases} \frac{1}{m}\sum_{i=1}^{m}\ell(w, X_i, Y_i) \\ \text{s.t.} \quad \delta - \frac{X_i \cdot X_i^0}{||X_i||||X_i^0||} \leq 0, \quad i = 1, \ldots, m \end{cases} \end{aligned}$$

Below, we introduce two applications of the adversarial model and their corresponding datasets. Both text-based datasets are embedded as vectors in the space $\mathbb{R}^{128}$ with Google's BERT [2].

spam_email

The spam_email dataset is constructed from the email corpora provided for the NIST Text Retrieval Conference [1]. Of these emails, 571 are spam and the remaining 1429 are legitimate. The goal is to train a classifier that detects spam emails while considering how an adversary might modify their spam emails in an attempt to evade detection.

fake_reviews

The fake_reviews datasets contains a collection of 2000 Amazon reviews for cellphones and their accessories [3]. However, 1269 of these are fake reviews generated by bots. The goal is to train a classifier to detect these fake reviews while considering how an adversary might modify them in an attempt to evade detection.

References

[1] Gordon Cormack. Trec 2006 spam track overview. In *Proc. Fifteenth Text REtrieval Conference (TREC-2006)*, 2006.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[3] Naveed Hussain, Hamid Turab Mirza, Ibrar Hussain, Faiza Iqbal, and Imran Memon. Spam review detection using the linguistic and spammer behavioral methods. *IEEE Access*, 8:53801–53816, 2020.