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.
«____»_____2017 року

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ АКТУАЛЬНІ ПРОБЛЕМИ «DATA MINING»

для студентів

галузь знань **12 «Інформаційні технології»**
(шифр і назва)
спеціальність **122 «Комп'ютерні науки»**
(шифр і назва спеціальності)
освітній рівень **магістр**
(молодший бакалавр, бакалавр, магістр)
освітня програма **«Інформатика»**
(назва освітньої програми)

вид дисципліни за вибором ВНЗ

Форма навчання	**денна**
Навчальний рік	**2017/2018**
Семестр	**2**
Кількість кредитів ECTS	**4**
Мова викладання, навчання та оцінювання	**українська**
Форма заключного контролю	**екзамен**

Викладач: **к.ф.-м.н., асистент Россада Т.В.** (лекції, лабораторні заняття)

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

КИЇВ – 2017

Робоча програма навчальної дисципліни «Актуальні проблеми «Data Mining» для студентів спеціальності 122 «Комп'ютерні науки» освітнього рівня магістр *о*світньої програми «Інформатика»

«31» серпня 2017 року – 12 с.

Розробник: Россада Т.В., к.ф.-м.н., асист.

Робоча програма дисципліни «Актуальні проблеми «Data Mining» затверджена на засіданні кафедри теорії та технології програмування
Протокол №___ від «____» _____ 2017 року
В.О. Завідувача кафедри _____ Панченко Т.В.

Схвалено науково-методичною комісією факультету
Протокол №___ від «____» _____ 2017 року
Голова НМК _____ Хусаїнов Д.Я.

Робоча програма дисципліни «Актуальні проблеми «Data Mining» затверджена на засіданні Вченої ради факультету кібернетики
Протокол Протокол №___ від «____» _____ 2017 року

Вступ

Навчальна дисципліна «Актуальні проблеми «Data Mining» є складовою циклу професійної підготовки фахівців спеціальності 122 «Комп'ютерні науки» освітнього рівня магістр освітньої програми «Інформатика».

Дисципліна «Актуальні проблеми «Data Mining» є дисципліною за вибором ВНЗ, що викладається на **1**-му курсі магістратури у **2**-му семестрі в обсязі **120** годин (**4** кредити ECTS), в тому числі **26** години лекцій, **12** годин лабораторних занять (**2** підгрупи) і **80** годин самостійної роботи. Завершується дисципліна **екзаменом**.

Метою і завданням навчальної дисципліни є знайомство з теоретичними та прикладними аспектами технології Data Mining, методами інтелектуального аналізу даних, можливостями їх застосування.

Структура курсу. В результаті вивчення навчальної дисципліни студент повинен **знати** теоретичні та прикладні аспекти методів Data Mining та **вміти** їх застосовувати для розв'язання широкого спектру задач.

Для освоєння курсу необхідні базові знання з *програмування*, *теорії баз даних*, *вищої математики*.

Контроль знань і розподіл балів, які отримують студенти

Контроль знань студентів здійснюється за модульно-рейтинговою системою. Результати навчальної діяльності студентів оцінюються за 100-бальною шкалою.

Робота в семестрі поділяється на 2 змістових модуля. При виставленні балів за змістовий модуль враховується: робота на лабораторних заняттях – 10 балів, оцінка за модульну контрольну роботу – 15 балів. Завершується дисципліна іспитом. Оцінювання за формами контролю:

	ЗМ1		*ЗМ2*	
	Min. – 18	*Max. – 30*	*Min. – 18*	*Max. – 30*
Лабораторні роботи	10	20	10	20
Модульні контрольні роботи	8	10	8	10

У випадку відсутності студента з поважних причин відпрацювання та перездачі МКР здійснюються у відповідності до «Положення про порядок оцінювання знань студентів при кредитно-модульній системі організації навчального процесу» від 1 жовтня 2010 року.

При простому розрахунку отримаємо:

	ЗМ1	ЗМ2	Екзамен	Підсумкова оцінка
Мінімум	18	18	24	60
Максимум	30	30	40	100

При цьому, кількість балів:

- 1-34 відповідає оцінці «незадовільно» з обов’язковим повторним вивченням дисципліни;
- 35-59 відповідає оцінці «незадовільно» з можливістю повторного складання;
- 60-64 відповідає оцінці «задовільно» («достатньо»);
- 65-74 відповідає оцінці «задовільно»;
- 75 - 84 відповідає оцінці «добре»;
- 85 - 89 відповідає оцінці «добре» («дуже добре»);
- 90 - 100 відповідає оцінці «відмінно».

Шкала відповідності

За 100 – бальною шкалою	За національною шкалою
90 – 100	Відмінно
85 – 89	Добре
75 – 84	
65 – 74	Задовільно
60 – 64	
1 – 59	Не задовільно

ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Змістовий модуль 1. Задачі та методи Data Mining – 36 год.
Тема 1. Задачі та методи Data Mining – 20 год.
Тема 2. Стандарти Data Mining – 16 год.

Змістовий модуль 2. Класифікація та кластеризація – 82 год.
Тема 3. Задача класифікації – 32 год.
Тема 4. Задачі пошуку асоціативних правил – 24 год.
Тема 5. Задача кластеризації – 26 год.

СТРУКТУРА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
ТЕМАТИЧНИЙ ПЛАН ЛЕКЦІЙ ТА ЛАБОРАТОРНИХ ЗАНЯТЬ

№ п/п	Назва лекції	Кількість годин		
		лекції	лаборат.	сам.
Змістовий модуль 1. Задачі та методи Data Mining				
1	Тема 1.Задачі та методи Data Mining	8	2+2	10
2	Тема 2. Стандарти Data Mining	4	2+2	10
	Модульна контрольна робота 1			
Змістовий модуль 2. Класифікація та кластеризація даних				
5	Тема 3. Задача класифікації	8	4+4	20
6	Тема 4. Задача пошуку асоціативних правил	2	2+2	20
7	Тема 5. Задача кластеризації	4	2+2	20
	Модульна контрольна робота 2			
	Всього	26	12+12	80

Загальний обсяг **120** год., в тому числі:

лекції – **26** год.

лабораторні заняття – **12+12** год. (**2** підгруппи по **12** годин)

самостійна робота – **80** год.

консультації - **2** год.

ЗМІСТОВИЙ МОДУЛЬ 1: ЗАДАЧІ КЛАСИФІКАЦІЇ ТА РЕГРЕСІЇ

Тема 1. ЗАДАЧІ ТА МЕТОДИ DATA MINING – 20 год

Лекція 1. Вступ до дисципіни. Поняття та задачі Data Mining – 2 год.

Лекція 2. Набір даних та їх атрибути. Вимірювання даних – 2 год.

Лекція 3. Особливості обробки даних. Модель Map Reduce – 2 год.

Лабораторне заняття 1. Виконання лабораторної робота № 1 – 2 год.

Лекція 4. Методи та засоби Data Mining – 2 год.

Самостійна робота. Виконання лабораторної роботи № 1– 10 год.

Тема 2. СТАНДАРТИ DATA MINING – 16 год

Лекція 1. Методи візуалізації – 2 год.

Лабораторне заняття 2. Здача лабораторної робота № 1 – 2 год.

Лекція 2. Стандарти Data Mining: CWM, CRISP, PMML – 2 год.

Самостійна робота. Виконання лабораторної роботи № 1– 10 год.

Контрольні запитання

1. Поняття Data Mining.
2. Поняття даних. Набір даних та їх атрибути.
3. Вимірювання даних.
4. Особливості обробки даних.
5. Модель Map Reduce
6. Візуальний аналіз даних. Методи візуалізації
7. Стандарти Data Mining: CWM
8. Стандарти Data Mining: CRISP
9. Стандарти Data Mining: PMML

Типове завдання модульної контрольної роботи № 1

1. Поняття даних. Набір даних та їх атрибути.
2. Стандарти Data Mining: CRISP
3. Описати функції map та reduce для знаходження найбільшого числа послідовності чисел.

Рекомендована література: [1-3]

ЗМІСТОВИЙ МОДУЛЬ 2: КЛАСИФІКАЦІЯ ТА КЛАСТЕРИЗАЦІЯ

Тема 3. ЗАДАЧА КЛАСИФІКАЦІЇ – 32 год.

Лекція 1. Постановка задачі класифікації. Точність класицікації. Оцінка рівня помилок. – 2 год.

Лабораторне заняття 1. Виконання лабораторної роботи № 2. - 2 год.

Лекція 2. Методи класифікації. Алгоритм побудови елементарних правил (1-rule), алгоритм Naive Bayes. – 2 год.

Лекція 3. Застосування нейронних мереж для задач класифікації – 2 год.

Лабораторне заняття 2. Виконнання лабораторної роботи № 2. - 2 год.

Лекція 4. Дерева прийняття рішень. Алгоритм найближчого сусіда – 2 год.

Самостійна робота. Виконання лабораторної роботи № 2 – 20 год.

Тема 4. ЗАДАЧА ПОШУКУ АСОЦІАТИВНИХ ПРАВИЛ – 24 год.

Лекція 1. Постановка задачі пошуку асоціативних правил, її різновиди. Представлення результатів. Алгоритм Apriori та його різновиди. – 2 год.

Лабораторне заняття 1. Виконання лабораторної роботи № 2 – 2 год.

Самостійна робота. Виконання лабораторної роботи № 2 – 20 год.

Тема 5. ЗАДАЧА КЛАСТЕРИЗАЦІЇ – 26 год.

Лекція 1. Задача кластерізації. Ієрархічні алгоритми кластеризації – 2 год.

Лекція 2. Алгоритм k-Means. Метод найближчого сусіда.– 2 год.

Лабораторне заняття 1. Здача лабораторної роботи № 2 – 2 год.

Самостійна робота. Виконання лабораторної роботи № 2. – 20 год.

Контрольні запитання

1. Задача класифікації
2. Точність класифікації. Оцінка рівня помилок
3. Постановка задачі пошуку асоціативних правил, її різновиди. Представлення результатів
4. Алгоритм Apriori та його різновиди
5. Постановка задачі кластерізації
6. Ієрархічні алгоритми кластеризації
7. Алгоритм k-Means (Hard-c-means)
8. Метод найближчого сусіда
9. Дерева прийняття рішень
10. Алгоритм найближчого сусіда
11. Адаптивні методи кластерізації.

Типове завдання модульної контрольної роботи № 2

1. Постановка задачі класифікації. Алгоритми побудови елементарних правил та Naive Bayes: теоретичні основи, основні етапи, застосування.
2. Застосування нейронних мереж для задач кластеризації.
3. Для даних з табл. 1 побудувати дерево прийняття рішень (S=1) використовуючи теоретико-інформаційний критерій вибору атрибута та знайти клас останнього об’єкта.

4. Використавши алгоритм Apriori (minsup=3), вкажіть асоціативні правила для набору даних з табл.2. Визначте для кожного правила достовірність та підтримку.
5. Розбити дані, представлені на рисунку 1, на кластери за допомогою алгоритму k-means та за допомогою методу найближчого сусіда (взявши порогове значення – 2).

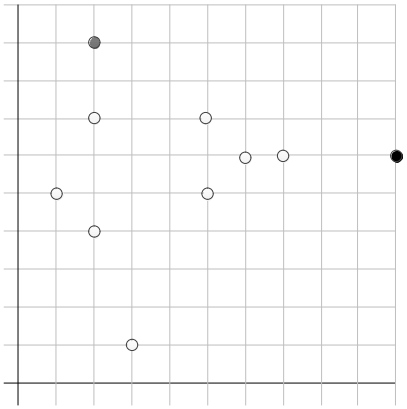
Таблиця 1

	q1	q2	q3	q4	S
1	1	0	1	0	1
2	1	1	0	1	0
3	0	0	0	1	0
4	0	0	1	0	1
5	0	1	1	1	0
6	0	2	1	0	1
7	0	1	1	0	1
8	0	2	1	0	1
9	0	1	0	0	1
10	1	0	0	1	0
11	*0*	*1*	*0*	*0*	*?*

Таблиця 2

1	0, 1, 2, 3
2	1, 2, 3
3	1, 4, 5, 6, 7
4	1, 2, 3, 4, 6, 9
5	1, 2, 3, 4, 5
6	0, 5, 6
7	0, 8, 9
8	0, 1, 4, 7
9	5, 6, 7, 8, 9
10	4, 5, 7

Рисунок 1



Питання на іспит

1. Поняття Data Mining.
2. Поняття даних. Набір даних та їх атрибути.
3. Вимірювання даних.
4. Особливості обробки даних.
5. Модель Map Reduce
6. Візуальний аналіз даних. Методи візуалізації
7. Стандарти Data Mining: CWM
8. Стандарти Data Mining: CRISP
9. Стандарти Data Mining: PMML
10. Задача класифікації
11. Точність класифікації. Оцінка рівня помилок
12. Постановка задачі пошуку асоціативних правил, її різновиди. Представлення результатів
13. Алгоритм Apriori та його різновиди
14. Постановка задачі кластерізації
15. Ієрархічні алгоритми кластеризації
16. Алгоритм k-Means (Hard-c-means)
17. Метод найближчого сусіда
18. Дерева прийняття рішень
19. Алгоритм найближчого сусіда
20. Адаптивні методи кластерізації.

Типове завдання екзаменаційного білету

1. Постановка задачі класифікації. Алгоритми побудови елементарних правил та Naive Bayes: теоретичні основи, основні етапи, застосування.

2. Застосування нейронних мереж для задач кластеризації.
3. Для даних з табл. 1 побудувати дерево прийняття рішень (S=1) використовуючи теоретико-інформаційний критерій вибору атрибута та знайти клас останнього об’єкта.
4. Використавши алгоритм Apriori (minsup=3), вкажіть асоціативні правила для набору даних з табл.2. Визначте для кожного правила достовірність та підтримку.
5. Розбити дані, представлені на рисунку 1, на кластери за допомогою алгоритму k-means та за допомогою методу найближчого сусіда (взявши порогове значення – 2).

Рекомендована література

1. Марченко О. О., Россада Т.В. Актуальні проблеми Data Mining: навчальний посібник для студентів факультету комп’ютерних наук та кібернетики. — Київ. — 2017. — 150 с.
2. Leskovec J. Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman // Stanford Univ. – 2010.
3. Bradley, P., Fayyad, U., Reina, C. Scaling Clustering Algorithms to Large Databases, Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, Calif., 1998.
4. Zhang, T., Ramakrishnan, R., Livny, M. Birch: An Efficient Data Clustering Method for Large Databases, Proc. ACM SIGMOD Int’l Conf. Management of Data, ACM Press, New York, 1996.
5. Paul S. Bradley, Usama M. Fayyad, Cory A. Reina Scaling EM (Expectation-Maximization) Clustering to Large Databases, Microsoft Research, 1999.
6. Z. Huang. Clustering large data sets with mixed numeric and categorical values. In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997.
7. Milenova, B., Campos, M. Clustering large databases with numeric and nominal values using orthogonal projections, Oracle Data Mining Technologies, 2002.
8. Z. Huang. A fast clustering algorithm to claster very large categorical data sets in Data Mining. Research Issues on on Data Mining and KDD, 1997.
9. Wang, K., Xu, C.. Liu, B. Clustering transactions using large items. In Proc. CIKM’99, Kansas, Missouri, 1999.
10. Guha S., Rastogi R.,Shim K. CURE: An Efficient Clustering Algorithm for Large Databases, Proc. ACM SIGMOD Int’l Conf. Management of Data, ACM Press, New York, 1998.
11. Ganti V., Gerhke J., Ramakrishan R. CACTUS – Clustering Categorical Data Using Summaries. In Proc KDD’99, 1999.
12. J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Tech. Report ICSI-TR-97-021, 1997.
13. Добыча данных в сверхбольших базах данных / В. Ганти, Й. Герке, Р. Рамакришнан // Открытые системы, №9-10, 1999.

14. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004
15. Berry, Michael J. A. “Data mining techniques: for marketing, sales, and customer relationship management “/ Michael J.A. Berry, Gordon Linoff. – 2nd ed.
16. Larose, Daniel T. “Discovering knowledge in data: an introduction to data mining” / Daniel T. Larose
17. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
18. S.Murthy. Automatic construction of decision trees from data: A Multi-disciplinary survey.1997.
19. W. Buntine. A theory of classification rules. 1992.
20. Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.
21. К. Шеннон. Работы по теории информации и кибернетике. М. Иностранная литература, 1963
22. С.А. Айвазян, В.С Мхитарян Прикладная статистика и основы эконометрики, М. Юнити, 1998
23. Dirk Emma Baestaens, Willem Max Van Den Bergh, Douglas Wood, "Neural Network Solution for Trading in Financial Markets", Pitman publishing
24. R. M. Hristev, "Artifical Neural Networks"
25. R. Agrawal, T. Imielinski, A. Swami. 1993. Mining Associations between Sets of Items in Massive Databases. In Proc. of the 1993 ACM-SIGMOD Int’l Conf. on Management of Data, 207-216.
26. R. Agrawal, R. Srikant. "Fast Discovery of Association Rules", In Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994.
27. R. Srikant, R. Agrawal. "Mining Generalized Association Rules", In Proc. of the 21th International Conference on VLDB, Zurich, Switzerland, 1995.
28. R. Srikant, R. Agrawal. "Mining quantitative association rules in large relational tables". In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.
29. Savasere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", In Proc. 21st Int’l Conf. Very Large Data Bases, Morgan Kaufmann, San Francisco, 1995.
30. 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.
31. 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.
32. J. Hipp, U. Guntzer, and G. Nakaeizadeh. Algorithms for Association Rule Mining – A General Survey and Comparison. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.

Електронні джерела

1. Курс лекций Николая Анохина (mail.ru DataGroup)
   https://www.youtube.com/playlist?list=PLrCZzMib1e9pyyrqknouMZbIPf4l3CwUP
2. Data is the New Oil By Michael Palmer
   http://ana.blogs.com/maestros/2006/11/data_is_the_new.html
3. Анализ данных как область знания
   http://postnauka.ru/video/34960
4. Материалы на тему анализа данных
   http://www.basegroup.ru/library/methodology/data_mining/
5. Наивный Байесовский классификатор в 25 строк кода
   http://habrahabr.ru/post/120194/
6. Фильтрация смс спама с помощью наивного байесовского классификатора
   http://habrahabr.ru/post/184574/
7. Лекции курса «Машинное обучение» от yandex
   https://yadi.sk/d/V9p7E6uAFjHcD
8. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования
   http://www.ccas.ru/voron/download/Clustering.pdf
9. Котов А., Красильников Н. Кластеризация данных. 2006
   http://logic.pdmi.ras.ru/~yura/internet/02ia-seminar-note.pdf
10. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных
   www.machinelearning.ru/
11. Н.Ю. Золотых Как обучаются машины? научно-популярная лекція
   http://www.uic.unn.ru/~zny/ml/Pop/ml_pop.pdf
12. Главы из книги на тему машинного обучения и презентации уроков Сергея Николенко
   http://logic.pdmi.ras.ru/~sergey/teaching/ml/

Бази данних

http://vincentarelbundock.github.io/Rdatasets/datasets.html

Додаток до робочої програми
Завдання для самостійної роботи з елементами дистанційного навчання з дисципліни «Актуальні проблеми Data Mining»
на період з 24 січня до 28 лютого 2018 р.
для студентів
1 курсу магістратури спеціальності 122 «Комп'ютерні науки»
освітньої програми «Інформатика»

Викладач: к.ф.-м..н., асист. Россада Т.В. (e-mail: trossada@knu.ua)

Види та форми контрольних заходів з перевірки самостійної роботи студентів
Контроль за виконанням самостійної роботи студентами здійснюється у двох формах: у січні-лютому за допомогою електронних засобів (електронною поштою), у березні – шляхом проведення письмової контрольної роботи. Виконання самостійної роботи є допуском до написання контрольної роботи у березні 2018 р.

Впродовж січня-лютого (24 січня – 20 лютого 2018 р.) студенти мають вивчити запропоновані питання визначених тем на базовому рівні. Для підтвердження виконання завдання студенти мають надіслати відповіді на 3 теоретичних питання та виконану лабораторну роботу із звітом не пізніше **15 лютого 2018 р**. Завдання першого етапу, які мають бути виконані та надіслані на електронну пошту викладача (trossada@knu.ua), подано **у додатку 1**.

На контрольну роботу за підсумками самостійної роботи виносяться всі зазначені нижче теоретичні питання.

Теоретичні питання для самостійного опрацювання
1. Поняття Data Mining.
2. Поняттяданих.Набірданихтаїхатрибути.
3. Вимірюванняданих.
4. Особливостіобробкиданих.
5. Модель Map Reduce
6. Візуальнийаналізданих.Методивізуалізації
7. СтандартиDataMining:CWM
8. СтандартиDataMining:CRISP
9. СтандартиDataMining:PMML

Критерії оцінювання
Викладач оцінює виконані завдання в категоріях «**зараховано**» або «**не зараховано**». Щоб отримати оцінку «зараховано» потрібно набрати 4 і більше балів за відповіді на теоретичні питання та виконати лабораторну роботу. Відповіді на теоретичні питання оцінюються наступним чином: відповіді немає — 0 балів; відповідь є, але не повна — 1 бал; відповідь повна — 2 бали.

Якщо студент отримає оцінку «не зараховано», у нього є час до **20 лютого** переробити завдання та надіслати їх викладачу повторно.

Контрольна робота оцінюється максимум в **10 балів**. Вона включає в себе 5 тестових питань з проблематики, винесеної на самостійну роботу, та одне практичне завдання. Правильна відповідь на кожне тестове завдання оцінюється в 1 бал. За розв'язання задачі студент може отримати від 1 до 5 балів.

Контрольна робота проводиться на першому лекційному занятті з курсу у березні 2018 р. Її тривалість – 1 академічна година.

Рекомендована література

1. Марченко О. О., Россада Т.В. Актуальні проблеми Data Mining: навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. — Київ. — 2017. — 150 с.
2. Leskovec J. Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman // Stanford Univ. – 2010.
3. Bradley, P., Fayyad, U., Reina, C. Scaling Clustering Algorithms to Large Databases, Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, Calif., 1998.
4. Zhang, T., Ramakrishnan, R., Livny, M. Birch: An Efficient Data Clustering Method for Large Databases, Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York, 1996.
5. Paul S. Bradley, Usama M. Fayyad, Cory A. Reina Scaling EM (Expectation-Maximization) Clustering to Large Databases, Microsoft Research, 1999.
6. Z. Huang. Clustering large data sets with mixed numeric and categorical values. In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1997.
7. Milenova, B., Campos, M. Clustering large databases with numeric and nominal values using orthogonal projections, Oracle Data Mining Technologies, 2002.
8. Z. Huang. A fast clustering algorithm to claster very large categorical data sets in Data Mining. Research Issues on on Data Mining and KDD, 1997.
9. Wang, K., Xu, C.. Liu, B. Clustering transactions using large items. In Proc. CIKM'99, Kansas, Missouri, 1999.
10. Guha S., Rastogi R.,Shim K. CURE: An Efficient Clustering Algorithm for Large Databases, Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York, 1998.
11. Ganti V., Gerhke J., Ramakrishan R. CACTUS – Clustering Categorical Data Using Summaries. In Proc KDD'99, 1999.
12. J. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Tech. Report ICSI-TR-97-021, 1997.
13. Добыча данных в сверхбольших базах данных / В. Ганти, Й. Герке, Р. Рамакришнан // Открытые системы, No9-10, 1999.
14. Барсегян и др. Методы и модели анализа данных: OLAP и Data Mining. – СПб., 2004
15. Berry, Michael J. A. "Data mining techniques: for marketing, sales, and customer relationship management "/ Michael J.A. Berry, Gordon Linoff. – 2nd ed.

16. Larose, Daniel T. “Discovering knowledge in data: an introduction to data mining” / Daniel T. Larose
17. J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.
18. S.Murthy. Automatic construction of decision trees from data: A Multi- disciplinary survey. 1997.
19. W. Buntine. A theory of classification rules. 1992.
20. Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.
21. К. Шеннон. Работы по теории информации и кибернетике. М. Иностранная литература, 1963
22. С.А. Айвазян, В.С Мхитарян Прикладная статистика и основы эконометрики, М. Юнити, 1998
23. Dirk Emma Baestaens, Willem Max Van Den Bergh, Douglas Wood, "Neural Network Solution for Trading in Financial Markets", Pitman publishing
24. R. M. Hristev, "Artifical Neural Networks"
25. R. Agrawal, T. Imielinski, A. Swami. 1993. Mining Associations between Sets of Items in Massive Databases. In Proc. of the 1993 ACM-SIGMOD Int’l Conf. on Management of Data, 207-216.
26. R. Agrawal, R. Srikant. "Fast Discovery of Association Rules", In Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994.
27. R. Srikant, R. Agrawal. "Mining Generalized Association Rules", In Proc. of the 21th International Conference on VLDB, Zurich, Switzerland, 1995.
28. R. Srikant, R. Agrawal. "Mining quantitative association rules in large relational tables". In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.
29. Savasere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association Rules in Large Databases", In Proc. 21st Int’l Conf. Very Large Data Bases, Morgan Kaufmann, San Francisco, 1995.
30. 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.
31. 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.
32. J. Hipp, U. Guntzer, and G. Nakaeizadeh. Algorithms for Association Rule Mining – A General Survey and Comparison. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.

Рекомендовані електронні джерела

1. Курс лекций Николая Анохина (mail.ru DataGroup)
https://www.youtube.com/playlist?list=PLrCZzMib1e9pyyrqknouMZbIPf4l3CwUP
2. Data is the New Oil By Michael Palmer http://ana.blogs.com/maestros/2006/11/data_is_the_new.html
3. Анализ данных как область знания
http://postnauka.ru/video/34960
4. Материалы на тему анализа данных http://www.basegroup.ru/library/methodology/data_mining/
5. Наивный Байесовский классификатор в 25 строк кода
http://habrahabr.ru/post/120194/

6. Фильтрация смс спама с помощью наивного байесовского классификатора
http://habrahabr.ru/post/184574/
7. Лекции курса «Машинное обучение» от yandex
https://yadi.sk/d/V9p7E6uAFjHcD
8.Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования
http://www.ccas.ru/voron/download/Clustering.pdf
9. Котов А., Красильников Н. Кластеризация данных. 2006
http://logic.pdmi.ras.ru/~yura/internet/02ia-seminar-note.pdf
10.Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных
www.machinelearning.ru/
11.Н.Ю. Золотых Как обучаются машины? научно-популярная лекція http://www.uic.unn.ru/~zny/ml/Pop/ml_pop.pdf
12.Главы из книги на тему машинного обучения и презентации уроков Сергея Николенко http://logic.pdmi.ras.ru/~sergey/teaching/ml/

Додаток 1. Завдання для самостійної роботи

Теоретичні питання

1. Дайте розгорнуті визначення поняттям: дані, база даних, шкала, класифікація даних, кластеризація даних, Data Mining.
2. Технологія Map Reduce 3. Стандарт CRISP-DM

Лабораторна робота No1

1. Завантажити abalone data set
Abalone Data Set: https://archive.ics.uci.edu/ml/datasets/abalone

2. Знайті і опрацювати наукові статті, присвячені abalone data set 3. Піготувати abalone data set до аналізу
Pyle D. Data Preparation for Data Mining: https://pdfs.semanticscholar.org/ 470a/ 828d5e3962f2917a0092cc6ba46ccfe41a2a.pdf

3. Застосувати кілька класичних класифікаторів scikit-learn на abalone datasets
Classifier comparison: http://scikit-learn.org/stable/auto_examples/ classification/ plot_classifier_comparison.html

4. Оцінити якість класифікації з використанням precision-score

5. Підготувати звіт