

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

ФАКУЛЬТЕТ КОМП'ЮТЕРНИХ НАУК ТА КІБЕРНЕТИКИ

Кафедра теорії та технології програмування



«ЗАТВЕРДЖУЮ»

Заступник декана
з навчальної роботи

Кашпур О.Ф.

«28» серпня 2020 року

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Актуальні проблеми «Data Mining»/Data Mining Actual Problems

для студентів / for students

галузь знань	12 – Інформаційні технології / Informational Technologies
спеціальність	122 – Комп'ютерні науки / Computer Science
освітній рівень	магістр / Master
освітня програма	Математичні Методи Штучного Інтелекту / Mathematical Methods of Artificial Intelligence
вид дисципліни	обов'язкова / mandatory

Форма навчання	денна
Навчальний рік	2020/2021
Семестр	2
Кількість кредитів ECTS	5
Мова викладання, навчання та оцінювання	англійська
Форма заключного контролю	залік

Викладачі: к.ф.-м.н., асистент Криволап А.В.

Пролонговано: на 2021/2022 н.р.  «11» 02 2021р.
(підпис, ПІБ, дата)

на 20__/20__ н.р. (_____) «__» __ 20__р.
(підпис, ПІБ, дата)

КИЇВ – 2020

Розробник: **Криволап Андрій Володимирович**, кандидат фізико-математичних наук,
асистент кафедри теорії та технології програмування

ЗАТВЕРДЖЕНО
Зав. кафедри теорії та технології програмування

 (Нікітченко М.С.)

Протокол № 1 від « 28 » 08 20 20 р.

Схвалено Гарантом освітньо-наукової програми «Математичні методи штучного
інтелекту»  (підпис) (Завадський І.О.)

« 28 » 08 20 20 р.

Схвалено науково-методичною комісією факультету комп'ютерних наук та кібернетики

Протокол від « 28 » 08 20 20 року № 1

Голова науково-методичної комісії  (Омельчук Л.Л.)

« 28 » 08 20 20 року

1. Мета дисципліни – поглиблення знань з інтелектуального аналізу даних та штучного інтелекту, вивчення основних підходів до розв’язання основних задач – це задачі класифікації, кластеризації, пошуку асоціативних правил.

Discipline aim. The purpose of the discipline is to broaden knowledge of data mining and artificial intelligence, studying the basic approaches to solving basic data analysis problems. These are the tasks of classification, clustering, search for associative rules.

2. Попередні вимоги до опанування або вибору навчальної дисципліни:

Знати: базову дисципліну «Кластеризація та класифікація інформації», базові поняття штучного інтелекту та методів оптимізації; мати сучасні уявлення про основні задачі, що вирішуються в рамках штучного інтелекту та аналізу даних.

Вміти: описувати задачу аналізу даних, визначати атрибути та тип задачі, будувати модель.

Preliminary demands to master or choice of the course discipline:

1. To know the basic discipline “Clasterization and classification of information”, concepts of artificial intelligence and optimization methods; have a modern understanding of the main problems that are solved with the methods of artificial intelligence and data analysis.

2. To be able to describe the task of data analysis, determine the attributes and type of problem, build a model.

3. Анотація навчальної дисципліни:

Навчальний курс присвячений основним задачам інтелектуального аналізу даних: класифікації, кластеризації, пошуку асоціативних правил. Розглядаються основні класи алгоритмів для розв’язку відповідних задач. Наводиться порівняльний аналіз підходів, та можливі модифікації. Зокрема вивчаються алгоритми побудови дерев прийняття рішень, питання адаптивної кластеризації. Досліджуються основні проблеми, пов’язані з використанням штучних нейронних мереж для вирішення задач аналізу даних.

Synopsis of the course:

The learning course is devoted to the main problems of data mining: classification, clustering, search for associative rules. The main classes of algorithms for solving the corresponding problems are considered. A comparative analysis of approaches and possible modifications are given. In particular, algorithms for constructing decision trees, adaptive clustering are studied. The main considerations associated with the use of artificial neural networks for data analysis problems are investigated.

4. Завдання (навчальні цілі):

Опанування курсу покликане забезпечити формування поданого нижче переліку компетентностей:

ЗК5: Здатність спілкуватися іноземною мовою;

СК3: Здатність до дослідження та аналізу надвеликих масивів даних із складною неоднорідною і/або невизначеною структурою для прийняття зважених бізнес-рішень;

СК10: Здатність передбачати довгострокові бізнес-вимоги, впливати на покращення ефективності організаційного процесу, ефективно керувати фінансовими, людськими, технічними та іншими проектними ресурсами задля забезпечення успішності проектів;

СК18: Здатність аналізувати та використовувати математичні методи штучного інтелекту.

Objectives of study:

Learning course intends to provide formation following competencies:

CC5. Ability to communicate in a foreign language;

SC3. Ability to study and analyze large data sets with a complex, nonuniform and / or indefinite structure for making informed business decisions;

SC10. Ability to predict long-term business requirements, to influence the effectiveness of the organizational process, to effectively manage financial, human, technical and other project resources in order to ensure the success of the projects;

SC18. Ability to analyze and use mathematical methods of artificial intelligence

5. Результати навчання за дисципліною / Results of learning:

Результат навчання (1. знати; 2. вміти; 3. комунікація; 4. автономність та відповідальність)		Форми (та/або методи і технології) викладання і навчання	Методи оцінювання та пороговий критерій оцінювання (за необхідності)	Відсоток у підсумковій оцінці з дисципліни
Код	Результат навчання			
PH1.1	Знати основні поняття, задачі та етапи інтелектуального аналізу даних. To know the basic concepts, tasks and stages of data mining.	Лекція, самостійна робота / Lecture, Individual work	Контрольна робота / Test	15%
PH1.2	Знати основні методи інтелектуального аналізу даних, штучних нейронних мереж. To know the basic methods of data mining, artificial neural networks.	Лекція, самостійна робота / Lecture, Individual work	Контрольна робота / Test	15%
PH2.1	Вміти аналізувати задачу та обирати адекватний метод аналізу даних, оцінювати точність застосованих методів, виділяти суттєві ознаки To be able to analyze the problem and choose an adequate data mining method, evaluate the accuracy of applied methods, select the important features	Лекція, самостійна робота, лабораторні заняття / Lecture, Individual work, laboratory classes.	Контрольна робота, захист лабораторних робіт / Test, defense of laboratory work	50%
PH3.1	Обґрунтовувати власний погляд на задачу та спосіб її розв'язання, спілкуватися з колегами з питань застосування методів інтелектуального аналізу даних To substantiate one's own view on the problem and the way of its solution, to communicate with colleagues on the application of	Лекція, самостійна робота / Lecture, Individual work	Захист реферату, поточне оцінювання / Defense of the course paper, current evaluation	10%

	the data mining methods			
РН4.1	Організовувати свою самостійну роботу для досягнення результату To organize your independent work to achieve results	Лекція, самостійна робота, лабораторні заняття / Lecture, Individual work, laboratory classes	Захист реферату, захист лабораторних робіт / Defense of the course paper, defense of the laboratory work	10%

6. Співвідношення результатів навчання дисципліни із програмними результатами навчання / Correspondence between learning results and program study results

Результати навчання дисципліни	РН 1.1	РН 1.2	РН 2.1	РН 3.1	РН 4.1
Програмні результати навчання					
(з опису освітньої програми)					
ПРН4. Аналізувати великі дані та моделювати високорівневі абстракції у великих наборах даних різної природи, проектувати сховища великих даних, для видобутку даних і знань, візуалізовувати великі дані, будувати і оцінювати регресивні моделі, що генеруються на основі великих даних. PLO4. To master new data tools by processing weblogs, text mining and machine learning, for forecasting business processes and situational management, sentimental analysis of reviews, development of advisory systems for the field of electronic commerce, media, social networks, banking, advertising, etc.	+	+	+	+	+

7. Схема формування оцінки / Evaluation scheme.

7.1 Форми оцінювання студентів / Forms of evaluation:

- семестрове оцінювання / semester evaluation:

1. Контрольна робота / Test: РН 1.1, РН 1.2, РН 2.1 – 40 балів / 15 балів
2. Лабораторні роботи / Laboratory works: РН 2.1, РН 4.1 – 25 балів / 15 балів
3. Реферат / Course paper: РН 3.1, РН 4.1 – 15 балів / 18 балів
4. Поточне оцінювання / Current evaluation: РН 2.1, РН 3.1, РН 4.1 – 20 балів / 12 балів

- підсумкове оцінювання (у формі заліку)/final evaluation (credit)

- Залікові бали визначаються як сума оцінок/балів за всіма успішно оціненими результатами навчання, передбаченими даною програмою.
- Оцінки нижче від мінімального порогового рівня не додаються.
- Мінімальний пороговий рівень для сумарної оцінки за всіма компонентами становить 60% від максимально можливої кількості балів.
- /
- Credit points are defined as the sum of grades / points for all successfully assessed learning outcomes provided by this program.
- Scores below the minimum threshold are not added.
- The minimum threshold for the total assessment of all components is 60% of the maximum possible number of points.

Види завдань / types of tasks:

Структура контрольної роботи та критерії оцінювання / Structure of test and evaluation criteria:

1. Теоретичне запитання / theoretical task (PH 1.1 – PH 1.2).
2. Теоретичне запитання / theoretical task (PH 1.1 – PH 1.2).
3. Задача / problem (PH 2.1).
4. Задача / problem (PH 2.1).

Критерії оцінювання контрольної роботи / Criteria for evaluating the test

Завдання	Вид завдання	Максимальний бал (відсоток)	Всього балів (відсотків)
Завдання 1 / Tasks 1	Теоретичне запитання / theoretical task	8 балів (20 %)	8 балів (20 %)
Завдання 2 / Tasks 2	Теоретичне запитання / theoretical task	8 балів (20 %)	8 балів (20 %)
Завдання 3 / Task 3	Задача / problem	12 балів (30 %)	12 балів (30 %)
Завдання 4 / Task 4	Задача / problem	12 балів (30 %)	12 балів (30 %)
Всього / total			40 балів (100%)

Питання до контрольної роботи / Test questions

1. Поняття даних. Набір даних і їх атрибутів.
2. Вимірювання. Шкали.
3. Типи наборів даних. Формати зберігання даних.
4. Поняття метаданих.
5. Задачі та методи Data Mining.
6. Класифікація та властивості методів Data Mining.
7. Задача класифікації.
8. Точність класифікації: оцінка рівня помилок.
9. Алгоритм побудови елементарних правил (1-rule).
10. Алгоритми класифікації. Наївний баєсівський класифікатор.
11. Застосування нейронних мереж для задач класифікації.
12. Методи побудови дерев прийняття рішень.
13. Алгоритм найближчого сусіда.
14. Постановка задачі пошуку асоціативних правил. Алгоритм Apriori та його

різновиди.

15. Постановка задачі кластеризації, загальна схема кластеризації.
16. Ієрархічні алгоритми кластеризації, алгоритм k-means та метод найближчого сусіда.
17. Застосування нейронних мереж для задач кластеризації (Карта Кохонена).
18. Адаптивні методи кластеризації.

1. The concept of data. Data set and their attributes.
2. Measurement. Scales.
3. Types of data sets. Data storage formats.
4. The concept of metadata.
5. Tasks and methods of Data Mining.
6. Classification and properties of Data Mining methods.
7. The problem of classification.
8. Accuracy of classification: estimation of the level of errors.
9. Algorithm for constructing elementary rules (1-rule).
10. Classification algorithms. Naive Bayesian classifier.
11. Application of neural networks for classification problems.
12. Methods of building decision trees.
13. Algorithm of the nearest neighbor.
14. Statement of the problem of searching for associative rules. Apriory algorithm and its modifications.
15. Statement of the clustering problem, the general scheme of clustering.
16. Hierarchical clustering algorithms, k-means algorithm and nearest method neighbor.
17. Application of neural networks for clustering problems (Kohonen map).
18. Adaptive methods of clustering.

7.2 Організація оцінювання:

Терміни проведення форм оцінювання:

1. *Лабораторні роботи: до 10 тижня семестру.*
2. *Захист реферату: до 14 тижня семестру.*
3. *Контрольна робота: до 14 тижня семестру.*
4. *Поточне оцінювання: протягом семестру.*

Студент має право на одне перескладання контрольної роботи із можливістю отримання максимально 12 балів. Термін перескладання визначається викладачем.

За відсутності студента з поважних причин перездача КР здійснюється відповідно до «Положення про організацію освітнього процесу».

7.3 Шкала відповідності оцінок

Зараховано / Passed	60-100
Не зараховано / Failed	0-59

8. Структура навчальної дисципліни. Тематичний план лекцій

№ лекції	Назва лекції	Кількість годин		
		Лекції	Лабораторні заняття	Самостійна робота
Частина 1. Основні поняття Data Mining. Задача класифікації даних Part 1. Basic concepts of Data Mining. The problem of data classification				
1	Тема 1. Поняття даних. Задачі Data Mining. Стандарти Data Mining. Методи Data Mining Theme 1. The concept of data. Data Mining Tasks. Data Mining Problems. Data Mining Methods Самостійна робота: Модель MapReduce. Класифікація методів DM. Властивості методів DM. Independent work: MapReduce model. Classification of DM methods. Properties of DM methods.	4		8
2	Тема 2. Постановка задачі класифікації. Точність класифікації. Оцінка рівня помилок Theme 2. Classification problem statement. Accuracy of classification. Estimation of the level of errors Самостійна робота: Оцінка рівня помилок за допомогою крос перевірки. Питання вибору тестової множини. Independent work: Assessing the level of errors using cross-checking. The problem of choosing a test set.	2		8
3	Тема 3. Алгоритм побудови елементарних правил (1-rule). Наївний баєсівський класифікатор Theme 3. Algorithm for constructing elementary rules (1-rule). Naive Bayesian classifier Самостійна робота: Обмеження наївного баєсівського класифікатора. Задача визначення спаму. Independent work: Limitations of the naive Bayesian classifier. The problem of detecting spam. Лабораторна робота: Застосування алгоритмів побудови елементарних правил та наївного баєсівського класифікатора для задачі класифікації даних. Practical work: Application of algorithms for constructing elementary rules and a naive Bayesian classifier for the problem of data classification.	2	2	10
4	Тема 4. Методи побудови дерев прийняття рішень. Алгоритм найближчого сусіда Theme 4. Methods of building decision trees. Algorithm of the nearest neighbor Самостійна робота: Проблема перенавчання для дерев прийняття рішень та способи її вирішення. Області застосувань дерев рішень. Основні проблеми алгоритму найближчого сусіда і шляхи їх вирішення. Independent work: The problem of retraining for decision-making trees and ways to solve it. Areas of application of decision trees. The main problems of the algorithm of the nearest neighbor and ways to solve them. Лабораторна робота: Застосування алгоритмів побудови дерева прийняття рішень та алгоритму найближчого сусіда для задачі класифікації даних. Practical work: Application of decision tree construction algorithms and nearest neighbor algorithm for data classification problem.	4	2	12

5	Тема 5. Застосування нейронних мереж для задач класифікації Theme 5. Application of neural networks for classification problems Самостійна робота: Подання вхідних даних для штучних нейронних мереж. Вибір архітектури мережі. Independent work: Presentation of input data for artificial neural networks. Choice of network architecture.	2		12
Всього за частиною 1		14	4	50
Частина 2. Задача пошуку асоціативних правил. Задача кластеризації даних Part 2. The problem of finding associative rules. The problem of data clustering				
6	Тема 6. Задача пошуку асоціативних правил. Алгоритм Apriori та його різновиди Theme 6. The problem of finding associative rules. Apriori algorithm and its variants Самостійна робота: Узагальнені асоціативні правила. Модифікації алгоритму Apriori. Алгоритм FP-Growth. Independent work: Generalized associative rules. Modifications of the Apriori algorithm. FP-Growth algorithm. Лабораторна робота: Застосування алгоритму Apriori для пошуку асоціативних правил. Practical work: Application of the Apriori algorithm to search for associative rules.	2	2	12
7	Тема 7. <u>Задачі кластеризації</u>. Виділення характеристик. Визначення метрики. Приклади метрик Theme 7. Data clustering problem. Selection of characteristics. Definition of metrics. Examples of metrics Самостійна робота: Стратегії кластеризації. Багатовимірні евклідові простори та «прокляттям вимірності». Independent work: Clustering strategies. Multidimensional Euclidean spaces and the "curse of dimension".	2		8
8	Тема 8. <u>Ієрархічні алгоритми кластеризації</u>. Дендрограми. Алгоритм Single-link. Алгоритм Complete-link. Theme 8. Hierarchical clustering algorithms. Dendrograms. Single-link algorithm. Complete-link algorithm. Самостійна робота: Ефективність ієрархічної кластеризації. Ієрархічна кластеризація у неевклідових просторах. Independent work: Efficiency of hierarchical clustering. Hierarchical clustering in non-Euclidean spaces. Лабораторна робота: Застосування ієрархічних алгоритмів для задачі кластеризації. Practical work: Application of hierarchical algorithms for the clustering problem.	2	2	10
8	Тема 9. Неієрархічні алгоритми кластеризації. <u>Алгоритм k-means</u>. <u>Алгоритм найближчого сусіда</u>. Theme 9. Non-hierarchical clustering algorithms. K-means Самостійна робота: <u>Алгоритм fuzzy k-means</u>. Алгоритм Бредлі, Файяда та Рейна. Алгоритм CURE. Independent work: Fuzzy k-means algorithm. Bradley, Fayyad and Rhine algorithm. CURE algorithm. Лабораторна робота: Застосування алгоритмів <u>k-means</u> та найближчого сусіда для задачі кластеризації. Practical work: Application of k-means and nearest neighbor	2	2	10

	algorithms for clustering problem.			
9	Тема 10. Самоорганізаційна Карта Кохонена . Алгоритм навчання. Відображення кластерів. Theme 10. Kohonen Self-Organizing Map. Learning algorithm. Clusters representation. Самостійна робота: Початкова ініціалізація карти. Питання вибору конфігурації сітки. Independent work: Initial map initialization. The choice of grid configuration.	2		12
10	Тема 11. Адаптивні методи кластеризації . Визначення якості кластеризації. Показники чіткості Theme 11. Adaptive clustering methods. Determining the quality of clustering. Clarity indicators Самостійна робота: Нечіткі алгоритми кластеризації. Independent work: Fuzzy clustering algorithms.	2		8
	Контрольна робота	2		
	Всього за частиною 2	14	6	60
	ВСЬОГО	28	10	110

Загальний обсяг 150 год., в тому числі:

Лекцій – **28 год.**

Лабораторні заняття – **10 год.**

Консультації – **2 год.**

Самостійна робота - **110 год.**

9. Рекомендовані джерела /References

Основні / Main:

1. Berry, Michael J. A. "DM techniques: for marketing, sales, and customer relationship management" / Michael J.A. Berry, Gordon Linoff. – 2nd ed.
2. Larose, Daniel T. "Discovering knowledge in data: an introduction to DM" / Daniel T. Larose
3. Leskovec J. Mining of Massive Datasets / Jure Leskovec [Anand Rajaraman](#), [Jeffrey David Ullman](#) // Stanford Univ. – 2010.
4. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
5. Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.
6. R. Agrawal, R. Srikant. "Fast Discovery of Association Rules", In Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994.

Додаткові / Additional:

7. G. Lee, U. Yun A new efficient approach for mining uncertain frequent patterns using minimum data structure without false positives. Future Generation Computational Systems 68:89–110p., 2017.
8. S. Rustogi, M. Sharma, S. Morwal Improved Parallel Apriori Algorithm for Multi-cores. IJ Inf Technol Comput Sci 4:18–23p., 2017.
9. M.K. Gupta, P. Chandra A comparative study of clustering algorithms. In: Proceedings of the 13th INDIACom-2019; IEEE Conference ID: 461816; 6th International Conference on "Computing for Sustainable Global Development", 2019.
10. W. Buntine. A theory of classification rules. 1992.
11. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
12. R. M. Hristev, "Artificial Neural Networks"
13. R. Srikant, R. Agrawal. "Mining Generalized Association Rules", In Proc. of the 21th International Conference on VLDB, Zurich, Switzerland, 1995.
14. J.S. Park, M.-S. Chen, and S.Y. Philip, "An Effective HashBased Algorithm for Mining Association Rules", In Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York, 1995.
15. S. Brin et al., "Dynamic Itemset Counting and Implication Rules for Market Basket Data", In Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York, 1997.