



Машинне навчання з малими даними: практичний підхід

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	Другий (магістерський)
Галузь знань	12 Інформаційні технології
Спеціальність	122 Комп'ютерні науки
Освітня програма	Цифрові технології в енергетиці
Статус дисципліни	Вибірковий
Форма навчання	Заочна
Рік підготовки, семестр	1 курс, 2 семестр
Обсяг дисципліни	4 кредитів (120 годин) 8 лек. 6 лаб. 106 СРС
Семестровий контроль/ контрольні заходи	Залік, МКР
Розклад занять	https://schedule.kpi.ua/
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: к.т.н., доц. Полягушко Любов Григорівна, ddpolytechnic2016@gmail.com Лабораторні: к.т.н., доц. Полягушко Любов Григорівна, ddpolytechnic2016@gmail.com
Розміщення курсу	Засоби Classroom та e-mail.

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Програму навчальної дисципліни «Машинне навчання з малими даними: практичний підхід» складено відповідно до освітньої програми освітньо-професійної програми «Цифрових технологій в енергетиці» другого рівня вищої освіти спеціальності 122 «Комп'ютерні науки».

Метою навчальної дисципліни є формування у студентів здатності:

- застосовувати методи машинного навчання для роботи з обмеженими наборами даних
- впроваджувати техніки підвищення ефективності моделей на малих даних
- проектувати та реалізовувати системи штучного інтелекту в умовах дефіциту навчальних прикладів

У результаті вивчення дисципліни студенти мають отримати:

ЗНАННЯ:

- сучасних методів машинного навчання, адаптованих для малих наборів даних
- технік попередньої обробки та аугментації даних
- методів регуляризації та ансамблювання моделей
- принципів роботи нейронних мереж та їх оптимізації
- підходів до генерації синтетичних даних
- методів передавального навчання та багатозадачного навчання

УМІННЯ:

- застосовувати методи попередньої обробки, аналізу та візуалізації малих наборів даних
- реалізовувати класичні та сучасні алгоритми машинного навчання
- проектувати архітектури нейронних мереж для специфічних задач
- використовувати техніки регуляризації для запобігання перенаванчання
- генерувати синтетичні дані для збільшення навчальних вибірок
- застосовувати методи передавального навчання для підвищення ефективності моделей
- оцінювати якість моделей за допомогою відповідних метрик

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

У структурно-логічній схемі навчальна дисципліна «Машинне навчання з малими даними: практичний підхід» розміщена тоді, коли студенти вже прослухали навчальні дисципліни «Обробка надвеликих масивів даних», «Технології розробки програмного забезпечення», що достатньо для виконання лабораторних робіт з даної дисципліни.

Отримані знання можна використовувати при виконанні магістерської дисертації.

3. Зміст навчальної дисципліни

Тема 1. Обробка та аналіз даних Завантаження та робота з даними. Пошук пропущених та аномальних значень. Аналіз класової рівності. Коефіцієнти кореляції та вибір ознак. Методи аугментації, масштабування та нормалізації даних.

Тема 2. Основи машинного навчання Класичні методи машинного навчання: лінійна та логістична регресія, наївний байєсівський класифікатор, k-найближчих сусідів, дерева рішень, метод опорних векторів. Методи регуляризації (Ridge, Lasso, Elastic Net). Методи ансамблювання моделей (Voting, Bagging, Boosting, Stacking).

Тема 3. Навчання та оцінка моделей Розбиття датасету на вибірки. Підбір гіперпараметрів. Процес навчання і тестування моделей. Метрики оцінки точності для задач регресії та класифікації. Перехресна валідація.

Тема 4. Нейронні мережі Архітектура нейронних мереж. Типи шарів (Dense, Convolutional, Pooling, Recurrent, Dropout, BatchNormalization). Алгоритми навчання. Функції втрат та оптимізатори. Функції активації. Практична побудова нейронних мереж.

Тема 5. Генерація синтетичних даних Статистичні методи генерації даних. Агентно-орієнтовані системи. Генеративно-змагальні мережі (GAN). Варіаційні автоенкодера (VAE).

Тема 6. Передавальне навчання Виділення ознак з попередньо навчених моделей. Тонке налаштування моделей. Багатозадачне навчання.

Тема 7. Нейронні мережі Колмогорова-Арнольда Теоретичні основи KAN. Архітектура та відмінності від класичних мереж. Практична реалізація та застосування.

4. Навчальні матеріали та ресурси

Базова:

1. McCulloch W. S., Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics. 1943. Vol. 5, no. 4. P. 115–133. DOI: <https://doi.org/10.1007/bf02478259>
2. Safonova A., Hamdi Z. M., Montzka C., Demarez V., Koetz B., Vereecken H. Ten deep learning techniques to address small data problems with remote sensing. International Journal of Applied Earth Observation and Geoinformation. 2023. Vol. 125. P. 103569. DOI: <https://doi.org/10.1016/j.jag.2023.103569>
3. Piffer S., Barzegar Khalilsarai M., Stølen-Bjerglund J., Eklund A., Gómez P. Tackling the small data problem in medical image classification with artificial intelligence: a systematic review. Progress in Biomedical Engineering. 2024. DOI: <https://doi.org/10.1088/2516-1091/ad525b>
4. Demir-Kavuk O., Kamada M., Akutsu T., Knapp E. W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. 2011. Vol. 12, no. 1. P. 412. DOI: <https://doi.org/10.1186/1471-2105-12-412>

5. Khan A. A., Chaudhari O., Chandra R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications. 2023. P. 122778. DOI: <https://doi.org/10.1016/j.eswa.2023.122778>
6. Slipchenko V., Poliahushko L., Rudyk V. The problem of small data in research based on machine learning methods. Collection of Scientific Papers with the Proceedings of the 1st International Scientific and Practical Conference «From Ideas to Solutions: Innovations in Science and Technology». London, United Kingdom, 17–19 February 2025. 2025. P. 40–42. DOI: <https://doi.org/10.70286/eoss-17.02.2025>
7. Ross P., Maynard K. Towards a 4th industrial revolution. Intelligent Buildings International. 2021. Vol. 13, no. 3. P. 159–161. DOI: <https://doi.org/10.1080/17508975.2021.1873625>
8. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1958. Vol. 65, no. 6. P. 386–408.
9. TensorFlow Documentation. TensorFlow Tutorials. URL: <https://www.tensorflow.org/tutorials>
10. Scikit-learn Documentation. Scikit-learn User Guide. URL: https://scikit-learn.org/stable/user_guide.html

Додаткова:

11. McCausland T. The Bad Data Problem. Research-Technology Management. 2021. Vol. 64, no. 1. P. 68–71. DOI: <https://doi.org/10.1080/08956308.2021.1844540>
12. Diranisha V., Agung Triayudi, Ratih Titi Komalasari. Implementation of K-Nearest Neighbour (KNN) Algorithm and Random Forest Algorithm in Identifying Diabetes. SAGA: Journal of Technology and Information System. 2024. Vol. 2, no. 2. P. 234–244.
13. Keras Documentation. Early Stopping Callback URL: https://keras.io/api/callbacks/early_stopping/
14. PyKAN Documentation. Kolmogorov Arnold Networks. URL: <https://kindxiaoming.github.io/pykan/>
15. Mesa Documentation. Agent-Based Modeling in Python. URL: https://mesa.readthedocs.io/latest/apis/api_main.html

Навчальний контент

5. Методика опанування навчальної дисципліни(освітнього компонента)

Лекція 1. Обробка даних та базові алгоритми машинного навчання

Завантаження та первинний аналіз даних з використанням pandas та numpy. Виявлення пропущених значень та аномалій. Лінійна та логістична регресія як фундаментальні алгоритми. Метод k-найближчих сусідів та найвний байєсівський класифікатор для малих даних.

Лекція 2. Регуляризація, ансамблювання та оцінка моделей

Методи регуляризації (Ridge, Lasso, Elastic Net) для боротьби з перенавчанням на малих даних. Основи ансамблювання: Voting, Bagging, Random Forest. Розбиття даних на вибірки, метрики оцінки якості, крос-валідація.

Лекція 3. Нейронні мережі для малих даних

Архітектура багатошарового перцептрону. Алгоритми прямого та зворотного поширення. Основні типи шарів (Dense, Dropout, BatchNormalization). Функції втрат та оптимізатори. Методи регуляризації в нейронних мережах.

Лекція 4. Сучасні підходи до роботи з малими даними

Огляд методів генерації синтетичних даних (статистичні методи, GAN). Основи передавального навчання та багатозадачного підходу. Нейронні мережі Колмогорова-Арнольда як альтернатива класичним архітектурам.

6. Самостійна робота студента/аспіранта

Тема 1. Обробка та аналіз даних

Використання бібліотек pandas та numpy для завантаження, огляду структури даних, створення нових ознак та базових операцій з DataFrame.

Виявлення та обробка пропущених значень, аномалій у даних. Аналіз збалансованості класів, обчислення кореляцій між ознаками. Методи масштабування, нормалізації та аугментації даних.

Тема 2. Основи машинного навчання

Лінійна та логістична регресія як базові алгоритми. Наївний байєсівський класифікатор та його застосування. Метод k-найближчих сусідів для задач класифікації та регресії.

Принципи побудови дерев рішень, критерії розбиття вузлів. Метод опорних векторів з жорсткою та м'якою межею, використання ядерних функцій для нелінійних задач.

Методи регуляризації Ridge, Lasso та Elastic Net регресія для боротьби з перенавчанням. Особливості застосування L1 та L2 регуляризації при роботі з малими наборами даних.

Методи ансамблювання. Voting як базовий метод поєднання моделей. Bagging та принципи Bootstrap Aggregating. Random Forest як найпопулярніший представник методів bagging. Boosting алгоритми для послідовного покращення моделей. Stacking та blending для складних ансамблів. Каскадні моделі та їх особливості.

Тема 3. Навчання та оцінка моделей

Правильне розбиття даних на тренувальну, валідаційну та тестову вибірки. Автоматизований підбір гіперпараметрів за допомогою GridSearchCV та RandomizedSearchCV.

Метрики для регресії та класифікації. Візуальні методи оцінки якості моделей. Крос-валідація як інструмент надійної оцінки на малих даних.

Тема 4. Нейронні мережі

Архітектура багатошарового перцептрону. Алгоритми прямого та зворотного поширення. Основні типи шарів та функції втрат для різних задач.

Згорткові та рекурентні шари для специфічних типів даних. Методи регуляризації в нейронних мережах через Dropout та нормалізацію.

Сучасні оптимізатори для навчання нейронних мереж. Функції активації та їх вплив на якість навчання. Практичне створення моделей у TensorFlow.

Тема 5. Генерація синтетичних даних

Статистичні підходи до створення штучних даних. Агентно-орієнтовані системи, генеративно-змагальні мережі та варіаційні автоенкодера для збільшення обсягу навчальних даних.

Тема 6. Передавальне навчання

Використання попередньо навчених моделей як екстракторів ознак. Тонке налаштування параметрів для адаптації до нових задач. Багатозадачне навчання для покращення узагальнення.

Тема 7. Нейронні мережі Колмогорова-Арнольда

Теоретичні основи KAN як альтернативи класичним нейронним мережам. Практична реалізація за допомогою бібліотеки PyKAN та переваги для малих наборів даних.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

- Відвідування лекцій та лабораторних занять, а також відсутність на них, не оцінюється.
- При захисті лабораторних робіт студент має продемонструвати розроблений програмний код та результати його виконання на тестах, як заздалегідь підготованих, так і запропонованих викладачем. У випадку дистанційної форми навчання захист відбувається на відповідній конференції шляхом демонстрації екрана.
- Політика та принципи академічної доброчесності визначені у розділі 3 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: <https://kpi.ua/code>.
- Норми етичної поведінки Норми етичної поведінки студентів і працівників визначені у розділі 2 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: <https://kpi.ua/code>.

8. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

1. Стартовий рейтинг (виконання робіт в семестрі) оцінюється у 100 балів. Розподіл балів наведено в таблиці:

Види робіт	Бали
------------	------

Лабораторна робота № 1. Аналіз та попередня обробка малих наборів даних	20
Лабораторна робота № 2. Порівняльний аналіз класичних алгоритмів машинного навчання та нейронних мереж для малих даних	30
Лабораторна робота № 3. Сучасні підходи до роботи з малими даними: генерація синтетичних даних, передавальне навчання та мережі Колмогорова-Арнольда	40
Модульна контрольна робота	10
Всього	100

Штрафні бали віднімаються за:

- 1) неоптимальний алгоритм – 10% від максимальної кількості балів;
- 2) неоптимальні структури представлення інформації – 10% від максимальної кількості балів;
- 3) ненадану або невірну відповідь на запитання – 20% від максимальної кількості балів при захисті лабораторної роботи або 100% - на контрольній роботі.

2. Умови допуску до заліку: відсутність заборгованостей з лабораторних робіт 1 - 3.

Студенти, які виконали умови допуску до заліку та мають рейтингову оцінку 60 і більше балів, отримують відповідну до набраного рейтингу оцінку без додаткових випробувань.

Зі студентами, які виконали умови допуску до заліку та мають рейтингову оцінку менше 60 балів, а також з тими студентами, хто бажає підвищити свою рейтингову оцінку, на останньому за розкладом занятті з дисципліни в семестрі проводиться семестровий контроль у вигляді заліку.

3. Відповідь на заліку оцінюється у 100 балів. Залікова робота складається з одного теоретичного питання та одного практичного завдання. Ваговий бал теоретичного питання – 50 балів, завдання – 50 балів.

Теоретична частина оцінюється таким чином:

- правильна чітко викладена, повна відповідь – (не менше 90% потрібної інформації) – 45-50 балів;
- достатньо повна відповідь (не менше 75% потрібної інформації) – 38-44 бали;
- неповна відповідь (не менше 60% потрібної інформації) – 30-37 балів;
- незадовільна відповідь – 0 балів.

Практичне завдання оцінюється таким чином:

- повне, безпомилкове розв'язування завдання – 45-50 балів;
- повне, розв'язування завдання із несуттєвими невідповідностями – 38-44 бали;
- завдання виконане з певними недоліками – 30-37 балів;
- завдання не виконано – 0 балів.

4. Рейтингова оцінка за освітній компонент за бажанням студента визначається одним з таких способів:

- 1) кількість балів, отриманих за поточний контроль, або
- 2) результат виконання залікової контрольної роботи (тоді не враховуються бали, отримані в семестрі).

Таблиця переведення рейтингових балів у оцінку за університетською шкалою:

Кількість балів	Оцінка
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

Робочу програму навчальної дисципліни (силабус):

Складено доцент каф. ЦТЕ, к.т.н., доц. Полягушко Любов Григорівна

Ухвалено кафедрою ЦТЕ (протокол № 22 від 25.06.2025)

Погоджено Методичною радою НН ІАТЕ КПІ ім. Ігоря Сікорського (протокол № 9 від 27.06.25)