Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Кафедра автоматизації проектування енергетичних процесів і систем

ОБРОБКА НАДВЕЛИКИХ МАСИВІВ ДАНИХ

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	Нормативна
Форма навчання	очна(денна)
Рік підготовки, семестр	1 курс осінній семестр
Обсяг дисципліни	На засвоєння дисципліни передбачено 120 год / 4 кредити ЄКТС, 36 лек, 18 лаб., 66 сам.роб.
Семестровий контроль/ контрольні заходи	Екзамен, МКР
Розклад занять	Науково-педагогічний працівник
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: PhD, Москаленко Юрій Володимирович, yuramuv@gmail.com Лабораторні: PhD, Москаленко Юрій Володимирович, yuramuv@gmail.com
Розміщення курсу	Кампус/студентські групи

Програма навчальної дисципліни

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

Знання операційних систем є основою успішної кар'єри в сфері програмування, сприяє становленню зрілого мислення програміста, знання мережевих технологій і протоколів, віртуальних машин, методів сучасного програмування. Дисципліна «Операційні системи» передбачає вивчення операційної системи Windows 10. На даний момент згідно сервісу StatCounter лідером серед користувачів операційних систем визнано Windows 10, а саме 87,03%. Отримані знання дозволяють виконувати роботи з адміністрування системи, розроблення програмних систем для вирішення різних задач, особливо для яких є критичним час швидкодії.

Метою кредитного модуля є формування у студентів компетентностей у відповідності до ОПП.

ЗК 1	Здатність до абстрактного мислення, аналізу та синтезу.
ЗК 2	Здатність застосовувати знання у практичних ситуаціях.
ЗК 5	Здатність вчитися й оволодівати сучасними знаннями.
ФК 1	Усвідомлення теоретичних засад комп’ютерних наук.
ФК 3	Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області.
ФК 4	Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проєктних рішень.
ФК 5	Здатність розробляти, описувати, аналізувати та оптимізувати архітектурні рішення інформаційних та комп’ютерних систем різного призначення.

ФК 6	Здатність застосовувати існуючі і розробляти нові алгоритми розв’язування задач у галузі комп’ютерних наук.
ФК 7	Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень.
ФК 8	Здатність розробляти і реалізовувати проєкти зі створення програмного забезпечення, у тому числі в непередбачуваних умовах, за нечітких вимог та необхідності застосовувати нові стратегічні підходи, використовувати програмні інструменти для організації командної роботи над проєктом.
ФК 9	Здатність розробляти та адмініструвати бази даних та знань.
ФК 15	Здатність до проектування та програмної реалізації методів комп'ютерної обробки надвеликих за обсягом даних в інформаційних середовищах різноманітного призначення.

Предмет навчальної дисципліни – серія підходів, інструментів і методів обробки структурованих і неструктурованих різноманітних даних великих розмірів для отримання результатів, які легко сприймаються людиною.

Результати навчання. В результаті вивчення дисципліни студенти повинні продемонструвати такі програмні результати навчання:

ПРН 1	Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп’ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп’ютерних наук та на межі галузей знань
ПРН 2	Мати спеціалізовані уміння/навички розв’язання проблем комп’ютерних наук, необхідні для проведення досліджень та/або провадження інноваційної діяльності з метою розвитку нових знань та процедур.
ПРН 4	Управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів.
ПРН 6	Розробляти концептуальну модель інформаційної або комп’ютерної системи
ПРН 7	Розробляти та застосовувати математичні методи для аналізу інформаційних моделей.
ПРН 8	Розробляти математичні моделі та методи аналізу даних (включно з великими)
ПРН 9	Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).
ПРН 11	Створювати нові алгоритми розв’язування задач у сфері комп’ютерних наук, оцінювати їх ефективність та обмеження на їх застосування.
ПРН 12	Проєктувати та супроводжувати бази даних та знань.
ПРН 14	Тестувати програмне забезпечення.
ПРН 15	Виявляти потреби потенційних замовників щодо автоматизації обробки інформації.

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

Пререквізити дисципліни.

Вивчення на попередньому рівні вищої освіти дисциплін з напрямів:

- бази даних,
- аналіз даних;
- зберігання даних.

3. Зміст навчальної дисципліни

Тема 1 Поняття “великих даних”. Основи машинного навчання.
Тема 2 Навчання з учителем (Supervised Learning).
Тема 3 Навчання без вчителя (Unsupervised Learning).
Тема 4 Навчання з підкріпленням (Reinforcement Learning).
Тема 5. Базові програмні засоби роботи з надвеликими масивами даних.
Тема 6. Технології зберігання “великих даних”
Тема 7. Інфраструктура програмних рішень Spark.

4. Навчальні матеріали та ресурси

1. Mayer-Schönberger V and Cukier K (2013) Big Data: The Essential Guide to Work, Life and Learning in the Age of Insight. London: John Murray. ISBN 9781473647206
2. Tom White Hadoop: The Definitive Guide, 4th Edition O'Reilly Media. 2015 – 756p.
3. Smolan, Rick, and Jennifer Erwitt. The Human Face of Big Data. Sausalito, Calif: Against All Odds Productions, 2012. Print.
4. Miner, Donald, and Adam Shook. Mapreduce Design Patterns. Beijing: O'Reilly, 2nd edition, 2015. Internet resource.
5. Lam, Chuck. Hadoop in Action. Greenwich, Conn: Manning Publications, 2011. Print.
6. Witten, I H, and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005. Internet resource.
7. Mitchell, Tom M. Machine Learning. New York: McGraw-Hill, 1997. Print.
8. Cherkassky, Vladimir S, and Filip Mulier. Learning from Data: Concepts, Theory, and Methods. Hoboken, N.J: IEEE Press, 2007. Internet resource.
9. Marsland, Stephen. Machine Learning: An Algorithmic Perspective. Boca Raton: CRC Press, 2009. Print.
10. Harrington, Peter. Machine Learning in Action. Shelter Island, NY: Manning Publications, 2012. Print.
11. Ланде Д. В. Оброблення надвеликих масивів даних (Big Data) [Електронний ресурс] : навчальний посібник для використання у навчальному процесі з підготовки фахівців другого (магістерського) рівня вищої освіти зі спеціальності 122 «Комп’ютерні науки» / Д. В. Ланде, І. Ю. Субач, А. Я. Гладун ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані. – Київ : КПІ ім. Ігоря Сікорського, 2021. – 168 с. - https://ela.kpi.ua/handle/123456789/46129

Додаткова література

1. Holmes, Alex. Hadoop in Practice. Shelter Island, NY: Manning, 2012. Print.
2. Alpaydin, Ethem. Introduction to Machine Learning. , 2014. Internet resource.
3. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. Cambridge, MA: MIT Press, 2012. Internet resource.
4. Shalev-Shwartz, Shai, and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. , 2014. Print.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Тема 1. Поняття “великих даних”

Лекція 1. Основні аспекти та складові елементи трактування поняття “Великі дані”. Сфери застосування надвеликих масивів даних. Характеристики “великих даних”. Проблема масштабування. Базові компоненти аналізу даних. Машинне навчання, його історичний розвиток і сучасний стан. Проблема навчання. Приклади прикладних задач, які використовують методи машинного навчання.

Тема 2. Навчання з учителем (Supervised Learning)

Лекція 2. Класифікація. Регресія. Постановка задачі класифікації. Постановка задачі регресії. Дерева рішень. Класифікація на основі дерев рішень. Лінійна регресія з однією змінною. Оціночна функція. Метод градієнтного спуску. Лінійна регресія з декількома змінними. Нормалізація. Нормальні рівняння. Самостійна робота студентів. Більш детальніше ознайомитися з поняттями класифікації та регресії. Ознайомитися з підходами до вирішення задач класифікації. Розглянути поняття лінійної регресії.

Лекція 3. Логістична регресія. Границі рішень. Оціночна функція. Регуляризація. Багатокласова класифікація. Метод One-vs-all.

Самостійна робота студентів. Метод градієнтного спуску в логістичній регресії. Вибір коефіцієнту навчання. Проблема недостатнього та надмірного навчання.

Лекція 4. Нейронні мережі. Презентація даних в нейронних мережах. Персептрон. Навчання нейронної мережі. Метод зворотного поширення похибки. Класифікація за допомогою нейронних мереж. Параметри та гіперпараметри моделі.

Самостійна робота студентів. Ознайомитися з програмним забезпеченням MatLab для побудови та навчання нейронних мереж. Більш детальніше ознайомитися з принципами побудови та навчання нейронних мереж.

Лекція 5. Метод опорних векторів. Цілі оптимізації. Постановка та формальний опис задачі. Лінійно неподільна вибірка. Застосування ядра.

Самостійна робота студентів. Ознайомитися з найбільш поширеними ядрами для методу опорних векторів. Ознайомитися з програмним забезпеченням реалізації методу опорних векторів.

Тема 3. Навчання без вчителя (Unsupervised Learning)

Лекція 6. Кластеризація. Базова задача кластеризації. Кластеризація методом k-середніх. Оптимізація методом k-середніх. Властивості кластеризації.

Самостійна робота студентів. Більш детально ознайомитися з базовими підходами до кластеризації. Ознайомитися з метриками відстаней.

Лекція 7. Метод головних компонент. Формальна постановка задачі. Діагоналізація коваріаційної матриці.Сингулярний розклад матриці даних Матриця перетворення до головних компонент. . Приклади використання.

Самостійна робота студентів. Властивості і обмеження методу головних компонент. Відбір головних компонент за правилом Кайзера. Нормування. Ефективність методу головних компонент.

Лекція 8. Метод колаборативної фільтрації. Постановка задачі. Типи колаборативної фільтрації. Обмеження на застосування методу. Рекомендаційні системи.

Самостійна робота студентів. Методи градієнтного спуску в алгоритмах колаборативної фільтрації. Використання методу в соціальних мережах.

Тема 4. Навчання з підкріпленням (Reinforcement Learning)

Лекція 9. Навчання з підкріпленням. Базова модель навчання з підкріпленням. Марковський процес вирішування. Метод часових різниць. Q-Learning, DQN.Вплив змінних на алгоритм.

Самостійна робота студентів. Алгоритми для навчання керуванню. Критерій оптимальності. Підходи функції цінності. Методи Монте-Карло. Зворотне навчання з підкріпленням.

6. Самостійна робота студента

Тема 1. Поняття “великих даних”

Розглянути: 1) генезис поняття “Великі дані”; 2) прикладні задачі, що потребують обробки надвеликих масивів даних. Більш детальніше ознайомлення з характеристиками “великих даних”.

Дослідити генезис поняття “машинне навчання”. Ознайомитися з основними напрямками машинного навчання (з вчителем та без). Ознайомитися з сучасними досягненнями в машинному навчанні на прикладі глибинного навчання.

Тема 2. Навчання з учителем (Supervised Learning)

Більш детальніше ознайомитися з поняттями класифікації та регресії. Ознайомитися з підходами до вирішення задач класифікації. Розглянути поняття лінійної регресії.

Логістична регресія. Метод градієнтного спуску в логістичній регресії. Вибір коефіцієнту навчання. Проблема недостатнього та надмірного навчання. Нейронні мережі. Ознайомитися з програмним забезпеченням MatLab для побудови та навчання нейронних мереж. Більш детальніше ознайомитися з принципами побудови та навчання нейронних мереж. Метод опорних векторів. Ознайомитися з найбільш поширеними ядрами для методу опорних векторів. Ознайомитися з програмним забезпеченням реалізації методу опорних векторів.

Тема 3. Навчання без вчителя (Unsupervised Learning)

Кластеризація. Більш детально ознайомитися з базовими підходами до кластеризації. Ознайомитися з метриками відстаней.

Метод головних компонент. Властивості і обмеження методу головних компонент. Відбір головних компонент за правилом Кайзера. Нормування. Ефективність методу головних компонент.

Метод колаборативної фільтрації. Методи градієнтного спуску в алгоритмах колаборативної фільтрації. Використання методу в соціальних мережах.

Тема 4. Навчання з підкріпленням (Reinforcement Learning)

Навчання з підкріпленням. Алгоритми для навчання керуванню. Критерій оптимальності. Підходи функції цінності. Методи Монте-Карло. Зворотне навчання з підкріпленням.

Тема 5. Базові програмні засоби роботи з надвеликими масивами даних

Поняття розподіленої файлової системи.. Розглянути особливості екосистеми Hadoop. Ознайомитися з структурою менеджера ресурсів YARN. Ознайомитися з шаблонами проектування MapReduce.

Тема 6. Технології зберігання “великих даних”

Виконати огляд технологій зберігання “великих даних”. Ознайомитися з принципами NoSQL баз даних. Ознайомитися з колоночною базою даних HBase.

Тема 7. Інфраструктура програмних рішень Spark.

Підходи до збереження та обробки надвеликих масивів даних в Spark. Вирішення задач машинного навчання з використанням бібліотеки MLlib.

Під час навчання застосовуються методи: пояснювально-ілюстративний, частково-пошуковий, репродуктивний, проблемний.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Відвідування лекційних та лабораторних занять є обов’язковим за винятком поважних причин (хвороби, форс-мажорних обставин).

В разі пропущення занять з поважних причин викладач надає можливість студенту виконати усі або деякі лабораторні завдання (винятком є виконання деяких завдань у зв’язку із закінченням навчального процесу).

В разі пропущення занять без поважних причин, а також через порушення граничного терміну виконання завдання (deadline) студент може отримати 80% від максимальної оцінки відповідне завдання.

Протягом семестру студенти:

- виконують та захищають лабораторні роботи у відповідні терміни (на кожну лабораторну роботу відводиться два тижні для здачі),
- пишуть модульну контрольну роботу,
- повинні позитивно закрити дві атестації (в кінці березня та в середині травня),
- по закінченні навчального процесу складають екзамен.

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 (РСО)

Система рейтингових (вагових) балів та критерії оцінювання

1) Робота на лекціях

На лекціях може бути проведено бліцопитування студентів. Такі опитування проводяться на довільних лекціях 5 разів протягом семестру, наприкінці лекції. Ваговий бал за вірну відповідь - 1. Максимальна кількість балів, що може отримати кожен студент за семестр - 5.

2) Лабораторні роботи

Максимальна кількість балів за усі виконані лабораторні роботи дорівнює 50 балів. Розподіл балів серед лабораторних робіт наступний:

№ з/п	Назва лабораторної роботи	Кількість балів
1	Вирішення задачі регресії	5
2	Вирішення задачі класифікації	5
3	Навчання нейронної мережі для вирішення задачі класифікації	7
4	Метод опорних векторів для вирішення задачі класифікації	5
5	Вирішення задачі кластеризації з використанням алгоритмів навчання без учителя	5
6	Реалізація методу головних компонент	5
7	Реалізація рекомендаційної системи з використанням методу колоборативної фільтрації	7
8	Реалізація класичних тестових задач алгоритмів навчання з підкріпленням (benchmark)	6
9	Реалізація базових програм Mapper та Reducer та відстеження процесу їх виконання	5
Разом		50

Критерії оцінювання:

Виконання лабораторної роботи:

- виконана своєчасно (протягом двох тижнів з моменту видачі), у повному обсязі – відповідний бал згідно номеру лабораторної роботи;
- виконана із запізненням – знімається 10 – 30% від максимальної кількості балів в залежності від терміну запізнення;
- виконана не самостійно, із запізненням – знімається 50% від максимальної кількості балів;
- невиконана протягом відведеного часу – 0 балів.

3) Модульна контрольна робота

Ваговий бал – 10. Максимальна кількість балів контрольну роботу дорівнює 10 балів.

Якість виконання роботи:

-всі відповіді вірні та повні –10 балів,

-у відповідях допущені несуттєві неточності – 8 балів,

-половина відповідей вірна – 5 балів,

-відповіді з суттєвими неточностями, але без критичних помилок – 2 бали,

-менше половини відповідей вірна – 0 балів.

4) Складання екзамену

Максимальний ваговий бал $r_{екз}$=35

Умови позитивної проміжної атестації.

Для отримання „зараховано" з першої проміжної атестації студент матиме не менше ніж 11 балів (за умови, що за 8 тижнів згідно з календарним планом контрольних заходів „ідеальний" студент має отримати 5+5+5+5 = 20 балів).

Для отримання „зараховано" з другої проміжної атестації студент матиме не менше ніж 25 балів (за умови, що за 14 тижнів згідно з календарним планом контрольних заходів „ідеальний" студент має отримати 20 + 5+5+8+6+6 =50 балів).

Умови допуску до екзамену.

Необхідною умовою допуску до екзамену є зарахування усіх лабораторних робіт та виконання модульної контрольної роботи, а також стартовий рейтинг (R_c) не менше 40 балів. Для отримання екзамену з кредитного модуля "автоматом" потрібно мати рейтинг не менш ніж 60 балів, а також зараховане виконання всіх завдань лабораторних робіт.

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру (шкала рейтингу) складає:

$$R = r_{лек} + r_{прак} + r_{мод} + r_{екз} = 5+50 + 10 + 35 = 100 \text{ балів.}$$

Стартовий рейтинг становить $R_c = r_{лек} + r_{лаб} + r_{мод} = 65$ балів.

Рейтинг екзамену дорівнює 35 балів.

Таким чином, рейтингова шкала з кредитного модуля складає

$$R = 65+35=100 \text{ балів.}$$

Для отримання студентом відповідних оцінок (ECTS та традиційних) його рейтингова оцінка RD переводиться згідно таблиці:

Бали	**Оцінка**
95 - 100	Відмінно
85 - 94	Дуже добре
75 - 84	Добре
65 - 74	Задовільно
60 - 64	Достатньо
R≤ 59	Незадовільно
R_c < 40 або не виконані інші умови допуску до екзамену	Не допущений

Робочу програму навчальної дисципліни (силабус):

Складено PhD., Москаленко Юрій Володимирович

Ухвалено кафедрою цифрових технологій в енергетиці (протокол № 1 від 01.07.2022)

Погоджено Методичною комісією інституту (протокол № 10 від 04.07.2022)