МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики
Кафедра програмування

Затверджено
На засіданні кафедри програмування факультету прикладної математики та інформатики
Львівського національного університету імені Івана Франка
(протокол № 1 від 29 серпня 2023 р.)

Завідувач кафедри Сергій ЯРОШКО

Силабус з навчальної дисципліни «Моделі статистичного навчання», що викладається в межах ОПП «Інформатика» другого (магістерського) рівня вищої освіти для здобувачів зі спеціальності 122 Комп'ютерні науки

Львів 2023 р.

Назва дисципліни	Моделі статистичного навчання
Адреса викладання дисципліни	Львівський національний університет імені Івана Франка, вул. Університетська 1, м. Львів, Україна, 79000
Факультет та кафедра, за якою закріплена дисципліна	Факультет прикладної математики та інформатики, кафедра програмування
Галузь знань, шифр та назва спеціальності	Галузь знань: 12 Інформаційні технології Спеціальність: 122 Комп'ютерні науки
Викладачі дисципліни	Заболоцький Тарас Миколайович, д. е. н., професор, професор кафедри програмування
Контактна інформація викладачів	Електронна пошта: taras.zabolotskyy@lnu.edu.ua, веб-сторінка: https://ami.lnu.edu.ua/employee/zabolotskyj-t-m
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/лабораторних занять (за попередньою домовленістю та за умови проведення аудиторних занять). В іншому випадку можливі он-лайн консультації через Zoom чи MSTeams. Для погодження часу он-лайн консультацій слід писати на електронну пошту викладача або дзвонити.
Сторінка курсу	https://ami.lnu.edu.ua/course/models-of-statistical-learning
Інформація про дисципліну	Дисципліна «Моделі статистичного навчання» є нормативною дисципліною освітньо-професійної програми «Інформатика» другого (магістерського) рівня вищої освіти для здобувачів зі спеціальності комп'ютерні науки, яка викладається в першому семестрі в обсязі 6 кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Мета та цілі дисципліни	Метою вивчення нормативної дисципліни «Моделі статистичного навчання» є: ознайомлення студентів з технічними деталями статистичних підходів до машинного навчання; набуття навичок використання тих чи інших моделей, візуалізації даних; розуміння ними сильних та слабких сторін різних підходів; набуття здатностей ефективно реалізовувати теоретичні знання у професійній діяльності.
Коротка анотація дисципліни	Курс розроблено таким чином, щоб надати студентам необхідні знання з методів статистичного навчання. Тому у курсі розглядаються як моделі навчання з вчителем (лінійні та нелінійні) так і моделі навчання без вчителя. Для кращого розуміння, усі теми доповнені прикладами з використанням як згенерованих так і реальних даних.
Література для вивчення дисципліни	*Основна література:* 1. An introduction to statistical learning with applications in R / James G., Witten D., Hastie T., Tibshirani R. New York : Springer , 2022. 622 p. 2. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York : Springer, 2016. 767 p. 3. Bruce P., Bruce A., Gedeck P. Practical statistics for data scientists: 50+ essential concepts using R and Python. Sebastopol : O'Reilly Media, 2020, 360 p. *Додаткова література* 4. McKinney W. Python for data analysis data wrangling with Pandas, NumPy, and IPython. Sebastopol : O'Reilly Media, 2017, 541 p. 5. Pierson L., Porway J. Data science for dummies. New Jersey : John Wiley & Sons, Inc., 2015. 411 p. 6. Сеньо П. С.Теорія ймовірностей та математична статистика. Підручник. / П. С. Сеньо. – К. : Знання, 2007. – 556 с.
Обсяг курсу	64 години аудиторних занять. З них 32 години лекцій, 32 години лабораторних занять та 116 годин самостійної роботи

Очікувані результати навчання	Після завершення цього курсу студент буде : • ***знати:*** 1) основні моделі навчання з вчителем; 2) основні моделі навчання без вчителя; 3) методи підбору «найкращої» моделі; 4) методи перетворення лінійних моделей на нелінійні; • ***вміти***: 1) здійснювати візуалізацію даних; 2) обґрунтовувати вибір моделі для опису даних та будувати на її основі прогнози; 3) оцінити якість вибраної моделі щодо опису конкретних даних; 4) визначати фактори, які мають найбільший вплив на результат; 5) знижувати розмірність даних та перебудовувати вибірку.
Компетентності	*Інтегральна*: Здатність розв'язувати задачі дослідницького та/або компетентність інноваційного характеру у сфері комп'ютерних наук *Загальні (ЗК):* 1. Здатність до абстрактного мислення, аналізу та синтезу. 2. Здатність застосовувати знання у практичних ситуаціях. 5. Здатність вчитися й оволодівати сучасними знаннями. *Спеціальні (фахові, предметні) компетентності (СК):* 3. Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області 4. Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проектних рішень. 12. Володіння методами подання знань і вміння застосовувати їх у системах штучного інтелекту.
Програмні результати навчання	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. 7. Розробляти та застосовувати математичні методи для аналізу інформаційних моделей. 8. Розробляти математичні моделі та методи аналізу даних (включно з великими). 20. Уміти працювати із задачами прийняття рішень та застосовувати в них алгоритми машинного навчання
Ключові слова	Лінійна регресія, метод К-найближчих сусідів, кластеризація, дерева рішень, метод основних компонент, логістична регресія.
Формат курсу	Очний
	Проведення лекцій, лабораторних робіт та консультації для кращого розуміння тем
Теми	Теми курсу наведено в схемі курсу нижче.
Підсумковий контроль, форма	екзамен в кінці семестру
Пререквізити	Для вивчення курсу студенти потребують базових знань з дисциплін «Програмування», «Теорія ймовірності та математична статистика», «Лінійна алгебра».
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Лекції, діалог, ілюстрація (з використанням презентації), лабораторний метод та практична робота, ілюстрація та дослідницький метод.
Необхідне обладнання	Для проведення лекцій: комп'ютер, проектор.

<table>
<tr><td></td><td>Для проведення лабораторних та виконання завдань: комп’ютер, ОС Windows, доступ до інтернету, R. За домовленістю з викладачем, індивідуальні завдання можуть виконуватися з використанням довільних прикладних програм чи мов програмування, зокрема Python, SAS тощо. Для оформлення звітів пропонується використовувати LibreOffice, чи MS Office 365, чи WinEdt.</td></tr>
<tr><td>Критерії оцінювання (окремо для кожного виду навчальної діяльності)</td><td>Оцінювання проводиться за 100-бальною шкалою.
<table>
<tr><th colspan="2" rowspan="2">Оцінка за шкалою ECTS</th><th rowspan="2">Оцінка в балах</th><th colspan="2">Екзамен</th></tr>
<tr></tr>
<tr><td>A</td><td>Відмінно</td><td>90-100</td><td>Відмінно</td><td>5</td></tr>
<tr><td>B</td><td>Дуже добре</td><td>81-89</td><td rowspan="2">Добре</td><td rowspan="2">4</td></tr>
<tr><td>C</td><td>Добре</td><td>71-80</td></tr>
<tr><td>D</td><td>Задовільно</td><td>61-70</td><td rowspan="2">Задовільно</td><td rowspan="2">3</td></tr>
<tr><td>E</td><td>Достатньо</td><td>51-60</td></tr>
<tr><td>F (FX)</td><td>Незадовільно</td><td>0-50</td><td>Незадовільно</td><td>2</td></tr>
</table>

Бали нараховуються за наступним співвідношенням:
• протягом семестру – виконання індивідуальних завдань за варіантами: 8 індивідуальних завдань (максимальна кількість балів за кожне 6,25). Максимальна кількість балів – 50. На протязі семестру необхідно виконати усі завдання. Для кожного завдання встановлено терміни здачі. Роботи, які здаються із порушенням термінів без поважних причин, оцінюються на нижчу оцінку. Запізнення до 7 днів –50%, від 8 до 14 днів –75 %, більше 14 днів – 90%.
Критерії оцінювання індивідуальних завдань
<table>
<tr><th>Кількість балів</th><th>Критерій оцінювання</th></tr>
<tr><td>6,25</td><td>студент повністю виконав умови завдання, алгоритми реалізовано правильно, при захисті роботи відповідає на всі запитання, пов'язані з тематикою завдання, проводить чіткий аналіз, порівняння та інтерпретацію отриманих результатів, пропонує інші підходи до вирішення поставленого завдання;</td></tr>
<tr><td>5-6</td><td>студент повністю виконав умови завдання, алгоритми реалізовано правильно, на деякі запитання, пов'язані з тематикою завдання відповідає з незначними неточностями, проводить аналіз, порівняння та інтерпретацію отриманих результатів з незначними неточностями;</td></tr>
<tr><td>3-4</td><td>студент виконав завдання з незначними помилками, проте самостійно їх виправляє та може пояснити, якщо на них вкаже викладач, на деякі запитання, пов'язані з тематикою завдання, відповідає з неточностями, проводить аналіз, порівняння та інтерпретацію отриманих результатів з неточностями;</td></tr>
<tr><td>2-3</td><td>студент виконав завдання частково, алгоритми реалізовано з помилками, які частково може виправити, якщо на них вкаже викладач, на запитання відповідає з помилками, проводить аналіз, порівняння та інтерпретацію отриманих результатів з помилками;</td></tr>
</table></td></tr>
</table>

	<table><tr><td>1-2</td><td>студент виконав завдання частково, алгоритм реалізовано з помилками, які самостійно не може виправити, переважно не відповідає на запитання, не здатний провести аналіз, порівняння та інтерпретацію отриманих результатів;</td></tr><tr><td>0</td><td>студент не володіє навчальним матеріалом і не виконав завдання</td></tr></table> • в кінці семестру – іспит: форма іспиту – тестування. Максимальна кількість балів 50. На іспиті студенту пропонується виконати 25 тестових завдань, кожне з яких оцінюється 2 балами (відповідь правильна) чи 0 балів (відповідь неправильна). На виконання завдань виділяється 30 хвилин. Підсумкова максимальна кількість балів 100 Очікується, що студенти виконають 8 письмових робіт у вигляді звітів. Очікується, що роботи студентів будуть їх оригінальними дослідженнями чи міркуваннями. Відсутність посилань на використані джерела, фабрикування джерел, списування, втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в письмовій роботі студента є підставою для її незарахуванння, незалежно від масштабів плагіату чи обману. **Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції і лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися усіх строків визначених для виконання усіх видів письмових робіт, передбачених курсом. При відсутності студента на лабораторному занятті без поважної причини, на наступному занятті відбувається захист індивідуальних завдань за темою пропущеного заняття. Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих. **Політика виставлення балів.** Враховуються бали набрані за індивідуальні завдання та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час лабораторного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін. Жодні форми порушення академічної доброчесності не толеруються.
Питання до екзамену.	1. Параметричні та непараметричні методи оцінювання параметрів. 2. Інтерпретованість та гнучкість моделі. 3. Навчання з вчителем та навчання без вчителя. 4. Якісні та кількісні дані. 5. Якість моделі. 6. Компроміс між зміщенням та дисперсією 7. Класифікатор Байеса. 8. Проста лінійна регресія. Оцінка коефіцієнтів. Оцінка точності оцінок коефіцієнтів. Оцінка точності моделі.

	9. Багатовимірна лінійна регресія. Оцінка коефіцієнтів. 10. Лінійна регресія. Якісні предиктори. 11. Розширення лінійної регресії. 12. Лінійна регресія. Потенційні проблеми. 13. Класифікація. Постновка задачі. 14. Логістична регресія. Модель. Оцінка коефіцієнтів регресії. Побудова прогнозів. 15. Множинна (багатовимірна) логістична регресія. 16. Логістична регресія для > 2 класів залежної змінної. 17. Лінійний дискримінаційний аналіз. Використання теореми Байєса для класифікації. 18. Лінійний дискримінантний аналіз для p = 1. 19. Лінійний дискримінантний аналіз для p> 1. 20. Квадратичний дискримінантний аналіз. 21. Перехресна перевірка. Множина перевірки. 22. Перехресна перевірка. LOOCV. 23. Перехресна перевірка. k-кратна перехресна перевірка. Компроміс між зміщенням і варіацією. 24. Перехресна перевірка у випадку класифікації. 25. Бутстрап. 26. Вибір лінійної моделі. Вибір підмножини предикторів. Вибір найкращої підмножини. 27. Вибір лінійної моделі. Вибір підмножини предикторів. Покроковий вибір. 28. Вибір лінійної моделі. Вибір підмножини предикторів. Вибір оптимальної моделі. 29. Методи стиснення. Гребенева регресія. 30. Методи стиснення. Регресія ласо. 31. Методи стиснення. Вибір параметра. 32. Методи зменшення розмірності. Метод головних компонент. 33. Методи зменшення розмірності. Метод часткових найменших квадратів. 34. Дані з високою розмірністю. Регресія у великих розмірностях. Інтерпретація результатів. 35. Нелінійні моделі. Поліноміальна регресія. 36. Нелінійні моделі. Східчасті функції. 36. Нелінійні моделі. Регресійні сплайни. Кускові многочлени. 37. Нелінійні моделі. Регресійні сплайни. Обмеження. 38. Нелінійні моделі. Регресійні сплайни. Кускові многочлени. Вибір кількості та місцезнаходження вузлів. 39. Нелінійні моделі. Згладжувальні сплайни. Вибір параметра згладжування λ. 40. Нелінійні моделі. Локальна регресія. 41. Нелінійні моделі. Узагальнені адитивні моделі для проблем регресії. 42. Нелінійні моделі. Узагальнені адитивні моделі для проблем для класифікаційних проблем. 43. Моделі на основі дерев. Дерева рішень для проблем регресії. Обрізка дерева. 44. Моделі на основі дерев. Дерева рішень для проблем класифікації. 45. Моделі на основі дерев. Переваги та недоліки. 46. Моделі на основі дерев. Бутстрап агрегація. 47. Моделі на основі дерев. Випадкові ліси. 48. Моделі на основі дерев. Підсилення.

	49. Метод опорних векторів (МОВ). Класифікація на основі розділювальної гіперплощини. 50. Лінійний МОВ, жорстке розділення. 51. Лінійний МОВ, м'ягке розділення. 52. Нелінійний МОВ, м'ягке розділення. 53. МОВ з більш ніж двома класами. Класифікація «Один проти одного». 54. МОВ з більш ніж двома класами. Класифікація «Один проти всіх». 55. Навчання без вчителя. Основні проблеми. 56. Навчання без вчителя. Метод головних компонент. 57. Навчання без вчителя. Кластеризація методом k-середніх. 58. Навчання без вчителя. Ієрархічна кластеризація. 59. Навчання без вчителя. Практичні проблеми кластеризації.
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

Тиж.	Тема, план, короткі тези	Форма діяльності (заняття)* *лекція, самостійна, дискусія, групова робота)	Завдання, год	Термін виконання
1-2	Вступ до статистичного навчання. Основи статистичного навчання. Параметричні та непараметричні методи оцінювання параметрів. Інтерпретованість та гнучкість моделі. Навчання з вчителем та навчання без вчителя. Якісні та кількісні дані. Якість моделі. Компроміс між зміщенням та дисперсією. Класифікатор Байеса.	лекція	4	
1	Вступ до R.	лабораторна робота	2	Наступне лабораторне заняття
3-4	Проста лінійна регресія. Оцінка коефіцієнтів. Оцінка точності оцінок коефіцієнтів. Оцінка точності моделі. Багатовимірна лінійна регресія. Оцінка коефіцієнтів. Лінійна регресія. Якісні предиктори. Розширення лінійної регресії. Лінійна регресія. Потенційні проблеми.	лекція	4	
2-3	Лінійна регресія	лабораторна робота	4	Наступне лабораторне заняття
5-6	Класифікація. Постновка задачі. Логістична регресія. Модель. Оцінка коефіцієнтів регресії. Побудова прогнозів. Множинна (багатовимірна) логістична регресія. Логістична регресія для > 2 класів залежної змінної. Лінійний дискримінаційний аналіз. Використання теореми Байєса для класифікації. Лінійний дискримінантний аналіз для p = 1. Лінійний дискримінантний аналіз для p> 1. Квадратичний дискримінантний аналіз.	лекція	4	

4-5	Логістична регресія, лінійний та квадратичний дискримінантний аналіз та метод К-найближчих сусідів.	лабораторна робота	4	Наступне лабораторне заняття
7	Перехресна перевірка. Множина перевірки. LOOCV. k-кратна перехресна перевірка. Компроміс між зміщенням і варіацією. Перехресна перевірка у випадку класифікації. Бутстрап.	лекція	2	
6-7	Перехресна перевірка, Бутстрап.	лабораторна робота	4	Наступне лабораторне заняття
8-9	Вибір лінійної моделі. Вибір підмножини предикторів. Вибір найкращої підмножини. Покроковий вибір. Вибір оптимальної моделі. Методи стиснення. Гребенева регресія. Регресія ласо. Вибір параметра. Методи зменшення розмірності. Метод головних компонент. Метод часткових найменших квадратів. Дані з високою розмірністю. Регресія у великих розмірностях. Інтерпретація результатів.	лекція	4	
8-9	Методи вибору підмножини, гребенева регресія та Лассо, метод головних компонент та часткових найменших квадратів.	лабораторна робота	4	Наступне лабораторне заняття
10-11	Нелінійні моделі. Поліноміальна регресія. Східчасті функції. Регресійні сплайни. Кускові многочлени. Обмеження. Кускові многочлени. Вибір кількості та місцезнаходження вузлів. Згладжувальні сплайни. Вибір параметра згладжування λ. Локальна регресія. Узагальнені адитивні моделі для проблем регресії. Узагальнені адитивні моделі для проблем для класифікаційних проблем.	лекція	4	
10-11	Нелінійні моделі.	лабораторна робота	4	Наступне лабораторне заняття
12	Моделі на основі дерев. Дерева рішень для проблем регресії. Обрізка дерева. Дерева рішень для проблем класифікації. Переваги та недоліки. Бутстрап агрегація. Випадкові ліси. Підсилення.	лекція	2	
12	Дерева рішень.	лабораторна робота	2	Наступне лабораторне заняття
13-14	Метод опорних векторів (МОВ). Класифікація на основі розділювальної гіперплощини. Лінійний МОВ, жорстке розділення.Лінійний МОВ, м’ягке розділення. Нелінійний МОВ, м’ягке розділення. МОВ з більш ніж двома класами. Класифікація «Один проти одного». Класифікація «Один проти всіх».	лекція	4	
13-14	Метод опорних векторів.	лабораторна робота	4	Наступне лабораторне заняття
15-16	Навчання без вчителя. Основні проблеми. Метод головних компонент. Кластеризація методом k-середніх. Ієрархічна кластеризація. Практичні проблеми кластеризації.	лекція	4	

15-16	Навчання без вчителя	лабораторна робота	4	