

Математичні Основи Штучних Нейронних Мереж

19 листопада 2024

Виконав: Захаров Дмитро Олегович¹

Науковий керівник: Ігнатович Світлана Юріївна²

¹Студент групи МП41 IV курсу (перший бакалаврський рівень), спеціальності 113 “Прикладна математика” освітньої програми “Прикладна математика”.

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

План

- 1** Вступ: Задачі Глибокого Навчання
 - Приклади
 - Проблема параметризації
- 2** Багатошарова Модель Персептронів
 - Теорема Цибенко (1989)
 - Універсальність апроксимації класифікатора
- 3** Мережі Колмогорова-Арнольда
 - Історична довідка: 13 проблема Гільберта
 - Мережа Колмогорова-Арнольда

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?

- На вхід подається певний набір даних \mathcal{D} . Найчастіше, це набір пар $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{1 \leq n \leq N}$ (*supervised learning*).
- Ми віримо, що є певна інформація, яку ми хочемо здобути з цього набору. Ми інкапсулюємо цю інформацію у вигляді функції $f(\mathbf{x})$. 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)	$f : \mathcal{I} \rightarrow \{0, 1\}^{128}$: Бінарний вектор характеристик	Набір зображень і ідентифікатор людей
[3]	2024, Computers & Security (Elsevier)	$f : \mathcal{I} \rightarrow \{0, 1\}$: Класифікатор живності	Набір зображень і біт, чи фейкова людина
[2]	2024, Engineering Applications of AI (Elsevier)	$f : \mathcal{I} \rightarrow \mathcal{I}$: "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 $\theta \in \Theta \subset \mathbb{R}^n$. Записуємо це як $f(\mathbf{x}|\theta)$.

Example

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

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

Example (Багатовимірна лінійна регресія)

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

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

Зауваження

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

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

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

Example (Багатовимірна лінійна регресія)

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

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

Візуалізація

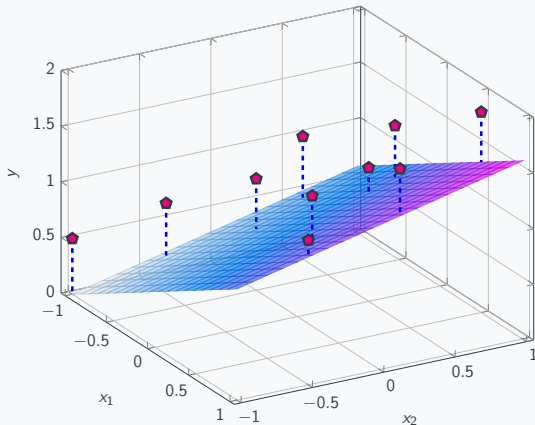


Рис.: Приклад багатовимірної лінійної регресії для $\mathbf{x}_n \in \mathbb{R}^2$, $y_n \in \mathbb{R}$. Ціль підібрати площину $f(\mathbf{x}|\beta, w_1, w_2) := \beta + w_1x_1 + w_2x_2$ так, щоб втрата $\mathcal{L}(\mathcal{D}|\beta, w_1, w_2) = \frac{1}{N} \sum_{n=1}^N \|f(\mathbf{x}_n) - y_n\|^2$ 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?
2. **Статистика:** як побудувати функцію втрати \mathcal{L} , щоб вона максимально відображала наші очікування від моделі?
3. **Апроксимація:** Ми хочемо зробити $\min_{\theta} \mathcal{L}(\theta)$ як можна меншим. Отже, $f(x|\theta)$ 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) $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ називається функція, що задовольняє двом умовам:

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

Example

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

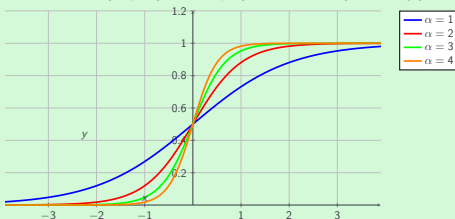


Рис.: Логістична функція з різними параметрами α .

Робота Цибенко (1989)

Апроксимація функцій лінійною комбінацією сігмоїдів

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

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

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

Питання

Який клас функцій може апроксимувати така модель?

Візуалізація Архітектури Цибенко

Вхідний шар

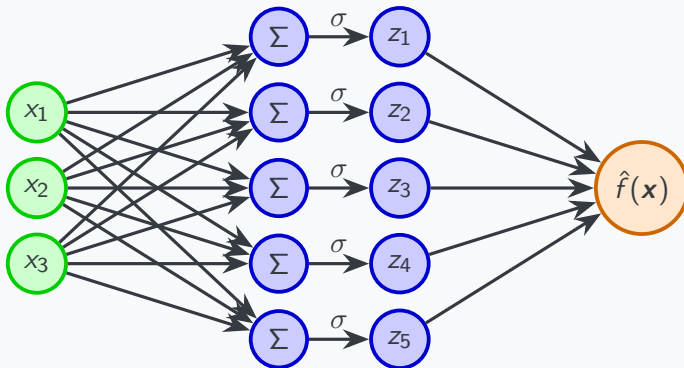
m нейронів

Прихований шар

n нейронів

Вихідний шар

1 нейрон



Theorem

Зауваження **Нейрон** — це просто значення у графі обчислень.

Теорема Цибенко

- $\mathcal{Q}_m = [0, 1]^m$ є m -вимірним одиничним гіперкубом.
- $\mathcal{C}(\mathcal{Q}_m)$ — простір неперервних функцій $f : \mathcal{Q}_m \rightarrow \mathbb{R}$.
- 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. $\int_{\mathcal{Q}_m} |\hat{f}(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} < \varepsilon$.

Висновок

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

Чи це все?

Питання

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

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

Чи може $\hat{f}(\mathbf{x}) := \sum_{j=1}^n \alpha_j \sigma(\mathbf{w}_j^\top \mathbf{x} + \beta_j)$ 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 $|\hat{f}(\mathbf{x}) - f(\mathbf{x})| < \varepsilon$ для всіх $\mathbf{x} \in \mathcal{D}$.

Ілюстрація

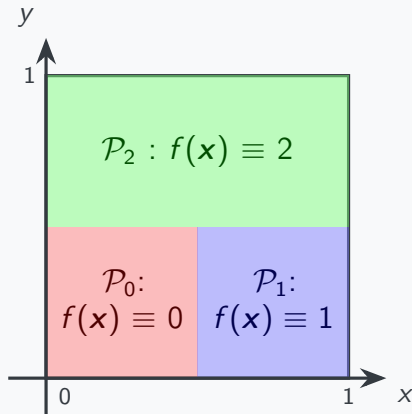


Рис.: Розбиття Q_2 на три класи P_0, P_1, P_2 (себто, $C = 3$).

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} .

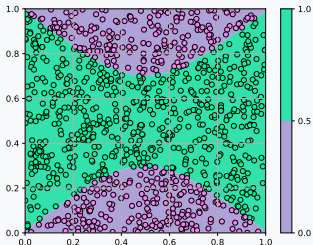


Рис.: Правильне розбиття $\mathcal{P}_0, \mathcal{P}_1$.

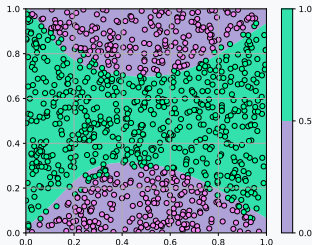


Рис.: Розбиття, знайдене класифікатором \hat{f} .

Узагальнення: MLP

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

Definition (Багатошарова модель персептронів (MLP))

Таким чином, узагальнена архітектура:

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

Таким чином, параметризація моделі є $\theta = \left\{ \mathbf{W}^{(j)}, \beta^{(j)} \right\}_{0 \leq j \leq \ell - 1}$.

Зауваження

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

Мережі Колмогорова-Арнольда

13 проблема Гільберта

Питання

Чи існують справжні неперервні функції від багатьох змінних?

Перефразоване питання

Чи можна будь-яку неперервну функцію $f : \mathcal{Q}_m \rightarrow \mathbb{R}$ 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
 $f(x, y) = \phi_1(x) + \phi_2(y)$.

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

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 $\Phi_1, \dots, \Phi_{2m+1} \in \mathcal{C}(\mathbb{R})$ такі, що

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

Мережа Колмогорова-Арнольда (KAN)

Зауваження

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

Definition (З'єднання KAN Мережі [4])

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

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

Мережа Колмогорова-Арнольда(KAN), cont.

Definition (Архітектура KAN [4])

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

$$\hat{f}_{\text{KAN}}(\mathbf{x}) = \Phi^{\langle \ell-1 \rangle} \circ \dots \circ \Phi^{\langle 1 \rangle} \circ \Phi^{\langle 0 \rangle} \circ \mathbf{x}.$$

Example (Формула Колмогорова)

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

$$\begin{aligned} \hat{f}_{\text{KAN}}(\mathbf{x}) &= [\Phi_1, \dots, \Phi_{2m+1}] \circ \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,m} \\ \vdots & \ddots & \vdots \\ \phi_{2m+1,1} & \cdots & \phi_{2m+1,m} \end{bmatrix} \circ \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \\ &= \sum_{q=1}^{2m+1} \Phi_q \left(\sum_{p=1}^m \phi_{q,p}(x_p) \right). \end{aligned}$$

Візуалізація

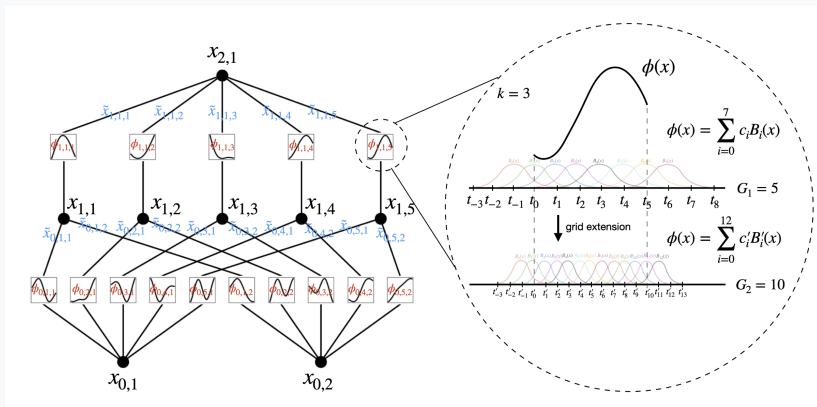


Рис.: Архітектура мережі Колмогорова-Арнольда з оригінальної роботи [4].

Література I

- [1] G. Cybenko. “Approximation by superpositions of a sigmoidal function”. B: *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”. B: *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!

