Силабус освітнього компонента
Програма навчальної дисципліни

Інтелектуальний аналіз даних

Шифр та назва спеціальності
122 – Комп’ютерні науки

Інститут
ННІ Комп'ютерного моделювання, прикладної фізики та математики

Освітня програма
Комп’ютерні науки. Моделювання, проектування та комп’ютерна графіка

Кафедра
Геометричного моделювання та комп'ютерної графіки (163)

Рівень освіти
БакалаврЇ

Тип дисципліни
Спеціальна (фахова), Обов'язкова

Семестр
6

Мова викладання
Українська

Викладачі, розробники

Дашкевич Андрій Олександрович
(відповідальний лектор)
Andrii.Dashkevych@khpi.edu.ua
Кандидат технічних наук, доцент

Автор понад 60 наукових та навчально-методичних праць. Провідний лектор з дисциплін: «Інтелектуальний аналіз даних», «Обчислювальна візуалізація»

Детальніше про викладача на сайті кафедри

Успенський Валерій Борисович
(асистент із лабораторних робіт)
valerii.uspenskyi@khpi.edu.ua
Доктор технічних наук, доцент

Детальніше про викладача на сайті кафедри

Вязовиченко Юлія Андріївна
(асистент із лабораторних робіт)
yuliia.viazovychenko@khpi.edu.ua
Кандидат технічних наук, доцент

Детальніше про викладача на сайті кафедри

Загальна інформація

Анотація

В рамках курсу студенти вивчають сучасні підходи до аналізу багатовимірних даних із використанням сучасних методів класифікації, кластерного аналізу, методів зниження розмірності даних та їх візуалізації.

Мета та цілі дисципліни

Навчити студентів методам класифікації, кластеризації та зниження розмірності даних із використанням бібліотек Pandas та sklearn для мови програмування Python.

Формат занять

Лекції, лабораторні заняття, самостійна робота, консультації. Підсумковий контроль – екзамен.

Компетентності

ЗК2: Здатність застосовувати знання у практичних ситуаціях.
ЗК3: Знання та розуміння предметної області та розуміння професійної діяльності
ЗК8: Здатність генерувати нові ідеї (креативність).
ЗК12: Здатність оцінювати та забезпечувати якість виконуваних робіт.
СК2: Здатність до виявлення статистичних закономірностей недетермінованих явищ, застосування методів обчислювального інтелекту, зокрема статистичної, нейромережевої та нечіткої обробки даних, методів машинного навчання та генетичного програмування тощо
СК7: Здатність застосовувати теоретичні та практичні основи методології та технології моделювання для дослідження характеристик і поведінки складних об'єктів і систем, проводити обчислювальні експерименти з обробкою й аналізом результатів
СК11: Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв’язування прикладних задач

Результати навчання

ПР1: Застосовувати знання основних форм і законів абстрактно-логічного мислення, основ методології наукового пізнання, форм і методів вилучення, аналізу, обробки та синтезу інформації в предметній області комп'ютерних наук
ПР4: Використовувати методи обчислювального інтелекту, машинного навчання, нейромережевої та нечіткої обробки даних, генетичного та еволюційного програмування для розв’язання задач розпізнавання, прогнозування, класифікації, ідентифікації об’єктів керування тощо
ПР12: Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використанням програмних інструментів підтримки багатовимірного аналізу даних на основі технологій DataMining, TextMining, WebMining

Обсяг дисципліни

Загальний обсяг дисципліни 120 год. (4 кредити ECTS): лекції – 16 год., лабораторні заняття – 32 год., самостійна робота – 72 год.

Передумови вивчення дисципліни (пререквізити)

Теоретичною і науковою основою дисципліни є знання методів теорії ймовірностей, лінійної алгебри, обчислювальних методів, об'єктно-орієнтованого програмування та проектування, методів оптимізації, математичнихметодів теорії штучного інтелекту.

Особливості дисципліни, методи та технології навчання

Практичні навички із застосування методів інтелектуального аналізу даних протягом лабораторних занять набуваються із використанням сучасних бібліотек інтелектуального аналізу

даних для мови програмування Python, одним із варіантів налаштованого середовища для виконання завдань є пакет прикладних програм Anaconda.
Лекції проводяться з використанням мультимедійних технологій, де використовується метод кейс-стаді, який дозволяє студентам аналізувати реальні ситуації з використанням теоретичних знань, сприяючи критичному мисленню та здатності застосовувати теорію на практиці. На лабораторних роботах використовуються методи випробувань та помилок, колективного аналізу, та інтерактивного моделювання, що сприяє розвитку практичних навичок, критичного мислення та ефективної командної взаємодії серед студентів.
Навчальні матеріали доступні студентам на Microsoft OneDrive

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Основні поняття інтелектуального аналізу даних. Сфери застосування, етапи та основні задачі інтелектуального аналізу даних. Інженерія ознак
1.1 Основні визначення та поняття
1.2 Приклади практичних задач
1.3 Етапи аналізу даних
1.4 Exploratory Data Analysis
1.5 Підходи до аналізу даних
1.6 Інструментальне забезпечення аналізу даних
1.7 Виміри, ознаки та класи
1.8 Етапи роботи із ознаками
1.9 Методи інженерії ознак

Тема 2. Кероване навчання. Задача класифікації та прогнозування. Методи класифікації. Вибір та оцінювання моделей
2.1 Задача класифікації та прогнозування
2.2 Поняття лінійної роздільності класів
2.3 Методи розв'язання задачі класифікації. Метод найближчого сусідства
2.4 Методи на основі дерев прийняття рішень
2.5 Методи побудови дерев
2.6 Ансамблеві методи покращення якості класифікації
2.7 Класифікація на основі методів теорії ймовірностей
2.8 Відбір та оцінювання моделей. Метрики оцінювання
2.9 Крос-валідація та оптимізація гіперпараметрів моделей

Тема 3. Некероване навчання. Задача кластеризації. Методи кластерного аналізу
3.1 Постановка задачі кластеризації
3.2 Алгоритми кластеризації
3.3 Ієрархічна кластеризація
3.4 Центроїдна кластеризація. Алгоритми k-середніх та Mean Shift
3.5 Щільнісні методи кластеризації
3.6 Графові методи кластеризації
3.7 Функціонали якості кластеризації

Тема 4. Методи інтелектуального аналізу даних високої розмірності
4.1 Проблеми обробки даних високої розмірності
4.2 Методи аналізу даних високої розмірності
4.3 Метод головних компонент
4.4 Метод випадкових проєкцій
4.5 Багатовимірне шкалювання
4.6 Ймовірнісні підходи до аналізу даних високої розмірності
4.7 Візуалізація даних високої розмірності

Тема 5. Методи інтелектуального аналізу даних високої розмірності. Навчання ознак

5.1 Підходи до навчання ознак
5.2 Навчання словнику
5.3 Методи розв'язання задачі словникового навчання
5.4 Навчання ознак методом k-середніх
5.5 Автокодувальники
5.6 Підхід Bag-of-Words

Тема 6. Методи інтелектуального аналізу для роботи із цифровими зображеннями
6.1 Способи представлення цифрових зображень для задач інтелектуального аналізу даних
6.2 Проблеми представлення цифрових зображень як ознакових векторів. Підхід Bag-of-Words для зображень
6.3 Методи словникового навчання для аналізу зображень
6.4 Алгоритми-детектори ключових точок
6.5 Методи на основі автокодувальників

Тема 7. Методи інтелектуального аналізу текстової інформації
7.1 Основні концепції Natural language processing
7.2 Проблеми представлення текстової інформації як ознакових векторів
7.3 Модель TF-IDF
7.4 Підвищення ефективності представлення текстової інформації. Hashing Trick та алгоритм MinHash

Тема 8. Методи інтелектуального аналізу послідовностей. Спеціалізовані методи підвищення ефективності алгоритмів обробки текстових даних та даних високої розмірності
8.1 Проблеми представлення послідовностей
8.2 Моделі на основі нейронних мереж
8.3 Large Language Models
8.4 Методи на основі алгоритмів хешування для обробки даних високої розмірності
8.5 Locality-sensitive hashing

Теми практичних занять

Не передбачено навчальним планом.

Теми лабораторних робіт

Тема 1. Підготовчі етапи інтелектуального аналізу даних. Exploratory Data Analysis
Завдання на роботу: вивчення методів первинної підготовки даних, методів дослідження та візуалізації наборів даних на основі технік Exploratory Data Analysis

Тема 2. Методи керованого навчання. Навчання класифікаторів. Вибір та оцінювання класифікаторів
Завдання на роботу: вивчення методів класифікації на основі алгоритму kNN та дерев прийняття рішень, застосування ансамблевого навчання для збільшення якості класифікації, відбір моделей класифікаторів на основі методів крос-валідації та оптимізації гіперпараметрів

Тема 3. Методи некерованого навчання. Основи кластерного аналізу. Агломеративна та центроїдна кластеризація
Завдання на роботу: вивчення базових алгоритмів кластеризації агломеративного та центроїдного типу

Тема 4. Методи некерованого навчання. Алгоритми DBSCAN та Affinity Propagation
Завдання на роботу: вивчення базових алгоритмів кластеризації щільнісного та графового типу

Тема 5. Зниження розмірності та візуалізація багатовимірних даних
Завдання на роботу: вивчення базових алгоритмів зниження розмірності для задач кластеризації та візуалізації даних

Тема 6. Методи інтелектуального аналізу даних для обробки цифрових зображень
Завдання на роботу: вивчення алгоритмів інтелектуального аналізу даних для обробки та аналізу цифрових зображень.

Тема 7. Методи інтелектуального аналізу даних для обробки текстової інформації
Завдання на роботу: вивчення алгоритмів інтелектуального аналізу даних для представлення та обробки текстової інформації.

Тема 8. Методи інтелектуального аналізу даних для обробки послідовностей. Виявлення аномалій у даних на основі методів просторового хешування
Завдання на роботу: вивчення алгоритмів інтелектуального аналізу даних для обробки та аналізу часових послідовностей та виявлення аномалій у послідовностях

Самостійна робота

Теми для самостійної підготовки під час виконання лабораторних занять:
1. Бібліотека Pandas: базові методи попередньої обробки даних
2. Основи класифікації, кластеризації та візуалізації засобами бібліотеки sklearn
3. Виділення ознак у даних. Конвеєр обробки даних в sklearn
4. Програмні засоби бібліотеки для крос-валідації та оптимізації гіперпараметрів
5. Засоби бібліотеки sklearn для кластеризації в метричних просторах
6. Бібліотеки для візуалізації даних для мови Python та базові способи візуалізації в них
7. Засоби бібліотеки sklearn для лінійного та нелінійного зниження розмірності даних
8. Програмні засоби реалізації словникового навчання для цифрових зображень та текстової інформації
9. Бібліотека NLTK для первинної обробки та виділення ознак з текстів.

Теми для самостійного опрацювання лекційного матеріалу:
1.Web Mining (проблеми аналізу інформації з Web, етапи Web Mining, Web Mining й інші інтернет-технології, категорії Web Mining).
2. Методи добування Web-контенту (добування Web-контенту в процесі інформаційного пошуку, добування Web-контенту для формування баз даних).
3.Добування Web-структур (подання Web-структур, оцінка важливості Web-структур, пошук Web-документів з урахуванням гіперпосилань, кластерізація Web-структур).
4.Дослідження використання Web-ресурсів (дослідницька інформація, етап препроцесінгу, етап добування шаблонів, етап аналізу шаблонів й їхнє застосування).

Індивідуальне завдання полягаэ в розробці проекту для збору, обробки та аналізу даних з інтернету для отримання корисної інформації або знань. Проект зосереджується на автоматизації процесу збору даних з веб-сайтів, їхньої очистки та структурування, застосуванні алгоритмів машинного навчання або статистичних методів для аналізу цих даних, а також на візуалізації отриманих результатів.

Література та навчальні матеріали

1. Data Mining : пошук знань в даних / А. Я. Гладун, Ю. В. Рогушина. – К. : ТОВ «ВД «АДЕФ-Україна», 2016. – 452 с.
2. Аналіз даних та знань : навч. посібник / В. В. Литвин, В. В. Пасічник, Ю. В. Нікольський. – Львів : «Магнолія 2006», 2018. – 276 с.
3. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки : навчальний посібник / Д. В. Ланде, І. Ю. Субач, Ю. Є. Бояринова. – К. : ІСЗЗІ КПІ ім. Ігоря Сікорського, 2018. – 297 с.
4. Kantarzic M. Data Mining. Concepts, Models, Methods and Algorithms / M. Kantarzic, 3rd Ed. – Publisher : Wiley, 2019. – 672 p.
5. Ситник В. Ф., Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг): Навч. посібник. — К.: КНЕУ, 2007. — 376 с.

6. Клименко В., Радченко В., Саєнко О. Дані великих обсягів: технології збору, зберігання та аналізу [Текст] / В. Клименко, В. Радченко, О. Саєнко. — К.: Видавничий дім «ІнЖЕК», 2020. — 320 с.

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

Оцінка з дисципліни складається із наступних компонентів:
1. Захист лабораторних робіт: до 74 балів, які розподіляються наступним чином:
- обов'язкові лабораторні заняття 1-8 складаються із базової (мінімально необхідної для виконання) частини, лабораторні заняття 6 та 7 містять додаткові завдання підвищеної складності:
- виконання та захист базової частини лабораторних робіт 1-8: 54 бали;
- виконання та захист додаткових завдань в лабораторних роботах 6 та 7: до 10 балів;
-виконання індивідуального завдання: до 10 балів;
2. Теоретична контрольна робота (тест), включаючи питання для самостійного опрацювання: 16 балів.
3. Практична контрольна робота (екзамен) - розв'язання практичної задачі із усними відповідями на запитання: 10 балів

Шкала оцінювання

Сума балів	Національна оцінка	ECTS
90–100	Відмінно	A
82–89	Добре	B
75–81	Добре	C
64–74	Задовільно	D
60–63	Задовільно	E
35–59	Незадовільно (потрібне додаткове вивчення)	FX
1–34	Незадовільно (потрібне повторне вивчення)	F

Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту. Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Погодження

Силабус погоджено 28.08.2023 Завідувач кафедри
Ольга ШОМАН

28.08.2023 Гарант ОП
Оксана ТАТАРІНОВА