

004.89
М 38

КОМП'ЮТИНГ

Т.М. БАСЮК, В.В.ЛИТВИН,
Л.М.ЗАХАРІЯ, Н.Е.КУНАНЕЦЬ

МАШИННЕ НАВЧАННЯ

Навчальний посібник



Computing

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»
ІНСТИТУТ КОМП'ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА МЕРЕЖ

Т. М. БАСЮК, В. В. ЛИТВИН, Л. М. ЗАХАРІЯ, Н. Е. КУНАНЕЦЬ

МАШИННЕ НАВЧАННЯ

НАВЧАЛЬНИЙ ПОСІБНИК

«Новий Світ-2000»

Львів

2019

*Рекомендувала Науково-методична рада
Національного університету «Львівська політехніка»
як навчальний посібник для студентів, що навчаються
за першим (бакалаврським) рівнем вищої освіти
за спеціальностями галузі знань 12 „Інформаційні технології”
(протокол №41 від 6.02.2019 р.)*

Рецензенти:

Мокін Віталій Борисович - д.т.н., професор, завідувач кафедри системного аналізу, комп'ютерного моніторингу та інженерної трафіки Вінницького національного технічного університету

Литвиненко Володимир Іванович -д.т.н., професор, завідувач кафедри інформатики та комп'ютерних наук Херсонського національного технічного університету

Філатов Валентин Олександрович - д.т.н., професор, завідувач кафедри штучного інтелекту Харківського національного університету радіоелектроніки

Теслюк Василь Миколайович - д.т.н., професор, професор кафедри систем автоматизованого проектування Національного університету «Львівська політехніка»

Т. М. Басюк, В. В. Литвин, Л. М. Захарія, Н. Е. Кунанець

Машинне навчання: Навчальний посібник призначений для студентів, що навчаються за першим (бакалаврським) рівнем вищої освіти за спеціальностями галузі знань 12 „Інформаційні технології”. Львів: Видавництво «Новий Світ - 2000», 2019. – 315 с.

Навчальний посібник „Машинне навчання” призначений для формування теоретичних, практичних навичок та їх використання при побудові аналітичних інформаційних систем та аналітичних модулів у системах підтримки прийняття рішень, коли необхідно спрогнозувати результат за вхідними даними. Він складається з двох частин. У першій частині начального посібника розглянуто класичні методи машинного навчання: навчання з вчителем та без вчителя. Методи навчання з вчителем призначені для розв'язування задач класифікації (прогнозування категорії об'єкта) та регресія (прогнозування місця на числовій прямій); без вчителя - для кластеризації, узагальнення та пошуку правил. Детально описано основні методи таких навчань. Друга частина присвячена іншим методам машинного навчання (навчання з підкріпленням та нейронним мережам).

ISBN 978-617-7519-42-2

© Т. М. Басюк, В. В. Литвин, Л. М. Захарія, Н. Е. Кунанець, 2019

© «Новий Світ-2000», ФОП ПічаС.В., 2019

Зміст

ВСТУП.....	3
ЧАСТИНА 1. КЛАСИЧНЕ НАВЧАННЯ.....	11
РОЗДІЛ 1. МАШИННЕ НАВЧАННЯ.....	11
1.1. Вступ в машинне навчання.....	11
1.1.1. Означення навчання.....	11
1.1.2. Програми, що навчаються.....	13
1.1.3. Мотивація до навчання.....	17
1.1.4. Таксономія машинного навчання.....	18
1.1.5. Споріднені галузі.....	20
1.1.6. Навчання як розділ штучного інтелекту.....	22
1.2. Загальне формулювання задач навчання за прецедентами.....	23
1.2.1. Основні поняття та означення.....	24
1.2.2. Типологія задач навчання за прецедентами.....	24
1.2.3. Задачі з описом об'єктів на основі ознак.....	27
1.3. Приклади задач машинного навчання.....	28
1.3.1. Задачі класифікації.....	28
1.3.2. Задача відновлення регресії.....	34
1.3.3. Задачі прогнозування та прийняття рішень.....	34
1.3.4. Задачі кластеризації.....	37
1.3.5. Задачі аналізу клієнтських середовищ.....	38
1.4. Навчання понять в штучному інтелекті.....	40
1.4.1. Задача навчання понять – пошук у просторі гіпотез.....	41
1.4.2. Упорядкування гіпотез „від загальної до конкретної”.....	44
1.4.3. Алгоритм Find-S пошуку максимально конкретної гіпотези.....	46
1.4.4. Алгоритм „вилучення кандидата”.....	49
1.5. Задачі для самостійного розв’язування.....	65
Питання для самоперевірки.....	70
РОЗДІЛ 2. НАВЧАННЯ З ВЧИТЕЛЕМ.....	71
2.1. Задача навчання із вчителем.....	71
2.1.1. Постановка задачі навчання з вчителем.....	71
2.1.2. Машина опорних векторів.....	73
2.2. Регресійний аналіз.....	82
2.2.1. Регресійна модель.....	82
2.2.2. Гребенева регресія.....	93
2.2.3. Логістична регресія.....	99
2.3. Стохастичний градієнтний спуск і онлайн-підхід до навчання.....	111

2.3.1.	Стохастичний градієнтний спуск.....	111
2.3.2.	Онлайн-підхід до навчання.....	112
2.3.3-	Робота з категоріальними ознаками Label Encoding, One-Hot Encoding, Hashing trick.....	113
2.3.4.	Навчання логістичній регресії.....	115
2.4.	Бібліотека Vowpal Wabbit.....	116
2.4.1.	Багатокласова класифікація.....	120
2.4.2.	Бінарна класифікація.....	123
2.4.3.	Класифікація понять.....	125
2.5.	Градієнтний бустинг.....	126
2.5.1.	Вступ та історія появи бустинга.....	126
2.5.2.	GBM алгоритм. Постановка задачі.....	127
2.5.3.	Функціональний градієнтний спуск.....	129
2.5.4.	Класичний GBM алгоритм Friedman-a.....	130
2.6.	Функції втрат.....	132
2.6.1.	Функції втрат регресії.....	133
2.6.2.	Класифікації функції втрат.....	135
2.6.3.	Застосування wаg.....	137
2.7.	Критерії порівняння алгоритмів машинного навчання.....	139
2.8.	Наївний баєсів класифікатор.....	141
2.8.1	Формулювання.....	141
2.8.2.	Робота з нульовими рахунками (дисконтування).....	144
	Питання для самоперевірки.....	145

РОЗДІЛ 3. ЛОГІЧНІ ТА МЕТРИЧНІ МЕТОДИ КЛАСИФІКАЦІЇ.

ДЕРЕВА РІШЕНЬ.....	147	
3.1.	Означення метода дерева рішень.....	147
3.2.	Алгоритми побудови дерева рішень.....	149
3.2.1.	Алгоритм побудови дерева рішень – ID3(D, C, A).....	150
3.2.2.	Методи оптимізації дерев рішень–рання зупинка. Усічення дерев рішень, проблеми вибору атрибута для процедури “розділяй і керуй”.....	155
3.2.3.	Алгоритм побудови дерева рішень – C4.5(D, C, A).....	159
3.2.4.	Алгоритм CART (Classification and Regression Trees).....	164
3.3.	Нечіткі дерева рішень.....	166
3.3.1	Основні принципи нечіткої логіки та її використання.....	166
3.3.2.	Алгоритм побудови нечіткого дерева рішень.....	168
3.3.3.	Випадковий ліс.....	173
3.3.4.	Градієнтний бустинг дерев рішень.....	175
3.4.	Метричні методи класифікації.....	176
3.4.1	Метод найближчого сусіда і його узагальнення.....	176
3.4.2.	Алгоритм k найближчих сусідів (k nearest neighbors, kNN).....	178

3.4.3. Алгоритм к зважених найближчих сусідів.....	178
3.4.4. Метод парзенівського вікна.....	179
3.4.5. Метод потенційних функцій.....	180
3.4.6. Відбір еталонних об'єктів.....	181
3.4.7. Алгоритм STOLP для відбору еталонних об'єктів.....	183
Питання для самоперевірки.....	185

РОЗДІЛ 4. НАВЧАННЯ БЕЗ ВЧИТЕЛЯ..... 186

4.1. Задача навчання без вчителя.....	186
4.2. Основні поняття кластерного аналізу.....	186
4.3. Застереження стосовно використання кластерного аналізу.....	188
4.4. Поняття подібності.....	189
4.5. Міри подібності.....	192
4.5.1. Коефіцієнти кореляції.....	192
4.5.2. Міри відстані.....	193
4.5.3. Коефіцієнти асоціативності.....	194
4.5.4. Імовірнісні коефіцієнти подібності.....	195
4.6. Методи кластерного аналізу.....	195
4.6.1. Ієрархічні агломеративні методи.....	196
4.6.2. Ітеративні методи групування.....	200
4.6.3. Інші методи.....	203
4.7. Модель кластерного аналізу.....	206
4.8. Підготовка даних до кластерного аналізу.....	208
4.8.1. Формування таблиці «об'єкт - ознака».....	209
4.8.2. Нормування даних.....	210
4.8.3. Побудова дендограм.....	212
4.9. Алгоритми кластеризації.....	219
4.9.1. Метод головних компонент.....	219
4.9.2. Метод к-середніх.....	220
4.9.3. Метод медіан.....	221
4.9.4. Алгоритм DBSCAN.....	221
4.9.5. Прийоми і метрики для кластеризації часових рядів.....	225
4.10. Метрики якості кластеризації.....	226
4.10.1. Метрика Adjusted Rand Index (ARI).....	226
4.10.2. Метрика Adjusted Mutual Information (AMI).....	227
4.10.3. Гомогенність, повнота, V-міра.....	227
4.10.4. Силует.....	227
Перелік контрольних запитань.....	229

РОЗДІЛ 5. ВІДСТАНІ ТА МЕРЕЖЕВІ МЕТОДИ..... 230

5.1. Вимірювання відстаней.....	230
5.1.1 Метрики відстаней.....	230

5.1.2. Метрика Лк відстаней.....	231
5.1.3. Робота у вищих вимірах.....	233
5.1.4. Просторовий егалітаризм.....	233
5.1.5. Точки та вектори.....	234
5.1.6. Відстані між розподілами ймовірностей.....	235
5.2. Класифікація сусідніх значень.....	237
5.3. Асоціативні правила.....	240
5.3.1. Поняття асоціативного правила.....	240
5.3.2. Алгоритм Аргіогі побудови асоціативних правил.....	244
5.3.3. Сфера застосування асоціативних правил.....	247
Питання для самоперевірки.....	247

ЧАСТИНА 2. ІНШІ МЕТОДИ НАВЧАННЯ..... 248

РОЗДІЛ 6. НАВЧАННЯ З ПІДКРІПЛЕННЯМ..... 248

6.1. Генетичні алгоритми.....	248
6.1.1. Схема генетичного алгоритму.....	248
6.1.2. Генетичні операції.....	249
6.1.3. Подання даних.....	251
6.1.4. Відбір.....	254
6.1.5. Підходи Дарвіна, Ламарка і Болдуїна.....	255
6.2. Стимулююче навчання.....	258
6.2.1. Оцінювання поведінки агента.....	259
6.2.2. Багаторуки бандити.....	261
6.2.3. Доказово оптимальні алгоритми.....	262
6.2.4. Евристичні стратегії.....	264
6.2.5. Модель агентів з декількома станами.....	266
6.2.6. Пошук оптимальних стратегій у відомій моделі.....	267
Питання для самоперевірки.....	269

РОЗДІЛ 7. НЕЙРОННІ МЕРЕЖІ..... 270

7.1. Навчання на основі зв'язків.....	270
7.1.1. Біологічні нейронні мережі.....	271
7.1.2. Модель штучного нейрона.....	271
7.1.3. Подання нейромереж та їх архітектура.....	275
7.1.4. Сучасні архітектури нейромереж.....	277
7.1.5. Навчання одношарових нейромереж прямого поширення.....	282
7.1.6. Навчання багатошарових нейромереж прямого поширення.....	289
7.2. Мережі, що самоорганізуються.....	299
7.2.1. Опис мереж, що самоорганізуються.....	299
7.2.2. Міри відстані між векторами.....	301
7.2.3. Проблема нормалізації векторів.....	302

7.2.4. Міра організації мережі.....	302
7.2.5. Механізм стомлення нейронів.....	303
7.2.6. Методи навчання мереж, що самоорганізуються.....	304
7.2.7. Самоорганізаційні карти Кохонена.....	306
7.2.8. Winner relaxation/winner enhancing.....	309
7.2.9. Cluster refinement phase.....	310
7.2.10. Візуалізація, Rectifying SOM.....	312
7.2.11. Розширюваний нейронний газ (Growing Neural Gas, GNG).....	313
Перелік контрольних запитань.....	317
Список рекомендованої літератури.....	318