



СТАТИСТИЧНІ МЕТОДИ МАШИННОГО НАВЧАННЯ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	<i>Перший (бакалаврський)</i>
Галузь знань	<i>12 Інформаційні технології</i>
Спеціальність	<i>123 Комп'ютерна інженерія</i>
Освітня програма	<i>Комп'ютерні системи та мережі</i>
Статус дисципліни	<i>Вибіркова</i>
Форма навчання	<i>Очна (денна)</i>
Рік підготовки, семестр	<i>4 курс, весняний семестр</i>
Обсяг дисципліни	<i>4 кредити, 120 годин</i>
Семестровий контроль/ контрольні заходи	<i>Залік</i>
Розклад занять	<i>Лекції – 36 годин. Лабораторні роботи - 36 годин. Самостійна робота студентів- 48 години</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	Лектор: д.т.н, професор, Новотарський Михайло Анатолійович novotar@gmail.com, http://novotarsky.pp.ua Лабораторні: д.т.н, професор, Новотарський Михайло Анатолійович novotar@gmail.com, http://novotarsky.pp.ua
Розміщення курсу	https://ecampus.kpi.ua

Програма навчальної дисципліни

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

Силабус дисципліни «Статистичні методи машинного навчання» складено відповідно до освітньо-професійної програми підготовки бакалаврів спеціальності 123 – «Комп'ютерна інженерія».

2. Навчальна дисципліна «Статистичні методи машинного навчання» є вибірковою дисципліною.

Програма навчальної дисципліни «Статистичні методи машинного навчання» встановлює вимоги до знань і вмінь студента і визначає зміст і види навчальних занять.

Вивчення навчальної дисципліни «Статистичні методи машинного навчання» базується на наступних дисциплінах:

- Математичний аналіз
- Лінійна алгебра та аналітична геометрія
- Теорія імовірностей та математична статистика

Для освоєння навчальної дисципліни студенти повинні володіти:

- знаннями основних визначень і теорем наведених вище дисциплін;
- навичками розв'язування типових задач цих дисциплін.

1.1. Мета та завдання навчальної дисципліни

1.1. Мета навчальної дисципліни

Метою освоєння дисципліни «Статистичні методи машинного навчання» є вивчення

основних методів машинного навчання для задач класифікації, кластеризації і регресії (прогнозування). В рамках даного курсу студенти отримують уявлення про завдання, які вирішують за допомогою даної теорії, і принципах побудови деяких основних класифікаторів. При вивченні дисципліни значна увага приділена алгоритмічним та обчислювальним аспектам, тому студенти додатково отримують знання з мови програмування Python та бібліотеки THEANO як інструментів відпрацювання практичних навичок.

1.2. Основні завдання навчальної дисципліни

Згідно з вимогами освітньо-професійної програми студенти після засвоєння навчальної дисципліни мають продемонструвати такі результати навчання:

Знання: принципів побудови векторів ознак, правил рішень і класифікації; основних видів класифікаторів; принципів побудови лінійних класифікаторів; принципів побудови нелінійних класифікаторів; особливостей вибору ознак класифікації і попередньої обробки даних.

Вміння: вибирати відповідний вид класифікатора залежно від розв'язуваної задачі; вибирати набір ознак для класифікації і проводити попередню обробку даних; вміти застосовувати алгоритми побудови і навчання класифікатора за вибіркою; виконувати обчислення, пов'язані з навчанням і роботою класифікатора, в середовищі Python з застосуванням бібліотеки THEANO.

Досвід: студент повинен мати навички: вибору, побудови, навчання і використання основних класифікаторів при вирішенні завдань; самостійної роботи в сучасних програмних комплексах; освоєння великого обсягу інформації; програмування для вирішення задач аналізу даних; мати культуру постановки завдань та проведення експерименту; володіти засобами візуалізації для демонстрації одержаних результатів.

2. Структура навчальної дисципліни

На вивчення навчальної дисципліни відводиться 120 годин/4 кредити ECTS.

Навчальна дисципліна містить 1 кредитний модуль: «Статистичні методи машинного навчання»

Розподіл навчального часу

Форма навчання	Кредитні модулі	Всього		Розподіл навчального часу за видами занять			Семестрова атестація
		Кредитів	Годин	Лекції	Лабораторні роботи	СРС	
Денна	1	4	120	36	36	48	Залік

3. Зміст навчальної дисципліни

Тема 1. Вступні положення теорії машинного навчання

Поняття про машинне навчання. Сфери застосування машинного навчання. Типи навчання. Дедуктивне та індуктивне навчання. Зв'язок з іншими областями науки. Символьне навчання. Сучасні реалізації символного навчання.

Тема 2. Постановка задачі і базові поняття про статистичне навчання

Об'єкти і ознаки. Типи задач. Модель алгоритмів та метод навчання. Приклади моделей. Функціонал якості. Мінімізація емпіричного ризику. Проблема перенавчання. Прикладні задачі навчання: задачі класифікації, задачі оцінки регресії, задачі ранжування; задачі кластеризації; задачі пошуку асоціацій. Методика тестування алгоритмів навчання. Прийоми генерації модельних даних.

Тема 3. Метричні методи класифікації

Метод найближчого сусіда і його узагальнення: узагальнений метричний класифікатор, метод k -найближчих сусідів, метод парзенівського вікна, метод потенційних функцій. Вибір еталонних об'єктів: поняття відступу об'єкта, алгоритм STOLP для відбору еталонних об'єктів.

Тема 4. Лінійні методи класифікації

Штучний нейрон як модель нервової клітини. Конективізм і нейронні мережі. Апроксимація й регуляризація емпіричного ризику. Лінійна модель класифікації. Метод стохастичного градієнта: класичні окремі випадки, евристики для поліпшення градієнтних методів навчання.

Тема 5. Логістична регресія. Метод опорних векторів

Обґрунтування логістичної регресії. Метод стохастичного градієнта для логістичної регресії: принцип максимуму правдоподібності, градієнтний крок, аналогія з правилом Хебба, переваги та недоліки логістичної регресії. Скоринг і оцінювання апіорних ймовірностей, імовірнісний вихід та оцінювання ризиків, імовірнісна калібровка. Метод опорних векторів. Вибірка з лінійною роздільністю: оптимальна розділяюча гіперплощина, нормування, ширина роздільної смуги. Лінійно не роздільна вибірка: регуляризація емпіричного ризику, двоїста задача. Ядра і спрямляючі простори: конструктивні способи побудови ядер. Двошарова нейронна мережа для реалізації методу опорних векторів. Переваги та недоліки методу опорних векторів.

Тема 6. Байєсівська теорія класифікації. Основні положення. Імовірнісна постановка задачі. Непараметрична класифікація.

Основні положення теорії ймовірностей, необхідні для використання в рамках курсу: функція розподілу ймовірностей дискретної випадкової величини та її властивості, неперервні випадкові величини, умовна ймовірність, повна ймовірність, формула Байєса. Імовірнісна постановка задачі класифікації. Функціонал середнього ризику. Аргументи максимізації та мінімізації. Оптимальний класифікатор Байєса. «Наївний» байєсівський класифікатор. Непараметрична класифікація: одновимірний неперервний випадок, багатовимірний неперервний випадок. Метод парзенівського вікна.

Тема 7. Байєсівська теорія класифікації. Нормальний дискримінантний аналіз. Розділення суміші розподілів.

Багатовимірний нормальний розподіл. Квадратичний дискримінант. Лінійний дискримінант Фішера. EM-алгоритм. Суміші багатовимірних нормальних розподілів.

Тема 8. Методи відновлення регресії (оцінки регресії)

Метод найменших квадратів. Непараметрична регресія: ядерне згладжування, формула Надарая-Ватсона, вибір ядра й ширини вікна, проблема викидів (робастна непараметрична регресія), алгоритм LOWESS (локально зважене згладжування), проблема крайових ефектів. Лінійна регресія: сингулярне розкладання, проблема мультиколінеарності, гребенева регресія. Нелінійні методи відновлення регресії: нелінійна модель регресії, нелінійні одновимірні перетворення ознак.

Тема 9. Логічні алгоритми класифікації

Постановка задачі для логічних методів класифікації. Статистичне визначення інформативності. Ентропійне визначення інформативності. Багатокласова інформативність. Зважена інформативність. Методи пошуку інформативних закономірностей: бінаризація кількісних ознак, «жадібний» алгоритм злиття зон, пошук закономірностей у формі кон'юнкцій, «градієнтний» алгоритм синтезу кон'юнкції, «жадібний» алгоритм синтезу кон'юнкції, стохастичний локальний пошук, емерджентний алгоритм синтезу кон'юнкцій, форми закономірностей.

Тема 10. Списки та дерева ухвалення рішень

Списки ухвалення рішень: жадібний алгоритм побудови списку ухвалення рішень, приклади списків ухвалення рішень. Дерева ухвалення рішень: синтез дерев ухвалення рішень, алгоритм побудови дерева ухвалення рішень ID3, обробка пропусків, оцінювання ймовірностей, трудомісткість алгоритму ID3, переваги і недоліки алгоритму ID3, редукція дерев ухвалення рішень. Перетворення дерева рішень у список рішень. Заглядання вперед.

Тема 11. Зважене голосування правил

Принципи голосування: алгоритм простого голосування, алгоритм зваженого голосування, налаштування ваг, диверсифікація правил, відмови від класифікації. Алгоритм КОРА: принципи побудови різних модифікацій алгоритму КОРА, переваги та недоліки алгоритму КОРА. Алгоритм ТЕМП: принципи роботи алгоритму, переваги та недоліки алгоритму ТЕМП. Алгоритм бустинга: експонентна апроксимація граничної функції втрат, принципи роботи алгоритму бустинга, переваги та недоліки алгоритму бустинга.

Тема 12. Штучні нейронні мережі

Проблема повноти та задача «XOR». Обчислювальні можливості нейронних мереж. Багатошарові нейронні мережі та метод зворотного поширення помилки. Переваги та недоліки методу зворотного поширення помилки. Евристики для покращення збіжності: вибір початкового наближення, вибір градієнтного методу оптимізації. Оптимізація структури нейронної мережі: вибір кількості шарів, вибір кількості нейронів у прихованому шарі, динамічне додавання нейронів, видалення надлишкових зв'язків.

Тема 13. Кластеризація. Алгоритми кластеризації

Цілі кластеризації. Типи кластерних структур. Евристичні графові алгоритми: алгоритм виділення зв'язних компонентів, алгоритм найкоротшого незамкнутого шляху, алгоритм FOREL. Функціонали якості кластеризації. Статистичні алгоритми: гіпотеза про простір об'єктів і форму кластерів, метод k -середніх, кластеризація з частковим навчанням. Ієрархічна кластеризація: властивість монотонності, теорема Міллігана, властивості розтягування і стискання, властивість редукованості, теорема Діде і Моро, визначення числа кластерів, переваги і недоліки кластеризації.

Тема 14. Мережі Кохонена

Моделі конкурентного навчання: правило жорсткої конкуренції WTA, правило справедливої конкуренції CWTA, правило м'якої конкуренції WTM. Карти Кохонена, що самоорганізуються, мистецтво інтерпретації карт Кохонена. Недоліки карт Кохонена. Гібридні мережі зустрічного поширення: кусочно-постійна апроксимація, гладка апроксимація. Багатовимірне шкакування: розміщення одного об'єкта методом Ньютона-

Рафсона, субквадратичний алгоритм багатовимірного шкалування, карта подібності, діаграма Шепарда.

Навчальні матеріали та ресурси

Базова:

1. Новотарський М.А. Лекції х курсу «Статистичні методи машинного навчання» // <https://cloud.comsys.kpi.ua/s/pjw4nAStZFPsDos>
2. Новотарський М.А. Методичні вказівки до виконання лабораторних робіт з курсу «Основи науки продані» // Новотарський М.А. Лекції х курсу «Основи науки продані» //
3. Asuncion A., Newman D. UCI machine learning repository: Tech. rep.: University of California, Irvine, School of Information and Computer Sciences, 2007. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
4. Bishop C. M. Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning.— Springer, 2001. — 533 pp. <http://http://www-stat.stanford.edu/~tibs/ElemStatLearn>.
6. Jordan M. I., Xu L. Convergence results for the EM algorithm to mixtures of experts architectures: Tech. Rep. A.I. Memo No. 1458: MIT, Cambridge, MA, 1993.
7. LeCun Y., Bottou L., Orr G. B., Muller K.-R. Efficient BackProp // Neural Networks: tricks of the trade. — Springer, 1998.
8. Smola A., Schoelkopf B. A tutorial on support vector regression: Tech. Rep. NeuroCOLT2 NC2-TR-1998-030: Royal Holloway College, London, UK, 1998. <http://citeseer.ist.psu.edu/smola98tutorial.html>.
9. Новотарський М.А., Нестеренко Б.Б. Штучні нейронні мережі: обчислення. К: Ін-т математики НАН України, 2004.—408 с.

Додаткова:

1. Келлехер Дж. Наука про дані: Базовий курс / Дж.Келлехер, Б. Тирни, Паблішер, 2020. — 157 с.
2. Wang L., Fu X. Data Mining with Computational Intelligence. —Springer, 2005. —280 p.

Навчальний контент

4. Методика опанування навчальної дисципліни (освітнього компонента)

Структура навчальної дисципліни «Статистичні методи машинного навчання» в таблиці 1 .

Таблиця 1

Структура навчальної дисципліни «Статистичні методи машинного навчання»

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Лабораторні роботи	СРС
Тема 1. Вступні положення теорії машинного навчання Поняття про машинне навчання. Сфери застосування машинного навчання. Типи навчання. Дедуктивне та індуктивне навчання. Зв'язок з іншими областями науки.	6	2	4	0

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Лабораторні роботи	СРС
Символьне навчання. Сучасні реалізації символічного навчання.				
Тема 2. Постановка задачі і базові поняття про статистичне навчання Об'єкти і ознаки. Типи задач. Модель алгоритмів та метод навчання. Приклади моделей. Функціонал якості. Мінімізація емпіричного ризику. Проблема перенавчання. Прикладні задачі навчання: задачі класифікації, задачі оцінки регресії, задачі ранжування; задачі кластеризації; задачі пошуку асоціацій. Методика тестування алгоритмів навчання. Прийоми генерації модельних даних.	3	2	0	1
Тема 3. Метричні методи класифікації Метод найближчого сусіда і його узагальнення: узагальнений метричний класифікатор, метод k -найближчих сусідів, метод парзенівського вікна, метод потенційних функцій. Відбір еталонних об'єктів: поняття відступу об'єкта, алгоритм STOLP для відбору еталонних об'єктів.	6	2	4	0
Тема 4. Лінійні методи класифікації Штучний нейрон як модель нервової клітини. Конективізм і нейронні мережі. Апроксимація й регуляризація емпіричного ризику. Лінійна модель класифікації. Метод стохастичного градієнта: класичні окремі випадки, евристики для поліпшення градієнтних методів навчання.	6	2	4	
Тема 5. Логістична регресія. Метод опорних векторів Обґрунтування логістичної регресії. Метод стохастичного градієнта для логістичної регресії: принцип максимуму правдоподібності, градієнтний крок, аналогія з правилом Хебба, переваги та недоліки логістичної регресії. Скоринг і оцінювання апріорних ймовірностей, імовірнісний вихід та оцінювання ризиків, імовірнісна калібровка. Метод опорних векторів. Вибірка з лінійною роздільністю: оптимальна розділяюча гіперплощина, нормування, ширина роздільної смуги. Лінійно не роздільна вибірка: регуляризація емпіричного ризику, двоїста задача. Ядра і спрямляючі простори: конструктивні способи побудови ядер. Двошарова нейронна мережа для реалізації методу опорних векторів. Переваги та недоліки методу опорних векторів.	8	4	4	0

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Лабораторні роботи	СРС
<p>Тема 6. Байєсівська теорія класифікації. Основні положення. Імовірнісна постановка задачі. Непараметрична класифікація.</p> <p>Основні положення теорії імовірностей, необхідні для використання в рамках курсу: функція розподілу імовірностей дискретної випадкової величини та її властивості, неперервні випадкові величини, умовна імовірність, повна імовірність, формула Байєса. Імовірнісна постановка задачі класифікації. Функціонал середнього ризику. Аргументи максимізації та мінімізації. Оптимальний класифікатор Байєса. «Наївний» байєсівський класифікатор. Непараметрична класифікація: одновимірний неперервний випадок, багатовимірний неперервний випадок. Метод парзенівського вікна.</p>	6	2	4	0
<p>Тема 7. Байєсівська теорія класифікації. Нормальний дискримінантний аналіз. Розділення суміші розподілів</p> <p>Багатовимірний нормальний розподіл. Квадратичний дискримінант. Лінійний дискримінант Фішера. EM-алгоритм. Суміші багатовимірних нормальних розподілів.</p>	10	2	8	0
<p>Тема 8. Методи відновлення регресії (оцінки регресії)</p> <p>Метод найменших квадратів. Непараметрична регресія: ядерне згладжування, формула Надарая-Ватсона, вибір ядра й ширини вікна, проблема викидів (робастна непараметрична регресія), алгоритм LOWESS (локально зважене згладжування), проблема крайових ефектів. Лінійна регресія: сингулярне розкладання, проблема мультиколінеарності, гребенева регресія. Нелінійні методи відновлення регресії: нелінійна модель регресії, нелінійні одновимірні перетворення ознак.</p>	4	2	0	2
<p>Тема 9. Логічні алгоритми класифікації</p> <p>Постановка задачі для логічних методів класифікації. Статистичне визначення інформативності. Ентропійне визначення інформативності. Багатокласова інформативність Зважена інформативність. Методи пошуку інформативних закономірностей: бінаризація кількісних ознак, «жадібний» алгоритм злиття зон, пошук закономірностей у формі кон'юнкцій,</p>	3	2	0	1

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Лабораторні роботи	СРС
«градієнтний» алгоритм синтезу кон'юнкції, «жадібний» алгоритм синтезу кон'юнкції, стохастичний локальний пошук, емерджентний алгоритм синтезу кон'юнкції, форми закономірностей.				
Тема 10. Списки та дерева ухвалення рішень Списки ухвалення рішень: жадібний алгоритм побудови списку ухвалення рішень, приклади списків ухвалення рішень. Дерева ухвалення рішень: синтез дерев ухвалення рішень, алгоритм побудови дерева ухвалення рішень ID3, обробка пропусків, оцінювання ймовірностей, трудомісткість алгоритму ID3, переваги і недоліки алгоритму ID3, редукція дерев ухвалення рішень. Перетворення дерева рішень у список рішень. Заглядання вперед.	8	4	4	0
Тема 11. Зважене голосування правил Принципи голосування: алгоритм простого голосування, алгоритм зваженого голосування, налаштування ваг, диверсифікація правил, відмови від класифікації. Алгоритм КОРА: принципи побудови різних модифікацій алгоритму КОРА, переваги та недоліки алгоритму КОРА. Алгоритм ТЕМП: принципи роботи алгоритму, переваги та недоліки алгоритму ТЕМП. Алгоритм бустинга: експонентна апроксимація граничної функції втрат, принципи роботи алгоритму бустинга, переваги та недоліки алгоритму бустинга.	5	4	0	1
Тема 12. Штучні нейронні мережі Проблема повноти та задача «XOR». Обчислювальні можливості нейронних мереж. Багатощарові нейронні мережі та метод зворотного поширення помилки. Переваги та недоліки методу зворотного поширення помилки. Евристики для покращення збіжності: вибір початкового наближення, вибір градієнтного методу оптимізації. Оптимізація структури нейронної мережі: вибір кількості шарів, вибір кількості нейронів у прихованому шарі, динамічне додавання нейронів, видалення надлишкових зв'язків.	4	2	0	2
Тема 13. Кластеризація. Алгоритми кластеризації	5	4	0	1

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Лабораторні роботи	СРС
Цілі кластеризації. Типи кластерних структур. Евристичні графові алгоритми: алгоритм виділення зв'язних компонентів, алгоритм найкоротшого незамкнутого шляху, алгоритм FOREL. Функціонали якості кластеризації. Статистичні алгоритми: гіпотеза про простір об'єктів і форму кластерів, метод k -середніх, кластеризація з частковим навчанням. Ієрархічна кластеризація: властивість монотонності, теорема Міллігана, властивості розтягування і стискання, властивість редукованості, теорема Діде і Моро, визначення числа кластерів, переваги і недоліки кластеризації.				
Тема 14. Мережі Кохонена Моделі конкурентного навчання: правило жорсткої конкуренції WTA, правило справедливої конкуренції CWTA, правило м'якої конкуренції WTM. Карти Кохонена, що самоорганізуються, мистецтво інтерпретації карт Кохонена. Недоліки карт Кохонена. Гібридні мережі зустрічного поширення: кусочно-постійна апроксимація, гладка апроксимація. Багатовимірне шкакування: розміщення одного об'єкта методом Ньютона-Рафсона, субквадратичний алгоритм багатовимірного шкакування, карта подібності, діаграма Шепарда.	8	2	4	2
Контрольна робота	2		0	2
Екзамен	36		0	36
Всього в семестрі:	120	36	36	48

Тематика лекційних занять, яка сформована відповідно до тем, що розглядаються у рамках навчальної дисципліни, наведена у таблиці 2.

Таблиця 2

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
1	Вступні положення теорії машинного навчання Поняття про машинне навчання. Сфери застосування машинного навчання. Типи навчання. Дедуктивне та індуктивне навчання. Зв'язок з іншими областями науки. Символьне навчання. Сучасні реалізації символьного навчання. Завдання на СРС. Задачі навчання по прецедентах.
2	Постановка задачі і базові поняття про статистичне навчання Об'єкти і ознаки. Типи задач. Модель алгоритмів та метод навчання. Приклади моделей. Функціонал якості. Мінімізація емпіричного ризику. Проблема перенавчання. Прикладні задачі навчання: задачі класифікації, задачі оцінки

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
	регресії, задачі ранжування; задачі кластеризації; задачі пошуку асоціацій. Методика тестування алгоритмів навчання. Прийоми генерації модельних даних. Завдання на СРС. Поняття узагальнюючої здатності.
3	Метричні методи класифікації Метод найближчого сусіда і його узагальнення: узагальнений метричний класифікатор, метод k -найближчих сусідів, метод парзенівського вікна, метод потенційних функцій. Відбір еталонних об'єктів: поняття відступу об'єкта, алгоритм STOLP для відбору еталонних об'єктів. Завдання на СРС. Недоліки найпростіших метричних алгоритмів типу kNN .
4	Лінійні методи класифікації Штучний нейрон як модель нервової клітини. Конективізм і нейронні мережі. Апроксимація й регуляризація емпіричного ризику. Лінійна модель класифікації. Метод стохастичного градієнта: класичні окремі випадки, евристики для поліпшення градієнтних методів навчання. Завдання на СРС. Незалежні параметри з нерівними дисперсіями.
5	Логістична регресія Обґрунтування логістичної регресії. Метод стохастичного градієнта для логістичної регресії: принцип максимуму правдоподібності, градієнтний крок, аналогія з правилом Хебба, переваги та недоліки логістичної регресії. Скоринг і оцінювання апіорних імовірностей, імовірнісний вихід та оцінювання ризиків, імовірнісна калібровка. Завдання на СРС. Переваги та недоліки логістичної регресії.
6	Метод опорних векторів. Вибірка з лінійною роздільністю: оптимальна розділяюча гіперплощина, нормування, ширина роздільної смуги. Лінійно не роздільна вибірка: регуляризація емпіричного ризику, двоїста задача. Ядра і спрямляючі простори: конструктивні способи побудови ядер. Двошарова нейронна мережа для реалізації методу опорних векторів. Переваги та недоліки методу опорних векторів. Завдання на СРС. Метод релевантних векторів (RVM).
7	Байєсівська теорія класифікації. Основні положення. Імовірнісна постановка задачі. Непараметрична класифікація. Основні положення теорії імовірностей, необхідні для використання в рамках курсу: функція розподілу імовірностей дискретної випадкової величини та її властивості, неперервні випадкові величини, умовна імовірність, повна імовірність, формула Байєса. Імовірнісна постановка задачі класифікації. Функціонал середнього ризику. Аргументи максимізації та мінімізації. Оптимальний класифікатор Байєса. «Наївний» байєсівський класифікатор. Непараметрична класифікація: одновимірний неперервний випадок, багатовимірний неперервний випадок. Метод парзенівського вікна. Завдання на СРС. Принцип максимуму апостеріорної імовірності.
8	Байєсівська теорія класифікації. Нормальний дискримінантний аналіз. Розділення суміші розподілів. Багатовимірний нормальний розподіл. Квадратичний дискримінант. Лінійний дискримінант Фішера. EM-алгоритм. Суміші багатовимірних нормальних розподілів. Завдання на СРС. Відстань Махаланобіса.

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
9	<p>Методи відновлення регресії (оцінки регресії) Метод найменших квадратів. Непараметрична регресія: ядерне згладжування, формула Надарая-Ватсона, вибір ядра й ширини вікна, проблема викидів (робастна непараметрична регресія), алгоритм LOWESS (локально зважене згладжування), проблема крайових ефектів. Лінійна регресія: сингулярне розкладання, проблема мультиколінеарності, гребенева регресія. Нелінійні методи відновлення регресії: нелінійна модель регресії, нелінійні одновимірні перетворення ознак. Завдання на СРС. Ласо Тибширані. Порівняння ласо і гребневої регресії.</p>
10	<p>Логічні алгоритми класифікації Постановка задачі для логічних методів класифікації. Статистичне визначення інформативності. Ентропійне визначення інформативності. Багатокласова інформативність Зважена інформативність. Методи пошуку інформативних закономірностей: бінаризація кількісних ознак, «жадібний» алгоритм злиття зон, пошук закономірностей у формі кон'юнкцій, «градієнтний» алгоритм синтезу кон'юнкції, «жадібний» алгоритм синтезу кон'юнкції, стохастичний локальний пошук, емерджентний алгоритм синтезу кон'юнкцій, форми закономірностей. Завдання на СРС. Співставлення двох критеріїв інформативності.</p>
11	<p>Списки та дерева ухвалення рішень Списки ухвалення рішень: жадібний алгоритм побудови списку ухвалення рішень, приклади списків ухвалення рішень. Дерева ухвалення рішень: синтез дерев ухвалення рішень. Завдання на СРС. Переваги та недоліки списків ухвалення рішень.</p>
12	<p>Списки та дерева ухвалення рішень (продовження) Алгоритм побудови дерева ухвалення рішень ID3, обробка пропусків, оцінювання ймовірностей, трудомісткість алгоритму ID3, переваги і недоліки алгоритму ID3, редукція дерев ухвалення рішень. Перетворення дерева рішень у список рішень. Заглядання вперед. Завдання на СРС. Трудомісткість задачі побудови дерева ухвалення рішень у загальному вигляді.</p>
13	<p>Зважене голосування правил Принципи голосування: алгоритм простого голосування, алгоритм зваженого голосування, налаштування ваг, диверсифікація правил, відмови від класифікації. Алгоритм КОРА: принципи побудови різних модифікацій алгоритму КОРА, переваги та недоліки алгоритму КОРА. Завдання на СРС. Деталі реалізації алгоритму КОРА.</p>
14	<p>Зважене голосування правил Алгоритм ТЕМП: принципи роботи алгоритму, переваги та недоліки алгоритму ТЕМП. Алгоритм бустинга: експонентна апроксимація граничної функції втрат, принципи роботи алгоритму бустинга, переваги та недоліки алгоритму бустинга. Завдання на СРС. Теорема про збіжність алгоритму бустинга. <i>разнотипных экспериментальных данных. – Новосибирск: Наука, 1981.</i></p>
15	<p>Штучні нейронні мережі Проблема повноти та задача «XOR». Обчислювальні можливості нейронних мереж. Багатошарові нейронні мережі та метод зворотного поширення помилки.</p>

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
	<p>Переваги та недоліки методу зворотного поширення помилки. Евристики для покращення збіжності: вибір початкового наближення, вибір градієнтного методу оптимізації. Оптимізація структури нейронної мережі: вибір кількості шарів, вибір кількості нейронів у прихованому шарі, динамічне додавання нейронів, видалення надлишкових зв'язків.</p> <p>Завдання на СРС. Теорема Колмогорова про представлення будь-якої неперервної функції n аргументів суперпозицією функцій одного аргументу.</p>
16	<p>Кластеризація. Алгоритми кластеризації</p> <p>Цілі кластеризації. Типи кластерних структур. Евристичні графові алгоритми: алгоритм виділення зв'язних компонентів, алгоритм найкоротшого незамкнутого шляху, алгоритм FOREL. Функціонали якості кластеризації.</p> <p>Завдання на СРС. Недоліки алгоритму виділення зв'язних компонентів.</p>
17	<p>Кластеризація. Алгоритми кластеризації (продовження)</p> <p>Статистичні алгоритми: гіпотеза про простір об'єктів і форму кластерів, метод k-середніх, кластеризація з частковим навчанням. Ієрархічна кластеризація: властивість монотонності, теорема Міллігана, властивості розтягування і стискання, властивість редукованості, теорема Діде і Моро, визначення числа кластерів, переваги і недоліки кластеризації.</p> <p>Завдання на СРС. Поняття ланцюгового ефекту.</p>
18	<p>Мережі Кохонена</p> <p>Моделі конкурентного навчання: правило жорсткої конкуренції WTA, правило справедливої конкуренції SWTA, правило м'якої конкуренції WTM. Карти Кохонена, що самоорганізуються, мистецтво інтерпретації карт Кохонена. Недоліки карт Кохонена. Гібридні мережі зустрічного поширення: кусочно-постійна апроксимація, гладка апроксимація. Багатовимірне шкакування: розміщення одного об'єкта методом Ньютона-Рафсона, субквадратичний алгоритм багатовимірного шкакування, карта подібності, діаграма Шепарда.</p> <p>Завдання на СРС. Ідеальна діаграма Шепарда.</p>

5. Самостійна робота студента

Самостійна робота студента складається з теоретичної та практичної складової. Теоретична складова передбачає вивчення додаткового матеріалу, що поглиблює знання, які отримані на лекції. Практична складова самостійної роботи студента полягає у виконанні лабораторних робіт, перелік яких наведено в таблиці 3.

Основне завдання циклу лабораторних занять полягає у набутті студентами необхідних практичних навичок розробки алгоритмів та програмного забезпечення для розв'язування задач обробки даних на ПК.

Метою проведення циклу лабораторних робіт є набуття студентами необхідних практичних навичок використання методів.

Лабораторна робота включає створення блок-схеми алгоритму, розробку програми на скриптовій мові програмування Python з застосуванням бібліотеки Theano, роздрукування результатів виконання програми та аналіз одержаних результатів.

За базове середовище розробки програм прийнято середовище розробки для Python – PyCharm Edu.

Перелік лабораторних робіт

№ з/п	Назва лабораторної роботи (комп'ютерного практикуму)	Кількість ауд. годин
1	Лабораторна робота № 1. Символьні обчислення та символічне навчання (Тема 1).	7
2	Лабораторна робота № 2. Метричні та лінійні методи навчання (Тема 2,3,4,5).	7
3	Лабораторна робота № 3. Байєсівські методи навчання та відновлення регресії. (Тема 6,7).	8
4	Лабораторна робота № 4. Логічні методи навчання. (Тема 9,10,11).	7
5	Лабораторна робота № 5. Нейромережеві методи навчання. (Тема 12,13,14).	7

Політика та контроль

6. Політика навчальної дисципліни (освітнього компонента)

Під час занять з навчальної дисципліни «Статистичні методи машинного навчання» студенти повинні дотримуватись певних дисциплінарних правил:

- забороняється запізнюватись на заняття;
- при вході викладача, на знак привітання, особи, які навчаються в КПІ ім. Ігоря Сікорського, повинні встати;
- не допускаються сторонні розмови або інший шум, що заважає проведенню занять;
- виходити з аудиторії під час заняття допускається лише з дозволу викладача;
- не допускається користування мобільними телефонами та іншими технічними засобами без дозволу викладача.

Лабораторні роботи здаються особисто з попередньою перевіркою теоретичних знань, які необхідні для виконання лабораторної роботи. Перевірка практичних результатів включає перевірку коду та виконання тестових завдань.

В процесі навчання викладач має право нарахувати до 5 заохочувальних балів за дострокове виконання лабораторної роботи, за проявлений творчий підхід при виконанні індивідуального завдання або за активну участь у обговоренні питань, що пов'язані з тематикою лекції або практичного заняття.

За виконання та здачу лабораторної роботи після зазначеного дедлайну, за значну кількість пропущених занять, або за порушення правил поведінки на заняттях викладач може призначити до 5 штрафних балів.

При проведенні контрольних заходів та при виконанні лабораторних робіт студенти мають дотримуватись правил академічної доброчесності. При виявленні значного відсотку списування або плагіату викладач може відмовити у прийнятті даної роботи та вимагати доброчесного виконання навчального плану.

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

Види контролю з навчальної дисципліни «Алгоритми та методи обчислень» включають:

Лабораторні роботи:

Заплановано самостійне виконання 5 лабораторних робіт.

Теми лабораторних робіт узгоджені у часі та за змістом з темами лекцій. Виконання лабораторних робіт у повному обсязі дозволяє набути практичних навичок застосування чисельних методів до розв'язування різних математичних задач та оволодіти сучасними технологіями програмування алгоритмів, які побудовані на основі даних методів.

Поточний контроль:

Передбачено 5 поточних тестувань закритими тестами у системі ТСЕХАМ, які повністю охоплюють тематику даної навчальної дисципліни. Кожний поточний закритий тест містить 10 питань та триває 10 хв. Загальний час на проведення становить 2 години та включає час тестування та час на вирішення організаційних питань. У випадку дистанційного навчання закритий поточний тест проводиться на початку лекції, яка слідує за лекцією, що завершує чергову тему. При очному навчанні час чергового поточного тестування призначається викладачем за узгодженням зі студентами.

Семестровий контроль:

Семестровий закритий тест проводиться в кінці семестру, триває 30 хвилин та складається з 25 питань.

Залік:

проводиться у вигляді співбесіди зі студентом для об'єктивного визначення рівня знань, умінь та практичних навичок, отриманих за семестр

Оскільки кредитний модуль має семестрову атестацію у вигляді заліку, рейтингова система оцінювання побудована за типом PCO – 1. Рейтинг студента з кредитного модуля складається з балів, які він отримує за види робіт відповідно до таблиці 4.

Таблиця 4

Оцінювання окремих видів навчальної роботи студента

4 курс	
Вид навчальної роботи	Максимальна кількість балів
Виконання та захист лабораторної роботи №1	10
Виконання та захист лабораторної роботи №2	10
Виконання та захист лабораторної роботи №3	10
Виконання та захист лабораторної роботи № 4	10
Виконання та захист лабораторної роботи № 5	10
Поточний контроль	20
Семестровий контроль закритими тестами	30
Усього за семестр	100

Індивідуальний семестровий рейтинг (**RD**) студента з навчальної дисципліни складається з балів, які він отримує за:

- 1) Поточний контроль: тематичні тестування закритими тестами у системі ТСЕХАМ (по 10 хвилин) – $6 \times 5 = 30$ балів.
- 2) Семестровий контроль, який включає тестування закритими тестами у системі ТСЕХАМ (60 хвилин) – 30 балів,

Разом за семестровий контроль та поточний контроль – 60 балів.

- 3) Виконання лабораторних робіт.

Протягом семестру студенти виконують 4 лабораторних роботи.

Максимальна кількість балів за кожну лабораторну роботу – 10.

Бали нараховуються за:

- своєчасність подання роботи до захисту 0 – 1 бал,
- оформлення протоколу лабораторної роботи 0 – 1 бал,
- виконання змістовного завдання на роботу 0 – 5 балів,

- перевірка теоретичних знань, необхідних для виконання лабораторної роботи
0 – 5 балів.

Разом за лабораторні роботи (максимальна кількість балів) – 40.

Виконання семестрового контрольного тестування.

Семестровий контрольний тест виконується на кафедральному сервері у системі тестування «ТСЕХАМ».

Випадковим чином з кожної теми вибирається однакова кількість питань, загальним числом 30 питань, 4 варіанти відповідей на кожне питання, час виконання тесту – 60 хвилин.

Виконання поточних контрольних тестів. Поточні контрольні тести виконуються на кафедральному сервері у системі тестування «ТСЕХАМ».

Кожний поточний контрольний тест складається з 10 запитань.

Повна ґрунтовна відповідь на запитання тесту 0.5 - бала. Відповідь, яка включає недолік неповноти – 0.375 бала.

Відповідь, що містить дві вибрані неправильні відповіді - 0.25 бала.

Відповідь, що містить дві вибрані неправильні відповіді - 0.125 бала.

Відповідь, що не містить правильних варіантів відповіді - 0 балів.

Розрахунок розміру шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру становить:

$$R = \sum_k r_k, \text{ де } r_k \text{ максимальний рейтинговий бал кожного з контрольних заходів}$$

(контрольне тестування, лабораторні роботи). Розмір рейтингової шкали з кредитного модуля становить:

$$R = 60 + 40 = 100 \text{ балів.}$$

У разі пропуску занять студентом без поважної причини нараховуються штрафні санкції у вигляді 1 бала з загальної суми балів за 1 годину пропуску (але не більше, ніж 0,1 R).

Індивідуальний семестровий рейтинг студента (підсумкова семестрова рейтингова оцінка **RD**) є сумою балів, отриманих студентом протягом семестру за участі у передбачених контрольних заходах (контрольні та лабораторні роботи).

Необхідною умовою допуску студента до заліку є його індивідуальний семестровий рейтинг (**RD**) не менший, ніж 30 балів, відсутність повної заборгованості з лабораторних робіт та не менше, ніж одна позитивна атестація. За невиконання хоча б однієї зі згаданих вимог студент до заліку не допускається.

Сума підсумкової семестрової (**RD**) та залікової рейтингових оцінок у балах становить підсумкову семестрову рейтингову оцінку, яка перераховується в оцінки за національною шкалою та шкалою ECTS (табл. 5).

Таблиця 5

Відповідність рейтингових балів оцінкам 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

8. Додаткова інформація з дисципліни (освітнього компонента)

Викладання дисципліни «Статистичні методи машинного навчання» для спеціальності «Комп'ютерна інженерія» має свою специфіку, яка пов'язана з тим, що розробка та експлуатація комп'ютерної техніки потребує знання правил побудови та аналізу алгоритмів. Знайомство з універсальними моделями алгоритмів дозволяє визначити складність задачі та можливість її розв'язання за допомогою комп'ютера. Значна увага повинна приділятися підходам, що сприяють засвоєнню теоретичного матеріалу та практичних методів алгоритмізації, обчислення математичних задач, вивченню особливостей застосування методів, створення високоефективних алгоритмів реалізації чисельних методів. Саме чисельні методи лежать в основі всіх розрахунків, які можна виконати з використанням комп'ютерів.

Робочу програму навчальної дисципліни (силабус):

Складено д.т.н., проф. Новотарський Михайло Анатолійович

Ухвалено кафедрою ОТ (протокол № 10 від 25.05.2022)

Погоджено: Методичною комісією ФІОТ (протокол № 10 від 09.06.2022)

Протокол засідання Вченої ради ФІОТ №10 від 13.06.2022