Математичні Основи Штучних Нейронних Мереж

25 листопада 2024

Виконав: Захаров Дмитро Олегович[1]
Науковий керівник: Ігнатович Світлана Юріївна[2]

[1]Студент групи МП41 IV курсу (перший бакалаврський рівень), спеціальності 113 "Прикладна математика" освітньої програми "Прикладна математика".

[2]Доктор фіз.-мат. наук, професор кафедри прикладної математики.

План

Вступ

Опис типової задачі

Проблема

Сучасний розвиток інструментів зводить розв'язок задач машинного навчання до вибору архітектури моделі, функції втрати та метрик якості. Часто, опускається фундаментальне питання: чому ці архітектури взагалі працюють?

Опис типової задачі

Проблема

Сучасний розвиток інструментів зводить розв'язок задач машинного навчання до вибору архітектури моделі, функції втрати та метрик якості. Часто, опускається фундаментальне питання: чому ці архітектури взагалі працюють?

- На вхід подається певний набір даних $\mathcal{D}$. Найчастіше, це набір пар $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{1\leq n\leq N}$ (*supervised learning*).

Опис типової задачі

Проблема

Сучасний розвиток інструментів зводить розв'язок задач машинного навчання до вибору архітектури моделі, функції втрати та метрик якості. Часто, опускається фундаментальне питання: чому ці архітектури взагалі працюють?

- На вхід подається певний набір даних $\mathcal{D}$. Найчастіше, це набір пар $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{1 \leq n \leq N}$ (*supervised learning*).
- Ми віримо, що є певна інформація, яку ми хочемо здобути з цього набору. Ми інкапсулюємо цю інформацію у вигляді функції $f(\boldsymbol{x})$. Це і є **модель**.

Опис типової задачі

Проблема

Сучасний розвиток інструментів зводить розв'язок задач машинного навчання до вибору архітектури моделі, функції втрати та метрик якості. Часто, опускається фундаментальне питання: чому ці архітектури взагалі працюють?

- На вхід подається певний набір даних $\mathcal{D}$. Найчастіше, це набір пар $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{1 \leq n \leq N}$ (*supervised learning*).
- Ми віримо, що є певна інформація, яку ми хочемо здобути з цього набору. Ми інкапсулюємо цю інформацію у вигляді функції $f(\boldsymbol{x})$. Це і є **модель**.
- Функцію f ми маємо підібрати з певного класу $\mathcal{F}$ так, щоб за неї досягався певний мінімум ($\hat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{L}(\mathcal{D}|f)$).

Приклади

У попередніх наших роботах [6, 3, 2] ми, зокрема, досліджували задачу кібербезпеки біометричних даних:

Робота	Рік, Журнал	Модель	Набір даних
[6]	2023, Multimedia Tools and Applications (Springer)	$f : \mathcal{I} \to \{0,1\}^{128}$: Бінарний вектор характеристик	Набір зображень і ідентифікатор людей
[3]	2024, Computers & Security (Elsevier)	$f : \mathcal{I} \to \{0,1\}$: Класифікатор живності	Набір зображень і біт, чи фейкова людина
[2]	2024, Engineering Applications of AI (Elsevier)	$f : \mathcal{I} \to \mathcal{I}$: "Геш" значення фотографії людини	Набір зображень і ідентифікатор людей

Табл.: Приклади наших робіт з біометрії. $\mathcal{I}$ — множина зображень.

Приклади

У попередніх наших роботах [6, 3, 2] ми, зокрема, досліджували задачу кібербезпеки біометричних даних:

Робота	Рік, Журнал	Модель	Набір даних
[6]	2023, Multimedia Tools and Applications (Springer)	$f : \mathcal{I} \to \{0,1\}^{128}$: Бінарний вектор характеристик	Набір зображень і ідентифікатор людей
[3]	2024, Computers & Security (Elsevier)	$f : \mathcal{I} \to \{0,1\}$: Класифікатор живності	Набір зображень і біт, чи фейкова людина
[2]	2024, Engineering Applications of AI (Elsevier)	$f : \mathcal{I} \to \mathcal{I}$: "Геш" значення фотографії людини	Набір зображень і ідентифікатор людей

Табл.: Приклади наших робіт з біометрії. $\mathcal{I}$ — множина зображень.

Проте...

Усі ці задачі містять багатовимірні дані (вимірність ≥ 100000), які важко апроксимувати класичними методами. Отже, ми використовуємо **глибоке навчання**.

Параметризація моделі

Зауваження

Як на практиці має виглядати $\mathcal{F}$? Зауважимо — це не може бути щось на кшталт $L^2(\mathbb{R})$. Тому, ми **параметризуємо** модель параметрами $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$. Записуємо це як $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Параметризація моделі

Зауваження

Як на практиці має виглядати $\mathcal{F}$? Зауважимо — це не може бути щось на кшталт $L^2(\mathbb{R})$. Тому, ми **параметризуємо** модель параметрами $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$. Записуємо це як $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Example

Якщо ми віримо, що $f : \mathbb{R} \to \mathbb{R}$ — квадратична, то шукаємо f як:

$$f(x|\boldsymbol{\theta}) = \theta_2 x^2 + \theta_1 x + \theta_0, \quad \boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3.$$

Параметризація моделі

Зауваження

Як на практиці має виглядати $\mathcal{F}$? Зауважимо — це не може бути щось на кшталт $L^2(\mathbb{R})$. Тому, ми **параметризуємо** модель параметрами $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$. Записуємо це як $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Example

Якщо ми віримо, що $f : \mathbb{R} \to \mathbb{R}$ — квадратична, то шукаємо f як:

$$f(x|\boldsymbol{\theta}) = \theta_2 x^2 + \theta_1 x + \theta_0, \quad \boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) \in \mathbb{R}^3.$$

Example (Багатовимірна лінійна регресія)

Нехай $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{1 \leq n \leq N} \subset \mathbb{R}^d \times \mathbb{R}$. Моделлю може бути наступна лінійна функція

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \boldsymbol{w}^\top \boldsymbol{x} + \beta, \quad \boldsymbol{\theta} = (\boldsymbol{w}, \beta) \in \mathbb{R}^{d+1}.$$

Зауваження

Модель — це не завжди функція, що повертає скаляр/вектор. Модель може повертати і зображення/аудіо/репрезентацію тексту/ймовірністний розподіл.

Зауваження

Модель — це не завжди функція, що повертає скаляр/вектор. Модель може повертати і зображення/аудіо/репрезентацію тексту/ймовірністний розподіл.

Так чи інакше, ми маємо функцію втрати $\mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$, що вимірює, наскільки добре модель описує дані. Правило вибору параметрів:

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}).$$

Зауваження

Модель — це не завжди функція, що повертає скаляр/вектор. Модель може повертати і зображення/аудіо/репрезентацію тексту/ймовірністний розподіл.

Так чи інакше, ми маємо функцію втрати $\mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$, що вимірює, наскільки добре модель описує дані. Правило вибору параметрів:

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}).$$

Example (Багатовимірна лінійна регресія)

Нехай $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{1\le n\le N} \subset \mathbb{R}^d \times \mathbb{R}^r$. Тоді, ми можемо обирати $f(\boldsymbol{x}|\boldsymbol{W}, \theta) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\beta}$ таким чином, щоб $f(\boldsymbol{x}_n) \approx \boldsymbol{y}_n$ для всіх n. Тому, в якості функції втрати можна взяти:

Зауваження

Модель — це не завжди функція, що повертає скаляр/вектор. Модель може повертати і зображення/аудіо/репрезентацію тексту/ймовірністний розподіл.

Так чи інакше, ми маємо функцію втрати $\mathcal{L}(\mathcal{D}|\boldsymbol{\theta})$, що вимірює, наскільки добре модель описує дані. Правило вибору параметрів:

$$\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\mathcal{D}|\boldsymbol{\theta}).$$

Example (Багатовимірна лінійна регресія)

Нехай $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{1\leq n\leq N} \subset \mathbb{R}^d \times \mathbb{R}^r$. Тоді, ми можемо обирати $f(\boldsymbol{x}|\boldsymbol{W}, \boldsymbol{\theta}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\beta}$ таким чином, щоб $f(\boldsymbol{x}_n) \approx \boldsymbol{y}_n$ для всіх n. Тому, в якості функції втрати можна взяти:

$$\mathcal{L}(\mathcal{D}|\boldsymbol{W}, \boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \|f(\boldsymbol{x}_n|\boldsymbol{W}, \boldsymbol{\theta}) - \boldsymbol{y}_n\|^2 .$$

Візуалізація

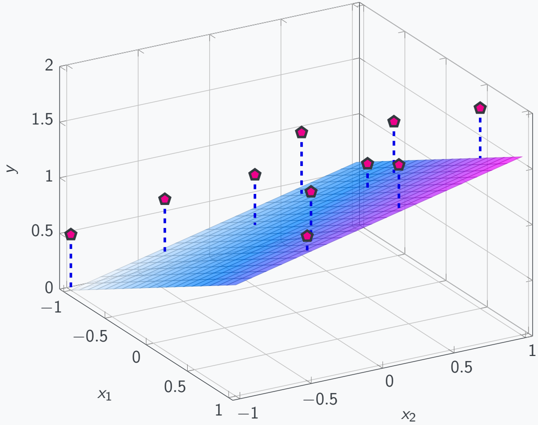


Рис.: Приклад багатовимірної лінійної регресії для $\boldsymbol{x}_n \in \mathbb{R}^2$, $y_n \in \mathbb{R}$. Ціль підібрати площину $f(\boldsymbol{x}|\beta, w_1, w_2) := \beta + w_1 x_1 + w_2 x_2$ так, щоб втрата $\mathcal{L}(\mathcal{D}|\beta, w_1, w_2) = \frac{1}{N}\sum_{n=1}^{N} \|f(\boldsymbol{x}_n) - y_n\|^2$ була мінімальною.

Проблеми машинного навчання

Що розв'язує машинне навчання?

Машинне навчання намагається розв'язати три основні проблеми:

1. **Оптимізація**: чи можна взагалі знайти $\hat{\theta}$? Які найкращі чисельні методи для цього?

Проблеми машинного навчання

Що розв'язує машинне навчання?

Машинне навчання намагається розв'язати три основні проблеми:

1. **Оптимізація**: чи можна взагалі знайти $\hat{\theta}$? Які найкращі чисельні методи для цього?
2. **Статистика**: як побудувати функцію втрати $\mathcal{L}$, щоб вона максимально відображала наші очікування від моделі?

Проблеми машинного навчання

Що розв'язує машинне навчання?

Машинне навчання намагається розв'язати три основні проблеми:

1. **Оптимізація**: чи можна взагалі знайти $\hat{\theta}$? Які найкращі чисельні методи для цього?
2. **Статистика**: як побудувати функцію втрати $\mathcal{L}$, щоб вона максимально відображала наші очікування від моделі?
3. **Апроксимація**: Ми хочемо зробити $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ як можна меншим. Отже, $f(\boldsymbol{x}|\boldsymbol{\theta})$ має описувати як можна більш широкий клас функцій.

Проблеми машинного навчання

Що розв'язує машинне навчання?

Машинне навчання намагається розв'язати три основні проблеми:

1. **Оптимізація**: чи можна взагалі знайти $\hat{\theta}$? Які найкращі чисельні методи для цього?
2. **Статистика**: як побудувати функцію втрати $\mathcal{L}$, щоб вона максимально відображала наші очікування від моделі?
3. **Апроксимація**: Ми хочемо зробити $\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ як можна меншим. Отже, $f(\boldsymbol{x}|\boldsymbol{\theta})$ має описувати як можна більш широкий клас функцій.

Зауваження

Сфокусуємось на третьому питанні, що і є темою нашої роботи. Отже, як побудувати влучну параметризацію?

Багатошарова Модель Персептронів

Сігмоїдальна Функція

Definition

Сігмоїдальною функцією (Сігмоїдом) $\sigma : \mathbb{R} \to \mathbb{R}$ називається функція, що задовольняє двом умовам:

$$\lim_{x \to +\infty} \sigma(x) = 1, \quad \lim_{x \to -\infty} \sigma(x) = 0.$$

Сігмоїдальна Функція

Definition

Сігмоїдальною функцією (Сігмоїдом) $\sigma : \mathbb{R} \to \mathbb{R}$ називається функція, що задовольняє двом умовам:

$$\lim_{x \to +\infty} \sigma(x) = 1, \quad \lim_{x \to -\infty} \sigma(x) = 0.$$

Example

Логістична функція $\sigma(x|\alpha) = 1/(1 + \exp(-\alpha x))$ є сігмоїдом.

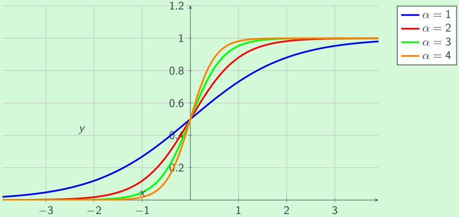


Рис.: Логістична функція з різними параметрами α.

Робота Цибенко (1989)

Апроксимація функцій лінійною комбінацією сігмоїдів

Робота Цибенко [1] присвячена на той час відомій апроксимації функції $f : \mathbb{R}^m \to \mathbb{R}$ за допомогою наступної суми:

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j), \quad \boldsymbol{w}_j \in \mathbb{R}^m, \quad \alpha_j, \beta_j \in \mathbb{R}.$$

- По суті, лінійна комбінація виразів $\{\sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)\}_{1 \leq j \leq n}$.
- n — кількість нейронів у прихованому шарі.
- Маємо рівно $(m+2)n$ параметрів.

Робота Цибенко (1989)

Апроксимація функцій лінійною комбінацією сігмоїдів

Робота Цибенко [1] присвячена на той час відомій апроксимації функції $f : \mathbb{R}^m \to \mathbb{R}$ за допомогою наступної суми:

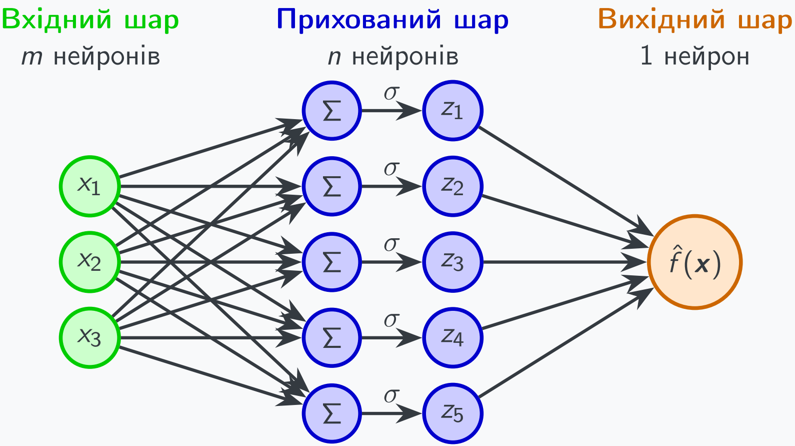
$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j), \quad \boldsymbol{w}_j \in \mathbb{R}^m, \quad \alpha_j, \beta_j \in \mathbb{R}.$$

- По суті, лінійна комбінація виразів $\{\sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)\}_{1 \leq j \leq n}$.
- n — кількість нейронів у прихованому шарі.
- Маємо рівно $(m+2)n$ параметрів.

Питання

Який клас функцій може апроксимувати така модель?

Візуалізація Архітектури Цибенко



Зауваження

Нейрон — це просто значення у графі обчислень.

Теорема Цибенко

- $\mathcal{Q}_m = [0, 1]^m$ є m-вимірним одиничним гіперкубом.

Теорема Цибенко

- $\mathcal{Q}_m = [0,1]^m$ є m-вимірним одиничним гіперкубом.
- $\mathcal{C}(\mathcal{Q}_m)$ — простір неперервних функцій $f : \mathcal{Q}_m \to \mathbb{R}$.

Теорема Цибенко

- $\mathcal{Q}_m = [0,1]^m$ є m-вимірним одиничним гіперкубом.
- $\mathcal{C}(\mathcal{Q}_m)$ — простір неперервних функцій $f : \mathcal{Q}_m \to \mathbb{R}$.
- Норма функції f на $\mathcal{C}(\mathcal{Q}_m)$: $\|f\|_{\mathcal{Q}_m} = \sup_{\boldsymbol{x} \in \mathcal{Q}_m} |f(\boldsymbol{x})|$.

Теорема Цибенко

- $\mathcal{Q}_m = [0,1]^m$ є m-вимірним одиничним гіперкубом.
- $\mathcal{C}(\mathcal{Q}_m)$ — простір неперервних функцій $f : \mathcal{Q}_m \to \mathbb{R}$.
- Норма функції f на $\mathcal{C}(\mathcal{Q}_m)$: $\|f\|_{\mathcal{Q}_m} = \sup_{\mathbf{x} \in \mathcal{Q}_m} |f(\mathbf{x})|$.

Theorem (Цибенко)

Нехай σ будь-яка неперервна сігмоїдальна функція. Суми вигляду $\hat{f}(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j \sigma(\mathbf{w}_j^\top \mathbf{x} + \beta_j)$ є щільними у $\mathcal{C}(\mathcal{Q}_m)$ та $L^1(\mathcal{Q}_m)$. Іншими словами, для будь-якої функції $f \in \mathcal{C}(\mathcal{Q}_m)$ та $\varepsilon > 0$, існує сума $\hat{f}(\mathbf{x})$ така, що:

1. $|\hat{f}(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$ для всіх $\mathbf{x} \in \mathcal{Q}_m$.

2. $\int_{\mathcal{Q}_m} |\hat{f}(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} < \varepsilon$.

Теорема Цибенко

- $\mathcal{Q}_m = [0,1]^m$ є m-вимірним одиничним гіперкубом.
- $\mathcal{C}(\mathcal{Q}_m)$ — простір неперервних функцій $f : \mathcal{Q}_m \to \mathbb{R}$.
- Норма функції f на $\mathcal{C}(\mathcal{Q}_m)$: $\|f\|_{\mathcal{Q}_m} = \sup_{\boldsymbol{x} \in \mathcal{Q}_m} |f(\boldsymbol{x})|$.

Theorem (Цибенко)

Нехай σ будь-яка неперервна сігмоїдальна функція. Суми вигляду $\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)$ є щільними у $\mathcal{C}(\mathcal{Q}_m)$ та $L^1(\mathcal{Q}_m)$. Іншими словами, для будь-якої функції $f \in \mathcal{C}(\mathcal{Q}_m)$ та $\varepsilon > 0$, існує сума $\hat{f}(\boldsymbol{x})$ така, що:

1. $|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})| < \varepsilon$ для всіх $\boldsymbol{x} \in \mathcal{Q}_m$.

2. $\int_{\mathcal{Q}_m} |\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})| d\boldsymbol{x} < \varepsilon$.

Висновок

$\sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)$ апроксимує довільну функцію на $\mathcal{C}(\mathcal{Q}_m)$.

Чи це все?

Питання

Нехай $\mathcal{P}_0, \ldots, \mathcal{P}_{C-1}$ — розбиття $\mathcal{Q}_m$ на C підмножин (що називають *класами*).

Чи це все?

Питання

Нехай $\mathcal{P}_0, \ldots, \mathcal{P}_{C-1}$ — розбиття $\mathcal{Q}_m$ на C підмножин (що називають *класами*).Нехай $f : \mathcal{Q}_m \to \{0, \ldots, C-1\}$ задана так:

$$f(\boldsymbol{x}) = j \iff \boldsymbol{x} \in \mathcal{P}_j.$$

Чи це все?

Питання

Нехай $\mathcal{P}_0, \ldots, \mathcal{P}_{C-1}$ — розбиття $\mathcal{Q}_m$ на C підмножин (що називають *класами*).Нехай $f : \mathcal{Q}_m \to \{0, \ldots, C-1\}$ задана так:

$$f(\boldsymbol{x}) = j \iff \boldsymbol{x} \in \mathcal{P}_j.$$

Чи може $\hat{f}(\boldsymbol{x}) := \sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)$ апроксимувати f?

Чи це все?

Питання

Нехай $\mathcal{P}_0, \ldots, \mathcal{P}_{C-1}$ — розбиття $\mathcal{Q}_m$ на C підмножин (що називають *класами*).Нехай $f : \mathcal{Q}_m \to \{0, \ldots, C-1\}$ задана так:

$$f(\boldsymbol{x}) = j \iff \boldsymbol{x} \in \mathcal{P}_j.$$

Чи може $\hat{f}(\boldsymbol{x}) := \sum_{j=1}^{n} \alpha_j \sigma(\boldsymbol{w}_j^\top \boldsymbol{x} + \beta_j)$ апроксимувати f?

Theorem (Цибенко про класифікатор)

Нехай σ будь-яка неперервна сігмоїдальна функція і функція f задана як вище. Тоді для будь-якої такої функції існує $\hat{f}$ та множина $\mathcal{D} \subseteq \mathcal{Q}_m$ така, що міра $\mu(\mathcal{D}) \geq 1 - \varepsilon$ та $|\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})| < \varepsilon$ для всіх $\boldsymbol{x} \in \mathcal{D}$.

Ілюстрація

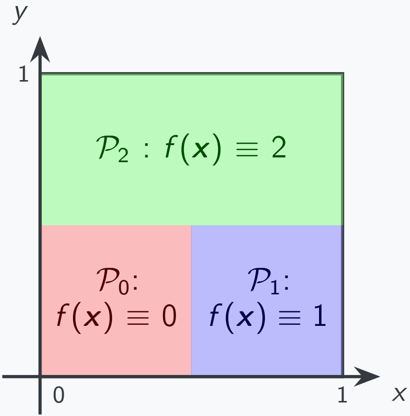


Рис.: Розбиття $\mathcal{Q}_2$ на три класи $\mathcal{P}_0, \mathcal{P}_1, \mathcal{P}_2$ (себто, $C = 3$).

Практична реалізація

Додаток

У курсовій роботі ми також написали програму, що для заданого бінарного розбиття $\mathcal{P}_0, \mathcal{P}_1$, знаходить класифікатор $\hat{f}$.

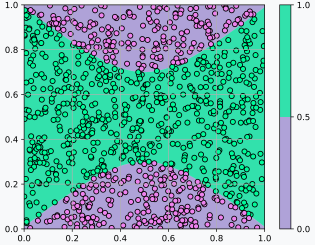


Рис.: Правильне розбиття $\mathcal{P}_0, \mathcal{P}_1$.

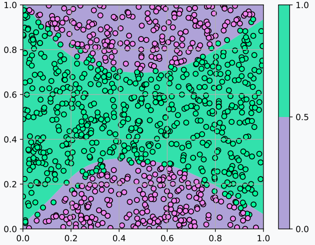


Рис.: Розбиття, знайдене класифікатором $\hat{f}$.

Узагальнення: MLP

1. Замість сігмоїду σ, використовуються інші **нелінійні** активаційні функції $\phi : \mathbb{R} \to \mathbb{R}$ (напр., $\phi(x) = \max\{0, x\}$).

Узагальнення: MLP

1. Замість сігмоїду σ, використовуються інші **нелінійні** активаційні функції $\phi : \mathbb{R} \to \mathbb{R}$ (напр., $\phi(x) = \max\{0, x\}$).
2. Замість двох шарів, може бути довільна кількість шарів.

Узагальнення: MLP

1. Замість сігмоїду σ, використовуються інші **нелінійні** активаційні функції $\phi : \mathbb{R} \to \mathbb{R}$ (напр., $\phi(x) = \max\{0, x\}$).
2. Замість двох шарів, може бути довільна кількість шарів.

Definition (Багатошарова модель персептронів (MLP))

Таким чином, узагальнена архітектура:

$$\boldsymbol{x}^{\langle j+1 \rangle} = \phi^{\langle j \rangle}(\boldsymbol{z}^{\langle j \rangle}), \quad \boldsymbol{z}^{\langle j \rangle} = \boldsymbol{W}^{\langle j \rangle}\boldsymbol{x}^{\langle j \rangle} + \boldsymbol{\beta}^{\langle j \rangle}, \quad j = 0, \dots, \ell - 1,$$

Таким чином, параметризація моделі є $\boldsymbol{\theta} = \left\{ \boldsymbol{W}^{\langle j \rangle}, \boldsymbol{\beta}^{\langle j \rangle} \right\}_{0 \le j \le \ell - 1}$.

Зауваження

Узагальнення: MLP

1. Замість сігмоїду σ, використовуються інші **нелінійні** активаційні функції $\phi : \mathbb{R} \to \mathbb{R}$ (напр., $\phi(x) = \max\{0, x\}$).
2. Замість двох шарів, може бути довільна кількість шарів.

Definition (Багатошарова модель персептронів (MLP))

Таким чином, узагальнена архітектура:

$$\boldsymbol{x}^{\langle j+1\rangle} = \phi^{\langle j\rangle}(\boldsymbol{z}^{\langle j\rangle}), \quad \boldsymbol{z}^{\langle j\rangle} = \boldsymbol{W}^{\langle j\rangle}\boldsymbol{x}^{\langle j\rangle} + \boldsymbol{\beta}^{\langle j\rangle}, \quad j = 0, \dots, \ell - 1,$$

Таким чином, параметризація моделі є $\boldsymbol{\theta} = \left\{\boldsymbol{W}^{\langle j\rangle}, \boldsymbol{\beta}^{\langle j\rangle}\right\}_{0 \le j \le \ell - 1}$.

Зауваження

У курсовій роботі, ми розглянули питання: (a) навіщо потрібно більше двох шарів, (b) які бувають узагальнення архітектури MLP та (c) навіщо інші активаційні функції.

Мережі Колмогорова-Арнольда

13 проблема Гільберта

Питання

Чи існують справжні неперервні функції від багатьох змінних?

13 проблема Гільберта

Питання

Чи існують справжні неперервні функції від багатьох змінних?

Перефразоване питання

Чи можна будь-яку неперервну функцію $f : \mathcal{Q}_m \to \mathbb{R}$ записати за допомогою суми та композицій $\phi_1, \ldots, \phi_N \in \mathcal{C}(\mathbb{R})$?

13 проблема Гільберта

Питання

Чи існують справжні неперервні функції від багатьох змінних?

Перефразоване питання

Чи можна будь-яку неперервну функцію $f : \mathcal{Q}_m \to \mathbb{R}$ записати за допомогою суми та композицій $\phi_1, \dots, \phi_N \in \mathcal{C}(\mathbb{R})$?

Example

$f(x, y) = 3x + 5y$. Якщо $\phi_1(x) = 3x$, $\phi_2 = 5y$, то $f(x, y) = \phi_1(x) + \phi_2(y)$.

13 проблема Гільберта

Питання

Чи існують справжні неперервні функції від багатьох змінних?

Перефразоване питання

Чи можна будь-яку неперервну функцію $f : \mathcal{Q}_m \to \mathbb{R}$ записати за допомогою суми та композицій $\phi_1, \ldots, \phi_N \in \mathcal{C}(\mathbb{R})$?

Example

$f(x, y) = 3x + 5y$. Якщо $\phi_1(x) = 3x$, $\phi_2 = 5y$, то $f(x, y) = \phi_1(x) + \phi_2(y)$.

Example

$f(x, y) = xy$. Оскільки $xy = \frac{(x+y)^2}{4} - \frac{(x-y)^2}{4}$, то якщо $\phi(x) = -x, \psi_+(x) = x^2/4, \psi_-(x) = -x^2/4$, то:

$$f(x, y) = \psi_+(x + y) + \psi_-(x + \phi(y)).$$

Теорема Колмогорова-Арнольда

Основна гіпотеза 13 проблеми Гільберта

Існує неперервна функція $f : \mathcal{Q}_3 \to \mathbb{R}$, що не може бути виражена як композиція та сума неперервних функцій $\phi_1, \ldots, \phi_N \in \mathcal{C}(\mathbb{R}^2)$.

Теорема Колмогорова-Арнольда

Основна гіпотеза 13 проблеми Гільберта

Існує неперервна функція $f : \mathcal{Q}_3 \to \mathbb{R}$, що не може бути виражена як композиція та сума неперервних функцій $\phi_1, \ldots, \phi_N \in \mathcal{C}(\mathbb{R}^2)$.

Definition (Теорема Колмогорова (1957, [5]))

Для будь-якого натурального $m \geq 2$, існують неперервні функції $\phi_{p,q} \in \mathcal{C}([0,1])$ такі, що для будь-якої функції $f \in \mathcal{C}(\mathcal{Q}_m)$ знайдуться неперервні функції $\Phi_1, \ldots, \Phi_{2m+1} \in \mathcal{C}(\mathbb{R})$ такі, що

$$f(x_1, \ldots, x_m) = \sum_{q=1}^{2m+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{p,q}(x_p) \right)$$

Мережа Колмогорова-Арнольда(KAN)

Зауваження

До роботи 2024 року [4], ідею такої репрезентації вважали недосяжною через "поганість" функцій $\phi_{p,q}$ та Φ_q.

Мережа Колмогорова-Арнольда(KAN)

Зауваження

До роботи 2024 року [4], ідею такої репрезентації вважали недосяжною через "поганість" функцій $\phi_{p,q}$ та Φ_q.

Definition (З'єднання KAN Мережі [4])

З'єднання KAN Мережі між шарами розміру n (вхід) та m (вихід) — це матриця $\mathbf{\Phi} = \{\phi_{q,p}\}_{1 \le p \le n, 1 \le q \le m}$, де кожна функція параметризована, а наступне значення активації $\boldsymbol{y} = \mathbf{\Phi} \circ \boldsymbol{x}$.

Мережа Колмогорова-Арнольда(KAN)

Зауваження

До роботи 2024 року [4], ідею такої репрезентації вважали недосяжною через "поганість" функцій $\phi_{p,q}$ та Φ_q.

Definition (З'єднання KAN Мережі [4])

З'єднання KAN Мережі між шарами розміру n (вхід) та m (вихід) — це матриця $\mathbf{\Phi} = \{\phi_{q,p}\}_{1\le p\le n, 1\le q\le m}$, де кожна функція параметризована, а наступне значення активації $\boldsymbol{y} = \mathbf{\Phi} \circ \boldsymbol{x}$.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,n} \\ \vdots & \ddots & \vdots \\ \phi_{m,1} & \cdots & \phi_{m,n} \end{bmatrix} \circ \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \triangleq \begin{bmatrix} \sum_{p=1}^{n} \phi_{1,p}(x_p) \\ \vdots \\ \sum_{p=1}^{n} \phi_{m,p}(x_p) \end{bmatrix}$$

Мережа Колмогорова-Арнольда(KAN), cont.

Definition (Архітектура KAN [4])

Мережа Колмогорова-Арнольда — це композиція ℓ з'єднань:

$$\hat{f}_{\mathrm{KAN}}(\boldsymbol{x}) = \boldsymbol{\Phi}^{\langle \ell-1 \rangle} \circ \cdots \circ \boldsymbol{\Phi}^{\langle 1 \rangle} \circ \boldsymbol{\Phi}^{\langle 0 \rangle} \circ \boldsymbol{x}.$$

Мережа Колмогорова-Арнольда(KAN), cont.

Definition (Архітектура KAN [4])

Мережа Колмогорова-Арнольда — це композиція ℓ з'єднань:

$$\hat{f}_{\mathrm{KAN}}(\boldsymbol{x}) = \boldsymbol{\Phi}^{\langle \ell-1 \rangle} \circ \cdots \circ \boldsymbol{\Phi}^{\langle 1 \rangle} \circ \boldsymbol{\Phi}^{\langle 0 \rangle} \circ \boldsymbol{x}.$$

Example (Формула Колмогорова)

Нехай $\boldsymbol{\Phi}^{\langle 0 \rangle} = \{\phi_{p,q}\}_{1\le p\le m, 1\le q\le 2m+1}$, $\boldsymbol{\Phi}^{\langle 1 \rangle} = \{\Phi_q\}_{1\le q\le 2m+1}$:

$$\hat{f}_{\mathrm{KAN}}(\boldsymbol{x}) = [\Phi_1, \ldots, \Phi_{2m+1}] \circ \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,m} \\ \vdots & \ddots & \vdots \\ \phi_{2m+1,1} & \cdots & \phi_{2m+1,m} \end{bmatrix} \circ \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

$$= [\Phi_1, \ldots, \Phi_{2m+1}] \circ \begin{bmatrix} \sum_{p=1}^{m} \phi_{1,p}(x_p) \\ \vdots \\ \sum_{p=1}^{m} \phi_{2m+1,p}(x_p) \end{bmatrix} = \sum_{q=1}^{2m+1} \Phi_q \left(\sum_{p=1}^{m} \phi_{q,p}(x_p) \right)$$

Візуалізація

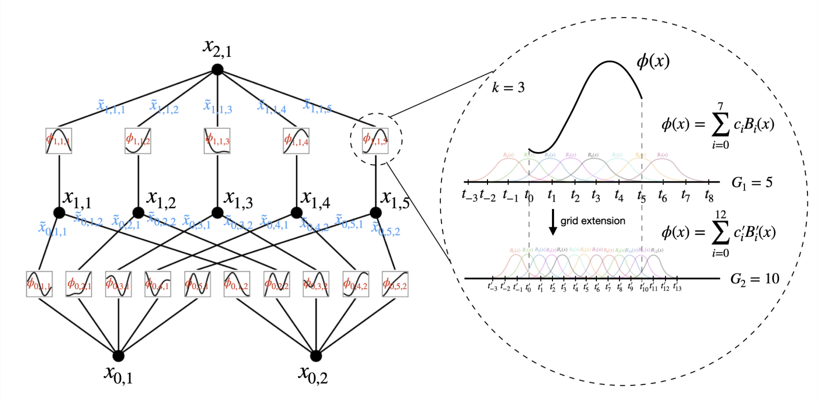


Рис.: Архітектура мережі Колмогорова-Арнольда з оригінальної роботи [4].

Література I

[1] G. Cybenko. "Approximation by superpositions of a sigmoidal function". В: *Mathematics of Control, Signals and Systems* 2.4 (1989), с. 303—314. DOI: 10.1007/BF02551274. URL: https://doi.org/10.1007/BF02551274.

[2] Oleksandr Kuznetsov, Dmytro Zakharov та Emanuele Frontoni. "Deep learning-based biometric cryptographic key generation with post-quantum security". В: *Multimedia Tools and Applications* 83.19 (2024), с. 56909—56938. DOI: 10.1007/s11042-023-17714-7. URL: https://doi.org/10.1007/s11042-023-17714-7.

Література II

[3] Oleksandr Kuznetsov та ін. "AttackNet: Enhancing biometric security via tailored convolutional neural network architectures for liveness detection". В: *Computers & Security* 141 (2024), с. 103828. ISSN: 0167-4048. DOI: https://doi.org/10.1016/j.cose.2024.103828. URL: https://www.sciencedirect.com/science/article/pii/S0167404824001299.

[4] Ziming Liu та ін. "Kan: Kolmogorov-arnold networks". В: *arXiv preprint arXiv:2404.19756* (2024).

[5] Kolmogorov A. N. "On the Representation of Continuous Functions of one Variable and Addition". В: *Doklady Akademii Nauk SSSR* 144 (1957), с. 679—681. URL: https://cir.nii.ac.jp/crid/1571980075616322176.

Література III

[6] Dmytro Zakharov, Oleksandr Kuznetsov та Emanuele Frontoni. "Unrecognizable yet identifiable: Image distortion with preserved embeddings". В: *Engineering Applications of Artificial Intelligence* 137 (2024), с. 109164. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2024.109164. URL: https://www.sciencedirect.com/science/article/pii/S0952197624013228.

Дякую за Вашу Увагу!

♥

Додаткові відомості

Definition (Міра)

Мірою μ називають невід'ємну σ-адитивну функцію множин, задана на півкільці $\mathcal{H}$:

- **Невід'ємна:** $\forall X \in \mathcal{H} : \mu(X) \geq 0$.
- σ**-адитивність:** $\forall \{X_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ таких, що $\{X_n\}_{n\in\mathbb{N}}$ є неперетинними та $\bigcup_{n\in\mathbb{N}} X_n \in \mathcal{H}$, справедливо:

$$\mu\left(\bigcup_{n\in\mathbb{N}} X_n\right) = \sum_{n\in\mathbb{N}} \mu(X_n).$$

Додаткові відомості

Definition (L^p простір)

L^p **простором** $(p \geq 1)$ над простором з мірою $(\Omega, \mathcal{F}, \mu)$ називають множину функцій $\mathcal{L}^p(\Omega, \mu)$, на яких інтеграл Лебега в p-ому степені модуля є скінченним:

$$\mathcal{L}^p(\Omega, \mu) = \left\{ f : \mathcal{F}\text{-вимірна} : \|f\|_p = \left(\int_\Omega |f|^p d\mu \right)^{1/p} < \infty \right\}.$$

Для $p = \infty$,
$\|f\|_\infty = \inf\{\gamma \in \mathbb{R}_{\geq 0} : |f(x)| \leq \gamma$ майже для всіх $x \in \Omega\}$.