

ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА
ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ ТА ІНФОРМАТИКИ
КАФЕДРА ОБЧИСЛЮВАЛЬНОЇ МАТЕМАТИКИ

КУРСОВА РОБОТА

на тему:

АНАЛІЗ АТАК НА ЛІНІЙНІ МОДЕЛІ МАШИННОГО НАВЧАННЯ

студента III курсу
групи ПМп-31
Середовича Віктора

Науковий керівник:
доцент Ю.А.Музичук

Завідуючий кафедри
обчислювальної математики
проф.

Львів — 2020

Зміст

1	Вступ	2
1.1	Постановка задачі	2
1.2	Класифікація атак	3
2	Тренування моделі	4
2.1	Модифікована логістична регресія	4
3	Методи атак	6
3.1	Метод Швидкого Градієнту	6
3.2	Нецілеспрямовані атаки	8
4	Альтернативні рішення	10
5	Висновок	11
	Література	12

Розділ 1

Вступ

>**TODO** Машинне навчання та штучний інтелект активно використовується у різних областях нашого життя, та допомагає у вирішенні таких задач як розпізнавання дорожніх знаків, облич, визначення ризику захворювання та багато іншого. А з поширенням його використання, також збільшується і ризик нападів злоумисників на ці алгоритми, що може привести, до трагічних наслідків. Тому варто досліджувати тему нападів на алгоритми машинного навчання, та знати як захистити свою модель.

В межах теми цієї роботи будуть розглядатись різні атаки на лінійні моделі машинного навчання, та методи їх захисту.

1.1 Постановка задачі

Мета даної роботи полягає у тому, щоб дослідити ефективність атак на лінійні моделі машинного навчання, та визначити методи захисту від них.

Виходячи з мети, визначені завдання роботи:

- Практична реалізація та дослідження методів атак
- Визначення методів захисту

1.2 Класифікація атак

Нехай існує класифікатор $f(x) : x \rightarrow y$, де $x \in X, y \in Y$, який передбачає значення y для вхідного x . Метою змагального прикладу є знайти такий x^* , який знаходиться в околі x , але хибно класифікується класифікатором. Зазвичай максимальний рівень шуму який може бути в змагальному прикладі може бути не більше за певну L_p норму $\|x^* - x\|_p < \epsilon$, де $p = 1, 2, \infty$.

Нецілеспрямовані атаки (Untargeted Attacks)

Нецілеспрямованими атаками є атаки метою яких є внести в приклад x , який правильно передбачається як $f(x) = y$ незначний шум так, щоб класифікатор зробив хибне передбачення $f(x^*) \neq y$.

Цілеспрямовані атаки (Targeted Attacks)

Цілеспрямованими атаками називають такі атаки, метою яких є внести у вхідний приклад x такі пертурбації щоб класифікатор передбачив якийсь конкретний клас $f(x^*) = y^*$, де y^* є заданою ціллю $y^* \neq y$.

Атаки на закриту модель (White Box attacks) - описує такий сценарій коли в нападника є повний доступ до параметрів і градієнтів моделі.

Атаки на відкриту модель (Black Box attacks) - це атаки при яких в нападника нема доступу до параметрів моделі.

Розділ 2

Тренування моделі

В якості прикладу лінійного методу машинного навчання, на який будуть здійснюватись атаки, буде використовуватись модифікований алгоритм логістичної регресії для мультикласової класифікації.

2.1 Модифікована логістична регресія

Задача прикладу $x \in R^{n_x}$, знайти $\hat{y} = P(y = 1 \mid x), 0 \leq \hat{y} \leq 1$. Шукатимемо у вигляді $\hat{y} = \sigma(\omega^T x + b)$, де

- параметри $\omega \in R^{n_x}, b \in R$ - невідомі, потрібно знайти оптимальні для даної задачі
- $\sigma(z) = \frac{1}{1+e^{-z}}$ - сигмоїд

Для кожного прикладу з тренувального датасету потрібно обчислити:

- $z^{(i)} = \omega^T x^{(i)} + b$
- $y^{(i)} = \sigma^{(i)}(z^{(i)})$

де $\sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}}$, так щоб $\hat{y}^{(i)} \approx y^{(i)}$

Для кожного прикладу визначена функція втрати:

$$L(\hat{y}^{(i)}, y^{(i)}) = -y^{(i)} \log(\hat{y}^{(i)}) - (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

На всіх прикладах обчислюємо штрафну функцію як:

$$J(w, b) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)})$$

Задача полягає в тому щоб знайти параметри $w \in R_x^n, b \in R$ що мінімізують функцію $J(\omega, b)$ Для цього будемо використовувати градієнтний спуск

Розділ 3

Методи атак

3.1 Атака Шумом

3.2 FGSM

Першим методом атаки який буде розглядатись є методу швидкого градієнту (Fast Gradient Sign Method).

Однокроковий метод градієнтного спуску, FGSM, щоб знайти такий adversarial приклад x^* який максимизую функцію втрати $J(x^*, y)$, де J є функцією кросс-ентропії.

$$X^* = X + \epsilon \cdot \text{sign}(\Delta_x J(x, y)),$$

де $\Delta_x J(x, y)$

Ідея полягає в тому ..=.

$$z_i = \omega^T x^{(i)} + b \quad \hat{y}_i = a_i = \text{softmax}(z_i) \quad \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^c e^{z_k}}$$

Похідна softmax функції

Якщо $i = j$,

$$\begin{aligned} \frac{\partial \hat{y}_i}{\partial z_j} &= \frac{\partial \frac{e^{z_i}}{\sum_{k=1}^c e^{z_k}}}{\partial z_j} = \frac{e^{z_i} \sum_{k=1}^c e^{z_k} - e^{z_j} e^{z_i}}{(\sum_{k=1}^c e^{z_k})^2} = \\ &= \frac{e^{z_j}}{\sum_{k=1}^c e^{z_k}} \times \frac{(\sum_{k=1}^c e^{z_k} - e^{z_j})}{\sum_{k=1}^c e^{z_k}} = y_i(1 - y_j) \end{aligned}$$

Якщо $i \neq j$,

$$\frac{\partial \hat{y}_i}{\partial z_j} = \frac{\partial \frac{e^{z_i}}{\sum_{k=1}^c e^{z_k}}}{\partial z_j} = \frac{0 - e^{z_j} e^{z_i}}{\sum_{k=1}^c e^{z_k}} \times \frac{e^{z_i}}{\sum_{k=1}^c e^{z_k}} = -y_i y_j$$

Функція кросс-ентропії

$$\xi(y, x) = - \sum_{i=1}^c y_i \log(\hat{y}_i)$$

Функція кросс-ентропії

$$\begin{aligned} \frac{\partial \xi}{\partial z_i} &= - \sum_{j=1}^c \frac{\partial y_j \log(\hat{y}_j)}{\partial z_i} = - \sum_{j=1}^c y_j \frac{\partial \log(\hat{y}_j)}{\partial z_i} = \\ &= - \sum_{j=1}^c y_j \frac{1}{\hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial z_i} = - \frac{y_i}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_i} - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_i} = \\ &= - \frac{y_i}{\hat{y}_i} \hat{y}_i (1 - \hat{y}_i) - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} (-\hat{y}_i \hat{y}_j) = -y_i + y_i \hat{y}_i + \sum_{j \neq i} y_j \hat{y}_i = \\ &= -y_i + \sum_{j=1}^c y_j \hat{y}_i = \hat{y}_i - y_i \end{aligned}$$

, $i = 1, \dots, C$

Обчислення урядіенту "має бути вище"

$$\frac{\partial \xi}{\partial x_i} = \frac{\partial \xi}{\partial z_i} \frac{\partial z_i}{\partial x_i} = (\hat{y}_i - y_i) \frac{\partial (\omega^T x_i + b)}{\partial x_i} = (\hat{y}_i - y_i) x_i = dz \cdot x$$

Обчислення урядіенту "має бути вище"

Algorithm 1: I-FGSM атака

Input: Приклад x , значення цілі y , класифікатор f .

Output: Adversarial x^* .

Ініціалізація $x^* \leftarrow x$.

Result: Write here the result

initialization;

while $\hat{k}(x_i) = \hat{k}(x^*)$ **do**

$x^* = x^* - \alpha \cdot \text{sign}(\Delta_x J(x, y))$

end

3.3 I-FGSM

3.4 MI-FGSM

3.5 Нецілеспрямовані атаки

Основна частина даної роботи полягала у написанні програми. Нижче наводимо основний алгоритм її роботи, на мові C:

```
1  % ===== %
2  #include <stdio.h>;
3  int main()
4  {
5  printf("Hello, world!\n");
6  return 0;
7  }
8
```

[language=Python]

```
1  def fit(self, X, w, b, y, alpha, max_iters, predict_func):
2      # Check that X and y have correct shape
3      self.w = w
4      self.b = b
5
6      self.y_ = np.expand_dims(y.T, axis=1)
7      self.X_ = X.T
8
9      self.num_iters = 0
10     self.X_ = self._gradient_descent(self.X_, self.y_, self.w, self.b,
11                                     alpha, max_iters, predict_func)
12
13     def _cost_function(self, X, Y, A):
14         m = X.shape[0]
15         if m == 0:
16             return None
17
18         J = (1 / m) * np.sum(-Y * np.log(A) - (1 - Y) * np.log(1 - A))
19         return J
```

Algorithm 2: How to write algorithms

Result: Write here the result

initialization;

while *While condition* **do**

 instructions;

if *condition* **then**

 instructions1;

 instructions2;

else

 instructions3;

end

end

Розділ 4

Альтернативні рішення

Деякі дослідники пишуть свої роботи в програмах типу Microsoft Word. Але то не є труйово[1].

Розділ 5

Висновок

Дана робота містить значний мій вклад, і перевершує попередні досягнення в багатьох напрямках. Окрім того, даний напрямок досліджень має значні перспективи подальшого розвитку. (Особливо добре було б, якби хтось вирішив проблему кирилиці в listings).

Бібліографія

- [1] Вікіпідручник *Як написати курсову?* (http://uk.wikibooks.org/wiki/%D0%AF%D0%BA_%D0%B2%D1%87%D0%B8%D1%82%D0%B8%D1%81%D1%8C_%D0%BA%D1%80%D0%B0%D1%89%D0%B5%3F/%D0%9A%D1%83%D1%80%D1%81%D0%BE%D0%B2%D1%96)