Міністерство освіти і науки України
Чернівецький національний університет
імені Юрія Федьковича

Системи машинного навчання

Методичні вказівки до лабораторних робіт

Чернівці
Чернівецький національний університет
2024

С-133
УДК

Рекомендовано до друку Вченою радою факультету математики та інформатики

Чернівецького національного університету імені Юрія Федьковича

(протокол № 8 від « 24 » квітня 2024 року)

Укладачі:

Дорошенко Ірина Вікторівна, кандидат фізико-математичних наук, доцент кафедри математичного моделювання;

С-133 **Системи машинного начання : методичні вказівки до лабораторних робіт** / Укл.: Дорошенко І.В. – Чернівці : Чернівецький нац. ун-т, 2024. – 112 с.

Методичні вказівки до лабораторних робіт з дисципліни «Системи машинного начання» містить завдання для лабораторних робіт за всіма розділами лекційного курсу.

УДК

Силабус

1. Анотація дисципліни (призначення навчальної дисципліни).

Призначення дисципліни – вивчення методів, що застосовуються для побудови складних моделей та алгоритмів із метою вирішення завдань класифікації, кластеризації та прогнозування. Ці аналітичні моделі дозволяють дослідникам, науковцям із даних, інженерам та аналітикам «виробляти надійні, повторювані рішення і результати» та розкривати «приховані розуміння» шляхом навчання з історичних співвідношень та тенденцій у даних.

2. Мета навчальної дисципліни: формування у студентів сучасного наукового світогляду в області методів машинного навчання; наукової уяви про задачі, що розв'язуються з допомогою методів машинного навчання, вивчення методів класифікації і регресії з вчителем, а також методи кластерного аналізу (без вчителя); знайомство студентів з сучасними технологіями машинного навчання та тенденціями розробки і застосування; подальше становлення і вдосконалення інформаційної та програмної культури майбутніх фахівців.

Основними завданнями: набуття практичних навичок і знань в області технологій машинного навчання. У результаті вивчення даної дисципліни студенти повинні освоїти основні методи навчання з вчителем (Байєсівськмй класифікатор, лінійний дискримінантний аналіз Фішера, логістична регресія, метод опорних векторів, дерева рішень, випадковий ліс) і без вчителя – кластеризація розбиттям, ієрархічна кластеризація та нечітка кластеризація. Також, в результаті освоєння матеріалу, студенти повинні вивчити основні практичні прийоми роботи з інформацією мовами R.

3. Пререквізити. Теорія ймовірностей та математична статистика, статистика, аналіз даних, програмування.

4. Результати навчання: студент повинен мати навички (набути досвід): - розробки інструментальних засобів аналізу даних методами машинного навчання.

Формування компетентностей, а саме

Загальні та фахові компетентності:

ЗК6. Здатність спілкуватися з представниками інших професійних груп різного рівня (з експертами з інших галузей знань/видів економічної діяльності).

Спеціальні компетентності (СК):

СК4. Здатність оцінювати ризики, розробляти алгоритми управління ризиками в складних системах різної природи.

СК5. Здатність моделювати, прогнозувати та проектувати складні системи і процеси на основі методів та інструментальних засобів системного аналізу.

СК6. Здатність застосовувати теорію і методи Data Science для здійснення інтелектуального аналізу даних з метою виявлення нових властивостей та генерації нових знань про складні системи.

СК7. Здатність управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів.

СК9. Здатність здійснювати захист прав інтелектуальної власності, комерціалізацію результатів досліджень та інновацій.

В результаті навчального курсу студенти повинні

знати: – основні задачі машинного навчання; основні типи даних та методи побудови матриць суміжності; основні методи машинного навчання; основні засоби мов R для розв'язання задач машинного навчання.

вміти:

- застосовувати методи машинного навчання для розв'язання складних задач системного аналізу;

- вільно презентувати та обговорювати усно і письмово результати досліджень та інновацій, інші питання професійної діяльності державною та англійською мовами;
- виконувати обчислення, пов'язані з навчанням і роботою моделей машинного навчання;
- вміти будувати різні типи алгоритмів машинного навчання;
- визначати оптимальний метод для кожної задачі;
- розв'язувати задачі машинного навчання засобами R.

Програмні результати навчання:

РН2. Будувати та досліджувати моделі складних систем і процесів застосовуючи методи системного аналізу, математичного, комп'ютерного та інформаційного моделювання.

РН3. Застосовувати методи розкриття невизначеностей в задачах системного аналізу, розкривати ситуаційні невизначеності та невизначеності в задачах взаємодії, протидії та конфлікту стратегій, знаходити компроміс при розкритті концептуальної невизначеності.

РН6 Застосовувати методи машинного навчання та інтелектуального аналізу даних, математичний апарат нечіткої логіки, теорії ігор та розподіленого штучного інтелекту для розв'язання складних задач системного аналізу.

РН11. Вільно презентувати та обговорювати усно і письмово результати досліджень та інновацій, інші питання професійної діяльності державною та англійською мовами

Зміст завдань для самостійної роботи

№ з/п	Назва теми	Кількість годин
1	Особливості роботи з реальними даними Пропуски в даних. Попередня обробка ознак. Чистка даних. Категорійні ознаки: кодування, хешування, лічильники. Робота з текстами. Розріджені ознаки: векторизація, хешування, TF-IDF. Косинусна метрика.	10
2	Машинне навчання в прикладних задачах. Етапи аналізу даних. Робота з числовими ознаками. Робота з категоріальними та текстовими ознаками. Підготовка даних. Оцінювання якості роботи алгоритму.	5
3	Навчання з учителем	5
4	Підходи до отримання ознак для складних даних Робота з зображеннями (фільтри, отримання ознак за допомогою нейромереж), текстами (word embeddings).	10
5	Колаборативна фільтрація. Задачі колаборативної фільтрації і матриця суб'єкти-об'єкти. Латентні методи на основі бі-кластеризації. Алгоритм Брегмана. Латентні методи на основі матричних розкладань. для розріджених даних.	10
6	Багатошарові нейронні мережі. Біологічний нейрон. Функції активації. Проблема повноти. Повнота двошарових мереж в просторі булевих функцій. Теореми Колмогорова, Стоуна, Горбаня (без доведення). Алгоритм зворотного поширення помилок. Метод пошарового налаштування мережі. Підбір структури мережі: методи поступового ускладнення мережі, оптимальне проріджування нейронних мереж.	10
7	Рекомендаційні системи Постановки задачі. Метрики якості. Методи, базовані на колаборативній фільтрації. Методи, базовані на матричних розкладах.	10

Освітні технології, методи навчання і викладання навчальної дисципліни

У викладання курсу використовуються такі освітні технології:

- Лекції та їх презентації.
- Онлайн-лекції.
- Лабороторні заняття.
- Групова робота, коли студенти розв'язують практичні завдання.
- Онлайн-тести та опитування: Використання системи MOODLE
  Електронні підручники і ресурси репозитарію ЧНУ

Методи навчання:

МН 1 - лекція-візуалізація;

МН 8 – робота з тестами;

МН 9 – робота в групах;

МН 12 – дистанційне навчання з використанням відповідних онлайн-платформ

Контроль та оцінювання результатів навчальних досягнень студентів з навчальної дисципліни

Види та форми контролю

1. Поточний (захист лабораторних робіт, опитування теоретичного матеріалу)

2. Модульний (тестування, виконання завдань)

3. Підсумковий (екзамен)

Засоби оцінювання

Засобами оцінювання та демонстрування результатів навчання можуть бути:

- перевірка викладачем та захист студентами письмових звітів про виконання кожної лабораторної роботи,
- експрес-опитування,
- тестові завдання.

Критерії оцінювання результатів навчання з навчальної дисципліни

Критерієм успішного проходження здобувачем освіти підсумкового оцінювання може бути досягнення ним мінімальних порогових рівнів оцінок за кожним запланованим результатом навчання навчальної дисципліни.

Оцінювання знань студента на екзамені, під час лабороторних занять та виконання індивідуальних завдань проводиться за такими критеріями:

- розуміння, ступінь засвоєння теорії та методології навчальної дисципліни для розв'язання проблем, що розглядаються;
- рівень знань з теорії дисципліни та понятійно-категоріального апарату, термінології, поняття і принципи предметної області навчальної дисципліни.
- повнота розкриття питання; уміння чітко формулювати визначення понять/термінів й пояснювати їх; здатність аргументувати відповідь;
- аналітичні міркування, порівняння, формулювання висновків; логічна послідовність, культура мови; емоційність та вміння переконувати.
- ступінь засвоєння фактичного матеріалу навчальної дисципліни;
- ознайомлення з рекомендованою літературою, а також із сучасною літературою з питань, що розглядаються;
- вміння поєднувати теорію з практикою при розгляді виробничих ситуацій, проведенні аналізу, розв'язанні задач, проведенні розрахунків у процесі виконання індивідуальних завдань;
- застосування аналітичних підходів;
- здатність проводити критичну та незалежну оцінку певних проблемних питань, бачити слабкі й сильні сторони організації, обґрунтовувати можливості і загрози, що існують у зовнішньому середовищі організації;
- вміння пояснювати альтернативні погляди та наявність власної точки зору, позиції на певне проблемне питання;
- якість і чіткість викладення міркувань;
- обґрунтованість висновків щодо розробки стратегії розвитку досліджуваного підприємства (організації).

Дедлайни та перескладання. Роботи, які здаються із порушенням термінів без поважних причин, оцінюються на нижчу оцінку. Перескладання модулів відбувається з дозволу деканату за наявності поважних причин (наприклад, лікарняний, участь у конференції, студентській олімпіаді).

Академічна доброчесність. Здобувачі вищої освіти самостійно виконують навчальні завдання, завдання поточного та підсумкового контролю результатів навчання. Обов'язковим є посилання на джерела інформації в разі використання ідей, розробок, тверджень.

Відвідування занять. Відвідування занять є обов'язковою умовою виконання навчального плану дисципліни. Форми навчання визначені затвердженим графіком освітнього процесу Чернівецького національного університету імені Юрія Федьковича.

Критерії оцінювання.

Оцінка знань здобувачів включає поточний та підсумковий контроль. Поточний контроль здійснюється впродовж семестру під час проведення лекційних та лабороторних занять. Підсумковий контроль має за мету – перевірку теоретичних знань здобувачів, виявлення навичок застосування перших при вирішенні практичних завдань, а також навиків самостійної роботи з навчальною і науковою літературою.

Загальна кількість балів, яку здобувач може отримати у процесі вивчення дисципліни становить 100 балів, з яких 60 балів (по 30 балів за перший та другий модуль) здобувач може одержати як суму результатів поточного контролю (контрольні, самостійні роботи та тестування) і 40 балів – на підсумковому модулі (екзамені).

Екзаменаційний білет містить чотири питання, з яких одне теоретичне, три практичних. Повна відповідь на кожне питання оцінюється 10 балами. За кожну помилку, яка допущена у відповіді, знімається певна кількість балів, а саме:

а) при відповіді на теоретичне питання у випадку неістотної помилки знімається 1-3 бали, а у випадку істотної 4-7 балів, якщо ж здобувач не опанував теоретичний матеріал дисципліни, плутається в означеннях, наводить логічно неправильні твердження, то знімається до 9 балів;

б) при оцінці практичного завдання за помилку, допущену при обчисленнях, знімається 1-2 бали, за істотну помилку, знімається 3-5 балів, якщо ж розв’язання задачі логічно неправильне, то знімається до 8 балів.

Підсумкова оцінка виставляється за результатами суми балів набраних за кожне питання екзаменаційного білета з додаванням сумарної кількості балів за перший та другий модуль. Процедура проведення екзамену (у дистанційній формі) вимагає обов’язкової ідентифікації/персоніфікації здобувача.

Шкала оцінювання: національна та ЄКТС

Оцінка за національною шкалою	**Оцінка за шкалою ECTS**	
	Оцінка (бали)	**Пояснення за розширеною шкалою**
Відмінно	A (90-100)	відмінно
Добре	B (80-89)	дуже добре
	C (70-79)	добре
Задовільно	D (60-69)	задовільно
	E (50-59)	достатньо
Незадовільно	FX (35-49)	(незадовільно) з можливістю повторного складання
	F (1-34)	(незадовільно) з обов'язковим повторним курсом

Розподіл балів, які отримують студенти

Поточне оцінювання (*аудиторна та самостійна робота*)		Сумарна к-ть балів

Змістовий модуль №1				Змістовий модуль № 2				Кількість балів (екзамен)	
Т1	Т2	Т3	Т4	Т1	Т2	Т3	Т4	40	100
-	10	10	10	10	-	10	10		

Т1, Т2 ... Т9 – теми змістових модулів.

Рекомендована література -основна

1. Гнатюк В. Вступ до R на прикладах: навчальний посібник.- Навчальний посібник. ХНЕУ, 2010, 107с.
2. David Barber. Bayesian Reasoning and Machine Learning. – Cambridge University Press, 2012. – 697 p.
3. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning. Data Mining, Inference and Predictio. – Springer, 2018. – 745 p.
4. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning (with applications in R). – Springer, 2018. – 426 p.
5. HastieT., TibshiraniR., FriedmanJ. The Elementsof Statistical Learning. Springer, 2014. — 739 p.
6. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.
7. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. MIT Press, 2012.
8. Murphy K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
9. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.

10. Doroshenko I.V. ., Knihnitska T.V. ., Deretorska T.I. Comparison of machine learning algorithms for predicting mortality from Covid-19 virus // Sworld Jornal Issue No11, Part 2 January 2022 – P. 72-77
11. Doroshenko I.V. ., Knihnitska T.V. ., Kreshtanovych M.A. Comparison of data clustering algorithms// Sworld Jornal Issue No23, Part 1 January 2024 – P. 116-127
12. Doroshenko, I., Knopov, O. & Vovk, L. Mathematical Models of Extreme Modes in Ecological Systems // // Cybernetics and Systems Analysis.– 2022.– Vol.58, N5.– P.764–779.

9. Інформаційні ресурси

1. http://cran.r-project.org/bin/windows/base/

Завдання до лабораторних робіт

Лабораторна робота №1.

Тема: «Знайомство з R»

Мета: полягає в освоєнні інтерфейсу RStudio, оволодінні навичками роботи в консольному режимі та написання скриптів, а також уміння підключати зовнішні пакети. Крім того, планується вивчення основних методів обробки статистичних даних.

Завдання:

1. Завантажити дані з пакету "datasets" у форматі data.frame та дізнатися інформацію про набір даних.

2. Створити вектор, який буде містити дані одного зі стовпців набору, та обчислити основні вибіркові характеристики, такі як середнє, дисперсія, мода, медіана, стандартне відхилення тощо.

3. Побудувати гістограму абсолютних частот та гістограму щільності розподілу.

4. Створити новий вектор, де кожен елемент дорівнює 1, якщо відповідний елемент початкового вектора більше за середнє значення, -1 – якщо менше та 0 – якщо рівне середньому

5. Згенерувати вибірку з розподілу (згідно з варіантом) розміром 200 елементів, побудувати графік емпіричної функції розподілу та оцінити параметри даного розподілу на підставі вибірки.

6. На тому ж графіку, що й в завданні 5, намалювати графік теоретичної функції розподілу, замінивши невідомі параметри їхніми оцінками

7. Показати на одній декартовій площині графіки щільності розподілу (відповідно до варіанту) з різними параметрами

*Створити таблицю з 30 записів, де буде вказано: Nrow – номер запису, Name – імя співробітника, BirthYear – рік народження, EmployYear – рік прийому на роботу, Salary – зарплата. Nrow – від 1 до 20, Name – довільно, BirthYear – рівномірно розподілене на відрізку [1960 ,1995], EmployYear – рівномірно розподілене на відрізку [BirthYear+11,2024], Salary=(*ln*(2024-EmployYear)+1)*10000. Підрахувати кількість співробітників з зарплатою понад 15000, додати поле, що відповідає сумарному податку на прибуток, що виплатив співробітник за період робити на підприємстві (13% річних).